A Survey of Spatial Crowdsourcing

S. Gummidi, X. Xie and T. Pedersen

<u>Presented By:</u> Rithika A Ramasesha Mehnaz Tabassum Mahin Shamali Shinde Urja Gaurav Parekh



Outline

- Spatial Crowdsourcing
 - Infrastructure
 - Classification
- Spatial Problems
- Constraints
 - Spatial
 - Quality
 - Budget
- Privacy Protection
- Truth Discovery and Crowdsourced Data Aggregation

 $e^{-1/2}$

Applications and Future Directions

SPATIAL CROWDSOURCING Introduction and Classification

Rithika A Ramasesha



Introduction

Spatial crowdsourcing advances the potential of a crowd to perform tasks related to real-world scenarios involving physical locations, which were not feasible with conventional crowdsourcing methods. The main feature of spatial crowdsourcing is the presence of spatial tasks that require workers to be physically present at a particular location for task fulfillment



Introduction

- SC has the potential for collecting information for a broad range of applications in domains such as environmental data collection (NoiseTube platform), transportation (Uber), journalism, and business intelligence (Gigwalk and TaskRabbit).
- SC can be extended to service complex spatial tasks such as employing a set of workers with diverse skills to renovate a house, hiring workers to perform household chores.





Spatial Crowdsourcing Infrastructure

Requester:

- a real-world entity:
- a person or an organization
- A requester designs the task and sets the conditions that need to be satisfied for performing the task
- Accepts the answers provided or data collected by the crowd and gives them feedback

Spatial Crowdsourcing Infrastructure

Worker:

- The main objective is to perform the assigned/selected spatial task
- physically move to a particular geographical location to perform a spatial task



Spatial Crowdsourcing Infrastructure

Worker definition:

$W = \langle I, R, maxT, Re, E, \theta \rangle$

- Current physical location of the worker I
- Region of interest R
- Maximum
- Expected reward Re
- Skillset E
- Worker's reputation score θ

Spatial Crowdsourcing Infrastructure

Spatial Task :

- A spatial task is requested by the requester and is fulfilled by the worker.
- It is a location-specific activity, such as answering a question about a local restaurant, taking pictures of a local tourist spot, or collecting noise pollution data
- The requirements might include the :
 - a. minimum level of skills required to fulfill a task
 - b. the number of answers needed
 - c. the expertise of the user, the task deadline
 - d. the user's experience in solving similar tasks.

Spatial Crowdsourcing Infrastructure

Spatial Task Definition:

t = <I, q, ti_i , ti_e , r, α , E, Cat>

- Physical location of the task I
- Query description q
- Issuing time ti_i
- Expiration time ti_e
- Associated reward r
- Minimum threshold for worker's reputation α
- Expected worker's skillset E
- Type/category of task Cat
- Maximum number of workers maxW

Spatial Crowdsourcing Infrastructure

Spatial Crowdsourcing Server:

- the communication between the requester and the worker of the spatial task
- facilitates the process to satisfy the task requirements
- assigns tasks to workers based on location/
- helps improve the quality of the outcome by executing different strategies
- identifies the anomalies and detects fraudulent responses
- protects the privacy of the involved stakeholders

Spatial Crowdsourcing Infrastructure

For the Crowd :

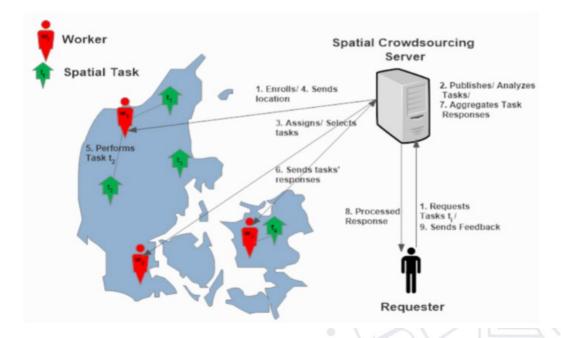
- The power to disclose or hide their locations to the SC server
- Ability to select tasks that need to be performed near their location
- Ability to specify a spatial region in which they wish to work, that is, the geographical extent to which they can travel for performing a spatial task
- Ability to provide location privacy protection of workers
- Ability to reject assigned tasks

Spatial Crowdsourcing Infrastructure

For Requested Tasks :

- Determining the maximum acceptable distance of the workers in the vicinity of the task
- Defining the geographical location where the requested task needs to be performed
- Ability to hide the location where the task needs to be performed or to be shown only to workers who accepted or selected the task.

Spatial Crowdsourcing Infrastructure



Classifications

Spatial Tasks :

- A) Spatial task complexity:
 - Atomic tasks: They are simple and can be performed by one worker

0

Complex tasks: Two or more related activities that need to performed

collaboratively for accomplishing the spatial task

B) Number of responses

Classifications

Spatial Tasks :

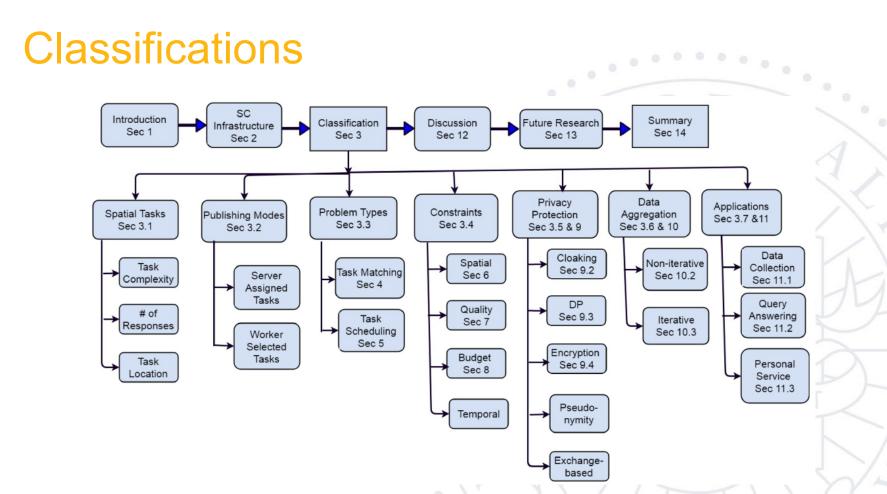
- C) Task's physical location:
 - Point task : The spatial task is required to be performed at that particular point location
 - Area/Region task : The spatial task is required to be performed by the worker in a region of geographical space instead of a particular location
 - Delivery Task: The spatial task has two parts: pick up the package from the point of interest A and deliver the package to the building B



Classifications

Modes of Publishing Spatial Tasks :

- Server Assigned Tasks: The SC server chooses available suitable workers for a given spatial task based on different parameters:
 - a) their proximity to the spatial task
 - b) availability to perform the task
 - c) abilities to match the requirements of the task
 - d) reliability of the workers to assure some degree of quality in task outcomes.
- Worker Selected Tasks: Workers select spatial tasks of their interest from a given list published by the SC server.



SPATIAL PROBLEMS

Mehnaz Tabassum Mahin



Spatial Problems

- Task assignment problems

 ✓A set of tasks and a set of workers
 ✓Arrangement between tasks and workers
 - ✓Optimization goal
 - Maximize the number of tasks assigned

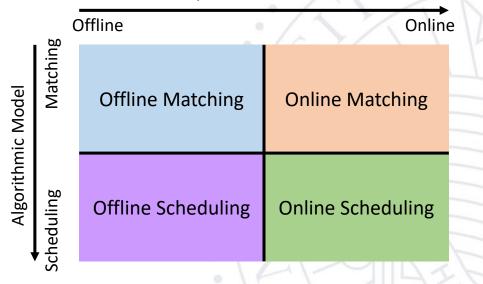
t4 1

Fig: Task assignment example

- Maximize the reward of the task
- Minimize the travel distance



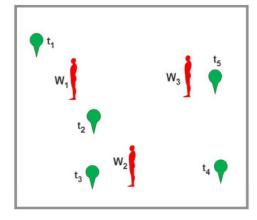
Task Assignment Problems



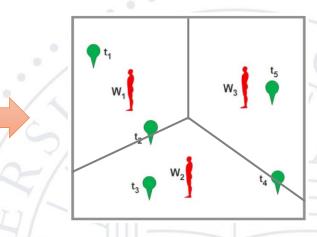
Input Arrival Scenario



Task Matching Problem



Worker ID	Preferred Spatial Task	Reward
W1	t1	5
	t2	3
W2	t2	6
	t3	2
W3	t5	8



۰.

Worker ID	Assigned Spatial Tasks	Rewards
W1	t1	5
W2	t2, t3, t4	8
W3	t5	8

Task Matching Problem: Offline MWBM

- Assumption: One worker can perform one task
- Objective: Maximize the rewards received by the workers

			5 11 /
Worker ID	Preferred Spatial Task	Reward	$W_1 \longrightarrow t_2$
	t1	5	6
W1	t2	3	W2 2 t3
W2	t2	6	— W3 t4
	t3	2	8 t5
W3	t5	8	ZILANNA
			A total reward of 19 is achieved in the offline scenario

Task Matching Problem: Online MWBM

- Arrival of tasks and workers are unknown.
 ✓ Partial bipartite graph: locally optimal solution
- Example:

<W1, t2, W2, t4, t1, W3, t3, t5>

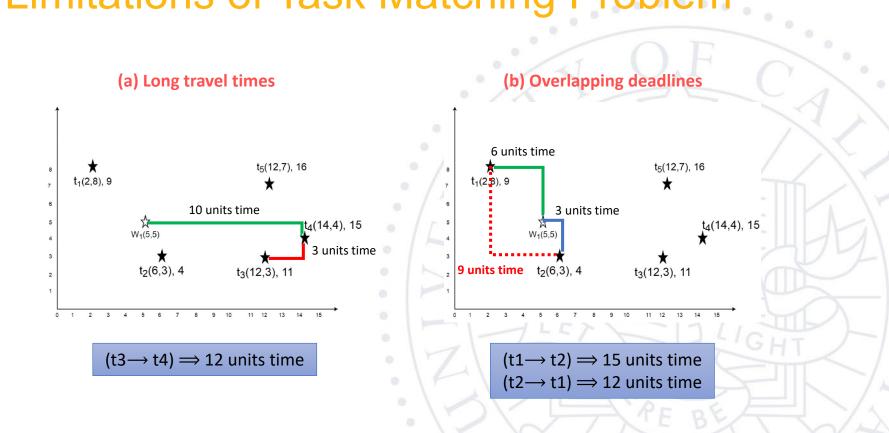
Worker ID	Preferred Spatial Task	Reward
\4/1	t1	5
W1	t2	3
W2	t2	6
W2	t3	2
W3	t5	8

Total reward is

reduced to 13

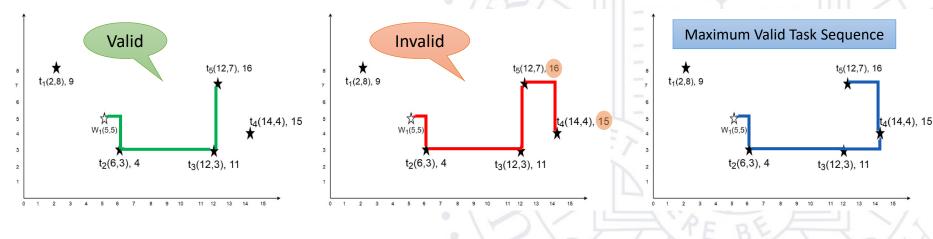
• Effectiveness depends on the arrival order of tasks and workers.

Limitations of Task Matching Problem



Task Scheduling Problem: Offline Model

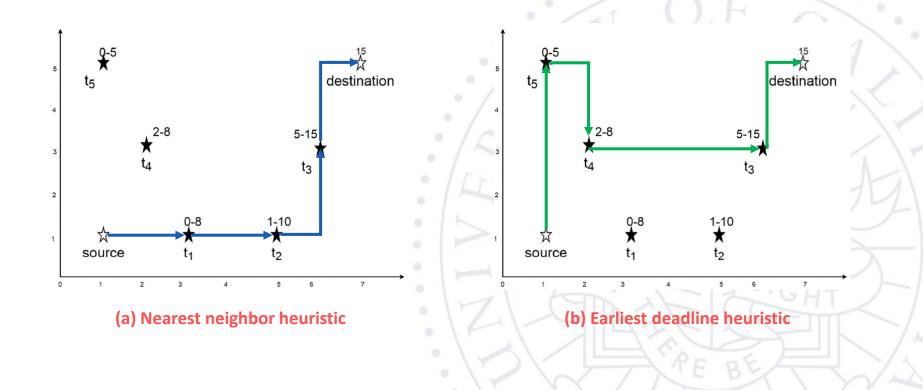
- Maximum Task Scheduling (MTS) Problem
 - ✓ Maximize Valid Task Sequence
 - ✓ NP-hard problem!
- Valid Task Sequence
 - ✓ All tasks in the sequence can be completed considering travel cost and deadlines



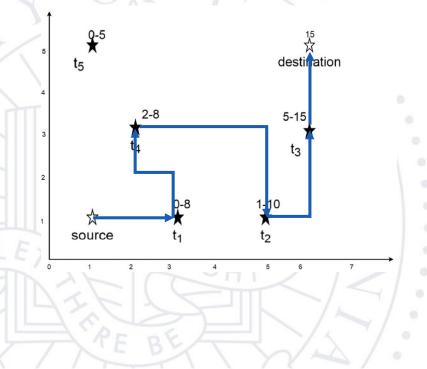
Task Scheduling Problem: Online Model

- Improvements of Task Scheduling Problem
 - ✓ Update task sequence based on arrival of new tasks
 - ✓Workers' preference
 - Tasks present on their route
- Online Route Recommendation (OnlineRR) Problem
 - ✓Workers choose their tasks
 - ✓ Solutions
 - Greedy approach
 - Rerouting method

OnlineRR Problem: Greedy Approach



OnlineRR Problem: Rerouting Method



<u>Time t</u>	Task sequence R		
0	(source, t1, t5, destination)		
2	(t1, t4, t2, destination)		
5	(t4, t2, t3, destination)		
10	(t2, t3, destination)		
13	(t3, destination)		

CONSTRAINTS & PRIVACY PROTECTION

Shamali Shinde



Spatial Constraints

Spatial constraints refer to the different spatial preferences of the worker or the task.

- Spatial region
- Maximum Travel Distance
- Direction of the worker's commute.

Spatial Constraints

Spatial constraints refer to the different spatial preferences of the worker or the task.

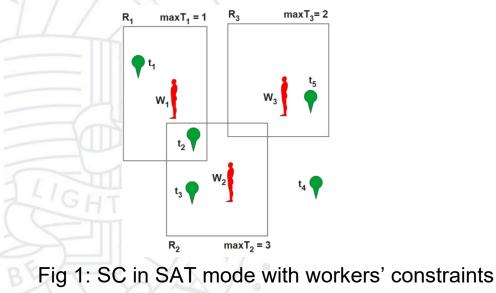
- Spatial region
- Maximum Travel Distance
- Direction of the worker's commute.

Spatial Region

- Workers define spatial region as a minimum bounding rectangle or as a bounding circle.
- Requester can specify spatial constraints on the task

0

. *



Maximum Travel Distance

- Worker can set a maximum traveling distance
- Server intuitively assign tasks to maximize the number of tasks performed and the overall rewards
- Server should not violate the maximum traveling distance threshold of individual workers

Direction of Worker's Commute

- Assigning tasks to a worker without a significant deviation from the worker's path
- Improves the diversity of spatial angles regarding the information queried by the task

.

Quality Constraints

Worker's Expertise:

.

- Expertise indicates the capabilities/skills of the worker in performing a specific type of task
- Worker's Reputation/ Reliability
 A worker's reputation/reliability refers to the probability that a task is completed correctly by the worker

Worker's Expertise

- Worker Expertise: E w
- Degree: D w
- Task Expertise: E t
- Degree D t
- Constraint for assigning a spatial expert task is given by:

 $E w = E t \wedge D w \ge D t$

The expertise E w or E t refers to a specific skill

Budget Constraints

- Reward Models and Incentive Mechanisms
- Platform-centric : SC server controls the allocation of rewards to the workers
 - Fixed reward for the task
 - Dynamic reward for the task
 - Fixed budget for time periods
 - Dynamic budget for time periods
- User-centric : Worker controls over the payment by denoting the price for which they are willing to perform the tasks
 Reward expected by the worker

Privacy Protection

Shared location information is highly sensitive and susceptible to privacy attacks from adversaries such as a potentially untrustworthy SC server.

Cloaking Techniques

0

- Differential Privacy-Based Techniques
- Encryption Techniques

Cloaking Technique

- The workers' locations are hidden in cloaked regions.
- Spatial k-anonymity: generates a cloaked region for each worker containing k 1 other workers
- Prone to homogeneity attacks when all of the k workers are present at the same location.

Differential Privacy-Based Techniques

- It refer to distortion of the original location information of the workers by addition of artificial noise.
- Addresses participant concerns regarding the leakage of personal information
- DP is the most used strategy to protect workers' locations during task assignment

Encryption Technique

- Task Service: uses onion encryption to hide the worker's identity and location through the anonymizing network, Tor.
- SC server performs the worker-task matching by communicating with the PSP in the encrypted domain
- Overhead computation cost is high compared with cloaking and DP-based techniques

TRUTH DISCOVERY & CROWDSOURCED DATA AGGREGATION

Urja Gaurav Parekh



WHAT IS TRUTH DISCOVERY?

•Non-Iterative Aggregation Identifying trustworthy information from the responses received from the workers. Truth discovery is relevant when a task is performed by multiple workers, wherein every worker provides uniform answers or conflicting answers.

•Majority Voting , Honeypot and ELICE EM, GLAD , SLME , and ITER are the non-iterative methods (techniques compute a single aggregated value for each question using heuristics such as the most common answer) to make the conclusions uniform and relevant.

•Iterative Aggregation : The iterative aggregation techniques consist of a sequence of iterations that compute probabilities of answers for each question in each iteration until convergence [54]. In this technique, the set of questions with known answers is not required.

•Expectation Maximization (EM): The EM technique simultaneously estimates all probabilities of answers for each question iteratively in two steps: expectation and maximization

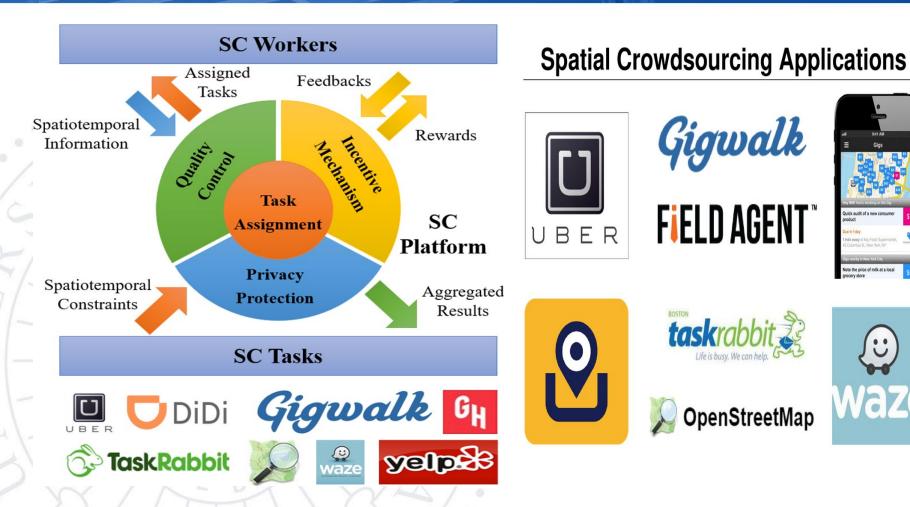
•Generative Model of Labels, Abilities, and Difficulties (GLAD): This technique is similar to EM except that it estimates toughness of the question along with workers' expertise, like the non-iterative aggregation technique ELICE

•Supervised Learning from Multiple Experts (SLME): This technique is similar to EM; however, instead of a confusion matrix, it characterizes the worker expertise by sensitivity, which is the ratio of positive answers that are correctly answered

•Iterative Learning (ITER): This technique is based on standard belief propagation. Similar to GLAD and ELICE, ITER also estimates the toughness of the question and the worker expertise

lote the price of milk at a local

waze



APPLICATIONS

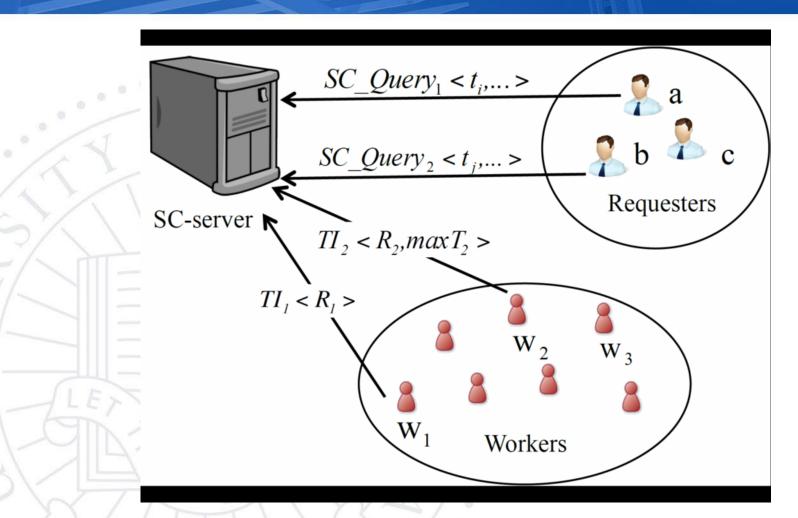
- Voting Systems, Information Sharing Systems, Game Systems, and Creative Systems.
- Inspired by these classification categories, we have grouped the SC applications into three broad categories based on the use of sensors, human knowledge, and human efforts: data collection, query answering, and personal service. Generally, the tasks involving quantifiable information fall under the data collection category and the ones involving qualitative information fall under the query answering category



Data Collection refers to the process of gathering information from the different sensors of workers' smartphones at particular locations without using the workers' knowledge capabilities. Most of the applications belonging to this category are labeled with "mobile crowdsensing" instead of SC. To describe the data collection process, let us take an example of building indoor floor plans from the traces of movement by the workers with the help of their smartphones [3]. These motion traces are based on the inertial sensors in the workers' smartphones, which are collected and processed later to build accurate indoor floor plan

Query Answering

- Query Answering is another class of SC application that involves harnessing the worker's knowledge to answer a group of questions related to a specific location or region. Unlike the case of data collection, the information-gathering process is not just limited to collecting sensor data of workers' smartphones. Query Answering also uses the worker's skill and the ability to answer the queries/tasks
- The query-answering applications can be further divided based on the type of queries answered. Some of the different types of queries are simple binary (Yes/No) queries, multipleoption queries, tagging the images with relevant tags, categorizing different images, describing images, and classifying the type of a spatial feature
- Data Collection refers to the process of gathering information from the different sensors of workers' smartphones at particular locations without using the workers' knowledge capabilities. Most of the applications belonging to this category are labeled with "mobile crowdsensing" instead of SC. To describe the data collection process, let us take an example of building indoor floor plans from the traces of movement by the workers with the help of their smartphones [3]. These motion traces are based on the inertial sensors in the workers' smartphones, which are collected and processed later to build accurate indoor floor plan.



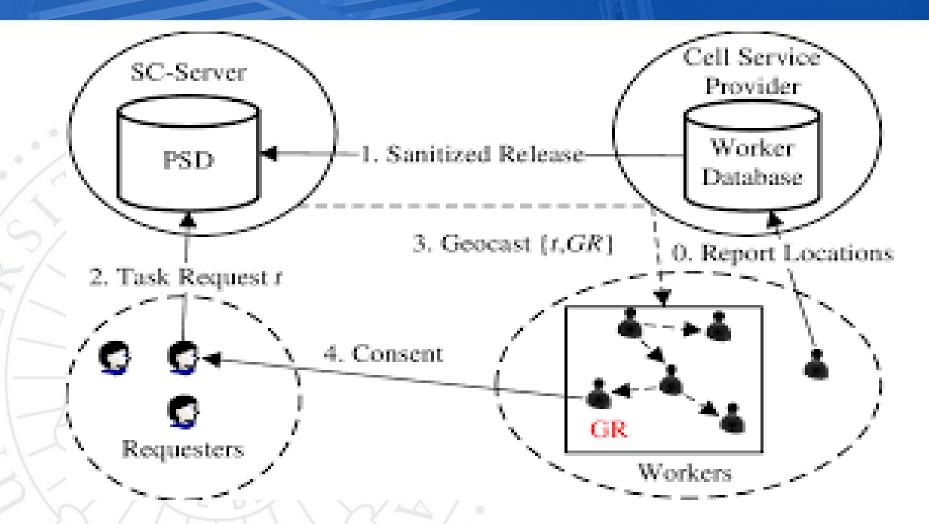
Personal Service

 Personal Service is another class of SC application that involves an additional human effort to perform the task, such as pick-up and delivery of a package/groceries/food order3, taxi calling and ride sharing4, and performing a task such as painting/cleaning/lawn mowing 5. The personalservice applications have to consider the different spatiotemporal, budget, and quality constraints for solving task matching and task scheduling problems. For example, the Ubereats3 application involves collecting food packages from a restaurant and delivery to customers. The Ubereats SC application matches the food-delivery tasks with the worker and plans the route for delivery.

Example SC Applications

TaskRabbit5

Mission

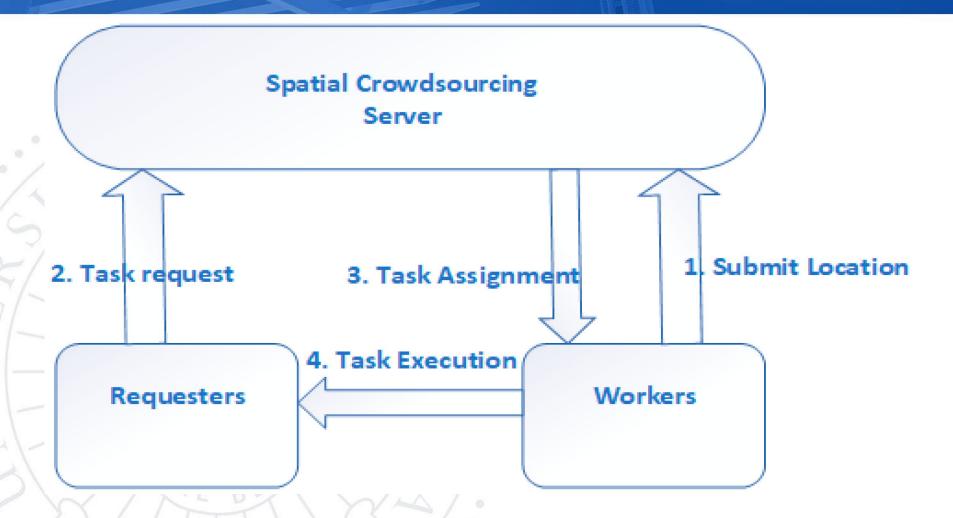


Task Matching and Task Scheduling Issues

- The task matching and task scheduling issues are related to the dynamic arrival of tasks and workers, the optimization goals, the immutability constraint, and the workers' movement patterns. The majority of the work done in SC does not account for the dynamic arrival of workers and tasks to the SC server during the task assignment phase [15, 26, 61] or task scheduling phase [29]. They work on the assumption that the SC server possesses the complete knowledge of the input sets of tasks and workers. This renders their solutions inefficient in dynamic real-time environments.
 - The current truth inference models of SC assume that the workers respond independently of each other, that is, there would not be any copying or sharing of answers among the workers. However, in reality, there could be a case of copying among the workers that could result in improper estimation of worker reliability, thereby affecting the quality of the task responses. Therefore, copy-detection methods are needed while inferring truth from the worker responses.

Privacy Issues

- Privacy concerns are one of the most fundamental problems of the worker. Although some literature in SC addressed this issue (as discussed in Section 9), there are many open issues that need to be addressed. Privacy concerns are fueled by the workers' lack of trust of the third-party SC server. To address this, some solutions use the concept of differential privacy to get the workers' aggregated data from worker-trusted cellular service providers to anonymize workers from the SC server.
- However, the SC server would still have the knowledge of the task locations and time intervals during which the assigned worker would visit task locations, resulting in a serious privacy breach. Furthermore, by anonymizing the worker to the SC server, the support for individual spatial constraints and quality constraints is hindered with the existing strategies.
- Although the authors of deal with the privacy-enabled quality assurance problem, they do not consider the differing requirements of the tasks and the differing spatial constraints of the workers.



Truth Inference Models

- In the existing truth inference models of SC, it is difficult to infer truth while considering worker's location privacy. For example, this is the case in location obfuscation, wherein the worker's location is generalized to protect the exact worker's location. In such cases, the existing truth inference models simply ignore the workers with obfuscated locations. However, considering the individual workers' location privacy concerns, new models are needed to tackle these issues.
- Similarly, the current truth inference models of SC assume that the workers respond independently of each other, that is, there would not be any copying or sharing of answers among the workers. However, in reality, there could be a case of copying among the workers that could result in improper estimation of worker reliability, thereby affecting the quality of the task responses. Therefore, copy-detection methods are needed while inferring truth from the worker responses. To prevent collusion between workers, it is important to quantify the probability that the workers collude with each other based on the quality of their responses to better assign the tasks. For example, in CC literature, the authors of have proposed a threestep θ -secure task assignment approach for task assignment avoiding collusion between workers. Similar approaches are needed in SC considering the spatial characteristics of workers and tasks. Furthermore, new approaches should be proposed similar to them for providing a global optimal solution in estimating the task ground truth and worker expertise in SC.

Lack of Real-World Datasets

- The primary challenge faced by all solutions proposed in the SC literature is to evaluate the strategies based on real-world datasets. Due to the lack of publicly available real-world datasets, SC algorithms are evaluated using synthetic datasets that are generated based on different distribution functions.
- Some SC works use few real-world datasets related to location-based social networks (LBSNs) such as Gowalla (http://snap.stanford.edu/data/loc-gowalla.html) and Bright-kite (http://snap.stanford.edu/data/loc-brightkite.html), in which users can check in to different POIs in their vicinity. These LBSN datasets are adapted to the SC scenario by assuming the check-in spots to be task locations, users to be workers, and a user checking into a spot is considered to be accepting the task.
- To advanced this strategy to generate synthetic SC datasets with realistic spatiotemporal properties and constraints adapted from geo-social datasets such as Gowalla and Yelp (http://yelp.com/dataset_challenge). The advantage of these datasets is that they exhibit the workers' nature of preferring to perform nearby tasks. However, there might be different geosocial phenomena in SC that are not observed with either synthetic or adapted datasets. There is a need to design an SC platform to collect real-world data for researchers to advance the research in SC.

Lacking User Participation

- For an SC application to be successful, it should be able to attract many task requesters and workers. Most of the existing applications are based on voluntary participation; as performing tasks involve spending time, effort, and resources, such as a smart phone's memory and battery, it is difficult to attract workers without offering rewards.
- Moreover, with the privacy concerns, the users might not be willing to participate in the SC application. To motivate users to participate in SC, incentive mechanisms have been developed that involve monetary rewards and virtual credits.
- Though incentive mechanisms have a positive impact on improving user participation, they are still limited to the users who are aware of the SC paradigm. To harness the true potential of SC, new methods should be developed to enable the general public to become aware of the SC application and to convince them to participate in SC applications.

FUTURE RESEARCH DIRECTIONS

- Improving Task-Assignment Protocols in Online Scenarios: Task assignment protocols should be improved to tackle the uncertainty of the dynamic real-time SC environment. Although different solutions, such as those presented in [113], are proposed to address this, further research needs be performed for improving the efficiency and for ascertaining the impact on different constraints, such as quality and privacy. In particular, strategies should be proposed to perform task assignments in an online scenario with privacy-enabled SC.
 - Assignment Protocols that Benefit Both Workers and Tasks: The existing task matching protocols attempt to assign tasks to workers based on workers' preferences or tasks' preferences. Optimizing the benefits for both tasks and workers would improve the success of the assignment protocol. A new assignment strategy could be proposed by adapting the solution proposed in the CC literature with spatiotemporal context. They have proposed a task-assignment framework in CC, called Task Assignment with Mutual Benefit Awareness (TAMBA), to offer mutual benefit to workers and tasks based on their preferences extracted from the historical data. Similarly, the authors of have proposed an auction-based framework, Auction-SC, that would benefit both workers and the SC server by allowing workers to bid on the arriving tasks.

FUTURE RESEARCH DIRECTIONS

- Integration of Task-Publishing Modes: There is a need for an effective framework that combines both task-publishing modes (WST and SAT) to provide effective solutions for improving the efficiency of SC and the task-acceptance rate. For instance, the workers can select their preferred tasks via WST mode. In the case in which multiple workers are opting for the same task, the SC server can employ the SAT mode to resolve the conflict.
 - **Privacy-Enabled Truth Inference Models**: The current truth inference models in SC can avoid real-time location tracking of the workers and exploit the historical information for inferring truth and to worker reliability models. However, in a privacy-protected SC, the individual worker locations would be unknown to the SC server, which the current inference models cannot support. Therefore, new truth inference models should be proposed for the different location-privacy models to ensure quality to the task requester. Furthermore, new techniques are needed to process complex textual or multimedia information to assess the trustworthiness of the responses.

FUTURE RESEARCH DIRECTIONS

- Quality Assurance-Privacy Preservation Trade-off: As discussed in Section 9.1, privacy preservation and quality assurance negatively affect each other, as privacy-protected SC might hinder the quality of the responses and vice versa, especially in the case of single-response tasks with differing requirements. Therefore, a trade-off mechanism is needed to balance the privacy requirements of workers and quality constraints of tasks set by requesters. For instance, with some cloaking techniques, such as k-anonymity, and quality constraint information, such as reliability and expertise skill set, can be aggregated on the client side of the workers. An aggregate query of k workers can be sent to SC server along with the cloaked region and quality constraints information for task assignment. Furthermore, the location information of tasks should also be protected from the SC server along with the worker information to avoid privacy breach.
 - **Improving User Participation in SC**: As discussed in Section 12, there is a need to devise better strategies to attract users to improve the participation in SC applications. The existing strategies, such as incentive mechanisms, are useful to an extent. However, they are still limited to the user base familiar with SC. The new strategies should expand the reach of the SC applications to users who are not as familiar with SC as well.

FUTURE RESEARCH DIRECTIONS

• Harnessing Geosocial Network Information: Current SC literature contains little work considering the geo-social relationships between workers that could be helpful in enriching the worker's profile. Location influence [93] of workers can be used to provide a partial ranking to workers in team-oriented task planning [42]. Resulting partial ranking helps the SC server to select the leaders for the teams. Furthermore, the location influence concept can be used in allocating rewards to the workers in a dynamic budget reward model. Workers with the highest location influence ranking would be attracted to perform the task by offering a higher reward. By attracting the highly ranked location influencers, their followers are attracted to the task location, thus increasing the worker diversity of the location and improving the chances for a task to be assigned in the neighborhood.



<u>SUMMARY</u>

- In this survey, we reviewed the existing literature related to SC from a technical perspective. We distinguished different topics in the research and proposed our taxonomy to organize them. We noticed that the architecture of SC adapts the structure of CC to serve spatiotemporal interests. Furthermore, we observed that the majority of the existing work focuses on task matching along with varying constraints since the SC server exerts more control on the task matching proceedings.
- Similarly, we observed a significant density of research focusing on offline scenarios. Our comparison study revealed the shortcomings of the different strategies and identified relationships between the various constraints of SC. The quality constraints are found to be negatively impacted by the privacy-protection approaches and positively correlated with the budget constraints. The identified shortcomings and challenges are related to task assignment in online scenarios, the dynamic movement of workers, the privacy-quality trade-off, and the geo-social relationships. We suggest future work to address these challenges and advance the application spectrum of SC.



NOW, ARE THERE ANY QUESTIONS?