Spatial Crowdsourcing: Challenges, Techniques, and Applications

Yongxin Tong† Lei Chen‡ Cyrus Shahabi§
†SKLSDE Lab, Beihang University, China
‡The Hong Kong University of Science and Technology, Hong Kong SAR, China
§University of Southern California, USA
†yxtong@buaa.edu.cn, ‡leichen@cse.ust.hk, §shahabi@usc.edu

ABSTRACT

Crowdsourcing is a new computing paradigm where humans are actively enrolled to participate in the procedure of computing, especially for tasks that are intrinsically easier for humans than for computers. The popularity of mobile computing and sharing economy has extended conventional web-based crowdsourcing to spatial crowdsourcing (SC), where spatial data such as location, mobility and the associated contextual information, plays a central role. In fact, spatial crowdsourcing has stimulated a series of recent industrial successes including Citizen Sensing (Waze), P2P ride-sharing (Uber) and Real-time Online-To-Offline (O2O) services (Instacart and Postmates).

In this tutorial, we review the paradigm shift from web-based crowdsourcing to spatial crowdsourcing. We dive deep into the challenges and techniques brought by the unique spatio-temporal characteristics of spatial crowdsourcing. Particularly, we survey new designs in task assignment, quality control, incentive mechanism design and privacy protection on spatial crowdsourcing platforms, as well as the new trend to incorporate crowdsourcing to enhance existing spatial data processing techniques. We also discuss case studies of representative spatial crowdsourcing systems and raise open questions and current challenges for the audience to easily comprehend the tutorial and to advance this important research area.

1. INTRODUCTION

Crowdsourcing is a new computing paradigm where humans actively participate in the procedure of computing, especially for tasks that are intrinsically easier for humans than for computers. It has attracted much attention of both the industrial and the academic communities with the blossom of successful crowdsourcing platforms [14, 20, 24, 31]. With the unprecedented development of mobile Internet and sharing economy, crowdsourcing platforms are shifting from traditional web-based crowdsourcing platforms, such as Amazon Mechanical Turks (AMT) [1] and oDesk [6], to spatial crowdsourcing (*a.k.a* mobile crowdsourcing) platforms [16], where *(i)* each crowd worker (worker for short) is considered as a mobile computing unit to complete tasks using their mobile devices [34] and *(ii)* spatial information such as location, mobility and the associated contexts plays a crucial role. Applications of spatial crowdsourcing cover a wide spectrum of ubiquitous computing in daily life, where the most representative include real-time taxi-calling service, *e.g.*, Uber [9] and DiDi [2], product placement checking in supermarkets, *e.g.*, Gigwalk [3] and TaskRabbit [8], on-wheel meal-ordering service, *e.g.*, GrubHub [4] and Instacart [5], and citizen sensing service, *e.g.*, Waze [10] and OpenStreetMap [7]. Despite the success of general-purposed crowdsourcing [14, 20, 24, 31], the unique spatio-temporal dynamics in spatial crowdsourcing calls for new designs in crowdsourcing theories and systems. The concept of crowdsourcing also brings in new opportunities to enhance existing research on spatial data processing.

In this tutorial, we review the state-of-the-art research on spatial crowdsourcing and point out future challenges and opportunities. The tutorial will act as an invitation to the database community to fill up and solve emerging and potential questions from all kinds of spatial crowdsourcing applications. The tutorial is divided into five parts. In the first part, we review a brief history of spatial crowdsourcing research and motivate the need for spatial crowdsourcing via typical real-world applications in daily life.The second part highlights four core issues in spatial crowdsourcing platforms including task assignment, quality control, incentive mechanism and privacy protection. The third part focuses on the new trend to incorporate crowdsourcing to enhance existing spatial data processing techniques. The fourth part shows various case studies of systems and applications in spatial crowdsourcing. Finally, we identify the open problems and current challenges of spatial crowdsourcing and conclude the tutorial in the fifth part.

2. TARGET AUDIENCE

The target audience for this tutorial is anyone who is interested in crowdsourced data processing in general, and spatial crowdsourcing in particular. The tutorial does not require any particular background or knowledge about crowdsourcing and spatial data processing techniques. However it requires basic database knowledge, which is assumed to be there for the typical database conference attendees including junior database students.

Proceedings of the VLDB Endowment, Vol. 10, No. 12

3. TUTORIAL OUTLINE

The tutorial consists of five parts as shown below and is explained in detail in the following.

- **Part 1: Motivations of Spatial Crowdsourcing (10 minutes)**
  - A brief history of spatial crowdsourcing
  - Typical applications of spatial crowdsourcing
- **Part 2: Science of Spatial Crowdsourcing (30 minutes)**
  - Task assignment
  - Quality control
  - Incentive mechanism
  - Privacy protection
- **Part 3: Crowdsourced Spatial Data Processing (30 minutes)**
  - Crowdsourced route recommendation
  - Crowdsourced map matching
  - Crowdsourced urban traffic speed estimation
  - Crowdsourced POI labelling
- **Part 4: Case Studies (10 minutes)**
  - General system case studies
  - Application case studies
- **Part 5: Open Problems (10 minutes)**
  - Theoretical guarantee of online task assignment
  - Dynamic indexing in spatial crowdsourcing
  - Lack of benchmarks about spatial crowdsourcing

3.1 Part 1: Motivations of Spatial Crowdsourcing

As introduced in the outline, the first part takes 10 minutes to explain the need for spatial crowdsourcing. Specifically, we first review the history from traditional web-based crowdsourcing to spatial crowdsourcing and then illustrate the importance of spatial crowdsourcing with representative real-world applications. Finally, we explain why existing web-based general-purpose crowdsourcing techniques cannot support the needs in spatial crowdsourcing [44, 45, 46].

3.2 Part 2: Science of Spatial Crowdsourcing

In the second part, as shown in the outline above, we spend 10 minutes to summarize the core issues in spatial crowdsourcing platforms, including *task assignment*, *quality control*, *incentive mechanisms* and *privacy protection*.

Task Assignment. Task assignment is important in spatial crowdsourcing [21, 22, 23, 28, 29, 46]. It aims to assign tasks to suitable workers such that the total number of assigned tasks or the total weighted value of the assigned pairs of tasks and workers is maximized. We categorize task assignment in spatial crowdsourcing into *(static) offline and (dynamic) online scenarios*. In offline scenarios most efforts unitize weighted bipartite matching approaches [21, 23, 28, 32, 44, 47, 49]. In online scenarios, the spatio-temporal information of all the tasks and workers is unknown beforehand. Existing solutions often develop two-sided online matching algorithms to adapt the subsequent unknown arrival objects [36, 39, 41, 45, 46, 48].

Quality Control. Results collected from mobile computing units can be of low quality and noisy, which makes quality control essential in spatial crowdsourcing. There are two types of quality control with either worker or spatio-temporal constraints. For worker constraints based quality control, traditional approaches such as majority voting [13] can be extended to spatial crowdsourcing [29]. For spatio-temporal constraints based quality control, research efforts not only use the spatio-temporal diversity to enhance the quality of aggregated results but also utilize different skills of workers to maximize the quality of aggregated results [18, 19].

Incentive Mechanism. An effective incentive mechanism is indispensable in spatial crowdsourcing. Workers need to move from one location to another to perform tasks and the platform needs to dynamically adjust the rewards to workers based on the spatial distribution of tasks and workers. Most solutions integrate online auction and game theory techniques for mechanism design in spatial crowdsourcing [37, 38].

Privacy Protection. Privacy protection in spatial crowdsourcing is related to location-based privacy [25], and focuses on the privacy of mobile workers in dynamic scenarios. Privacy protection is an emerging issue in spatial crowdsourcing [35, 42, 43], which often builds specific index structures to satisfy differential privacy.

3.3 Part 3: Crowdsourced Spatial Data Processing

In addition to the science of spatial crowdsourcing, we also introduce the efforts to enhance spatial data processing via crowdsourcing. First, we demonstrate how to integrate the suggestions from taxi drivers (crowds) for better route recommendation [40, 50]. Second, we show the benefits of crowdsourcing in map matching, a well-known difficult problem in spatial database research [12, 33]. Finally, we illustrate recent research on urban traffic speed estimation and POI labelling leveraging crowdsourcing [26, 27].

3.4 Part 4: Case Studies

In the fourth part, we spend 10 minutes on various case studies in spatial crowdsourcing. We first explain how full-fledged systems such as gMission [17] and MediaQ [30] are tailored to support the four core issues in spatial crowdsourcing. Then we demonstrate case studies of industrial applications to solve real-world problems in daily life, such as taxi dispatching [45] and ride-sharing [11].

3.5 Part 5: Open Problems in Spatial Crowdsourcing

In the fifth part, we discuss cutting-edge open problems in spatial crowdsourcing, including but not limited to the following three problems. First, most real-world applications of spatial crowdsourcing are under online scenarios [45, 46]. An interesting open question is whether there are online task assignment techniques that have theoretical guarantee for the assignment results [45]. The second open issue is whether existing spatial indexes, which support moving object queries, can be extended to support the online data processing in spatial crowdsourcing. Third, well-defined benchmarks to test and compare different spatial crowdsourcing data processing techniques are desired.

4. RELEVANCE TO VLDB

The research of spatial data processing has always been one of the most important parts of the database community. Crowdsourced data processing is becoming a hot topic in recent database conferences and journals. Particularly, with the blossom of successful large-scale spatial crowdsourcing platforms such as Uber, Didi, Gigwalk, etc., a tutorial about large-scale spatial crowdsourcing is desired. Many of the research efforts covered in this tutorial have been published in highly refereed database and geographic information systems conferences, such as VLDB, SIGMOD, ICDE, GIS. In addition, this tutorial introduces how to extend and integrate the key techniques of crowdsourced data processing and spatial data management to satisfy the requirements of the novel large-scale spatial platforms, which makes it directly relevant to the database community.

5. PREVIOUS TUTORIALS

Lei Chen has already delivered two successful tutorials about crowdsourcing-based data processing in ICDE 2015 and CIKM 2014, and two about uncertain data processing in VLDB 2015 and DASFAA 2012.

This tutorial is spiritually connected to two recent tutorials, "Data-driven Crowdsourcing: Management, Mining, and Applications" [14] and "Crowdsourcing in Information and Knowledge Management" [15], which are presented by Lei Chen with Dongwon Lee, Tova Milo and Meihui Zhang at ICDE 2015 and CIKM 2014, respectively. However, this tutorial substantially differs from the above two tutorials in content. All techniques and approaches described in the current tutorial focus on spatial crowdsourcing scenarios such as Uber and Gigwalk, rather than general web-based crowdsourcing scenarios, such as Amazon Mechanical Turks (AMT) and oDesk, which are covered in the previous two tutorials. Therefore, this tutorial is substantially different from the aforementioned two tutorial in ICDE 2015 and CIKM 2014, respectively.

6. CONCLUSION

The rapid development of mobile Internet and the Online-To-Offline (O2O) business model has stimulated the boom of all kinds of spatial crowdsourcing platforms in daily life. This tutorial aims to not only raise awareness of this topic in the database community but also invite the database researchers to advance this promising area. We survey the state-of-the-art techniques for spatial crowdsourcing, with comprehensive comparisons among the challenges and techniques in spatial crowdsourcing, traditional spatial data processing, and general-purposed crowdsourcing. We also highlight open questions for future research in this active research area. We envision this tutorial as a bridge to link the research in the database community and in other disciplines to develop more comprehensive techniques for spatial crowdsourcing.

7. BIOGRAPHIES OF PRESENTERS

7.1 Yongxin Tong

Yongxin Tong is an associate professor at the School of Computer Science and Engineering, Beihang University, China. He received his Ph.D. degree in Computer Science from the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST). Before joining Beihang Unversity, he served as a research assistant professor and a postdoctoral fellow at HKUST. His research interests include crowdsourcing, uncertain data processing and social network analysis. He has published more than 20 papers in highly refereed database and data mining journals and conferences such as SIGMOD, VLDB, ICDE, SIGKDD and TKDE. He received the Excellent Demonstration award and the Best Paper award conferred by the VLDB 2014 and WAIM 2016 conferences, respectively. He has also served in the program committees of some international conferences and workshops, e.g. VLDB, ICDE, SIGKDD, IJCAI, etc.

7.2 Lei Chen

Lei Chen is a full professor at the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. He received the BS degree in computer science and engineering from Tianjin University, China, in 1994, the M.A. degree from the Asian Institute of Technology, Thailand, in 1997, and the PhD degree in computer science from the University of Waterloo, Canada, in 2005. His current research interests include crowdsourcing-based data processing, uncertain and probabilistic databases, Web data management, multimedia databases, and privacy-preserved data publishing. He authored one book and more than two hundred research papers in highly refereed database and data mining journals and conferences. He got the SIGMOD Test-of-Time Award in SIGMOD 2015 and the Excellent Demonstration award in VLDB 2014. Currently, he serves as an editor-in-chief for VLDB Journal, an associate editor-in-chief for IEEE Transaction on Data and Knowledge Engineering (TKDE) and a Trustee Board Member of VLDB Endowment. He has served regularly in the organization committees and the program committees of many international conferences and workshops. In particular, he was PC track chair for SIGMOD 2014, VLDB 2014, ICDE 2017/2012, CIKM 2012, SIGMM 2011.

7.3 Cyrus Shahabi

Cyrus Shahabi is a professor of Computer Science and Electrical Engineering and the Director of the Information Laboratory (InfoLAB) at the Computer Science Department and also the Director of the NSF's Integrated Media Systems Center (IMSC) at the University of Southern California (USC). He received his B.S. in Computer Engineering from Sharif University of Technology in 1989 and then his M.S. and Ph.D. Degrees in Computer Science from the University of Southern California in May 1993 and August 1996, respectively. He authored two books and more than two hundred research papers in the areas of databases, GIS and multimedia with more than 12 US Patents. He received the ACM Distinguished Scientist Award in 2009, the 2003 US Presidential Early Career Awards for Scientists, and Engineers (PECASE) and the US NSF CAREER award in 2002. He is a fellow of the IEEE. He was an Associate Editor of IEEE Transactions on Parallel and Distributed Systems (TPDS) from 2004 to 2009, IEEE Transactions on Knowledge and Data Engineering (TKDE) from 2010-2013 and VLDB Journal from 2009-2015. He is currently on the editorial board of the ACM Transactions on Spatial Algorithms and Systems (TSAS) and ACM Computers in Entertainment. He is the general co-chair of SSTD'15, ACM GIS 2007, 2008 and 2009. He chaired the nomination committee of ACM SIGSPATIAL for the 2011-2014 terms. He has been a PC co-Chair of BigComp 2016, MDM 2016, DASFAA 2015, IEEE MDM 2013 and IEEE BigData 2013, and regularly serves on the program committee of major conferences.

Acknowledgment

We are grateful to anonymous reviewers for their constructive comments on this work. Yongxin Tong is supported in part by National Grand Fundamental Research 973 Program of China under Grant 2014CB340300, NSFC Grant No. 61502021 and 71531001, and SKLSDE (BUAA) Open Program SKLSDE-2016ZX-13. Lei Chen is supported in part by Hong Kong RGC Project N HKUST637/13, NSFC Grant No. 61328202, NSFC Guang Dong Grant No. U1301253, HKUST-SSTSP FP305, Microsoft Research Asia Gift Grant, Google Faculty Award 2013 and ITS170. Cyrus Shahabis work has been funded in part by NSF grants IIS-1320149 and CNS-1461963.

8. REFERENCES

[1] Amazon's mechanical turk. *https://www.mturk.com.*
[2] Didi. *http://www.xiaojukeji.com/news/newslisten.*
[3] Gigwalk. *http://www.gigwalk.com.*
[4] Grubhub. *https://www.grubhub.com/.*
[5] Instacart. *https://www.instacart.com/.*
[6] odesk. *http://www.odesk.com/.*
[7] Openstreetmap. *http://www.openstreetmap.org/.*
[8] Taskrabbit. *http://www.taskrabbit.com.*
[9] Uber. *https://www.uber.com/.*
[10] Waze. *http://www.waze.com/.*
[11] M. Asghari, D. Deng, C. Shahabi, U. Demiryurek, and Y. Li. Price-aware real-time ride-sharing at scale: an auction-based approach. In *GIS 2016*.
[12] S. Brakatsoulas, D. Pfoser, R. Salas, and C. Wenk. On map-matching vehicle tracking data. In *VLDB 2005*.
[13] C. C. Cao, J. She, Y. Tong, and L. Chen. Whom to ask? jury selection for decision making tasks on micro-blog services. *PVLDB 2012*.
[14] L. Chen, D. Lee, and T. Milo. Data-driven crowdsourcing: Management, mining, and applications. In *ICDE 2015*.
[15] L. Chen, D. Lee, and M. Zhang. Crowdsourcing in information and knowledge management. In *CIKM 2014*.
[16] L. Chen and C. Shahabi. Spatial crowdsourcing: Challenges and opportunities. *IEEE Data Engineering Bulletin*, 2016.
[17] Z. Chen, R. Fu, Z. Zhao, Z. Liu, L. Xia, L. Chen, P. Cheng, C. C. Cao, Y. Tong, and C. J. Zhang. gmission: A general spatial crowdsourcing platform. *PVLDB 2014*.
[18] P. Cheng, X. Lian, L. Chen, J. Han, and J. Zhao. Task assignment on multi-skill oriented spatial crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 2016.
[19] P. Cheng, X. Lian, Z. Chen, R. Fu, L. Chen, J. Han, and J. Zhao. Reliable diversity-based spatial crowdsourcing by moving workers. *PVLDB 2015*.
[20] A. I. Chittilappilly, L. Chen, and S. Amer-Yahia. A survey of general-purpose crowdsourcing techniques. *IEEE Transactions on Knowledge and Data Engineering*, 2016.
[21] D. Deng, C. Shahabi, and U. Demiryurek. Maximizing the number of worker's self-selected tasks in spatial crowdsourcing. In *GIS 2013*.
[22] D. Deng, C. Shahabi, U. Demiryurek, and L. Zhu. Task selection in spatial crowdsourcing from worker's perspective. *GeoInformatica 2016*.
[23] D. Deng, C. Shahabi, and L. Zhu. Task matching and scheduling for multiple workers in spatial crowdsourcing. In *GIS 2015*.
[24] H. Garcia-Molina, M. Joglekar, A. Marcus, A. G. Parameswaran, and V. Verroios. Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 2016.
[25] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K. Tan. Private queries in location based services: anonymizers are not necessary. In *SIGMOD 2008*.
[26] H. Hu, G. Li, Z. Bao, Y. Cui, and J. Feng. Crowdsourcing-based real-time urban traffic speed estimation: From trends to speeds. In *ICDE 2016*.
[27] H. Hu, Y. Zheng, Z. Bao, G. Li, J. Feng, and R. Cheng. Crowdsourced poi labelling: Location-aware result inference and task assignment. In *ICDE 2016*.
[28] L. Kazemi and C. Shahabi. Geocrowd: enabling query answering with spatial crowdsourcing. In *GIS 2012*.
[29] L. Kazemi, C. Shahabi, and L. Chen. Geotrucrowd: trustworthy query answering with spatial crowdsourcing. In *GIS 2013*.
[30] S. H. Kim, Y. Lu, G. Constantinou, C. Shahabi, G. Wang, and R. Zimmermann. Mediaq: Mobile multimedia management system. In *MMSys 2014*.
[31] G. Li, J. Wang, Y. Zheng, and M. Franklin. Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2016.
[32] C. Long, R. C.-W. Wong, P. S. Yu, and M. Jiang. On optimal worst-case matching. In *SIGMOD 2013*.
[33] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang. Map-matching for low-sampling-rate GPS trajectories. In *GIS 2009*.
[34] M. Musthag and D. Ganesan. Labor dynamics in a mobile micro-task market. In *CHI 2013*.
[35] L. Pournajaf, L. Xiong, V. Sunderam, and S. Goryczka. Spatial task assignment for crowd sensing with cloaked locations. In *MDM 2014*.
[36] J. She, Y. Tong, L. Chen, and C. C. Cao. Conflict-aware event-participant arrangement and its variant for online setting. *IEEE Transactions on Knowledge and Data Engineering*, 2016.
[37] Y. Singer and M. Mittal. Pricing mechanisms for crowdsourcing markets. In *WWW 2013*.
[38] A. Singla and A. Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *WWW 2013*.
[39] T. Song, Y. Tong, L. Wang, J. She, B. Yao, L. Chen, and K. Xu. Trichromatic online matching in real-time spatial crowdsourcing. In *ICDE 2017*.
[40] H. Su, K. Zheng, J. Huang, H. Jeung, L. Chen, and X. Zhou. Crowdplanner: A crowd-based route recommendation system. In *ICDE 2014*.
[41] H. To, L. Fan, L. Tran, and C. Shahabi. Real-time task assignment in hyperlocal spatial crowdsourcing under budget constraints. In *PerCom 2016*.
[42] H. To, G. Ghinita, L. Fan, and C. Shahabi. Differentially private location protection for worker datasets in spatial crowdsourcing. *IEEE Transactions on Mobile Computing*, 2016.
[43] H. To, G. Ghinita, and C. Shahabi. A framework for protecting worker location privacy in spatial crowdsourcing. *PVLDB 2014*.
[44] H. To, C. Shahabi, and L. Kazemi. A server-assigned spatial crowdsourcing framework. *ACM Transactions on Spatial Algorithms and Systems*, 2015.
[45] Y. Tong, J. She, B. Ding, L. Chen, T. Wo, and K. Xu. Online minimum matching in real-time spatial data: Experiments and analysis. *PVLDB 2016*.
[46] Y. Tong, J. She, B. Ding, L. Wang, and L. Chen. Online mobile micro-task allocation in spatial crowdsourcing. In *ICDE 2016*.
[47] L. U, M. Yiu, K. Mouratidis, and N. Mamoulis. Capacity constrained assignment in spatial databases. In *SIGMOD 2008*.
[48] U. ul Hassan and E. Curry. A multi-armed bandit approach to online spatial task assignment. In *UIC 2014*.
[49] R. C.-W. Wong, Y. Tao, A. W.-C. Fu, and X. Xiao. On efficient spatial matching. In *VLDB 2007*.
[50] C. J. Zhang, Y. Tong, and L. Chen. Where to: Crowd-aided path selection. *PVLDB 2014*.