Privacy-Preserving Mobile Crowdsourcing for Maximum Coverage via Randomized Response

ABSTRACT

We propose a novel mechanism of recruiting participants in order to maximize data coverage of a mobile crowdsourcing task with strong privacy guarantees. The mechanism is mainly based on randomized response, which preserves location privacy in a local way but provides low utility for values with low frequencies. Also, the original maximum coverage problem is NP-hard. We address these challenges by modeling the posterior of visit for each location using aggregated data and use it as the objective to find the optimal subset in polynomial time. We describe the design of the whole mechanism and present preliminary results.

CCS CONCEPTS

• **Security and privacy → Data anonymization and sanitization**;

KEYWORDS

privacy, crowdsourcing, maximum coverage

ACM Reference Format:

. 2018. Privacy-Preserving Mobile Crowdsourcing for Maximum Coverage via Randomized Response. In *Proceedings of ACM Woodstock conference (WOODSTOCK'97).* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Motivation. With the significant growth of mobile smart phone users, crowdsourcing has emerged as a feasible platform to collect data by using mobile phones as censors, especially under urban sensing scenarios. However, due to limited budget or resources, it's only feasible to recruit a subset of participant from the candidates pool. A subset which maximizes data coverage on given targets is the optimal choice for task allocation. At the meantime, participants will be reluctant to participate in a crowdsourcing program if their location privacy will be violated during the data collecting procedure. When the crowdsourcing task is first released, each candidate can choose several preferable locations from the given target location set. Since the crowdsourcing server may not be trusted and the chosen locations are usually most frequently visited places which may imply a candidate's daily travel plan, precise location leakage may lead to a broad spectrum of attack, e.g. physical surveillance and stalking. So the main concern here is how to provide strong privacy guarantees for participants while preserving high coverage of collected data. Here we propose a novel mechanism called randomized response to protect location privacy of participants with a rigorous privacy bound proof provided in the notation of differential privacy.

Randomized Response Developed in 1960s, the randomized response was proposed to collect statistics on sensitive topics with confidentiality. The basic idea is a respondent may decide whether to report the truth depending on the result of flipping a coin in secret, which achieves strong deniability.

Differential Privacy Introduced by Dwork et al [1], the rigorous notion of privacy has been widely adopted, which aims to ensure that the randomized mechanism will provide the same output with almost equal probabilities for two adjacent datasets with only one different record and therefore protect individual privacy. The formulation of differential privacy can be described as $P(K(v_1) \in R) \leq e^{\epsilon} P(K(v_2) \in R), \forall R \subseteq Range(K)$, where K is our mechanism and v_1 and v_2 denote all pairs of client's values. A smaller constant ϵ denotes a higher level of privacy protection. It's been proved in [2] that the randomized response mechanism satisfies the definition of differential privacy.

Challenges. First proposed in [2], the randomized response mechanism called RAPPOR can only detect those values with high frequencies, which may lead to inaccuracy in finding optimal subset for maximum data coverage. So the first challenge here is how to detect those individuals having traveled to those locations with lower frequency and add them to our participant set. The second challenge comes from the recruitment process. Since the original maximum coverage problem is NP-hard, novel algorithm should be proposed to make the whole process in an online manner.

Proposed Solution. Instead of focusing on estimation of aggregate number of visiting number of people in a specific area, here we plan to estimate the posterior of whether a location having visit record or not, given a subset of candidates. Then by summing up the posterior over location, we adopt a heuristic way of finding the optimal set of candidates with maximum estimated coverage.

2 SYSTEM DESIGN

The whole system for privacy-preserving mobile crowdsourcing task allocation can be illustrated in a client-server manner in Fig.1. Since all the real locations are kept secret on the client side and the server can only do task allocation based on obfuscated reports, the privacy is guaranteed.

Phase 1: *(Client Side)* Obfuscate Reports

The crowdsourcing server should first provide a set of $\mathcal{L}$ target locations and a budget $\mathcal{N}$ for the size of participant set. Then this task requirement is sent to client APP of every interested candidate. Every candidate can choose k preferable locations from target locations. Using one-hot encoding, each location is mapped to a

WOODSTOCK'97, July 1997, El Paso, Texas USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

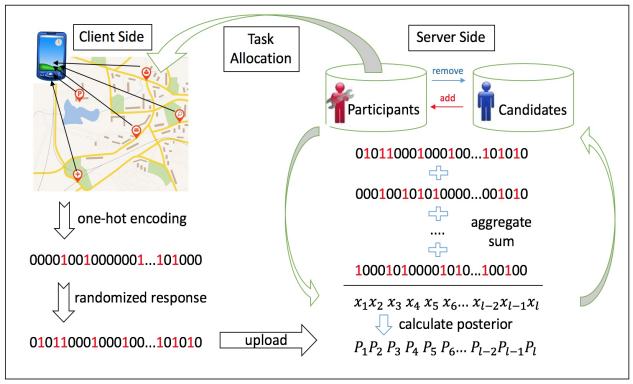

Figure 1: An overview of system design in a server-client manner

bit in the bit array, i.e. a bit array with k bits set to 1 while others are zero. Then the randomized response mechanism embedded in the client APP can help obfuscate each bit using the same way as flipping coin and report this obfuscated version back to the server. The formal formulation of obfuscation is shown in Eqn.1, where S_i denotes i-th bit in the report and B_i denotes i-th bit in the real bit array.

$$P(S_i = 1) = \begin{cases} p_{11}, & \text{if } B_i{=}1 \\ p_{01}, & \text{if } B_i{=}0 \end{cases} \tag{1}$$

Phase 2: *(Server Side)* Task Allocation

After waiting for a certain time period or receiving enough number of candidates, the server can start task allocation. In order to find optimal subset of candidates, the server first randomly choose $\mathcal{N}$ candidates as initial choice. Then these $\mathcal{N}$ obfuscated reports are used to calculate aggregate sum X_l for each location l. Also the posterior of having no visit record among the subset, namely $\overline{\mathcal{P}_l}$, can be calculated. Here we assume the prior of each location having visit records can be obtained using previous records. The formulation of posterior in a specific location l is listed as follows, where we omit the proof for brevity.

$$\overline{\mathcal{P}_l} = \frac{B(X_l; N, p_{01})(1-\xi_l)^N}{\sum_{i=0}^{N} \sum_{m=max(0, X_l+i-N)}^{min(i, X_l)} B(i; N, \xi_l) B(i; m, p_{11}) B(N-i; X_l - m, p_{01})} \tag{2}$$

where $B(m; N, p) = \binom{N}{m} p^m (1-p)^{N-m}$ denotes the probability mass function of binomial distribution and ξ_l denotes the prior for location l.

Note that the posterior varies only according to different X_l and the maximum data coverage is equivalent to the minimum of $\sum_{l=1}^{\mathcal{L}} \overline{\mathcal{P}_l}$. During each iteration, a candidate from those unchosen ones which maximizes the increment of $\sum_{l=1}^{\mathcal{L}} \overline{\mathcal{P}_l}$ will be added and one that minimizes the reduction of $\sum_{l=1}^{\mathcal{L}} \overline{\mathcal{P}_l}$ will be removed. Using a way similar to gradient descendant, the final optimal subset can be found after a certain times of iteration. Then only those chosen candidates are treated as participants and will finish the job at the locations which they marked in Phase 1.

Following the proof in [2], the privacy bound ϵ can be derived from Eqn.3 using p_{11} and p_{01}. So given a privacy bound ϵ and either of the transfer probability, the mechanism can be fixed.

$$\epsilon = \log\left(\frac{p_{11}(1-p_{01})}{p_{01}(1-p_{11})}\right) \tag{3}$$

3 EVALUATION

Dataset Description We perform experiments on real-trajectory dataset collected in Beijing among 10K users. The whole area of Beijing is divided into grids to simulate target locations. On each trajectory, the top-k frequent locations are chosen as preferable locations to simulate the process of crowdsourcing.

Results The performance of our mechanism is illustrated in 2, where we compare our model with three baselines. *No-noise* uses real locations, which preserves the highest utility but lowest privacy. *noisy* uses the obfuscated locations directly and *random* just sample a random subset from candidates. The random baseline is used as the unit one for comparison. Preliminary results show that our model outperform the latter two baselines when using obfuscated reports with different parameter settings.

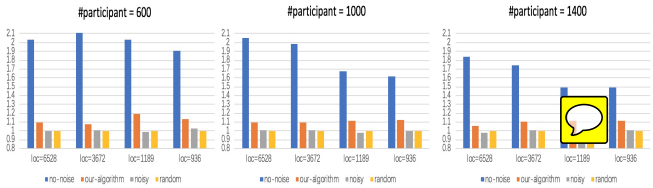

Figure 2: Preliminary results varying $\mathcal{N}$ and $\mathcal{L}$, given $\epsilon = 4$ and $p_{01} = 0.02$. Here the varying granularity of grid gives different number of location.

Future Work We plan to conduct more experiments to find out the relationship between data coverage and parameters, especially the number of participants $\mathcal{N}$ and number of locations $\mathcal{L}$.

REFERENCES

[1] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*. Springer, 1–19.

[2] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. ACM, 1054–1067.