Improvement of Estimate Distribution with Local Differential Privacy

Hikaru Horigome[1] and Hiroaki Kikuchi[1][0000−0002−0903−8430]

Gradulate School of Advanced Mathematical Science, Meiji University, 4-21-1 Nakano, Tokyo 164-8525, Japan {cs212030,kikn}@meiji.ac.jp

Abstract. Personalized high quality services including route finding and the nearest shops and restaurants are provided based on current location of owner of the smart device. However, location trace are very sensitive data for privacy. It allows to estimate our home residence or office. Hence, privacy preservation is required for reporting current location traces from smart devices.

This paper studies the privacy preservation of time-series location trace using LDP algorithm RAPPOR. Location trace is independently randomized according to given procedures and then is sent to service provider who aggregates data with noise. To discard noise and estimate true statistics, the maximum likelihood estimation is used in RAPPOR. But, MLE could fail if data distribution is skewed or data contains extraordinary values. To address the problem, we propose the expected maximization for estimate of true distributions. The proposed algorithm iteratively improves estimated posterior probabilities based on Bays' theorem until the difference converged for all elements. Our experiment using 6,528 individuals' location trace in Tokyo provided from Nightley Inc. demonstrates that the proposed algorithm performs better than the original MLE used in RAPPOR for every special ward in Tokyo in one day. We found that the accuracy is improved as privacy budget ϵ is smaller, and as many population is provided.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

Smart devices allow us to have better personalized service in the era of IoTs. For example, commercial services provide route finding and the nearest shops and restaurants based on current location of owner of the smart device. As for measure against Covid-19, time-series population distributions provided from cellphone providers plays important role for evaluation of restriction of people's movement.

However, location traces are very sensitive data for privacy. It allows to estimate our home residence or office. The correlation between any given location traces may reveal the personal relationship between them. Location service

provider may be compromised by malicious third party. The platforms may contain potential insider who can steal and disclose private customer's data. Therefore, privacy enhancement for location trace is necessary for preventing privacy threats.

Differential privacy has been studied to guarantee privacy preservation. With Laplacian mechanism, the statistics is perturbed so that no one can distinguish two neighboring datasets that differ only one individual. Erlingsson et al at Google proposed a LDP algorithm, Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR)[1]. It is well known that RAPPOR permits collecting over large number of devices without revealing private attributes such as frequencies, categories, and statistics of the devices. RAPPRO is based on randomized response[3] and estimates true attribute in the most likelihood value (MLE).

MLE used in RAPPOR does not always work well. It could fail to estimate true data if the distribution of data is biased or data contains extraordinary high/low values. We find that unbalanced distribution yields significant error in estimate in this paper. Since it estimate at the most likelihood value, even if only one illegal value can spoil the overall accuracy. Unfortunately, this could happen for use-case of location-based services where movement of people are unpredicted.

In this paper, we propose an iterative approach to improve estimate accuracy of perturbed data in LDP algorithm. Our idea is based on Bays' theorem and the Expected Maximization (EM) algorithm[5]. It estimates the posterior probability that are most consistent with given perturbed data. Due to iterative processes, the estimate is improved repeatedly. Hence it is more stable and more robust against unexpected behavior of hazard records.

We conduct an experiment using SNS-based location trace to demonstrate a feasibility of the proposed algorithm and to clarify accuracy improvement in the real location data. Our data contains 6,528 individuals' location trace in Tokyo provided from Nightley Inc. that are classified into several smaller special wards. We show the comparison between our proposed estimate (EM) and the MLE used in RAPPOR for several privacy budgets ϵ.

Our contribution has two folds.

- We propose a new algorithm to estimate the distribution of private data from perturbed data in RAPPOR. Our proposed algorithm improves accuracy of estimate based on iterative process of Bayesian posterior probabilities.
- We show the experimental results using a large scale location trace data in Tokyo with several smaller wards. The result shows that the proposed one performs better than the MLE used in RAPPOR for significant improvements.

2 Local Differential Privacy

2.1 Fundamental Definition

Suppose that users periodically submit their location data to a service provider. Differential privacy guarantees that the randomized data does not reveal any privacy disclosure from them. On the other hand, Local Differentially privacy (LDP) needs no trusted party. The private location data are randomized by users before submitting to the service provider. LDP is defined as follows.

Definition 1. *A randomized algorithm Q satisfies ϵ-local differential privacy if for all pairs of values v and v' of range V and for all $S \subset Z$ and for $\epsilon \geq 0$,*

$$Pr[Q(v) \in S] \leq e^{\epsilon} Pr[Q(v') \in S]. \tag{1}$$

2.2 RAPPOR[1]

Erlingsson et al at Google proposed a LDP algorithm, Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR)[1]. It is motivated by an application to track the distribution of users' browser configuration in Chrome.

Let v_i be element of V and be flipped according to randomized mechanism Q. Output z_i is set to be 1 for $v_i = 1$ with probability p, and 0 with probability $q = 1 - p$ as,

$$z_i = \begin{cases} v_i & w/p\ p \\ 1 - v_i & w/p\ q \end{cases}$$

In RAPPOR input v_i is so called "one-hot" encoded as a d-bit vector that contains exact 1 one and $d - 1$ zeros. Sensitivity Δf, the maximum influence that a single individual can have on the result of a randomized response, is 2 bits. For instance, suppose user 1 and 2 have $v = (0, 1, 0, 0)$ and $v' = (0, 0, 1, 0)$, respectively. A probability that randomized algorithm Q outputs $z = (0, 1, 0, 1)$ for v is

$$Pr[Q(v_1) = (0, 1, 0, 1) \,|\, v = (0, 1, 0, 0)] = (1 - q)p(1 - q)q.$$

Similarly, user 2 has the same output z with probability of $Pr[Q(v') = [0, 1, 0, 1]|v'] = (1 - q)q(1 - p)q$. If we set

$$p = \frac{e^{\frac{\epsilon}{\Delta f}}}{1 + e^{\frac{\epsilon}{\Delta f}}} = \frac{e^{\frac{\epsilon}{2}}}{1 + e^{\frac{\epsilon}{2}}} \text{ and } q = \frac{1}{1 + e^{\frac{\epsilon}{2}}}$$

then, it satisfies ϵ-local differential privacy as follow

$$\frac{Pr[Q(v) = z \,|\, v]}{Pr[Q(v') = Z \,|\, v']} = \frac{(1 - q)p(1 - q)q}{(1 - q)q(1 - p)q} \leq e^{\epsilon}.$$

Intuitively, no one can distinguish v and v' for users from the randomized output Z and hence the local (value) privacy is preserved. The privacy budget ϵ controls the degree of privacy and improves privacy as it is close to 0.

Generally, n-bit vectors v and v' have sensitivity $\Delta f = \sum_{k=1}^{n} ||v_i - v'_i|| \leq 2$. Letting r and r' be numbers of inconsistent bits in v and v', respectively, we have $Pr[Q(v) = z \in S|v] = p^{n-r}q^r$ and $Pr[Q(v') = z \in S|v'] = p^{n-r'}q^{r'}$. After all, we confirm that Equation (1) holds as

$$\frac{Pr[Q(v) = z \,|\, v]}{Pr[Q(v') = z \,|\, v']} = \frac{p^{n-r}q^r}{p^{n-r'}q^{r'}} = \left(\frac{p}{q}\right)^{h'-h} = e^{\frac{\epsilon f}{2}} \leq e^{\epsilon}$$

where $\Delta f = h' - h$.

3 Improvement of Estimate

3.1 Most Likelihood Estimate

We consider a problem of estimating population distribution from the randomized in RAPPOR in this section.

Let n be a population of city x at a time. Suppose that people who live in the city move in their daily life. So, the current population is dynamic quantity ranging from 0 to the maximum ℓ, say the total population in the state that city x belongs. According to RAPPOR algorithm, Let n' be the randomized population of x according to the RAPPOR algorithm. With probabilities p (true) and q (flipped), the expected value of binomial distribution gives $n' = np + (\ell - n)q$, which leads the most likely estimation (MLE) of n as

$$L[n] = \frac{n' - \ell q}{p - q}.$$

Letting h be the number of n individuals who submit 1, we have probability distribution of n by addition of binomial distributions as

$$p(n) = \binom{n}{h} p^h q^{n-h} + \binom{\ell - n}{n' - h} p^{\ell - n - n' + h} q^{n' - h}$$

Fig. 1 shows the probability distribution of population of Shinjuku city at 14:00. The randomized population is increased around 2100.

3.2 Iterative Estimate

MLE used in RAPPOR algorithm works well for most cases but could be suffered low estimate accuracy for biased distribution. Instead, we consider an iterative estimate approach known as Expectation Maximization (EM) algorithm.

EM algorithm performs an iterative process for which posterior probabilities are updated through Bayer's theorem. Each iteration estimate the best probabilities θ_i for all cities i that are consistent with the given randomized outputs Z computed in RAPPOR. Hence, it is more robust against unbalanced distribution than the MLE is.

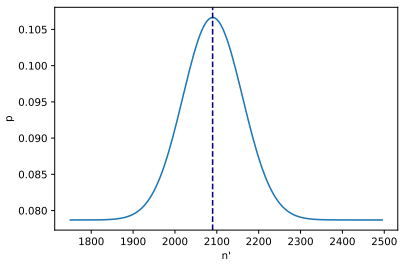

Fig. 1. probability distribuion $Pr(n'|n = 505)$

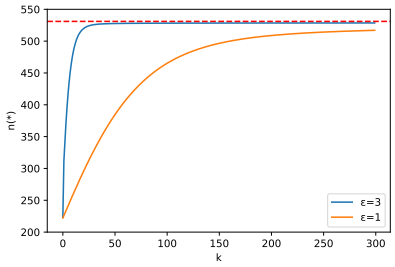

Fig. 2. Estimate population n^* with regards to number of iterations k

Algorithm 1 EM algorithm for RAPPOR

$\theta_i^{(0)} \leftarrow$ a population of city i.
repeat(E-step)
  $k \leftarrow 1$
  Estimate posterior probability $Pr[V_i = 1|Z]$ in Eq. (3).
  (M-step) Update marginal probability $\theta_i^{(k+1)}$ in Eq. (4).
until $|\theta_i^{(k+1)} - \theta_i^{(k)}| \leq \epsilon'$ **return** $n_i = \ell\theta_i^*$

Algorithm 1 shows the proposed Em algorithm for estimating true distribution from randomized data according to RAPPOR. It has two steps; Expectation (E-Step) and Maximization (M-Step). In the E-Step, Bays' theorem plays an significant role in estimating as the following ways.

Let V_i be random variable of population of i-th city in m cities in the state and Z be that of randomized one in RAPPOR. Conditional probability given $V_i = 1$ is

$$Pr[Z_i|V_i = 1] = \frac{Pr[Z_i, V_i = 1]}{Pr[Z_i = 1]}. \tag{2}$$

Bayes ' theorem gives the posterior probability of $V_i = 1$ given the randomized value Z as

$$Pr[V_i = 1|Z] = \frac{Pr[Z|V_i = 1]Pr[V_i = 1]}{\sum_{j=1}^{m} Pr[Z|V_j = 1]Pr[V_j = 1]} = \frac{Pr[V_i = 1|Z]\theta_i}{\sum_{j=1}^{m} Pr[Z|V_j = 1]\theta_j}, \tag{3}$$

where θ_i is the estimated probability of i-th city.

In the EM algorithm, the above Bayes' estimate is iterated to improve accuracy. For every iteration, a marginal distribution $\theta_i^{(k)}$ is replaced by the mean of posterior probability as

$$\theta_i^{(k+1)} = \sum_{j \in m} Pr[V_i = 1 \,|\, V_j = 1)\theta_i^{(k)}. \tag{4}$$

It continues until the estimate converges for all cities. Let θ_i^* be the converged probability $\theta_i^{(k+1)}$ if $|\theta_i^{(k+1)} - \theta_i^{(k)}|$ is less than a threshold of iteration. The final estimate is $n_i^* = \ell\theta_i^*$ for city $i \leq \ell$.

Fig. 2 shows the improvement of estimated population $n^{(*)}$ with regards to the number of iteration k. Estimated population in RAPPOR with $\epsilon = 1$ and 3 are plotted the figure, where the dotted line indicates the true population. Obviously, the accuracy is improved as k increases. We find the estimate is converged around $k = 200$ even for very strong privacy budget $\epsilon = 1$.

4 Experiment

4.1 Objective

Objective of the experiment is to explore the accuracy improvement using open location data and to demonstrate that the proposed algorithm works better than MLE in RAPPOR.

4.2 Data

Our experiment uses the time-series location data published from Nightley Inc. [4]. It is a synthetic data based on tweets of Social Networking Service. Table 2 shows specification of the dataset. We use one from the Nightley dataset that contains location trace for 6,258 individuals for a day. The populations are changed as people move from home to office or shops as shown in Table 4.

The city of Tokyo consists of 23 special wards. For each wards, we identify how many individuals stay for every three hours based on latitude and longitude provided with the trace (used by Google Map API). The time-series population for some major wards are shown in Table 3 and Fig. 3. We observe two typical behaviors;

(a) residential area, where populations are higher in morning and night, e.g., Nakano and Koto wards
(b) office area, where many schools, office and shops are located and population in the daytime is higher than morning and night, e.g., Tyuo and Bunkyo wards.

See Fig. 5 and 6 for heat-maps of population at 8:00 and 14:00. We find the dense area is at the center of Tokyo at 14:00, where is not crowded in the morning (at 8:00). This is a typical people's behavior in metropolitan city and our target to estimate from randomized location traces.

Fig. 4 illustrates how location traces are processed in our scheme, where n users move independently and belong one of 23 special wards. Their current memberships to one of 23 wards are encoded as 23-dimension vector $\boldsymbol{v}$. They perturb their location in RAPPOR $Q(\boldsymbol{v})$ before sending to a service provider. The service provider simply sums all perturbed locations and publishes timely a distribution of population Z_i for i-th ward. With either MLE or EM algorithms, we estimate the true distribution of population N_i.

Table 1. Time-series population Nightray in Tokyo

time	population
8:00	2,957
11:00	3,922
14:00	4,640
17:00	4,793
20:00	4,300
23:00	3,283

Table 2. Specification of Nightley dataset

Surveyed value	POI, Timestamp (SNS post), The road network.
Estimate value	Movement courses, A place of residence, Work location, Stay time, Gender, What he do during a stay (including shopping and the leisure)
Area	Tokyo Metropolitan Area
Target time	July, 2013, October, December
A time unit	As for every five minutes
Geodetic datum	WGS84 (EPSG:4326)
Records	It is approximately 70,000 cases with each csv file
File size	Approximately 100MB

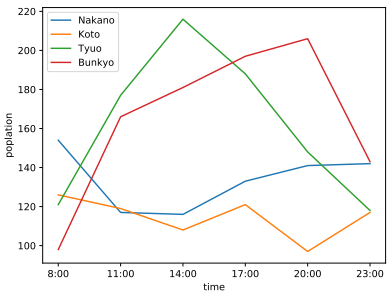

Fig. 3. Time-serise Population in Tokyo

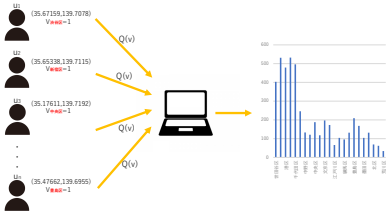

Fig. 4. System flow

4.3 Method

We apply RAPPOR algorithm to the Nightlay data for several privacy budgets ϵ.

最尤推定法と EM アルゴリズムを用いて RAPPOR アルゴリズムで収集したデータから推定した人口と実際の人口との誤差を ϵ の値を変化させて求めた. 誤差を平均絶対誤差 (MAE) として, 次のようにして求める.

1. Each user perturbs his/her current location $\boldsymbol{v}$ in RAPPOR for privacy budgets $\epsilon = 0.5, 1, 1.5, \ldots, 5.0$ to have $Q(\boldsymbol{v})$.
2. Service provider collects users' location data for every time and publishes the distribution of populations for 23 wards in Tokyo $Z_1, \ldots, Z_m$.
3. We estimate populations in MLE and the proposed (EM) algorithms, denote them by N_{MLE} and N_{EM}, respectively.
4. Evaluate accuracy for two estimates in mean absolute error (MAE), that is, $\sum_{i=1\ldots,m} |N_i - N_{MLE}|$, where N_i is true population of i-th ward.
5. Repeat the above steps for ten times.

Table 3. Time-series population in major wards in Tokyo

time	8:00	11:00	14:00	17:00	20:00	23:00
渋谷区	262	394	533	532	479	351
新宿区	278	414	505	531	454	304
港区	267	393	509	479	416	284
千代田区	186	381	506	496	476	248
世田谷区	295	331	367	403	368	317
杉並区	165	209	227	246	187	188
中央区	121	177	216	188	148	118
文京区	98	166	181	197	206	143
品川区	98	147	182	173	147	99
中野区	154	117	116	133	141	142

Table 4. Statistics of time-series population for 23 wards in Tokyo

time	8:00	11:00	14:00	17:00	20:00	23:00
mean	192.4	272.9	334.2	337.8	302.2	219.4
max	295	414	533	532	479	351
min	98	117	116	133	141	99

Fig. 5. heat-map of Tokyo at 8:00

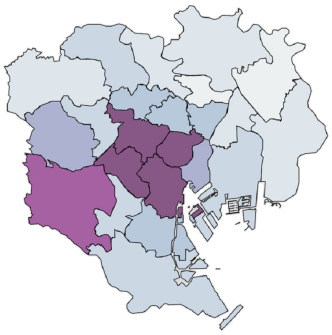

Fig. 6. heat-map of Tokyo at 14:00

4.4 Result

Fig. 7 shows the distributions of estimated populations for m special wards in Tokyo at 14:00. We perturb location data in RAPPOR with privacy budget $\epsilon = 0.5$ (very safety). We find that the proposed estimates (labeled as EM, colored in orange) are close to the true population (blue) for almost all wards. While, the ML estimates sometime are suffered with significant error, e.g., -200 and -400 at 12-th, 14-th, 23-rd, and 18-th wards.

The accuracy depends on wards and privacy budgets. So, we evaluate MAE and depict the MAEs at time 8:00, 11:00, 14:00, 17:00, 20:00 and 23:00 in Fig. 8, ..., 13. In this plot, we show MAE for both estimate algorithms with respect to privacy budgets $\epsilon = 0.5, 1, 1.5, \ldots, 5.0$. The results show that the proposed algorithm performs better than the MLE used in [1] for every time. The difference between two estimates maximizes as privacy budget decreases (randomized greatly and privacy is higher). We also note that the improvement of accuracy is higher at 8:00 than that at 14:00. This error of MLE is caused because the many people are in home in the morning, as shown in Fig. 3 and Fig. 5, and the unbalanced distributions of populations spoils the maximum likelihood. We

observe low accuracy for estimating population of small wards in Fig. 7. This must be happen with the same reason.

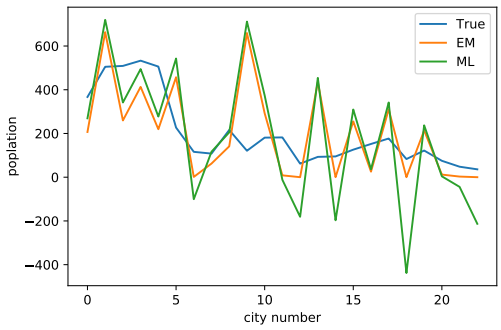

Fig. 7. Estimate populations for m wards in Tokyo at 14:00, $\epsilon = 0.5$

5 Conclusion

We have studied the privacy preservation of time-series location trace using LDP algorithm RAPPOR and proposed the expected maximization for estimate of true distributions. The proposed algorithm iteratively improves estimated posterior probabilities based on Bays' theorem until the difference converged for all elements. Our experiment using 6,528 individuals' location trace in Tokyo provided from Nightley Inc. demonstrates that the proposed algorithm performs better than the original MLE used in RAPPOR for every special ward in Tokyo in one day. We found that the accuracy is improved as privacy budget ϵ is smaller, and as many population is provided. We conclude that the iterative approach works well for data perturbed in LDP algorithm.

References

1. Úlfar Erlingsson,Vasyl Pihur,Aleksandra Korolova,"RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response",ACM Conference on Computer and Communications Security, 2014, pp.1054-1067.
2. "Learning with Privacy at Scale"(https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html, accessed in 2019)
3. Anna Mochizuki, Hiroaki Kikuchi,''Privacy-preserving Collaborative Filtering Using Randomized Response ' , ' Journal of information Processing Vol.21 No.4 617-623.

4. Nightlay Inc., "SNS-based People Flow Data is now available for free download" https://nightley.jp/archives/1954/ (accessed in Octorber 2019).
5. M. Miyagawa, "EM algorithm and marginal applications", Advanced Statistics, Vol. 16, No. 1, pp. 1-19 (in Japanese).
6. H. Ono, K. Fukuchi, and J. Sakuma, "Detection of Heavy hitters in restriction of local differential privacy", DEIM Forum 2018 E1-3 (in Japanese).
7. Zhan Qin,Yin Yang,Ting Yu, "Heavy Hitter Estimation over Set-Valued Data with Local Differential Privacy"
8. Xuebin Ren, Chia-Mu Yu, Weiren Yu, Shusen Yang, Xinyu Yang, Julie A. McCann, and Philip S. Yu," High-Dimensional Crowdsourced Data Publication with Local Differential Privacy"
9. Hiroaki Kikuchi, jin Akiyama, Howard Gobioff, "Stochastic Voting Protocol To Protect Voters Privacy"
10. R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining", ACM SIGMOD 2000, pp. 439-450, 2000.
11. H. Polat and W. Du, "Privacy-Preserving Collaborative Filtering using Randomized Perturbation Techniques", ICDM 2003, pp. 1-15, 2003.
12. Z. Huang, W. Du and B. Chen, "Deriving Private Information from Randomized Data", ACM SIGMOD 2005, pp. 37-48, 2005.
13. S. Zhang, J. Ford, and F. Makedon, "Deriving private information from randomly perturbed ratings, SIAM-Data Mining Conference, 2006.
14. S. Zhang, J. ford, F. Makedon, "A Privacy-preserving Collaborative Filtering Scheme with Two-way Communication", ACM EC'06, pp. 316-323, 2006.

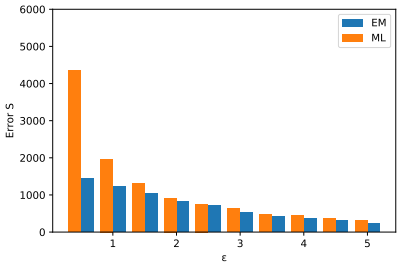

Fig. 8. MLE at 8:00 for privacy budgets ϵ

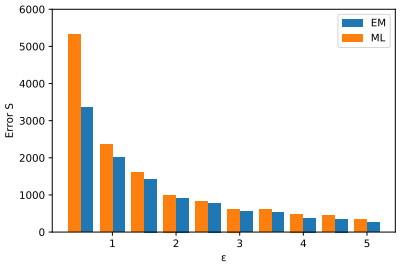

Fig. 9. MLE at 11:00 for privacy budgets ϵ

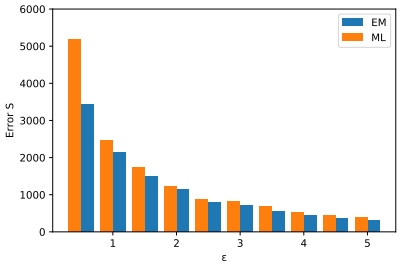

Fig. 10. MLE at 14:00 for privacy budgets ϵ

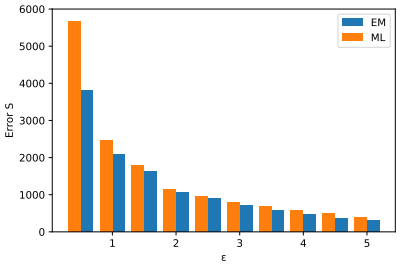

Fig. 11. MLE at 17:00 for privacy budgets ϵ

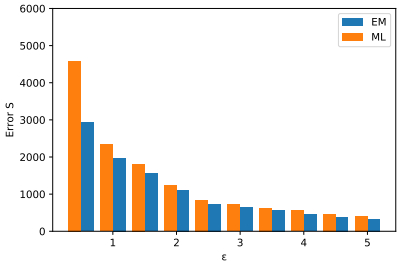

Fig. 12. MLE at 20:00 for privacy budgets ϵ

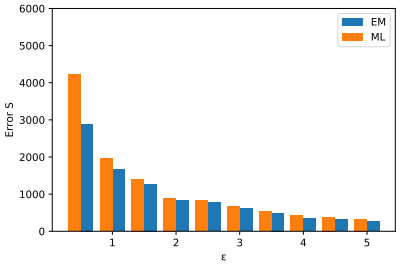

Fig. 13. MLE at 23:00 for privacy budgets ϵ