SCIENCE CHINA
Information Sciences

December 2020, Vol. 63 220303:1–220303:12
https://doi.org/10.1007/s11432-019-2834-x

• RESEARCH PAPER •
Special Focus on Challenges and New Insights for Network Security in 5G-and-Beyond

Generative adversarial networks enhanced location privacy in 5G networks

Youyang QU[1], Jingwen ZHANG[2], Ruidong LI[3], Xiaoning ZHANG[4], Xuemeng ZHAI[4] & Shui YU[5,2*]

[1]*School of Information Technology, Deakin University, Burwood* VIC 3125, *Australia;*
[2]*School of Computer Science, University of Technology Sydney, Ultimo* NSW 2007, *Australia;*
[3]*National Institute of Information and Communications Technology (NICT), Tokyo* 1840015, *Japan;*
[4]*School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu* 610051, *China;*
[5]*Center of AI and Big Data, Southeast Digital Economic Development Institute, Quzhou* 324000, *China*

Received 21 October 2019/Accepted 12 March 2020/Published online 4 November 2020

Abstract 5G networks, as the up-to-date communication platforms, are experiencing fast booming. Meanwhile, increasing volumes of sensitive data, especially location information, are being generated and shared using 5G networks for various purposes ceaselessly. Location and trajectory information in the published data has always been and will keep courting risks and attacks by malicious adversaries. Therefore, there are still privacy leakage threats by simply sharing the original data, especially data with location information, due to the short cover range of 5G signal tower. To better address these issues, we proposed a generative adversarial networks (GAN) enhanced location privacy protection model to cloak the location and even trajectory information. We use posterior sampling to generate a subset of data, which is proved complying with differential privacy requirements from the end device side. After that, a data augmentation algorithm modified from classic GAN is devised to generate a series of privacy-preserving full-sized synthetic data from the central server side. With the synthetic data generated from a real-world dataset, we demonstrate the superiority of the proposed model in terms of location privacy protection, data utility, and prediction accuracy.

Keywords 5G, privacy preservation, generative adversarial nets, differential privacy

Citation Qu Y Y, Zhang J W, Li R D, et al. Generative adversarial networks enhanced location privacy in 5G networks. Sci China Inf Sci, 2020, 63(12): 220303, https://doi.org/10.1007/s11432-019-2834-x

1 Introduction

Fast popularization of 5G networks and intelligent devices reshape this big data era. Rapidly increasing volume of data, including the sensitive data, is generated and shared in real-time among 5G networks [1]. According to [2], there is over 2.5 quintillion bytes of generated data in every single day, and it is only going to grow from there. With the help of the released data, significant convenience is provided from various perspectives of daily life [3].

However, the sharing of data also raises great concerns on location privacy leakage issues in 5G networks, which is because of the potential commercial value of the data [4]. Adversaries launch continuous attacks to steal sensitive information with all available techniques, in particular, machine learning [5]. Most current studies in this domain publish the real-time data stream with some specific privacy protection mechanisms such as differential privacy [6,7] and location cloaking [8]. These are inclusive risks due to the correlation among the shared data and the injected noise [9]. Since the range of a signal tower is

* Corresponding author (email: Shui.Yu@uts.edu.au)

significantly reduced, the location privacy is put under greater threats if the adjacent signal towers are compromised at the same time [10].

In addition to privacy preservation, data utility is another major concern because over-protection will lead to poor data utility [3]. Inaccurate prediction or optimization of data will eliminate the incentive of big data analysis and degrade the quality of service, which is unacceptable to the public and the 5G service providers [9, 11]. Furthermore, the accuracy of data has a great impact on the services provided by 5G networks [12]. This poses further challenges to the trade-off between privacy protection and data utility while taking the efficiency into consideration.

To address the aforementioned issues and derive an optimized trade-off, we propose a generative adversarial networks driven location privacy protection model. A subset of differentially private data is generated by posterior sampling, to limit the size of the data to be transmitted. The sampled subset also complies with differential privacy. Then, we use a modified generative adversarial networks (GAN) to implement data argumentation. By generating the new synthetic data, the location privacy is preserved while the data utility and prediction accuracy are enhanced, which is testified by extensive evaluation results on a real-world dataset.

The contributions of this work are summarized as follows.

• We devise a generative adversarial networks enhanced location privacy protection model in 5G networks. The proposed model only requires a subset of the raw data, which significantly prevents sensitive information from privacy leakage.

• We use a modified generative adversarial networks algorithm to achieve data argumentation. The argumentation of data restores a series of full-sized data. The newly generated synthetic dataset complies with differential privacy while enabling accurate prediction results.

• The efficiency could be improved due to a generated subset is transmitted between nodes in 5G networks. The experimental results show the superiority of the proposed model form the perspectives of privacy protection, data utility, and prediction accuracy.

The remainder of this work is organized as follows. In Section 2, we list the related work regarding location privacy in 5G, differential private machine learning, and GAN. We then describe the framework of the generative adversarial network enhanced location privacy protection model in Section 3. In Section 4, we present the convergence analysis, which is followed by the evaluation results on a real-world data set in Section 5. At last, we summarize and conclude this paper in Section 6 .

2 Related work

In this section, we present the related work regarding privacy-preserving data sharing, location privacy in 5G networks, and differentially private machine learning.

2.1 Privacy-preserving data sharing (PPDS)

In the PPDS domain, there are two mainstream branches, which are clustering-based methods and differential privacy with its extensions [13]. The clustering-based methods starts from K-anonymity, in which the data sets are grouped into several clusters consisting of at least k pieces of records. In this way, the probability of being re-identified is $1/k$ regarding each piece of the data [14]. Based on K-anonymity, an extension called l-diversity is proposed to guarantee at least l different diversities of data is involved in each cluster [15]. To further improve, T-closeness is devised, which ensures the distribution of every cluster and the distribution of the whole data set are almost the same [16]. A threshold could be defined to control the similarity. Clustering-based methods are proved practical to deal with micro-data privacy protection, but lack of scalability and insufficient theoretical foundation limits its development in this big data era [17].

The arising of differential privacy is regarded as a landmark in the privacy preservation domain. The primary advantage is to provide solid mathematical foundation to privacy preservation [6, 7]. This theory could guarantee that the adversaries will fail to re-identify any specific piece of record and the

performance is controlled by an index called privacy budget ϵ. The feasibility of differential privacy and its extensions is proved in various real-world scenarios, for example, crowdsourcing [18], location privacy in social networks [19], and machine learning [20].

2.2 Location privacy in 5G networks

Privacy issues have been widely concerned in recent few years [21]. Duan et al. [22] devised an authentication method to provide privacy protection in 5G networks based on SDN. In [23], the authors proposed a novel scheme to achieve security-aware and privacy-preserving device-to-device communications in 5G networks. Ni et al. [24] proposed a service-oriented authentication model to ensure privacy and efficiency in 5G-enabled Internet of Things. Liao et al. [25] leveraged the dynamic group division algorithm to protection both location and trajectory privacy. However, this method suffers from efficiency and background knowledge attacks. Qu et al. [26,27] devised efficient and privacy-preserving content sharing model based on blockchain. This mode can guarantee data integrity and non-tempering but fails to provide privacy protection to malicious insiders. In addition, 5G networks suffer from DDoS attacks [28, 29], which will be further discussed in our future work.

2.3 Differentially private machine learning

With the increasing demand of privacy protection in machine learning, the research direction of differentially private machine learning comes into being. There exists a fundamental difference between the proposed model and this topic. The proposed model focuses on improving the privacy preservation by means of machine learning while this topic discusses how to add privacy-preserving features into machine learning process. The key component in our model, GAN is firstly proposed by Goodfellow et al. [30]. They established a game between two perceptrons, which are generator and discriminator, to derive the Nash equilibrium. Based on this, Ajovisky et al. [31] devised Wasserstein generative adversarial networks (WGAN) by replacing KL divergence with Wasserstein-1 distance. Li et al. [32] proposed a new method, which is triple GAN, by adding an extra identifier to improve the performance.

In this field, one of the pioneering work is conducted in [33]. It discusses how to achieve differentially private updates during a stochastic gradient descent process to adding extra privacy protection. Moreover, Lee and Kifer [34] devised a novel model which enables differentially private stochastic gradient descent which adds adaptive privacy budget in each iteration process. Qu et al. [35,36] proposed a GAN-enabled privacy to achieve optimized trade off. Dwork et al. [37] presented some insights on how to recycle the holdout data set through using differentially private learning on holdout data set. Shokri et al. [38] and Abadi et al. [39] presented the idea of privacy-preserving deep learning in computer and communication systems (CCS), respectively. Built upon the previous studies, Ács et al. [40] proposed a differentially private mixture of generative neural networks. This model requires a mixture of existing generative models to achieve differentially private learning.

3 System modeling

In this section, we show the details of the proposed model, which is referred to as the generative adversarial nets enhanced location privacy protection model in 5G networks. We firstly use posterior sampling to sample the differentially private raw data and obtain differentially private a subset. Secondly, we use an improved GAN to generate a full-size synthetic dataset by data augmentation by using the subset. At last, the newly-generated synthetic location dataset is used to conduct prediction or other analysis instead of the raw data or even differentially private raw data.

3.1 5G heterogeneous network modelling

To model the 5G networks, we consider heterogeneous structures to make it more feasible in real-world scenarios. The main idea is to decouple downlink and uplink, in which the downlink received power

decides the downlink cell association while the path-loss is leveraged to deal with the uplink. This is built upon the basic structure presented in [41].

In this work, the traditional way of uplink and downlink cell association is extended without simply using downlink received power. The assumption here is that the downlink received power is remained as the base of the downlink association, while path-loss is introduced into the model to decide the uplink association. One potential issue exists in this approach. If a piece of user equipment has an uplink or downlink to a node, there is a requirement of a mechanism to execute the channel estimation, ACK (acknowledgement) process, and so on. This may result in major design changes, which is not practical. Therefore, the target is to study whether if the decoupling of the downlink and uplink (DDU) can justify these various major changes.

The DDU will bring about various cell boundaries in a HetNet's uplink and downlink. In this case, a piece of user equipment between the cell boundaries in this region will be associated to the corresponding Mcell and Scell in the downlink and uplink. The gains of the uplink provide the incentive to employ this model. At the same time, the capacities of the downlink are not impacted because the association keeps fixed. In Subsection 3.2, we will move on to the privacy protection model based on the DDU structure.

3.2 Differential privacy with posterior sampling

We use Laplacian mechanism achieve ϵ-differentially private processing through injecting controllable noise to numeral data.

The random mechanism $\{\mathcal{M} : R^n \to \Delta(R^n)\}$ adds noises $\mathcal{N}$ complying with a Laplacian distribution as follows:

$$\begin{aligned} &\mathcal{M}(\mathcal{D}) = \mathcal{D} + \mathcal{N}, \\ &\text{s.t. } N \sim \text{Lap}(b) \sim d\Pr[N = n] = \exp\left(-\frac{\|n\|_2}{b}\right), \end{aligned} \tag{1}$$

where N is the injected noise complying with Laplace mechanism and $d\Pr[N = n]$ is the density of Lap(b). Based on this, we regard $\mathcal{M}$ as a ϵ-differentially private mechanism to two adjacent datasets.

Based on Laplacian mechanism, we reformulate differential privacy as following.

Let ϵ be a small positive value as the privacy budget constrained by the protection level. We also define D_1 and D_2 as two datasets with an adjacent relationship, and Lap as a randomized algorithm that sanitizes the two datasets. The ϵ-differential privacy algorithm on D_1 and D_2 is

$$\Pr[\text{Lap}(D) \in \Omega] = \exp(\epsilon) \times \Pr[\text{Lap}(D') \in \Omega], \tag{2}$$

where the probability space Ω is taken over the randomness used by $\mathcal{M}$.

In order to provide further privacy protection, we use a posterior sampling method to generate a subset of the raw data. The generated subset is intrinsically differentially private as proved in [42]. The rationale behind the usage of posterior sampling is that we hope to improve the efficiency while maintaining privacy protection. By using posterior sampling, we can guarantee the data still complies with differential privacy. Then, the sampled data is transmitted to other parties. Since we are going to use a GAN model for the recipients, they are able to restore the data by data augmentation while the the privacy is still preserved.

Theorem 1. If $\sup_{x\in\text{chi},\theta\in\Theta}|\log p(x|\theta)| \leqslant B$, the published one sample can preserves $4B$-differential privacy generated from the posterior distribution $p(\theta|X^n)$. In addition, if χ is a constrained domain while $\log p(x|\theta)$ is an L-Lipschitz function in $\|\cdot\|_*$ for any $\theta \in \Theta$, we can conclude that the generated one sample preserves $4RL$-differential privacy.

We have the observation that this is a specific instance of the exponential mechanism of differential privacy, which is shown in Algorithm 1. It is a general mechanism that provide privacy protection while giving exponentially more consistent outputs with higher data utility. If the utility function is set to be a log-likelihood one and the privacy budget is initialized as $4B$, we can derive the differentially private posterior sample as expected. From the perspective of exponential mechanism, it can be easily extended to where we can specify ϵ by scaling the log-likelihood utility function. The advantage of this algorithm

is that zero deployment cost is required to extend the proposed model to all posterior sampling-based learning models with any specified ϵ-differentially private protection.

Algorithm 1 Differentially private posterior sample estimator

Input: Dataset D, privacy budget ϵ, log-likelihood function $f(\cdot|\cdot)$ satisfying $\sup_{x,\theta}\|f(x|\theta\| \leqslant B)$, a prior $\pi(\cdot)$.
Output: $\theta' \sim P(\theta|X) \propto (\exp)(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{f}'(\theta|\mathrm{x_i}))\pi'(\theta)$.
1: Set $\rho = \min 1, \frac{\epsilon}{4B}$;
2: Formulate log-likelihood function and the prior $f'(\cdot|\cdot) = \rho f(\cdot|\cdot)$ and $\pi' = (\pi(\cdot))^{\rho}$;
3: **Return** θ'.

To evaluate the privacy protection level of the model, we use ϵ as the measurement index. ϵ's value range is all nonnegative real number, namely, $[0, +\infty]$. If the value of ϵ decreases, the privacy protection level upgrades with it, and vice visa.

From the aspect of data utility, the root mean square error (RMSE) is one of the most effective and popular measuring matrix. Through the qualitative analysis, it can be conclude that the data utility and privacy protection and negatively correlated. Since data utility degrades with the upgrades of privacy protection level, it upgrades with the increase of ϵ's value. The calculation of RMSE is normally denoted by $\mathrm{RMSE} = \sqrt{\sum |\hat{y} - y|^2} = \sqrt{\sum |G(z, \theta_g)|^2}$. Moreover, for different weighted models are ultilized to conduct prediction so that the data utility can be further evaluated in term of the feasibility of the proposed model.

3.3 Generative adversarial networks

For purpose of pursuing higher performance, we introduce WGAN net into the proposed model. To improve the training stability, WGAN uses another more optimized distribution measurement to replace the traditional method which is Kullback-Leibler divergence.

We use $\{x_i|i \in I\}$ to denote the real data examples and $\{P_\theta|\theta \in R_d\}$ to represent a group of densities. The target is to solve a maximization problem as $\max_{\theta \in R_d} \frac{1}{m} \sum_i^m \log P_\theta(x_i)$. If real data distribution $\mathbb{P}_r(x_i)$ admits a density while $\mathbb{P}_\theta(x_i)$ denotes the distribution of $P_\theta(x_i)$, we target on minimizing the KL-divergence in the form of $\mathrm{KL}(\mathbb{P}_r\|(P)_\theta)$.

In this case, pre-modelling of the density P_θ is essential to make KL-divergence work. Nevertheless, the density P_θ is normally not pre-defined when dealing with distributions with low dimensions. In a specific situation where two distributions are P_θ and P_r, they do not share an intersection with a relatively high probability. KL-divergence may fail to function well. Thus, Wasserstein-1 distance is proposed to address the aforementioned issues. The Wasserstein-1 distance is formulated as

$$W(\mathbb{P}_\theta(x_i), \mathbb{P}_r(x_i)) = \inf_{\lambda \in \prod(\mathbb{P}_\theta(x_i), \mathbb{P}_r(x_i))} \mathbb{E}_{(x,y)\sim\lambda}[\|x - y\|], \tag{3}$$

where $\prod(\mathbb{P}_\theta(x_i), \mathbb{P}_r(x_i))$ is the set of all joint distributions $\lambda(x, y)$. The marginals of the joint distributions are $\mathbb{P}_\theta(x_i)$ and $\mathbb{P}_r(x_i)$, respectively. It can be concluded that $\lambda(x, y)$ shows how much is necessary to be transported from x to y so as to transform the distributions $\mathbb{P}_\theta(x_i)$ into the distribution $\mathbb{P}_r(x_i)$.

3.4 GAN-based data augmentation

Our synthetic dataset generation algorithm is modified from AugGAN [43]. The generation procedures include segmentation subtask, weight hybrid sharing for multi-task network, cycle consistency, and generative adversarial learning, which are shown in Algorithm 2.

Let D_r and D_n be two data spaces, $\hat{D}_r$ and $\hat{D}_n$ be the corresponding segmentation masks, and F denote the encoded feature space. There are two encoders ($\mathrm{EN}_r : D_r \to F$ and $\mathrm{EN}_n : D_N \to F$), two generators ($G_r : F \to \bar{D}_n$ and $G_n : F \to \bar{D}_r$), two segmentation generators ($\mathrm{SG}_r : F \to \hat{D}_n$ and $\mathrm{SG}_n : F \to \hat{D}_r$), two discriminators ($\mathrm{DIS}_r$ and DIS_n) for two data spaces, respectively. We use se to be short for cross-entropy in following sections.

Algorithm 2 Differentially private data augmentation

Input: Sampled dataset D_r, size of the original dataset.
Output: Full-sized differentially private synthetic dataset.
1: Establish encoder EN and segmentation loss $\mathfrak{L}_{\text{seg}}$;
2: Establish hybrid weight sharing loss $\mathfrak{L}_{w_{D_r}}$ and $\mathfrak{L}_{w_{D_n}}$;
3: Establish circle consistence loss $\mathfrak{L}_{\text{cc}}$;
4: Establish adversarial learning loss $\mathfrak{L}_{\text{GAN}_1}$ and $\mathfrak{L}_{\text{GAN}_2}$;
5: Integrate all loss to the final objective function $\mathfrak{L}$;
6: Return the full-sized differentially private synthetic dataset.

We use the encoder networks to extract the features via segmentation subtasks so that the intricate fine-grained semantic feature are contained. These features are valuable in terms of data utility and accuracy of prediction. We formulate the segmentation loss regarding D_r and D_n as follows:

$$\begin{aligned}\mathfrak{L}_{\text{seg}}^{D_r}(\text{SG}_{D_r}, \text{EN}_{D_r}, D_r, \hat{D}_r) = {} & \lambda_{\text{seg}}^{L1}\mathbb{E}_{x\sim p_{\text{data}}D_r}[\|P_{D_r}(E_{D_r}(D_r)) - \hat{D}_r\|_1] \\ & + \lambda_{\text{seg}}^{\text{ce}}\mathbb{E}_{x\sim p_{\text{data}}D_r}[\|\log(P_{D_r}(E_{D_r}(D_r)) - \hat{D}_r)\|_1],\end{aligned} \tag{4}$$

$$\begin{aligned}\mathfrak{L}_{\text{seg}}^{D_n}(\text{SG}_{D_n}, \text{EN}_{D_n}, D_r, \hat{D}_n) = {} & \lambda_{\text{seg}}^{L1}\mathbb{E}_{x\sim p_{\text{data}}D_n}[\|P_{D_n}(E_{D_n}(D_n)) - \hat{D}_n\|_1] \\ & + \lambda_{\text{seg}}^{\text{ce}}\mathbb{E}_{x\sim p_{\text{data}}D_n}[\|\log(P_{D_n}(E_{D_n}(D_n)) - \hat{D}_n)\|_1].\end{aligned} \tag{5}$$

Sharing the weights between generators and the parsing network enables the generators to fully ultilize the features. We use a hybrid weights sharing method in which 6 residual blocks are hard shared while 2 deconvolution blocks are soft shared. This sharing method is proved superior experimentally. To calculate the weight difference, the deconvolution layers of the two networks are used. Thus, we model the difference as a loss function in which the target of the mean square error is a zero matrix. Based on this, we formulate the weight sharing loss as

$$\mathfrak{L}_w(w_G, w_{\text{SG}}) = -\log\left(\frac{w_G^{D_r}\cdot w_{\text{SG}}^{D_r}}{\|w_G^{D_r}\|_2\|w_{\text{SG}}^{D_r}\|_2}\right)^2, \tag{6}$$

where w_G and w_{SG} denote the corresponding weight vectors decided by the deconvolution layers of both the generators and parsing networks.

In addition to weight sharing, cycle consistency is also considered fairly feasible for purpose of preventing GAN from generating random dataset in the target data space. Therefore, we also take cycle consistency in this model to further formulate this unsupervised synthetic dataset generation problem. The loss function of cycle consistency is

$$\begin{aligned}\mathfrak{L}_{\text{cc}}(\text{EN}_{D_r}, G_{D_r}, \text{EN}_{D_n}, G_{D_n}, D_r, D_n) = {} & \mathbb{E}_{D_r\sim p_{\text{data}}(D_r)}[\|G_{D_n}(\text{EN}_{D_n}(G_{D_r}(\text{EN}_{D_r}))) - D_r\|_1] \\ & + \mathbb{E}_{D_n\sim p_{\text{data}}(D_n)}[\|G_{D_r}(\text{EN}_{D_r}(G_{D_n}(\text{EN}_{D_n}))) - D_n\|_1].\end{aligned} \tag{7}$$

In this model, there are two generative adversarial networks functioning simultaneously, which are $\text{GAN}_1 : \text{EN}_{D_r}, G_{D_r}, \text{DIS}_{D_r}$ and $\text{GAN}_2 : \text{EN}_{D_n}, G_{D_n}, \text{DIS}_{D_n}$. Therefore, the adversarial losses function can be formulated as follows:

$$\begin{aligned}\mathfrak{L}_{\text{GAN}_1}(\text{EN}_{D_r}, G_{D_r}, \text{DIS}_{D_r}, D_r, D_n) = {} & \mathbb{E}_{D_n\sim p_{\text{data}}(D_n)}[\log\text{DIS}_{D_r}(D_n)] \\ & + \mathbb{E}_{D_r\sim p_{\text{data}}(D_r)}[\log(1 - \text{DIS}_{D_r}(G_{D_r}(\text{EN}_{D_r}(D_r))))],\end{aligned} \tag{8}$$

$$\begin{aligned}\mathfrak{L}_{\text{GAN}_2}(\text{EN}_{D_n}, G_{D_n}, \text{DIS}_{D_n}, D_n, D_r) = {} & \mathbb{E}_{D_r\sim p_{\text{data}}(D_r)}[\log\text{DIS}_{D_n}(D_r)] \\ & + \mathbb{E}_{D_n\sim p_{\text{data}}(D_n)}[\log(1 - \text{DIS}_{D_n}(G_{D_n}(\text{EN}_{D_n}(D_n))))].\end{aligned} \tag{9}$$

3.5 Privacy-preserving synthetic dataset generation

To generate the privacy-preserving synthetic dataset, we jointly address the learning problems for the data-translation streams EN_1, G_1 and EN_2, G_2, the data-parsing streams $\mathrm{EN}_1, \mathrm{SG}_1$ and $\mathrm{EN}_2, \mathrm{SG}_2$, and two generative adversarial networks GAN_1 and GAN_2. The global objective function after integration is as follows:

$$\begin{aligned}\mathfrak{L} = &\mathfrak{L}_{\mathrm{GAN}_1}(\mathrm{EN}_{D_r}, G_{D_r}, \mathrm{DIS}_{D_r}, D_r, D_n) + \mathfrak{L}_{\mathrm{GAN}_2}(\cdot)\\ &+ \lambda_{\mathrm{cc}} \times \mathfrak{L}_{\mathrm{cc}}(\mathrm{EN}_{D_r}, G_{D_r}, \mathrm{EN}_{D_n}, G_{D_n}, D_r, D_n)\\ &+ \lambda_{\mathrm{seg}} \times (\mathfrak{L}_{\mathrm{seg}}(\mathrm{SG}_{D_r}, \mathrm{EN}_{D_r}, D_r, \hat{D}_r) + \mathfrak{L}_{\mathrm{seg}}(\cdot))\\ &+ \lambda_w \times (\mathfrak{L}_{w_{G_r}}(w_{G_{G_r}}, w_{\mathrm{SG}_{G_r}}) + \mathfrak{L}_{w_{G_n}}(w_{G_{G_n}}, w_{\mathrm{SG}_{G_n}})).\end{aligned} \tag{10}$$

4 System analysis

We demonstrate the system analysis in this section, in particular, the global optimality and convergence. Under the constraint of $z \sim p_z$, p_g is trained as the synthetic probability distribution of the generator $G(z, l)$. If we hold the assumption that the computing resources is unlimited, the joint objective of the generator $G(z, l)$ and the structure $\mathrm{DIS}(x, l)$ is to reach the convergence of p_g to an accuracy approximation of p_d. The game between them will result in a Nash Equilibrium where $p_g = \Pr(\mathrm{GAN}(p_x))$.

Regarding the proposed model, we start with discussing the global optimality where $p_g = \Pr(\mathrm{GAN}(p_x))$. The right side of the equation is regarded as the gaming output derived from discriminator and differential privacy identifier. Then, we take the discriminator and encoder structure $S(\mathrm{DIS}; \mathrm{EN})$ into consideration in regards to any $G(z, l)$. In term of a certain generator G, the optimal $S^*(\mathrm{DIS}; \mathrm{EN})$ is formulated as

$$S^*(\mathrm{DIS}; \mathrm{EN}) = \frac{\Pr(\mathrm{GAN}(p_x))}{p_g + \Pr(\mathrm{GAN}(p_x))}. \tag{11}$$

Let the objective of training DIS be the maximized conditional probability $P(Y = y|x)$, which is the logarithmic function for estimating. The Y in the formulas shows if the variable x meets all the constraints of the game structure. Y describes whether variable x satisfies the constraints of gaming structure $S^*(\mathrm{DIS}; \mathrm{EN})$. In the case of $y = 1$, it happens only when x comes from the distribution p_x while meeting constraints of differential privacy. In the contrary, $y = 0$ will be obtained if only one conditions is satisfied. Based on the above analysis, the proposed Min-Max problem can be reshaped as

$$\begin{aligned}C(G) &= \max V(G; \mathrm{DIS}; \mathrm{EN})\\ &= \mathbb{E}_{x\sim p_d}[\log S^*(x|\mathrm{DIS}; \mathrm{EN})] + \mathbb{E}_{z\sim p_z}[\log(1 - S^*((G(z_i))|\mathrm{DIS}; \mathrm{EN}))]\\ &= \mathbb{E}_{x\sim p_d}[\log S^*(x|\mathrm{DIS}; \mathrm{EN})] + \mathbb{E}_{z\sim p_g}[\log(1 - S^*(x|\mathrm{DIS}; \mathrm{EN}))]\\ &= \mathbb{E}_{x\sim p_d}\left[\log\frac{\Pr(\mathrm{GAN}(p_x))}{p_g(x) + \Pr(\mathrm{GAN}(p_x))}\right] + \mathbb{E}_{x\sim p_g}\left[\log\frac{p_g(x)}{p_g(x) + \Pr(\mathrm{GAN}(p_x))}\right].\end{aligned} \tag{12}$$

Through Eq. (12), we know that the global minimum of the Min-Max objective only exists when $p_g = p_d$ is satisfied. Thus, we will have $\mathbb{E}_{x\sim p_d}[-\log 2] + \mathbb{E}_{x\sim p_g}[-\log 2] = -\log 4$. By subtracting the results back to $C(G) = V(\mathrm{DIS}^*G, G, \mathrm{EN})$, the function can be extended to $C(G) = -\log(4) + \frac{1}{2}[\mathrm{KL}(\Pr(\mathrm{GAN}(p_x))\|(\Pr(\mathrm{GAN}(p_x)) + p_g))/2 + \mathrm{KL}(p_g\|(\Pr(\mathrm{GAN}(p_x)) + p_g))/2]$. The symbol KL here denotes the KL-divergence of information theory, which is used to measure the distance of two distributions. As the value of $[\mathrm{KL}(\Pr(\mathrm{GAN}(p_x))\|(\Pr(\mathrm{GAN}(p_x)) + p_g))/2 + \mathrm{KL}(p_g\|(\Pr(\mathrm{GAN}(p_x)) + p_g))/2]$ is a non-negative real number, it can be leveraged to represent the distance of two distributions while equaling to zero when $p_g = \Pr(\mathrm{GAN}(p_x))$ is met. Thus, the global minimum in this case is derived as $C^* = -\log 4$ iff $p_g = \Pr(\mathrm{GAN}(p_x))$.

Table 1 Preliminary of the real-wolrd data set

Property	*s*-longitude1	*s*-latitude1	*s*-longitude2	*s*-latitude2
Count	150.00	150.00	150.00	150.00
Mean	5.84	3.05	3.76	1.20
Std	0.83	0.43	1.76	0.76
Min	4.30	2.00	1.00	0.10
25%	5.10	2.80	1.60	0.30
50%	5.80	3.00	4.35	1.30
75%	6.40	3.30	5.10	1.80
Max	7.90	4.40	6.90	2.50

5 Performance evaluation

We demonstrate our evaluation results on a real-world dataset. First, we show the setting up of experiments and the preliminaries of data sets. Second, we illustrate the evaluation results from the perspectives of privacy protection, data utility (prediction accuracy), as well as efficiency. With the simulation on real-world data set, the evaluation results confirm the performance superiority of the proposed location privacy protection model.

The algorithms are implemented by Python while we plot the figures with Matlab R2015b. The simulations codes are executed on a Mac laptop outfitting a Intel Core I5 2.7 GHz CPU and a 8 G RAM. We use a real-world data set which is from UCI machine learning repository [44]. The preliminaries of this real-world data set is shown in Table 1 while some statistic features such as the correlations and distributions are demonstrated in Figure 1(a).

We compare the proposed model with classic GAN, Laplace mechanism [45], and dummy-based methods in the following subsections. GAN generates the synthetic data by the game between two perceptrons which are generator and discriminator. Laplace mechanism will add controllable noise to the raw data, in which the noise complies with Laplace distribution. The dummy-based method, sometimes referred to as location cloaking method, is an extension of k-anonymity. It will generate some dummy data to conceal sensitive location information.

5.1 Generalized data utility

From Figure 1, we demonstrate the similarity of the raw data and the synthetic data. As indicated above, Figure 1(a) denotes the raw data while Figure 1(b) represents the synthetic data. It can be concluded that the data after sampling and argumentation is still valid in terms of relationship, distribution, and other statistic features. The correlations among each attribute are perfectly maintained shown by the scatter plots in both figures. The distribution changes slightly with a minor impact on the data utility. It is noticeable that the skewness of all attributes turns out to be a bit more positive. This can provide insights to us to further optimize the parameters or the model itself to better improve.

5.2 Data utility evaluation

In Figure 2, we demonstrate how data utility changes according to different iteration times when the privacy protection level is fixed. In this scenario, we set the privacy protection level of ϵ as 3 while the iteration times are 1000 and 10000, respectively.

The solid curve denotes the raw data. The dashed curve is the generated data by GAN. The dotted line is the generated data by the Laplace mechanism under differential privacy. The solid line with the triangle marker shows the generated data of the dummy-based method. At last, the solid line with the marker is the generated data by the proposed model. The following figures share the same legend.

In Figure 2(a) and (b), the only difference of parameters is the iteration times. From these two figures, we can tell the dummy-based method and the Laplace mechanism outputs random results regardless of the iteration times. The reason is that the dummy-based method in this work randomly selects some dummy data points while Laplace mechanism chooses random noise complying with Laplace distribution. These

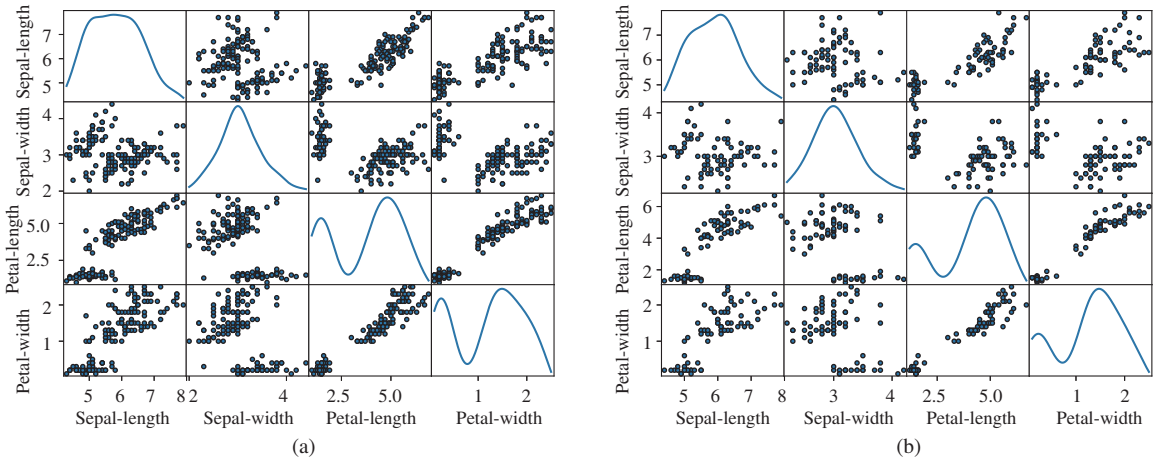

Figure 1 (Color online) Comparison between raw dataset (a) and generated synthetic dataset (b). The two figures show the correlations and distributions of each attribute before and after processing. The overall statistic features are quite similar, which guarantees high-level data utility.

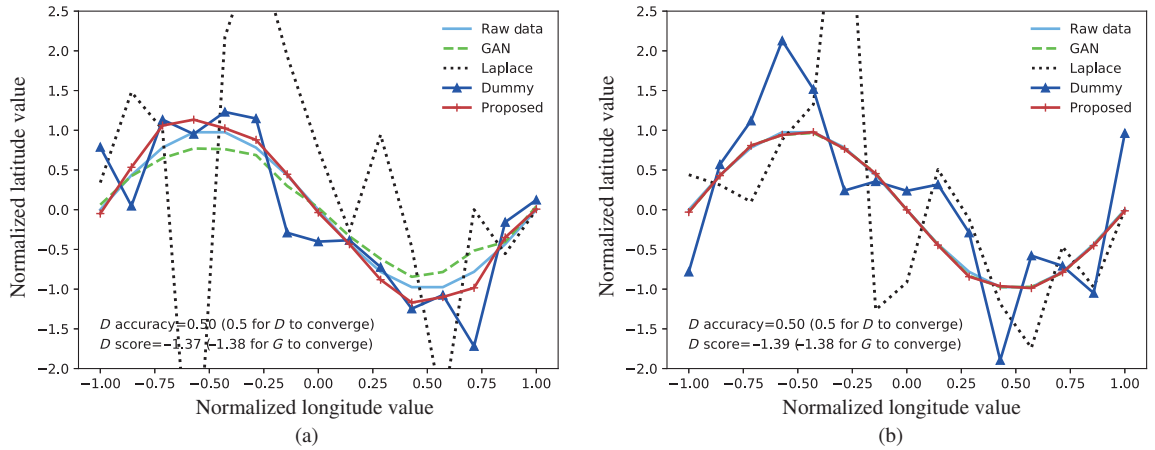

Figure 2 (Color online) Data utility evaluation. In the two figures, we compare the data utility in two cases where the iteration times are 1000 and 10000, respectively. The improvement of accuracy is significant with the increase of the iteration times. (a) $\epsilon = 3$, it $= 1000$; (b) $\epsilon = 1$, it $= 10000$.

random selections are independent and the iteration times have an impact on this stochastic process. In terms of GAN and the proposed model, the performance has noticeable upgradation regarding data utility. In Figure 2(a), we can tell both models have good approximations of the raw data. At the same time, the performance of the proposed model is slightly lower than the GAN due to the privacy requirements. However, when the iteration increases to 10000, the two models have perfect approximations of the raw data. In this case, the proposed model has no obvious performance degradation comparing to GAN although privacy requirements maintain the same. Therefore, the proposed model can achieve a highest data utility while satisfying all privacy requirements.

5.3 Privacy protection level

To better differentiate GAN and the proposed model in Figure 3, we select 1000 as the iteration times for privacy protection evaluation. When the iteration increases to 10000, the two models still have perfect approximations of the raw data as indicated above. In addition, the privacy protection levels are denoted by $\epsilon = 1$ and $\epsilon = 3$, respectively. The smaller the ϵ is, the higher the privacy protection level is.

Similarly, we can still see a random performance of the dummy-based method. However, we can observe that Laplace mechanism upgrades if the value of ϵ increases from 1 to 3, because Laplace mechanism is within the framework of differential privacy. In terms of GAN, the performance maintains the same

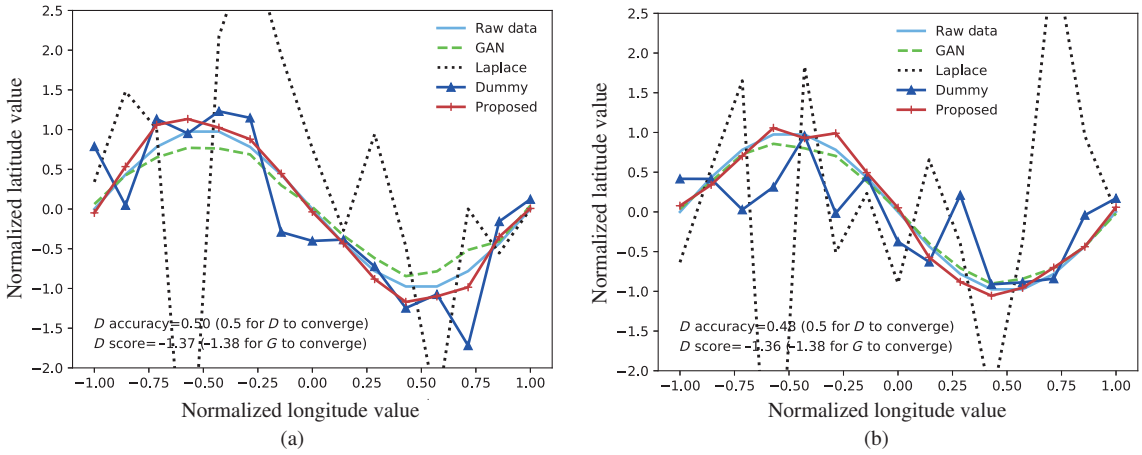

Figure 3 (Color online) Privacy protection evaluation. This figure reveals the privacy protection level upgrades with the decrease of ϵ, which is the index of privacy protection level under differential privacy framework. (a) $\epsilon = 3$, it $= 1000$; (b) $\epsilon = 1$, it $= 10000$.

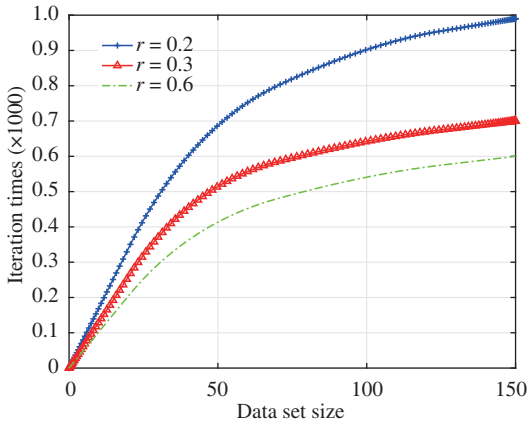

Figure 4 (Color online) Convergence v.s. learning rate. The learning rates are 0.2, 0.3, and 0.6, respectively. In all cases, the proposed model can achieve an convergence while greater learning rates accelerates the convergence.

although the generated data are not exactly the same. The rationale behind this is that GAN will generate an approximate synthetic data with the same level of privacy if the convergence is reached.

In both Figure 3(a) and (b), the proposed model has good performances in privacy protection. It is clear that the performance shows an upgradation when the value of ϵ increases from 1 to 3. No matter what the value of ϵ is, the proposed model has better performances comparing to the Laplace mechanism although both of them satisfy the requirements of differential privacy. Furthermore, the proposed model can achieve personalized privacy if the index is properly selected such as effective distance.

5.4 Efficiency

In terms of learning rates, the proposed model converges at different iteration times. In this subsection, we use three learning rates to evaluate the performance, which are 0.2, 0.3, and 0.6, respectively. We sample the size of the dataset at five data points intervals and present the fitted curve using the interpolation method.

As illustrated in Figure 4, greater learning rate results in faster convergence. The slope value of the curve keeps decreasing with the increasing of data size, which indicates there is a marginal effect on the iteration times. This also means after a certain threshold, such as 100 in this case, the iteration times grow slowly even if the data size keeps increasing. This feature enables the high scalability of the proposed model. Therefore, the proposed model fully meets the requirements of big data scenarios in the 5G era.

6 Summary and future work

In this paper, we start by identifying the weaknesses of directly publishing all the raw data even with privacy protection. This will cause either privacy leakage threats or data utility degradation due to over-protection. Motivated by this, we propose a generative adversarial networks enhanced location privacy protection model. We firstly generate a subset of differentially private data by posterior sampling. The subset is then leveraged in data argumentation based on generative adversarial networks. The generated synthetic dataset is proved to be differentially private and prediction-accurate comparing to existing relevant work. The extensive evaluation results on real-world datasets confirm the significance of the proposed model.

In the case of future work, we plan to further testify the model with various datasets and multiple fitting functions. In addition, we intend to extend this model to the scenario of continuous data publishing, in particular, trajectory privacy protection, which is more practical and feasible. Moreover, further optimization could be achieved with cross-discipline subjects such as convex composite optimization and Markov decision process.

Acknowledgements This work was partly supported by JSPS KAKENHI (Grant No. JP19H04105).

References

1 Ahmad I, Kumar T, Liyanage M, et al. Overview of 5G security challenges and solutions. IEEE Commun Stand Mag, 2018, 2: 36–43
2 Domo. Data never sleeps. 2018. https://www.domo.com/solution/data-never-sleeps-6
3 Qu Y Y, Yu S, Gao L X, et al. A hybrid privacy protection scheme in cyber-physical social networks. IEEE Trans Comput Soc Syst, 2018, 5: 773–784
4 Ji X S, Huang K Z, Jin L, et al. Overview of 5G security technology. Sci China Inf Sci, 2018, 61: 081301
5 Finlayson S G, Bowers J D, Ito J, et al. Adversarial attacks on medical machine learning. Science, 2019, 363: 1287–1289
6 Dwork C. Differential privacy. In: Proceedings of International Colloquium on Automata, Languages, and Programming, Venice, 2006
7 Dwork C, Kenthapadi K, McSherry F, et al. Our data, ourselves: privacy via distributed noise generation. In: Proceedings of the 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, 2006. 486–503
8 Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the 1st International Conference on Mobile Systems, Applications, and Services, San Francisco, 2003
9 Qu Y Y, Yu S, Zhou W L, et al. Privacy of things: emerging challenges and opportunities in wireless Internet of things. IEEE Wirel Commun, 2018, 25: 91–97
10 Chaudhary R, Kumar N, Zeadally S. Network service chaining in fog and cloud computing for the 5G environment: data management and security challenges. IEEE Commun Mag, 2017, 55: 114–122
11 Zeng D Z, Gu L, Guo S, et al. Joint optimization of task scheduling and image placement in fog computing supported software-defined embedded system. IEEE Trans Comput, 2016, 65: 3702–3712
12 Eiza M H, Ni Q, Shi Q. Secure and privacy-aware cloud-assisted video reporting service in 5G-enabled vehicular networks. IEEE Trans Veh Technol, 2016, 65: 7868–7881
13 Yu S. Big privacy: challenges and opportunities of privacy study in the age of big data. IEEE Access, 2016, 4: 2751–2763
14 Samarati P. Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng, 2001, 13: 1010–1027
15 Machanavajjhala A, Kifer D, Gehrke J, et al. *L*-diversity: privacy beyond *k*-anonymity. IEEE Trans Knowl Data Eng, 2007. doi: 10.1109/ICDE.2006.1
16 Li N H, Li T C, Venkatasubramanian S. Closeness: a new privacy measure for data publishing. IEEE Trans Knowl Data Eng, 2010, 22: 943–956
17 Qu Y Y, Yu S, Gao L X, et al. Big data set privacy preserving through sensitive attribute-based grouping. In: Proceedings of IEEE International Conference on Communications (ICC), 2017
18 Ma L C, Pei Q Q, Qu Y Y, et al. Decentralized privacy-preserving reputation management for mobile crowdsensing. In: Proceedings of International Conference on Security and Privacy in Communication Systems, 2019. 532–548
19 Xu C Q, Zhu L, Liu Y, et al. DP-LTOD: differential privacy latent trajectory community discovering services over location-based social networks. IEEE Trans Serv Comput, 2019. doi: 10.1109/tsc.2018.2855740
20 Gu B S, Gao L X, Wang X D, et al. Privacy on the edge: customizable privacy-preserving context sharing in hierarchical edge computing. IEEE Trans Netw Sci Eng, 2019. doi:10.1109/tnse.2019.2933639
21 Qu Y Y, Nosouhi M R, Cui L, et al. Privacy preservation in smart cities. In: Smart Cities Cybersecurity Privacy. Amsterdam: Elsevier, 2019. 75–88
22 Duan X Y, Wang X B. Authentication handover and privacy protection in 5G hetnets using software-defined networking.

IEEE Commun Mag, 2015, 53: 28–35
23 Zhang A Q, Lin X D. Security-Aware and privacy-preserving D2D communications in 5G. IEEE Netw, 2017, 31: 70–77
24 Ni J B, Lin X D, Shen X S. Efficient and secure service-oriented authentication supporting network slicing for 5G-enabled IoT. IEEE J Sel Areas Commun, 2018, 36: 644–657
25 Liao D, Li H, Sun G, et al. Location and trajectory privacy preservation in 5G-Enabled vehicle social network services. J Netw Comput Appl, 2018, 110: 108–118
26 Qu Y Y, Pokhrel S R, Garg S, et al. A blockchained federated learning framework for cognitive computing in industry 4.0 networks. IEEE Trans Ind Inf, 2020. doi:10.1109/TII.2020.3007817
27 Qu Y Y, Gao L X, Luan T H, et al. Decentralized privacy using blockchain-enabled federated learning in fog computing. IEEE Int Things J, 2020, 7: 5171–5183
28 Yu S, Zhou W L, Jia W J, et al. Discriminating DDoS attacks from flash crowds using flow correlation coefficient. IEEE Trans Parallel Distrib Syst, 2012, 23: 1073–1080
29 Yu S, Zhou W L, Doss R, et al. Traceback of DDoS attacks using entropy variations. IEEE Trans Parallel Distrib Syst, 2011, 22: 412–425
30 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Annual Conference on Neural Information Processing Systems, Montreal, 2014. 2672–2680
31 Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, 2017. 214–223
32 Li C X, Xu T, Zhu J, et al. Triple generative adversarial nets. In: Proceedings of Annual Conference on Neural Information Processing Systems, Long Beach, 2017. 4091–4101
33 Song S, Chaudhuri K, Sarwate A D. Stochastic gradient descent with differentially private updates. In: Proceedigns of IEEE Global Conference on Signal and Information Processing, Austin, 2013. 245–248
34 Lee J, Kifer D. Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, 2018. 1656–1665
35 Qu Y Y, Yu S, Zhang J W, et al. GAN-DP: generative adversarial net driven differentially privacy-preserving big data publishing. In: Proceedings of IEEE International Conference on Communications (ICC), 2019
36 Qu Y Y, Yu S, Zhou W L, et al. Gan-driven personalized spatial-temporal private data sharing in cyber-physical social systems. IEEE Trans Netw Sci Eng, 2020. doi: 10.1109/TNSE.2020.3001061
37 Dwork C, Feldman V, Hardt M, et al. The reusable holdout: preserving validity in adaptive data analysis. Science, 2015, 349: 636–638
38 Shokri R, Shmatikov V. Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, 2015. 1310–1321
39 Abadi M, Chu A, Goodfellow I J, et al. Deep learning with differential privacy. In: Proceedings of ACM SIGSAC Conference on Computer and Communications Security, Vienna, 2016. 308–318
40 Ács G, Melis L, Castelluccia C, et al. Differentially private mixture of generative neural networks. In: Proceedings of IEEE International Conference on Data Mining, New Orleans, 2017. 715–720
41 Elshaer H, Boccardi F, Dohler M, et al. Downlink and uplink decoupling: a disruptive architectural design for 5G networks. In: Proceedings of IEEE Global Communications Conference, Austin, 2014. 1798–1803
42 Wang Y X, Fienberg S E, Smola A J. Privacy for free: posterior sampling and stochastic gradient monte carlo. In: Proceedings of the 32nd International Conference on Machine Learning, Lille, 2015. 2493–2502
43 Huang S W, Lin C T, Chen S P, et al. Auggan: cross domain adaptation with gan-based data augmentation. In: Proceedings of the 15th European Conference, Munich, 2018. 731–744
44 Dheeru D, Taniskidou E K. UCI machine learning repository. 2017. http://archive.ics.uci.edu/ml
45 Dwork C. Differential privacy. In: Encyclopedia of Cryptography and Security. 2nd ed. 2011. 338–340