Privacy Enhancing Techniques in the Internet of Things Using Data Anonymisation

Wang Ren^{1,2} · Xin Tong² · Jing Du² · Na Wang² · Shancang Li³ · Geyong Min⁴ · Zhiwei Zhao⁵

Accepted: 4 February 2021
© The Author(s) 2021

Abstract

The Internet of Things (IoT) and Industrial 4.0 bring enormous potential benefits by enabling highly customised services and applications, which create huge volume and variety of data. However, preserving the privacy in IoT and Industrial 4.0 against re-identification attacks is very challenging. In this work, we considered three main data types generated in IoT: *context data*, *continuous data*, and *media data*. We first proposed a stream data anonymisation method based on k -anonymity for data collected by IoT devices; and then privacy enhancing techniques for both continuous data and media data were proposed for different IoT scenarios. The experiment results show that the proposed techniques can well preserve privacy without significantly affecting the utility of the data.

Keywords Privacy preserving · Internet of things (IoT) · Industrial 4.0 · Data anonymisation

1 Introduction

The Internet of things (IoT) provides great promising in support a variety of IoT applications, such as healthcare, smart manufacturing, industrial 4.0, smart home, etc., which create huge volume of data that need to be processed and shared that might contain sensitive information needs to be protected before share with others (Zhao et al. 2019). With the great potentials, there are however significant security and privacy concerns and legal issues to be aware of Zhang et al. (2015). The EU's General Data Protection Regulation (GDPR) outlines data protection and privacy and addresses the transfer of personal data and give users more control

over their personal data. In IoT, the data anonymisation techniques are wildly used to protect sensitive information and privacy related to personally identifiable information by erasing or encryption identifiers that connect an individual to stored data. It aims to simply the regulatory environment for business so both individuals and businesses can fully benefit from the digital economy. The GDPR also permits businesses to collect anonymised data without consent, use it for any purpose, and store it for an indefinite time (Otgonbayar et al. 2016; Da Xu et al. 2014).

The most recent data link technologies, such as *big data analytic*, etc., are able to establish links between dataset created by different components in IoT (Zhang et al. 2015; Zhang and Chen 2020). In an IoT environment, massive number of smart devices are connected that are used to gather data, this comes with security and privacy risks and the sensitive data breach might can be caused by malicious behaviours or careless. In addition, resource limited IoT devices compromise through vulnerability can cause disclosure of critical data (*e.g.*, in *healthcare*, *smart home*, etc.) (Li et al. 2019; Viriyasitavat et al. 2019).

In IoT scenarios, massive data being gathered and much of this data is digitised and stored online (cloud server, etc.) Although much of this is not made public, authorised accesses and the threat of hackers looking to steal data, often with malicious intent. Virtual everything related to user, both online/offline can be tracked in the form of data, such as doctor visits, interaction with companies,

✉ Shancang Li
shancang.li@ieee.org

¹ College of Electronics and Information Engineering,
Sichuan University, Chengdu, 610065, China

² China Information Technology Security Evaluation Center,
Beijing, China

³ Department of Computer Science, University of the West
of England, Bristol, BS16 1QY, UK

⁴ Department of Computer Science, University of Exeter,
Exeter, EX4 4PY, UK

⁵ School of Computer Science and Engineering,
University of Electronic Science and Technology of China,
Chengdu, China

browsing habits, app use, etc. If you live in a ‘smart home’, then daily actions, like the use of digit kitchen, temperature, clothes, etc. Some of this data might not directly linked to a user, privacy problems arise when information is tied to personally identifiable information (PII), including *email, name, ip address, location*, etc. (Ma et al. 2020; Zhou et al. 2019). Data anonymisation does not mean it is always impossible to discover the identity of the subject. Many anonymisation techniques can be reversible. E.g., hashed data could be de-anonymised by guessing the data until a matching hash was found. Even using irreversible suppression, probably the most fail-safe method of anonymisation, the remaining data could be cross-referenced with other data sets to identify the source (Yao et al. 2019).

Data anonymisation is a valuable privacy-preserving tool that can anonymise identifiers using *removing, substituting, distorting, generalisation or aggregation, etc.* Basically, user identifiers include *direct identifiers* and *indirect identifiers*. The direct identifiers are the attributes that can directly identify a user, such as *names, address, photo, etc.*, while the *indirect identifiers* are attributes that can identify a user by linking with other available dataset or information, such as *ages, salary, occupations, etc.* Many IoT applications collect as much as data in order to improve their service and develop new products. However, this will significantly increase the risk of data loss or accidental data breaches. Many exciting applications in IoT, such as *Facebook, Twitters, Tictok, Skyscanner, Hungryhouse, Trainline, Smart home, etc.*, collects key information such as user’s name, address, credit card, timeline, behaviour patterns of using apps, which may be re-associated with the data at a later time to identify personally identifying information. Among the arsenal of data analysis tools available, data anonymisation is highly recommended to satisfy the requirements of GDPR on the basis of the data is identifiable before sharing to third party or public.

In this work, we considered three main data types generated in IoT: *context data, continuous data, and media data*, and the main contributions are:

1. A stream data anonymisation method for IoT device is proposed based on *k*-anonymity, which is able to continuously facilitate *k*-anonymity over data stream in IoT scenarios.
2. Multiple modal based data anonymisation techniques for continuous dataset and media data (image and video) were proposed in IoT scenarios.
3. Experimental results show the effectiveness of proposed privacy enhancing techniques.

The remains of this paper is organised as: Section 2 discusses the state-of-the-art of data anonymisation in IoT; Section 3.3 proposed anonymisation schemes for *attributes*

data, continuous data, and visual data generated by IoT systems; Section 4 use the real data to evaluate the proposed schemes; and Section 5 concludes this paper.

2 Related Works

Suffering from data breach may significantly cause organisations loss on both finance and reputation. In the past decade, the IoT security has attracted lots of research attention and a number of privacy-preserving techniques have been developed for protecting data generated by IoT devices (Ouazzani and Bakkali 2018), including data *anonymisation, pseudonymisation, de-identification, etc.*

2.1 Data Anonymisation Techniques

In the past decades, a number of privacy-preserving mechanisms have been implemented through *data masking, pseudonymisation, generalisation, swapping, perturbation, synthetic, etc.* Key features of these techniques are summarised as:

- *Data masking*, using character modification (such as shuffling, substitution, encryption, etc.) to hide data, which against reverse-engineering or detection attacks (Gope and Sikdar 2019);
- *Pseudonymization*, replace private identifier with fake identifier or pseudonyms to hide key identifiable information, which preserves statistical accuracy and data integrity (Somolinos et al. 2015; Faldum 2007);
- *Generalization*, remove part of data to make it unidentifiable, e.g., data can be altered into a set of ranges (Deldar and Abadi 2019; Yaseen et al. 2018);
- *Perturbation*, hiding sensitive data patterns by adding a crafted random noise to prevent privacy data mining attacks (Amar et al. 2018);
- *Synthetic*, using algorithms to create artificial dataset with specific statistical patterns or models instead of altering the original dataset (El Emam 2020).

Specifically, a number of *k*-anonymisation based data anonymisation techniques have been developed. Gionis et al. improved the *k*-anonymisation by proposing a three-measure method for capturing the amount of information that is lost during the anonymisation (Gionis and Tassa 2009).

However, there are still many challenges in data anonymisation. In IoT applications, a huge number of smart sensors, devices are connected that continuously generate data for monitoring activities, status, etc. IoT applications can be designed to access these data. As discussed above, these data may includes sensitive data or specific data patterns that the data owners hope to keep private. To preserve the privacy in the complicated IoT environments, we need

to develop intelligent and automatic data anonymisation techniques that can hide key data attributes and potential patterns in raw data.

2.2 Recent Advances in Data Anonymisation

In the past decades, k -anonymity, l -diversity, and t -closeness based methods have been widely used in cloud-based application to protect sensitive information. In a dataset, some attributes themselves are not unique identifiers, but could create a unique identifier by correlating with other QIs, which we call quasi-identifiers (QIs). In dataset anonymisation, both direct identifiers and QIs should be anonymised. An attacker could be able to identify the individuals by linking QI attributes with external dataset(s) containing direct identifiers. For $\mathcal{D}$, if each combinations of QI attributes to be shared by at least k records, we say $\mathcal{D}$ has the k -anonymity property. To achieve l -diversity safe, $\mathcal{D}$ needs to ‘well present’ at least l values for each equivalence class; and $\mathcal{D}$ satisfies t -closeness if the distance (e.g. Earth Mover’s distance (EMD)) between the distribution of the sensitive attributions in a class and the distribution is rather than a threshold t .

Microaggregation is a perturbative data protection method (Shi et al. 2018), in which small clusters in a dataset can be replaced each original record by the centred of the corresponding cluster (each cluster should have between k and $2k$ elements), the larger the k , the larger the information loss and the lower the disclosure risk. It ensures k -anonymity only when multivariate microaggregation is applied processing all the variables of the dataset (Mahawaga Arachchige et al. 2020; Du et al. 2020).

Li et al. proposed a stream k -anonymity scheme to continuously facilitate k -anonymity on data streams (Li et al. 2008). However, it cannot well process data in diverse IoT scenarios. In Pervaiz et al. (2015), a data stream scheme was proposed for privacy protection in access control, in which k -anonymity or l -diversity is used to generalise stream data. Domingo et al. explored a unified and conceptually stream data anonymisation approach using the microaggregation (Domingo-Ferrer et al. 2019), which is a fine-grained data aggregation that supports both static dataset and dynamic data stream (Domingo-Ferrer et al. 2019). Khavkin et al. proposed a stream anonymisation scheme based on microaggregation against differential privacy attacks, which includes an algorithm that satisfies k -anonymity and recursive (c, l) -diversity aiming at minimising information lost and reducing data disclosure risks (Khavkin and Last 2018; Soria-Comas et al. 2017).

Differential privacy (DP) a strong privacy protection scheme aims to guarantee bounds on how much information can be revealed by the participation of an individual component in a database (Soria-Comas et al. 2017). Furthermore, ϵ -Differential privacy (DP) was proposed the measure

relative privacy: for datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ and a randomised algorithm κ , for all $S \in \text{Range}(\kappa)$, Eq. 1 holds (Huang et al. 2020; Wang et al. 2018).

$$\Pr(\kappa(\mathcal{D}_1) \in S) \leq \exp(\epsilon) \times \Pr(\kappa(\mathcal{D}_2) \in S) \quad (1)$$

In recent, the DP is widely used to protect the privacy in machine learning (Phan et al. 2017; Zhang et al. 2020). Actually in many IoT scenarios, anonymising personal data is not enough to protect privacy because heavily incomplete dataset will increase the risk to re-identify a specific individual. It is noticed that even anonymised dataset can be tracked back to individual using machine learning (Rocher et al. 2019; Yang et al. 2020; Lu and Ning 2020).

Actually, in recent the blockchain technologies were introduced to address the security and privacy issues in both the IoT and Industrial 4.0 (Gorkhali et al. 2020; Aceto et al. 2020; Yli-Ojanperä et al. 2019; Xu et al. 2018). This work will focus on the stream data anonymisation in IoT environments by using existing k -anonymity based techniques. The k -anonymity can guarantee the privacy of data from following aspects and have to follow conditions: (1) The sensitive data must not reveal information that was redacted in the generalised columns; (2) The value of sensitive columns are not all the same for a particular group of k ; (3) The dimensional of the data must be sufficiently low. Data generated by applications or IoT devices may contain sensitive PII, the new big data techniques, such as big data, etc. are able to learning key PII from this data.

3 Data Anonymisation in IoT

In most IoT scenarios, the data collected mainly fails into three categories; (1) normalised context data collected by IoT sensor & device, such as temperature, flow, pressure, and humidity, data collected using proprietary formats and protocols depending on the source; (2) continuous data gained via sensors, which is collected using appropriate communication protocols, such as MQTT, etc. and that keep the features of real-time; (3) media data, many IoT devices, such as IP camera, can collect media data, including audio, image, video, and text data in real-time.

In the past decades, many research efforts have been conducted to make trade-off between data utility and security of this information. (1) In statistical data analysis, adding noise is the most common way to maintain some statistical invariant, however, it may compromise the integrity of the data/dataset. Actually, it is not easy to add random noise without significantly reducing the utility; (2) data usually is stored at devices that may have different security levels, which can cause the risk of data disclose due to the fact that we cannot consider every potential attacks an IoT system may face. Another issue is that many data holders

share same data, however they may have different security or privacy concerns which makes it very challenges to balance these concerns; (3) In IoT systems, sophisticated users/devices authentication will ensure proper access controls and access control policies need to be well designed to protect privacy information.

In this work, we present an IoT data anonymisation scheme as shown in Fig. 1. The data generated by sensors in IoT will be aggregated into data stream, which needs to be analysed to identify sensitive information, and then the data will be anonymised using secure anonymisation algorithms before sending to apps or release to public.

3.1 IoT Data Pre-processing

We first introduce the data pre-processing before analysing attributes that might be sensitive information. In an IoT system, the data collected by a sensor may be intermittent, which needs to be leverage the time span gaps before further processing. For a sensor node N_i received n number of input data every time window w_j in t second with each data is of d size and one data requires p seconds to process. Then, the total output time needed will be $n \times p$ seconds, if $n \times p < t$, then there is no issue. However, if $n \times p > t$, then we need multiple windows, i.e., when time needed to process the input greater than time between two consecutive batches of data. In this work, we use a variable ϵ for multiple windows, the problem can be simplified to $[(n \times p)/\epsilon] < t$.

For an IoT gateway G , we introduce a factor ζ to denote the number of threads, which can leverage the CPU and RAM utilisation. The larger the ζ is, then more CPU computation resources will be used. When receives data, the IoT devices needs ζ windows to process. Hence, there will be ζ active windows and $n - \zeta$ pending items in queue. In this work, we use r denote the number of batches in memory, one batch can be processed by w windows at a time. The size of a batch is $\zeta \times d$ Then, we have

$$[(n \times p)/(\epsilon \times \zeta)] < t \quad (2)$$

$$r = S_{RAM}/(\epsilon \times \zeta) \quad (3)$$

here we use r to balance the data stream speed and bandwidth. For a fixed-length time-window of data, $\mathbf{X} = (\mathbf{x}_{sj}) \in \mathbb{R}^{M \times W}$, it contains some attributes together with specific patters that can be utilised to identify a specific user.

Definition 1 Data Streaming $DS = \{pid, X_t\}$ is an IoT data steam generated by smart sensor, in which pid denotes the identifier, and X_t denotes dataset that include both quasi-identifier attribute set and other attribute set, and t is the time window. In this work, the t is adjustable and quasi-identifier attributes in X_t need to be anonymised before publishing.

Definition 2 Anonymisation algorithm ($\mathcal{A}$) can be used to conduct anonymisation of a dataset or stream:

$$\mathcal{A} : DS \rightarrow DS' \quad (4)$$

in which DS' is the anonymised data stream that could be shared with third party or released publicly.

The collected data can be clustered into multiple clusters using specific algorithms. Following method can be used to form a data cluster C from data stream DS (Otgonbayar et al. 2016).

Definition 3 A cluster C built from DS satisfies k -anonymity if the distinct number $|C|$ in a cluster less than k ($|C| \geq k$), then we say C is generalised and the cluster C satisfies k -anonymity.

Definition 4 For two tuples in the same cluster C , $t_1 \in C$ and $t_2 \in C$, the distance between t_1 and t_2 can be calculated by

$$Distance(t_1, t_2) = \frac{1}{|X_1^q \cap X_2^q|} \sum_{q_i \in X_1 \cap X_2} d_i(q_i) \quad (5)$$

where $|X_1^q|(|X_2^q|)$ is the number of quasi-identifier attributes in $X_1^q(X_2^q)$, d_i is the normalised distance between the numerical values.

In data anonymisation, the quality of anonymisation algorithm is usually measured by average information loss defined in Eq. 6.

Definition 5 Average information Loss. The average information loss of first N tuples from DS is

$$L_{i-avg} = \frac{1}{N} \sum_{i=1}^N L_{info}(t_i, g_i) \quad (6)$$

in which g_i is generalisation of tuple t_i .

Fig. 1 IoT data stream anonymisation architecture

3.2 Data Stream Anonymisation

3.2.1 k -Anonymity for Data Streams

For an original data stream $S_{org}^t = \{ \langle pid_1, X_1 \rangle, \langle pid_2, X_2 \rangle, \dots, \langle pid_n, X_n \rangle \}$; Each tuple t_i comprise a vector of $|pid_i|$ identifier attributes and values X_i , S_{out}^t satisfies k -anonymity property with respect to QI is produced. Moreover, the S_{out} order deviates from the input stream.

Each $t_i = (id, q_1, \dots, q_m, z)$ includes the identity id , QI $q_1, \dots, q_m$, and sensitive attribute z . $s_i' = (q_1, \dots, q_m, z)$ is the anonymised s_i , where the id has been pruned. Stream anonymisation may cause delay and extra temperately data that can degrade the performance of the data processing.

3.2.2 Identifier Group

If a set of $\langle pid_i, X_i \rangle \in S_{org}$ have the same values on QI attributes, these tuples can form a QI group g_i . For a specific group g_i , we use a QI detection algorithm to check if these QI attributes match the k -anonymity requirements. If $|t_i| \geq k$, then it means the dataset $X_i \in g_i$ can be k -anonymised. If $|t_i| < k$, it means tuple $t_i \in S_{org}$ is k -anonymised already.

3.2.3 Classification of Attributes

In an IoT applications, the key attributes in an dataset could be used to identify subject, such as name, address, mobile number, which can uniquely identify an individual directly and should be hidden before releases. In healthcare application, key attributes could be *DoB*, *sex*, *postcode*, *etc.* Sensitive attributes includes *medical record*, *wage*, *home address*, *bank account*, *etc.*

3.3 Proposed IoT Based Anonymisation Algorithm

In is work, we classify the IoT data into three categories: (1) attribute dataset, which includes different attributes in each record; (2) Continuous dataset, which includes continuous data samples, such as monitoring values, motion data, etc.; (3) image dataset, includes images and videos (Fig. 2).

Fig. 2 Continuous data anonymisation architecture

At first, the data captured by an IoT devices will be clustered into partitions and find $k - 1$ nearest neighbors of expiring tuple t according to the distances in Eq. 5; and create cluster over the identified tuples; Then, anonymises expiring tuple using reusable k -anonymity cluster defined over the same partition which covert t , k -anonymity is based on information loss calculated by Eq. 6.

3.3.1 Attributes Dataset Anonymisation

In this scenarios, most existing anonymisation techniques, (k -anonymity, l -diversity, ϵ -closeness, *etc.*) can be used to anonymise identifiable attributes. Following main steps are involved:

1. Determine the release model, define how the anonymised dataset will be released;
2. Determine the acceptable re-identification risk threshold and utility, which is used to define the anonymisation parameters in the algorithms;
3. classify IoT data attributes. In this process, attributes need to be defined as *direct identifiers*, *indirect identifiers*, *non-identifiers*;
4. Remove unused attributes, since in IoT some attributes could be missed or anomaly data will be collected, this process will remove all unused data attributes;
5. Anonymise direct and indirect identifiers by applying techniques such as k -anonymity, l -diversity, t -closeness or the combination of these techniques;
6. Evaluate the risk or anonymisation quality, if need, adjust the parameters and repeat 5) and 6).
7. Examining the utility of anonymised dataset, if the utility is sufficient, then it can be released; if not, the anonymisation process need to be redesigned.

An IoT system possesses data anonymisation using an anonymise engine to define data anonymisation parameter, which hosts the anonymise algorithms with the parameters to perform the data anonymisation processes. Figure 1 provides the main architecture of proposed data anonymisation scheme in IoT environment.

To anonymise the data, we first need to use the anonymisation engine to define the privacy. The anonymisation can be conducted using Algorithm 1, in which we assume that the device owner and the device are data owner. Let $\mathbf{s}$ and $\mathbf{v}$ denote the data owner and data user respectively. The input is the data created in $\mathcal{D}$ and the output $\overline{\mathcal{D}}$ is the anonymised data.

Algorithm 1 Data anonymisation algorithm in IoT.

Input: Original dataset $\mathcal{D}$

Output: Anonymised dataset $\overline{\mathcal{D}}$

```

1:  $\mathbf{v} \leftarrow \text{video\_capture}(V)$ 
2:  $\mathbf{v}_g \leftarrow \text{video\_group}(\mathbf{v})$ 
3: for each frame  $f_i \in \mathbf{v}_g$  do
4:    $f_{hsv} \leftarrow \text{covColor}(f_i, \text{COLOR\_RGB2HSV})$ 
5:    $f_{shsv} \leftarrow \text{split}(f_{hsv})$ 
6:   for each channel  $c \in \{\text{hue}, \text{saturatin}, \text{value}\}$  do
7:      $\text{threshld}(f_{shsv}[c], d_{th}[c], \text{low}[c], \text{upper}[c])$ 
8:   end for
9:    $f_{result} \leftarrow \text{bitwise\_and}(d_{th}[0], d_{th}[1], d_{th}[2])$ 
10: end for
11:  $\overline{V} \leftarrow \text{update}(\mathbf{v}, \overline{\mathbf{v}}_g)$ 

```

Streaming IoT data often involves a large amount of data and a number of devices, which needs to be carefully processed. The data rate might be changeable depends the change of environment. The data stream may need to be buffered or aggregated before transmitting to cloud. Many existing stream processors, such as *Apache Spark*, *Storm*, *Flink*, *etc.*, have been developed perform MapReduce in real time on data streaming. Some tools, like *Kafka Stream*, *Apache Samza*, *etc.* can aggregate multiple data streams into a large or complex stream. The *InfluxDB*, *TimescaleDB*, *Cassandra*

1. Aggregate process. An aggregate perform aggregation by reading time series value from the short term buffer,

writing the sums to aggregate storage, and then deleting the aggregation time, together with the reservation;

2. Rule engine. The rule engine defines and maintains rules and store them in the same database as the short term buffer. *E.g.*, “triggerInterval”: “1m”,

3.3.2 Continuous Dataset Anonymisation

In many IoT applications, including *e-healthcare*, *body sensor networks*, *location-based services*, *etc.*, IoT devices can capture continuous motion data, location data, and continuous bio-signals, which may includes private information about users health status, behaviours, activities, locations, *etc.* that might be identifiable. To protect this kind of information, this section will discuss the ways to anonymise continuous data stream.

One of challenges is to classify the user-identifiable patterns from the dataset, in recent researchers have proposed that data patterns can be used to create fine-grained behavioral profiles of users that can reveal their identity (Neverova et al. 2016; Malekzadeh et al. 2019). For the continuous data anonymisation, we propose a continuous data anonymisation algorithm as Algorithm 2.

In continuous data anonymisation, adding Laplace noise to continuous or unbounded data is an effective way to improve randomness of the data against patterns learning, as shown in Eq. 7. The Laplace distribution noise can preserves differential privacy by adding the randomness but can keep the utility of the data. Adding Gaussian noise is also an effective way to protect privacy by adding randomness.

$$f(\mathbf{x}; \mu, \lambda) = \frac{1}{2\lambda} e^{\left(-\frac{|\mathbf{x}-\mu|}{\lambda}\right)} \quad (7)$$

in which μ the position of the distribution peak; the non-negative λ is the exponential decay. In practice, the μ and λ needs to be carefully defined depends the model.

Fig. 3 Privacy anonymisation in video/image using Yolov3 Object Detection Techniques

Algorithm 2 Continuous numerical data anonymisation algorithm in IoT.

Input: Original continuous dataset $\mathcal{X}$, trained patterns $\mathcal{P}$

Output: Anonymised dataset $\mathcal{X}'$

```

1:  $\mathbf{c} \leftarrow \text{cluster\_capture}(\mathcal{X})$ 
2:  $\mathbf{p}_g \leftarrow \text{pattern\_group}(\mathbf{c})$ 
3: if ( $\mathbf{p}_g \in \mathcal{P}$ ) then
4:   Anonymise  $\mathbf{g}' \leftarrow \mathbf{g}$  using Eq. 7
5: end if
6: replace  $\mathbf{g}$  using  $\mathbf{g}'$ ;
7: update  $\mathbf{c}, \mathbf{g}$ ;
8: return  $\mathcal{X}'$ 

```

As an example, in this work we use the data extracted using a wearable IoT system, which includes human behaviours detection, we first extract the continuous data of ‘walking’, ‘jogging’, ‘upstairs’, and using a deep learning network to train and extract features of above activities (Wang et al. 2020). In this works, we directly use the models trained in our previous work, and the use Algorithm 1 to anonymise the ‘walking’ data series, and the results can be found in Fig. 6.

3.3.3 Visual Data Anonymisation

In many IoT applications, such as *smart cities*, *intelligent transport system*, *etc.*, devices often collect visual data that may include personal data. In this section, we developed a deep learning based video anonymisation scheme that can remove all private data from images, videos, in IoT applications. This can also enable IoT applications compliance with GDPR (Xiao et al. 2019; Li et al. 2019).

The idea is to perform privacy information detection using Yolov3 from images or videos and the conduct image anonymisation. The privacy information detection is a task involves identifying the presence, location, and type of one or more privacy data items. E.g., in a road surveillance system, the images/video from a surveillance camera can capture *vehicles*, *pedestrian*, *buildings*, *etc.*, the register number, face of pedestrian, or building location/number may the information that the owner do not want to share with public.

Figure 3 shows the image anonymisation process. It can be seen that the first process is to identify the private information to use Yolov3. The Yolov3 is a popular image object detection algorithm that can perform real-time, accurate detection. It uses the Darknet-53 architecture including 53 layer network trained on ImageNet. The Yolo is a full convolutional network and its eventual output is

generated by applying a 1×1 kernel on a feature map. In V3, the detection is done by applying 1×1 detection kernels on feature maps of three different sizes at three different places in the network.

Algorithm 3 Visual data anonymisation algorithm in IoT.

Input: Original video $\mathcal{V}$, trained objective model $\mathcal{M}$

Output: Anonymised dataset $\mathcal{V}'$

```

1:  $\mathcal{I} \leftarrow \text{video\_capture}(\mathcal{V})$ 
2:  $\mathcal{O} \leftarrow \text{image\_obj\_detection}(\mathcal{I}, \mathcal{M})$ 
3: if ( $\mathcal{O} \neq \emptyset$ ) then
4:   for each object  $o_i \in \mathcal{O}$  do
5:      $o'_i \leftarrow \text{image\_Anonymise}(o_i)$ 
6:      $o_i \leftarrow o'_i$ 
7:   end for
8: end if
9:  $\text{update}(\mathcal{O})$ 
10:  $\text{update}(\mathcal{I})$ 
11:  $\text{update}(\mathcal{V})$ 
12:  $\overline{\mathcal{V}} \leftarrow \text{update}(\mathcal{O}, \mathcal{I})$ 

```

The video anonymisation can be performed using Algorithm 3, in which the $\mathcal{M}$ is a trained objects model using Yolov3 *darknet*. In practice, depending the specific privacy information or object (E.g., facial features for human, register number for vehicle, house number in street review), pre-trained model could be used to detect objects with the private information, object in an image or video.

4 Evaluation

As mentioned above, this work considers three types of data: (1) context data; (2) continuous data; and (3) media data. Based on these three scenarios, we evaluated the performance in different angles.

4.1 Dynamic k-anonymity Algorithm

In this section, we will evaluate the performance of dynamic k -anonymity algorithms over a context database. We use the *Adult* dataset¹ to simulate the attribute dataset collected in an IoT system, which includes 48842 records with 15 attributes: ‘age’, ‘workclass’, ‘fnlwtgt’, ‘education’, ‘education-num’, ‘marital-status’, ‘occupation’, ‘relationship’, ‘race’, ‘sex’, ‘capital-gain’, ‘capital-loss’, ‘hours-per-week’, ‘native-country’, ‘income’, all these attributes can be group into 9 categories: ‘workclass’, ‘education’,

¹<http://archive.ics.uci.edu/ml/datasets/Adult>

 (a) $k=l=2$

 (b) $k=l=3$

 (c) $k=l=4$

 (d) $k=l=5$

 (e) $k=l=6$

 (f) $k=l=2$
Fig. 4 Performance for k -anonymity (blue), l -diversity (green), t -closeness (red) algorithms

Fig. 5 Elapsed time when $QIs = 4$ for k -anonymity (blue), l -diversity (green), t -closeness (red) algorithms

'marital-status', 'occupation', 'relationship', 'sex', 'native-country', 'race', 'income'. In this work, we use the general k -anonymity, l -closeness anonymisation and the result as follows.

Figure 4a-f shows the anonymisation performance for when the k and l change from 2 to 7, and the number of QIs change from 2 to 5, in which the blue line is k -anonymity, and green line shows the performance of l -diversity, and red line represents the t -closeness algorithms.

To evaluate the performance of anonymisation, we fixed the QIs as {age, education-num, sex, race} and use $k = l$ ranges from 2 to 10, the elapsed time (sec) can be found from Fig. 5

Also,

4.2 Continuous Dataset Anonymisation

To evaluate the effectiveness of anonymisation for continuous data, we use a piece of walking motion accelerometer

Fig. 6 Walking Raw accelerometer data vs Anonymised accelerometer data

(a) acc_x

(b) anonymised acc_x

(c) acc_y

(d) anonymised acc_y

(e) acc_z

(f) anonymised acc_z

(a) Plate number anonymisation

(b) Face anonymisation

Fig. 7 Visual information anonymisation in video clips

raw data (x, y, z) we acquired using wearable body sensor networks (Wang et al. 2020), as shown in Fig. 6. The Fig. 6a, c, and e show the raw data in x , y , and z direction, respectively. Figure 6b, d, and f show the anonymised raw data of Fig. 6a, c, and e. In the anonymisation procedure, we use laplace noise to hide the pattern information and the parameters in the noise generation are designed depends the specific patterns. In this work, the noise for different direction were derived based on the walking pattern.

4.3 Visual Data Anonymisation

For media data, as discussed in Section 3.3, in this work, we use an example that use proposed algorithm to anonymise a video clip from a transport surveillance system. In this work, we first use the VOC and car reg-number dataset trained a model that can be used in the system for recognising vehicle number, pedestrian, etc.

The Yolov3 darknet can conduct object detection. In this section we defined vehicle ‘**registration number**’ and ‘**facial feature**’ are two key privacy information that the data owner might do not like to share with others. Then, the darknet is used to detect ‘register plates’ or ‘face’ and then to run an image anonymisation algorithm to anonymise the bounding box on the ‘license plate’ or ‘face’. We use the proposed model to conduct anonymisation for a video clip, and Fig. 7 shows the anonymised plate number and facial features.

5 Conclusion and Discussion

The IoT systems are increasingly gaining importance in both daily life and industrial applications, which create substantial opportunities for users together with huge volume of variety data. This work introduced a data anonymisation schemes that can conduct data stream anonymisation before store or share data to other systems or organisation without leaking user privacy or other confidentiality. The experimental shows the proposed scheme can effectively anonymise data stream in IoT.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aceto, G., Persico, V., Pescapé, A. (2020). Industry 4.0 and health: internet of things, big data, and cloud computing for healthcare 4.0. *Journal of Industrial Information Integration*, 18, 100129.
- Amar, Y., Haddadi, H., Mortier, R. (2018). An information-theoretic approach to time-series data privacy. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems* (pp. 1–6).
- Da Xu, L., He, W., Li, S. (2014). Internet of things in industries: a survey. *IEEE Transactions on Industrial Informatics*, 10(4), 2233.
- Deldar, F., & Abadi, M. (2019). PDP-SAG: personalized privacy protection in moving objects databases by combining differential privacy and sensitive attribute generalization. *IEEE Access*, 7, 85887.
- Domingo-Ferrer, J., Soria-Comas, J., Mulero-Vellido, R. (2019). Steered microaggregation as a unified primitive to anonymize data sets and data streams. *IEEE Transactions on Information Forensics and Security*, 14(12), 3298.
- Du, M., Wang, K., Xia, Z., Zhang, Y. (2020). Differential privacy preserving of training model in wireless big data with edge computing. *IEEE Transactions on Big Data*, 6(2), 283.
- El Emam, K. (2020). Seven ways to evaluate the utility of synthetic data. *IEEE Security Privacy*, 18(4), 56.
- Faldum, A. (2007). On the trustworthiness of error-correcting codes. *IEEE Transactions on Information Theory*, 53(12), 4777.
- Gionis, A., & Tassa, T. (2009). k-Anonymization with minimal loss of information. *IEEE Transactions on Knowledge and Data Engineering*, 21(2), 206.
- Gope, P., & Sikdar, B. (2019). Lightweight and privacy-friendly spatial data aggregation for secure power supply and demand management in smart grids. *IEEE Transactions on Information Forensics and Security*, 14(6), 1554.
- Gorkhali, A., Li, L., Shrestha, A. (2020). Blockchain: a literature review. *Journal of Management Analytics*, 7(3), 321.

- Huang, H., Zhang, D., Xiao, F., Wang, K., Gu, J., Wang, R. (2020). Privacy-preserving approach PBCN in social network with differential privacy. *IEEE Transactions on Network and Service Management*, 17(2), 931.
- Khavkin, M., & Last, M. (2018). Preserving differential privacy and utility of non-stationary data streams. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 29–34).
- Li, J., Ooi, B.C., Wang, W. (2008). Anonymizing streaming data for privacy protection. In *2008 IEEE 24th International Conference on Data Engineering* (pp. 1367–1369).
- Li, S., Choo, K.R., Sun, Q., Buchanan, W.J., Cao, J. (2019). IoT forensics: amazon echo as a use case. *IEEE Internet of Things Journal*, 6(4), 6487.
- Li, S., Zhao, S., Yang, P., Andriotis, P., Xu, L., Sun, Q. (2019). Distributed consensus algorithm for events detection in cyber-physical systems. *IEEE Internet of Things Journal*, 6(2), 2299.
- Lu, Y., & Ning, X. (2020). A vision of 6G–5G's successor. *Journal of Management Analytics*, 7(3), 301.
- Ma, Y., Wu, Y., Li, J., Ge, J. (2020). APCN: a scalable architecture for balancing accountability and privacy in large-scale content-based networks. *Information Sciences*, 527, 511.
- Mahawaga Arachchige, P.C., Bertok, P., Khalil, I., Liu, D., Camtepe, S., Atiquzzaman, M. (2020). Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 7(7), 5827.
- Malekzadeh, M., Clegg, R.G., Cavallaro, A., Haddadi, H. (2019). Mobile sensor data anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation* (pp. 49–58).
- Neverova, N., Wolf, C., Lacey, G., Fridman, L., Chandra, D., Barbello, B., Taylor, G. (2016). Learning human identity from motion patterns. *IEEE Access*, 4, 1810.
- Otgonbayar, A., Pervez, Z., Dahal, K. (2016). Toward anonymizing IoT data streams via partitioning. In *2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)* (pp. 331–336).
- Ouazzani, Z.E., & Bakkali, H.E. (2018). A new technique ensuring privacy in big data: K-anonymity without prior value of the threshold k. *Procedia Computer Science*, 127, 52. <https://doi.org/10.1016/j.procs.2018.01.097>. <http://www.sciencedirect.com/science/article/pii/S187705091830108X>. Proceedings of the First International Conference on Intelligent Computing in Data Sciences, ICDS2017.
- Pervaiz, Z., Ghafoor, A., Aref, W.G. (2015). Precision-bounded access control using sliding-window query views for privacy-preserving data streams. *IEEE Transactions on Knowledge and Data Engineering*, 27(7), 1992.
- Phan, N., Wu, X., Hu, H., Dou, D. (2017). Adaptive laplace mechanism: Differential privacy preservation in deep learning. In *2017 IEEE International Conference on Data Mining (ICDM)* (pp. 385–394): IEEE.
- Rocher, L., Hendrickx, J.M., De Montjoye, Y.A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1), 1.
- Shi, Y., Zhang, Z., Chao, H.C., Shen, B. (2018). Data privacy protection based on micro aggregation with dynamic sensitive attribute updating. *Sensors*, 18(7), 2307.
- Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., Megías, D. (2017). Individual differential privacy: a utility-preserving formulation of differential privacy guarantees. *IEEE Transactions on Information Forensics and Security*, 12(6), 1418.
- Somolinos, R., Muñoz, A., Hernando, M.E., Pascual, M., Cáceres, J., Sánchez-de-Madariaga, R., Fragua, J.A., Serrano, P., Salvador, C.H. (2015). Service for the Pseudonymization of electronic healthcare records based on ISO/EN 13606 for the secondary use of information. *IEEE Journal of Biomedical and Health Informatics*, 19, 1937.
- Viriyasitavat, W., Da Xu, L., Bi, Z., Hoonsopon, D. (2019). Blockchain technology for applications in internet of Thing's mapping from system design perspective. *IEEE Internet of Things Journal*, 6(5), 8155.
- Wang, Y., Huang, M., Jin, Q., Ma, J. (2018). DP3: a differential privacy-based privacy-preserving indoor localization mechanism. *IEEE Communications Letters*, 22(12), 2547.
- Wang, H., Zhao, J., Li, J., Tian, L., Tu, P., Cao, T., An, Y., Wang, K., Li, S. (2020). Wearable sensor-based human activity recognition using hybrid deep learning techniques. *Security and Communication Networks*, 2020, 1–12.
- Xiao, J., Li, S., Xu, Q. (2019). Video-based evidence analysis and extraction in digital forensic investigation. *IEEE Access*, 7, 55432.
- Xu, L.D., Xu, E.L., Li, L. (2018). Industry 4.0: state of the art and future trends. *International Journal of Production Research*, 56(8), 2941.
- Yang, Y., Huang, S., Huang, W., Chang, X. (2020). Privacy-preserving cost-sensitive learning *IEEE Transactions on Neural Networks and Learning Systems*, 1–12.
- Yao, Z., Ge, J., Wu, Y., Jian, L. (2019). A privacy preserved and credible network protocol. *Journal of Parallel and Distributed Computing*, 132, 150.
- Yaseen, S., Abbas, S.M.A., Anjum, A., Saba, T., Khan, A., Malik, S.U.R., Ahmad, N., Shahzad, B., Bashir, A.K. (2018). Improved generalization for secure data publishing. *IEEE Access*, 6, 27156.
- Yli-Ojanperä, M., Sierla, S., Papakonstantinou, N., Vyatkin, V. (2019). Adapting an agile manufacturing concept to the reference architecture model industry 4.0: a survey and case study. *Journal of Industrial Information Integration*, 15, 147.
- Zhang, X., Dou, W., Pei, J., Nepal, S., Yang, C., Liu, C., Chen, J. (2015). Proximity-aware local-recoding anonymization with MapReduce for scalable big data privacy preservation in cloud. *IEEE Transactions on Computers*, 64(8), 2293.
- Zhang, C., & Chen, Y. (2020). A review of research relevant to the emerging industry trends: industry 4.0, IoT, blockchain, and business analytics. *Journal of Industrial Integration and Management*, 5(01), 165.
- Zhang, T., Zhu, T., Xiong, P., Huo, H., Tari, Z., Zhou, W. (2020). Correlated differential privacy: feature selection in machine learning. *IEEE Transactions on Industrial Informatics*, 16(3), 2115.
- Zhao, S., Li, S., Yao, Y. (2019). Blockchain enabled industrial internet of things technology. *IEEE Transactions on Computational Social Systems*, 6(6), 1442.
- Zhou, R., Zhang, X., Wang, X., Yang, G., Wang, H., Wu, Y. (2019). Privacy-preserving data search with fine-grained dynamic search right management in fog-assisted Internet of Things. *Information Sciences*, 491, 251.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Wang Ren he received the M.S. degree in Software Engineering from Sichuan University, Chengdu, China, in 2009, where he is currently pursuing the Ph.D. degree in communication and information system. His current research interests include information security and risk assessment.

Xin Tong she received the Ph.D. degree in information security from Beijing University of Posts and Telecommunications University in 2008. She is now a researcher of China Information Technology Security Evaluation Center. Her interests are network security risk evaluation and communication security.

Jing Du Associate Researcher, China Information Security Evaluation Center. Graduated from Xi'an University of Electronic Science and Technology with a master's degree. She is mainly engaged in party and government offices' and national important information systems' risk assessment, data security assessment and network security inspection of critical information infrastructure.

Na Wang Research Assistant, China Information Technology Security Evaluation Center. She has a bachelor's degree in computer science and technology. She is mainly engaged in risk assessment, product evaluation, network security inspection of critical information infrastructure and quality management of information system.

Shancang Li is an associate professor at the Department of Computer Science and Creative Technology at the University of the West of England, Bristol, UK. His research interests include IoT security, lightweight crypto, digital forensics, and cyber security.

Professor Geyong Min is the Chair and Director of High Performance Computing and Networking (HPCN) Research Group at the University of Exeter, UK. He received the PhD degree in Computing Science from the University of Glasgow, UK, in 2003, and the B.Sc. degree in Computer Science from Huazhong University of Science and Technology, China, in 1995. He joined the University of Bradford as a Lecturer in 2002, became a Senior Lecturer in 2005 and a Reader in 2007, and was promoted to a Professor in Computer Science in 2012. His main research interests include Next-Generation Internet, Analytical Modelling, Cloud Computing, and Big Data.

Zhiwei Zhao is currently an associate professor at College of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC). He got his PhD degree in College of Computer Science, Zhejiang University, China. His research interests include wireless and mobile computing, low power and networked systems, edge computing, embedded intelligence, future networks, etc. He is a member of CCF, ACM and IEEE.