UWS Academic Portal

X-BAND

Otgonbayar, Ankbayar; Pervez, Zeeshan; Dahal, Keshav

Published in:
IEEE Internet of Things Journal

DOI:
[10.1109/JIOT.2019.2955435](https://doi.org/10.1109/JIOT.2019.2955435)

Published: 22/11/2019

Document Version
Peer reviewed version

[Link to publication on the UWS Academic Portal](#)

Citation for published version (APA):
Otgonbayar, A., Pervez, Z., & Dahal, K. (2019). X-BAND: Expiration-band for anonymizing varied data streams. *IEEE Internet of Things Journal*, 7(2), 1438-1450. [8910431]. <https://doi.org/10.1109/JIOT.2019.2955435>

General rights

Copyright and moral rights for the publications made accessible in the UWS Academic Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please contact pure@uws.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

“© © 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

Otgonbayar, A., Pervez, Z., & Dahal, K. (2019). X-BAND: Expiration-band for anonymizing varied data streams. *IEEE Internet of Things Journal*, 7(2), 1438-1450. <https://doi.org/10.1109/JIOT.2019.2955435>

X-BAND: Expiration-band for anonymizing varied data streams

Ankhubayar Otgonbayar, Zeeshan Pervez, *Senior Member IEEE* and Keshav Dahal, *Senior Member IEEE*

Abstract—The Internet-of-Things (*IoT*) has formed a whole new layer of the world built on internet, reaching every connected devices, actuators and sensors. Many organizations utilize *IoT* data streams for research and development purposes. To make value out of these data streams, the data handling party must ensure the privacy of the individuals. The most common approach to provide privacy preservation is anonymization. *IoT* data provides varied data streams due to the nature of the individual's preference and versatile devices pool. The conventional single tuple expiration driven sliding window method is not adequate to provide efficient anonymization. Furthermore, minimization of missingness has to be considered for the varied data stream anonymization. Therefore, we propose *X-BAND* algorithm that utilizes the new expiration-band mechanism for handling varied data streams to achieve efficient anonymization, and we introduce weighted distance function for *X-BAND* to reduce missingness of published data. Our experiment on real datasets shows that *X-BAND* is effective and efficient compared to famous conventional anonymization algorithm *FADS*. *X-BAND* demonstrated 5% to 11% and 1% to 3% less information loss on real dataset *Adult* and *PM2.5* respectively while performing similar on clustering, comparable to re-using suppression and runtime. Also, the new weighted distance function is effective for reducing missingness for anonymization.

Index Terms—Anonymization, Data streams, Internet-of-Things, Missing data, *K*-anonymity

I. INTRODUCTION

RECENT developments of information technology and electronics has expanded the capability of automation to the whole new level called Internet-of-things (*IoT*). This innovation created an enormous network of internet enabled devices which are used to monitor and manage factories, buildings, machines and almost anything that can be combined with sensors, actuators and embedded computers [1]. Tremendous amounts of information are transferred in the form of data streams in the *IoT* data-sphere. Moreover, harvesting and utilizing data from this extensive data rich environment can help organizations to improve their performance [2]. For example, smart thermostat data can be analyzed to provide efficient heating plan for homes [3], smart car data to optimize traffic control of smart cities and safety of the drivers [4], [5].

Nevertheless, profiting from these versatile and private information without violating personal privacy is one of the primary concerns for *IoT* [6], [7]. Each individual's privacy is at risk when their identity is disclosed to service providers for analytical purposes, a malicious service provider can exploit

this information to covertly learn personal information by combining the data with other available or already comprised sources e.g., electoral register, information shared on deep web etc [8]. Therefore, data sanitization is compulsory for *IoT* data utilization [7], and the data anonymization is the well-known data sanitization approach to achieve privacy preservation [8]. This technique replaces or removes the information which can be exploited by an attacker to breach the privacy of individual, thus, ensuring the information of individual is published without the risk of privacy disclosure [9]–[13]. Confidential or identifier information which must not be published to the public domain is called sensitive information such as social security number, bank details and phone number. However, removing sensitive information does not make publishing data privacy preserved, there is a type of non-sensitive information called quasi-identifier which can lead to privacy leak. Quasi-identifiers (*QID* hereafter) are not sensitive on their own, however, can be exploited by attackers by combining or correlating with other quasi-identifiers to create unique identifier for privacy attack [14], [15]. Therefore, *QID* must be anonymized before publication. For example: in healthcare dataset, address, age, race, gender and disease are regarded as *QID*.

Anonymization replaces or removes the values of *QIDs* and creates an uncertainty to ensure the privacy of individuals [9]–[11]. The quality of anonymization is measured by information loss which measures the amount of uncertainty in the published information [16]. There are two major classes of anonymization; dataset anonymization and data stream anonymization [17]–[21]. Dataset anonymization is performed on pre-recorded data and the main objective of anonymizing dataset is to publish the dataset under specified privacy constraints with minimum information loss [22]. On the other hand, data streams are received in sequential order and required to be processed dynamically [21], [23]. However, the task of anonymizing data streams is to find the balance where the publication delay is minimum, while anonymizing with minimum information loss. The quality of anonymizing data streams is defined by a tradeoff between the information loss and the publication delay [24], [25]. Therefore, the data stream is required to be anonymized in a dynamic environment with fast and consistent publication [18], [19], [21], [26], [27].

The sliding window is the most adequate technique for handling such operations [17], [19]. This technique keeps data stream anonymization algorithms consistent and memory tolerant when dealing with fast and high dimensional data streams. There are two types of sliding window: count-based [17], [20] and time-based [18], [20], [21], [26], [27]. In count-based sliding window, data are accumulated in the size

A.Otgonbayar, Z.Pervez and K.Dahal were with the Artificial Intelligence, Visual communications and Network research centre, University of the West of Scotland, Paisley, Scotland, UK.
E-mail: {Ankhubayar.Otgonbayar, Zeeshan.Pervez, Keshav.Dahal}@uws.ac.uk

constrained window and the anonymization is performed when the count-window is full. In contrast, for time-based sliding window, data are accumulated in the time constrained window and the anonymization is performed when the oldest tuple in the sliding window is expiring.

In IoT, every participating individual has one or more number of internet enabled sensing or actuating devices (e.g., fitness watch, smart speakers and smart car) those are used at any time for any duration which are not bounded to globally defined usage pattern. Therefore, data streams from IoT contain missing values, and there are four main factors that cause data missingness in IoT data streams:

- **Individual's choice of devices:** individuals have different collection of devices depending of their preferences.
- **Different usage pattern:** individuals can choose to use any devices at any time.
- **User-defined information sharing control:** individuals can define the data sharing preferences of their devices.
- **Uncertain environment condition:** malfunction or loss of connection can occur under harsh uncontrollable condition of environment.

Therefore, we call IoT data stream as varied data stream due to the varying pattern of their QIDs those appear on each tuples of missing data streams generated from IoT [21], [26], [28]. Anonymizing data with missing values is an interesting area for researchers [28]–[32]. Analytics must identify the cause of missingness before processing the data for anonymization. There are three main types of missingness [29], [33], [34]:

- **Not Missing At Random (NMAR):** the missingness is occurring based on incidents that are evidential.
- **Missing At Random (MAR):** the missingness is occurring based on incidents that are occurred randomly.
- **Missing Completely At Random (MCAR):** the missingness is occurring at absolute random.

Moreover, researchers found three main technique to handle missing values for anonymization:

- **Imputation:** missing values are replaced with counterfeit values for anonymization [28], [32], [33].
- **Marginalization:** missing values are ignored for anonymization [26], [34].
- **Partitioning:** data is separated to independent partitions regarding their description [21], [35].

The biggest challenge of anonymizing varied data stream is to deal with the missing values while publishing with minimum, information loss, delay and missingness. Standard sliding window is an adequate solution for regular data stream anonymization which scans the data once then anonymizes considering only the data distributions [19], [20], [25], [36]. On the other hand, for varied data stream, the difficulty of anonymization is more challenging due to missing values. However, sliding window is designed to run an anonymization iteration when tuples are expiring. Each iteration finds the best anonymization for the single expiring tuple, and this type of method can result high information loss if the expiring tuple is an outlier, wasting valuable computational time for inefficient anonymization i.e., an anonymized tuple with high information loss. Therefore, we extend the sliding window

anonymization by introducing the expiration-band and propose framework for anonymizing varied data streams under time-based sliding window called *X-BAND* and introduce new distance function for anonymizing varied data streams with minimum missingness.

Fig. 1: General scheme of *X-BAND* framework

A general scheme of the *X-BAND* framework is illustrated in Fig. 1. The basic idea of *X-BAND* is to anonymize a cluster that results in a least information loss among the clusters that are created for each tuple in expiration-band under time based sliding window. This allows varied data stream to overcome the deficiency of the standard sliding window and optimize the anonymization of varied data streams. For clustering, we utilize K -nearest neighbor (*KNN* hereafter) [37] with weighted distance function that considers the *QIDs* similarity and attributes distance to minimize the missingness of clusters. As shown in Fig. 1, Γ is the size of expiration-band. When tuple in the sliding window expires, it runs clustering for all tuples in expiration-band using *KNN*. Then, the cluster generating the least information loss is anonymized and published. After publication, tuples of expiration-band those are received before any published tuple in the anonymization round is transferred to the time constrained temporary set called the pocket - this mechanism helps to publish varied data stream without distorting the arrival order of the tuples. Also, pocket gives more chance of publication for the expired tuples which are not selected in the clustering anonymization from the expiration-band. However, all anonymization round tries to utilize tuples of the pocket for anonymization. For efficiency, similar to other data stream anonymization approaches, tuples of pocket is attempted for anonymization using recently published K -anonymous clusters for each round. However, if the tuple expires in the pocket according to the time constraint of a pocket, then, it is suppressed and published. More details of the *X-BAND* algorithm, expiration band, pocket mechanism and distance function are explained in Section IV.

The experiments on real datasets *Adult* and *PM2.5* demonstrated the effectiveness of the expiration-band mechanism applied for varied data stream anonymization. Expiration-band significantly helped data streams to be anonymized with

less information loss while performing in comparable time compared to conventional varied data stream anonymization approach. Also, the effect of the new normalized weighted distance function for varied data streams are proving to be accurate in terms of forming clusters with less missing values.

The rest of the paper is organized as follows: Section II reviews the related work. Section III introduces the basic concept of varied data stream along with the definition of the expiration-band. Section IV explains the *X-BAND* algorithm in detail. Section V demonstrates the experiment result. Finally, paper is concluded in Section VI.

II. RELATED WORK

A. Data stream anonymization

Data stream is received in sequential order and it is to be anonymized on the fly. Therefore, researchers implemented dynamic technique called sliding window to perform data stream anonymization. There are two types of sliding window; count-based and time-based. The early attempts of the data stream anonymization exploited specialization tree. Wang et al. [38], propose the algorithm *SWAF* that builds specialization tree when tuple is received and run top-down greedy approach to find the best K -anonymous cluster for expiring tuple under time based sliding window. Also, in [39] Jianzhong et al., presented another specialization tree based algorithm *SKY*. This algorithm searches the most specific specialization tree node for newly arriving tuple, then it checks if the tree node is eligible to anonymize the newly arrived tuple. Otherwise, the tuple is held until their corresponding specialization tree node accumulates enough tuples to satisfy K -anonymity.

Zhou et al. [36], introduce three step generalization method for anonymizing data streams. In the first step, they have propose an algorithm with a randomization feature which makes publication decisions based on the information loss of cluster generalizations. In the second step, they enhance the publication decision by considering the distribution of the data stream. Finally, to further reduce the information loss, they examine the potential of publishing with future tuples for every tuple arriving. The authors consider that uncertainty based publication decision may be ineffective. In addition, they highlight that utilization of tuples distribution is vital for data stream anonymization and proposed a new feature to the algorithm that considers distribution of tuples. Their new feature prioritizes to publish tuples from sparsely populated area before tuples from densely populated area.

Cao et al., propose *CASTLE* data stream anonymization algorithm in [17]. *CASTLE* handles data stream under count-based sliding window. Newly arriving tuples are assigned to new or existing clusters depending on the cluster enlargement cost. The cluster enlargement is measured on the gain of information loss that occurred when adding new tuples to an existing cluster. In *CASTLE*, information loss of any cluster must not exceed τ , therefore, a new cluster is created if none of the clusters found generate less than τ information loss to accommodate the tuple. According to count-based sliding window, tuple expires when the size of sliding window reaches δ . Each expiring tuple is published instantly if its residing cluster satisfies K -anonymity, otherwise merge and split operation

is called to create K -anonymous cluster for publication of the expiring tuples. Moreover, to reduce anonymization cost, *CASTLE* utilizes re-using strategy to anonymize newly arriving tuples using generalization information of recently published K -anonymous clusters.

Hessam and Sylvia introduce *FAANST* [20], a count-based sliding window algorithm for numeric data streams. The main purpose of this algorithm is to enhance data quality. Therefore, they introduced Δ , information loss constraint of clusters. When sliding window is full, *FAANST* uses K -means algorithm to create clusters then it publishes K -anonymous clusters those having less than Δ information loss. Re-using strategy is used to reduce information loss. *FAANST* outperform *CASTLE* in terms of time complexity and information loss.

Wang et al. [18], identify that *CASTLE* [17] created few huge clusters when applied on data streams, causing frequent split operations. The merge and split operations are costly, in terms of time complexity and information loss. To address this issue, they propose *B-CASTLE*, they set a threshold on cluster size and applying correlation distance to select merging clusters. *B-CASTLE* shows higher quality anonymization in shorter time.

Guo et al. [19], propose time based sliding window algorithm *FADS* to resolve the issues of *CASTLE*. They identify overload of clusters in *CASTLE* when receiving homogeneous data streams having non-negligible time difference between arriving tuples. Also, they mention that complicated merge and split operations of *CASTLE* are not essential for data stream anonymization since the size of the generated cluster is fixed by K . They solve these issues by setting time constraints for sliding window, utilizing re-usable clusters in the memory, and by running *KNN* on expiring tuples to form a K -anonymous cluster.

Adorenke et al. [40], mention that the high information loss can occur when data stream is intermittent. They propose an adaptive buffer resizing anonymization scheme for resource constrained environment to anonymize streaming crime data. They predict the data arrival rate to resize the buffer to minimize the information loss. For the calculation of data arrival rate they utilize Poisson probability model [41] by assuming the data streams as a sequence of events occurred in a fixed time interval.

B. Missing data handling

Missing data handling is one of the most important topics of data scientists [28], [29], [33], [34].

Ciglic et al. [35], mentioned that the term indistinct matching for the missing (*NULL*) values are not explicitly defined in the context of the K -anonymity. Therefore, they defined three *NULL* matching schemes for the dataset containing missing values which are:

- **Basic match:** *NULL* values do not match with *NULL* values, nor with any other value.
- **Extended match:** *NULL* values match with *NULL* values.
- **Maybe match:** *NULL* values match with any value, including *NULL* values.

Moreover, there are three main problems that missing data causes: creates substantial amount of information bias, makes data handling and analysis formidable, and inefficiency [42]. Imputation, marginalization and partitioning approaches are predominantly used to address the aforementioned issues.

Imputation: Imputation is a well-known method for handling missing values in statistical analysis. During imputation, missing values of varied data stream are replaced by pre-calculated representative values. Imputation is classified into two types: single imputation [43], [44] and multiple imputation [45], [46]. Single imputation replaces missing data with other values once. In contrast, multiple imputation consolidates the analysis of the multiple different imputations of the missing values [46]. Imputation helps to repair the missing values of the data for the anonymization and amplifies the uncertainty of anonymized data creating more information loss. If the imputation amount for a cluster is less compared to its size, then anonymization with imputation results in more secure privacy preservation while repairing the missing values. However, imputation based anonymization cannot tolerate high amount of missingness in the data stream. Because, the uncertainty of anonymized data makes more challenging to analyze due to the high amount of imputation noise.

Marginalization: In marginalization, missing values are ignored and processed as a *NULL*, and it is anonymized as part of range attributes and node of the generalization hierarchy tree of categorical attributes [35]. The major problem of the marginalization based approaches is the excessive amount of missing values in the published clusters. However, this method helps to publish data without adding or removing any information to the original data [34], therefore, moderate use of this method is the best option for handling missing values. To achieve better anonymization for varied data streams, the focus should be on minimizing information loss and missingness.

Partitioning: Dataset with missing values can be divided to multiple complete datasets, and each dataset can be anonymized by conventional anonymization approaches. Partitioning method is not cost efficient when dataset has excessive amounts of missing values compared to their size [21]. Nevertheless, this technique helps us to publish solid clusters without any missing values while ensuring a privacy preservation. Ciglic et al. [35], utilized *Basic-match* to *K*-anonymize dataset containing *NULL* values, and proposed a partitioning based anonymization algorithm *ANON*. In this algorithm, dataset is split into separate partitions based on tuples' attribute description, and then best-first-search [47] is executed to find the optimal anonymization for each partition.

Moreover, suppression-based methods can be called as anonymization algorithm which suppresses tuples' attribute to provide privacy preservation. Wang et al. [48], propose two transactional data stream anonymization approach based on generalization and suppression to satisfy ρ -uncertainty [49]. They identify that, addition or deletion of transaction record may break the ρ -uncertainty [49] for the sliding window. Therefore, they introduce affected sensitive rule trie to overcome the above mentioned issue.

III. ANONYMIZING VARIED DATA STREAM

In this section, we introduce the concept of varied data stream anonymization under sliding window with expiration-band. In conventional data streaming, each tuple has same fixed attributes with no missing values; whereas, in varied data stream one or more attributes can be missing. Therefore, tuples of varied data streams contain varying combination of *QIDs*.

Definition 1 (Quasi-identifiers of data stream) Suppose $q_1, q_2, \dots, q_n$ are the known quasi-identifiers attributes of data stream. Then, quasi-identifiers set of data stream is defined as follows: $Q = \{q_1, q_2, \dots, q_n\}$.

Definition 2 (Tuple of data stream) Assume, id_t is identity of an individual who is feeding their information to the data stream, $Q_t = \{q_1, q_2, \dots, q_m\}$ be a set of quasi-identifiers' information arrived from id_t and ts_t be a arrival timestamp of the information. Therefore, tuple of data stream received at timestamp ts_t from id_t is defined as $t(id_t, Q_t, ts_t)$.

In traditional data streams, tuples are received without any missing values. In contrast, for varied data streams, tuples are received with one or more missing values. Therefore, varied data stream is defined in the following definition.

Definition 3 (Varied data stream) Let quasi-identifier set $Q = \{q_1, q_2, \dots, q_n\}$ be the total pool of quasi-identifiers those value might be received in the tuple. The definition of the varied data stream is defined as $VS(id, Q_t, ts_t)$ where id is the identity information of individual, Q_t ($Q_t \subseteq Q$) is the quasi-identifier set of tuple received at the timestamp ts_t .

Definition 4 (Cluster of varied data streams) Let S_c be a set of tuples of the varied data stream VS and Q_c denotes the total set of quasi-identifiers those appeared in the tuples of S_c . Therefore, a cluster C of varied data stream VS is $C(Q_c) = \{t(id_t, Q_t, ts_t) \mid t \in S_c \wedge Q_t \subseteq Q_c\}$

In [50], the definition of the *K*-anonymity is stated as: Each release of data must be such that every combination of values of *QIDs* can be indistinctly matched to at least *K* individuals. Based on the original definition of the *K*-anonymity, we define the *K*-anonymous cluster of the varied data streams as follows:

Definition 5 (*K*-Anonymous cluster) Let $C(Q_c)$ be a cluster generated from a varied data stream VS and let $\sim$ be a match predicate on $C(Q_c)$. $C(Q_c)$ is called a *K*-anonymous cluster with respect to $\sim$ when $\forall t \in C(Q_c) : |\{t' \mid t \sim t'\}| \geq K$.

Definition 6 (*K*-anonymized varied data stream) Let $VS(id, Q_t, ts)$ be a varied data stream, and VS_{out} be an anonymization of varied data stream VS . If following conditions are met, then we call VS_{out} is *K*-anonymized varied data stream: a)

- 1) For $\forall t \in VS, \exists t' \in VS_{out}$ corresponds to t .
- 2) For $\forall t' \in VS_{out}, DI(C(Q'_t)) \geq k$, when $C(Q'_t)$ is a *K*-anonymous cluster containing t' which belongs to

VS_{out} . DI counts the number of distinct values of the tuples' ids in $C(Q'_t)$.

Definition 7 (Maybe-match) Let $t_1(pid, Q_1)$ and $t_2(pid, Q_2)$ be a tuple of varied data stream VS , and $Q_m = Q_1 \cup Q_2$. Then the maybe-match is defined as:

$$t_1 \sim_m t_2 \iff \forall q \in Q_m : t_1[q] = t_2[q] \vee (t_1[q] \text{ is } NULL \vee t_2[q] \text{ is } NULL) \quad (1)$$

The generalization of varied data stream clusters must consider the different set of $QIDs$ of its tuples, whereas traditional data stream anonymization only considers values of its $QIDs$. Therefore, cluster generalization of varied data stream must be able to handle missingness while maintaining the anonymity. Therefore, we define following cluster generalization for varied data stream clusters based on the *Maybe-match* (see definition 7) defined in 8. As discussed in [35], maybe match based K -anonymity is vulnerable against hampering reconstruction and $NULL$ identifier attack; however, in varied data stream, it is impossible to identify the reason of the missingness of each receiving tuples due the causes explained in the Section I.

TABLE I: Tuples' selection to create a 3-anonymous cluster

No	Age	Gender	Height	Weight
t_1	30	Male	172	75
t_2	26	Male	165	null
t_3	null	Female	171	69

Definition 8 (Maybe-match based cluster generalization)

Suppose $G_c^*(G_1, G_2, \dots, G_m)$ is a generalization of the varied data stream cluster $C(Q_m)$, then, we define the maybe match based generalization of each QID of Q_m as follows.

- 1) $g_i = [r_{i.min}, r_{i.max}]$, where $r_{i.min}(r_{i.max})$ is the minimum(maximum) of the values of all tuples in C that have attribute values on q_i . Where q_i is a numerical attribute.
- 2) $g_i = H_{i.lowest}$ where $H_{i.lowest}$ is the lowest common ancestor of the v_{q_i} values of the tuples in cluster C that have values on q_i . If q_i is a categorical attribute.

TABLE II: Maybe match based 3-anonymous cluster after generalization

No	Age	Gender	Height	Weight
t_1	[26-30]	Gender	[165-172]	[69-75]
t_2	[26-30]	Gender	[165-172]	null
t_3	null	Gender	[165-172]	[69-75]

Let us explain the cluster generalization for tuples with missing data. Table I contains three tuples that are grouped into a 3-anonymous cluster, *null* denotes the missing value. t_1 has no missing value, t_2 and t_3 have missing value on Age and Weight $QIDs$ respectively. The generalization hierarchy of Gender is defined in [16]. According to definition 8, generalization of each QID is calculated as follows. $G_{Age} = [26 - 30]$, $G_{Gender} = \text{Gender}$, $G_{Height} = [165 - 172]$ and $G_{Weight} = [69 - 75]$. Therefore, cluster generalization of t_1, t_2

and t_3 is $G_{TableI}([26 - 30], \text{Gender}, [165 - 172], [69 - 75])$. In the Table II, generalized version of Table I is illustrated.

Definition 9 (Information loss of tuple) The calculation of information loss occurred during generalization of $t(pid, Q_t)$ to $G_t(g_1, g_2, \dots, g_m)$ is defined as:

$$InfoLoss(t, G_t) = \frac{1}{|G_t|} \left(\sum_{q_i \in Q_t} Loss(q_i^g) \right) \quad (2)$$

Where $Loss(q_i^g)$ is the information loss of quasi-identifier q_i caused by the generalization g_i , which is defined as:

$$Loss(q_i^g) = \begin{cases} \frac{r_{i.u} - r_{i.l}}{R_{i,u} - R_{i,l}} & \text{if } g_i \in [r_{i.l}, r_{i.u}] \\ \frac{|leaves(H_i)| - 1}{|leaves(DGH_i)| - 1} & \text{if } g_i = H \end{cases} \quad (3)$$

Where $[r_{i.l}, r_{i.u}]$ is the value domain of a numeric attribute q_i , DGH_i is the domain graph hierarchy (DGH) of a categorical attribute q_i , $|leaves(H_i)|$ and $|leaves(DGH_i)|$ is the number of nodes of a tree rooted on H_i and DGH_i .

The quality of the anonymization algorithm is measured by the average information loss caused by the anonymization of the data stream. We define it as:

Definition 10 (Average information loss) The average information loss of anonymizing first N tuples of varied data stream is defined as:

$$AverageInfoLoss(N) = \frac{1}{N} \sum_{i=1}^N InfoLoss(t_i, G_i) \quad (4)$$

Where G_i is generalization of a tuple t_i .

IV. X-BAND: EXPIRATION-BAND FOR ANONYMIZING VARIED DATA STREAMS

Anonymizing varied data streams is more challenging than anonymization of regular data streams. Regular data streams anonymization approaches are only performed to publish the cluster with the least information loss, and this is easier to accomplish when only distribution of data is considered for the clustering. Therefore, single scanning of tuples is enough for the processing. In contrast, for varied data streams, distortion of missing values can affect the performance of clustering and multiple scanning of the tuples is required to achieve efficient anonymization. The main idea of expiration band is to extend the mechanism of expiration tuple in sliding window into expiration band to create an opportunity of multiple scanning to find the best cluster for the anonymization. The size of expiration band is controlled by the constraint Γ . The expiration band trades small amount of calculation time for better anonymization. Unlike traditional sliding window approaches, to avoid extensive number of suppressions, pocket structure is utilized to temporary hold the expired tuples from expiration window for the future clustering. This pocket-ing strategy gives a second chance to any expired tuple of expiration band. In the following, the inner workings of expiration band and pocket are explained.

Definition 11 (Expiration band Γ) . Let δ be a time constraint for sliding window. When the oldest tuple of the sliding window is expiring according to δ , then Γ number of the oldest tuples are automatically become the tuples of the expiration band. For example, if $\Gamma=1$ then only the expiring tuple is considered as the expiration band.

A. Distance function

As we discussed earlier, multiple scanning is required to perform better anonymization. Moreover, only measuring attribute value distance between two tuples is not adequate for the varied data stream anonymization. Therefore, *QIDs*' dissimilarity must be also used for measuring distance for *KNN* to minimize the missingness of *K*-anonymous clusters. We utilize the Jaccard distance [51] for calculation of the dissimilarity of two tuples. In the following definition, the distance function of two tuples in varied data stream anonymization is defined using both, attribute value distance and Jaccard distance. We used normalized weight function to control and balance the influence of the core metrics of distance function.

Definition 12 (Attribute values distance) The attribute value distance between two tuples $t_1(pid, Q_1)$ and $t_2(pid, Q_2)$ is calculated on common *QIDs* both t_1 and t_2 posses.

$$AttributeDistance(t_1, t_2) = \frac{\sum_{q_i \in |Q_1 \cap Q_2|} d_i(q_i)}{|Q_1 \cap Q_2|} \quad (5)$$

$$d_i(q_i) = \begin{cases} \frac{|r_{i,1} - r_{i,2}|}{|R_{i,u} - R_{i,l}|} & \text{if } q_i \text{ is numerical} \\ \frac{|\text{leaves}(H_i)| - 1}{|\text{leaves}(DGH_i)| - 1} & \text{if } q_i \text{ is categorical} \end{cases} \quad (6)$$

Definition 13 (Jaccard distance) Let t_1 and t_2 be the tuples having attributes on Q_1 and Q_2 respectively. Their Jaccard distance is calculated as:

$$Jaccard(t_1, t_2) = \frac{|Q_1 \cup Q_2| - |Q_1 \cap Q_2|}{|Q_1 \cup Q_2|} \quad (7)$$

Definition 14 (Distance between two tuples) The distance between two tuples $t_1(pid, Q_1)$ and $t_2(pid, Q_2)$ is defined as follows:

$$Distance(t_1, t_2) = \alpha \times AttributeDistance(t_1, t_2) + \beta \times Jaccard(t_1, t_2) \quad (8)$$

where $\alpha + \beta = 1$, α and β are user defined weights.

B. X-BAND algorithm

The details of *X-BAND* are given in Algorithm 1. *X-BAND* has five parameters: the varied data stream *VS*, *K*-anonymity degree *K*, sliding window time constraint δ , time constraint for re-using *K*-anonymized clusters ω and the size of expiration-band Γ . Firstly, *X-BAND* reads tuples from *VS* and stores in the buffer S_t with received timestamp attached and updates each range of numeric *QIDs*. Then, if the oldest tuple of S_t is expiring, *X-BAND* removes the generalization information of *K*-anonymous clusters those existed more than ω from the recently published re-usable *K*-anonymous cluster

set S_k , and invokes the procedure *TriggerPublish()* for publication. The details of procedure *TriggerPublish()* are described in Algorithm 2. When no tuples are received, *X-BAND* calls *TriggerPublish()* to publish remaining tuples in S_t . Finally, the remaining tuples of the pocket P_t are processed by *SuppressAnonymization(t)*. The details of procedure *SuppressAnonymization(t)* are shown in Algorithm 3.

Algorithm 1 *X - BAND*(*VS*, *K*, δ , ω , Γ)

```

1: Let  $S_t$  be a set of tuples that acts as buffer, initialized empty;
2: Let  $S_k$  be a set of K-anonymous re-usable clusters which are expiring in  $\omega$ , initialized empty;
3: Let  $P_t$  be a pocket of tuples expires in  $\delta$ , initialized empty;
4: while VS  $\neq$  NULL do
5:   Read a tuple  $t_n$  from VS and insert it to  $S_t$ ;
6:   Remove re-usable clusters that exist longer than or equal to  $\omega$ ;
7:   if Oldest tuple in buffer is expiring then
8:     TriggerPublish();
9:   end if
10: end while
11: while  $S_p \neq$  NULL do
12:   TriggerPublish();
13: end while
14: while  $P_t \neq$  NULL do
15:    $t$  be the oldest tuple of  $P_t$ .
16:   SuppressAnonymization(t);
17: end while

```

The details of procedure *TriggerPublish()* are presented in Algorithm 2. If the size of S_t is not less than $K + \Gamma$ then it is possible to find *K*-1 neighbours for each tuple of the expiration-band (see Definition 11). Therefore it forms *K*-anonymous cluster for each tuple of expiration-band using *KNN* and then chooses the cluster with minimum information loss for publication. In contrast, if size of the S_t is within the range of $[K, K + \Gamma]$ then it forms *K*-anonymous clusters for the oldest $|S_t| - K$ number of expiring tuples using *KNN*, then publishes the cluster that results in a minimum information loss. At the end of any new *K*-anonymous cluster's publication, the tuples those are received before any anonymized tuples of expiration-band are transferred from S_t into P_t which will act as a pocket. This helps to regulate anonymization of data stream by preventing multiple expiration on a single tuple and maintains the sequential order in approximate deviation. To reduce information loss, generalization information of each published *K*-anonymous cluster is inserted to S_k for re-using purposes.

Finally, if the size of S_t is not eligible to produce *K*-anonymous cluster for expiring tuple - this occurs when the data stream is ended or interrupted for uncertain duration. *SuppressAnonymization()* (see Algorithm 3) is called to publish the expiring tuple by re-using generalization information of recently anonymized clusters, otherwise, the expiring tuple is suppressed. The suppression is the extreme form of anonymization that removes all the values of a tuple - suppression means that the values of a tuple are completely

Algorithm 2 *TriggerPublish()*

```

1:  $C_{gen}$  be the set of cluster
2: if  $|S_t| \geq K + \Gamma$  then
3:   for each tuple  $t'$  of expiration-band do
4:     Create cluster  $C_{new}$  by using KNN on both  $S_t$  and
        $C_t$  and add it to  $C_{gen}$ 
5:   end for
6: else if  $K \leq |S_t| < K + \Gamma$  then
7:   for each tuple  $t'$  of expiration-band do
8:     Create cluster  $C_{new}$  by using KNN on both  $S_t$  and
        $C_t$  and add it to  $C_{gen}$ 
9:   end for
10:  $C_{best}$  be the cluster from  $C_{gen}$  that has the minimum
    information loss
11: Publish  $C_{best}$  and add it to  $S_k$ .
12: Remove tuples of  $C_{best}$  from  $S_t$ .
13: Move tuples from expiration-band those are received
    before any published tuple to  $P_g$ .
14: else
15:   SuppressAnonymization(t)
16: end if
17: if  $|P_t| \neq NULL$  & oldest tuple  $t$  stayed more than  $\delta$  in
     $P_t$  then
18:   SuppressAnonymization(t).
19: end if

```

unknown to the data handler. For *X-BAND*, it only performed on the expired tuples of the pocket that did not get published by re-use anonymization.

Algorithm 3 *SuppressAnonymization(t)*

```

1: Find  $K$ -anonymous cluster  $C_k$  from  $S_k$  which covers  $t$ 
   with minimum information loss
2: if  $C_k$  found then
3:   Use cluster generalization of  $C_k$  to publish  $t'$ ;
4: else
5:   Suppress  $t'$  and publish
6: end if
7: Remove  $t'$  from  $P'$ 

```

V. EXPERIMENTAL EVALUATION

In order to measure the performance of *X - BAND* algorithms, we compared it with *FADS* [19]. The *FADS* is the fastest and well-known algorithm for anonymizing data streams. To run *FADS* on varied data streams, we implemented marginalization for *FADS* for handling missing values of varied data streams. We used *Adult*¹ and *PM2.5 Data of Five Chinese cities*² datasets for evaluation. *Adult* dataset is widely used to evaluate the anonymization algorithms [17]–[21] and it has fourteen *QID*s containing eight categorical and six numeric attributes which are: *education*, *marital-status*, *work-class*, *occupation*, *relationship*, *race*, *gender*, *country* and *age*, *final-weight*, *education-number*,

capital-gain, *capital-loss*. The generalization hierarchy of eight categorical attributes are defined in [16], and summary of *Adult* dataset is described in Table III. We used 30,000 tuples of *Adult* dataset for the experiment.

TABLE III: *QID* descriptions of *Adult* dataset

Attribute name	Type	Range	
		Min	Max
Age	Numeric	17	90
Final-weight	Numeric	13769	1484705
Education-number	Numeric	1	16
Capital-gain	Numeric	0	99999
Capital-loss	Numeric	0	4356
Hours-per-week	Numeric	1	99
		Hierarchy tree	
		Height	Nodes
Education	Categorical	5	26
Marital-status	Categorical	4	11
Work-class	Categorical	5	13
Country	Categorical	4	62
Occupation	Categorical	3	15
Relationship	Categorical	3	7
Rage	Categorical	3	6
Gender	Categorical	2	3

The *PM2.5* Data of Five Chinese cities (*PM2.5* hereafter) is a meteorological data recorder in Chinese Five cities using IoT sensors. We used each dataset of five cities for the experiment. The selected *QID*s of *PM2.5* dataset consist of two categorical and ten numerical attributes which are: *season*, *wind-direction* (*combined-wind-direction*) and *first-post*(*PM2.5*), *second-post*(*PM2.5*), *third-post*(*PM2.5*), *dew-point*, *temperature*, *humidity*, *pressure*, *wind-speed*(*cumulated-wind-speed*), *h-precipitation*(*hourly-precipitation*), *c-precipitation*(*cumulated-precipitation*). The generalization hierarchy of two categorical attributes are defined in [21], a summary of *PM2.5* dataset is described in Table IV. For the experiment on *PM2.5* data, we utilized the original data consist of over 250,000 tuples across the five separate files which contains naturally occurred missing values.

TABLE IV: *QID* descriptions of *PM2.5* dataset

Attribute name	Type	Range	
		Min	Max
First-post	Numeric	1	1528
Second-post	Numeric	1	940
Third-post	Numeric	1	968
Dew-point	Numeric	-40	28
Temperature	Numeric	-25	41
Humidity	Numeric	2	100
Pressure	Numeric	975	1042
Wind-speed	Numeric	0	608
H-precipitation	Numeric	0	61.6
C-precipitation	Numeric	0	226.4
		Tree	
		Height	Nodes
Season	Categorical	3	8
Wind-Direction	Categorical	2	5

To imitate consistent and continues data flow of data streams each tuple is retrieved from dataset with the delay of 500ms. All algorithms are implemented in *Java*. The experiments are conducted on a PC with Intel® Core™ i7-6700HQ with 16GB RAM and *Windows 10x64* with *JDK8.0*.

¹<https://archive.ics.uci.edu/ml/datasets/adult>

²<https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities>

Along the lines of the previous works [19], [21], [26], we have selected the following parameters for the experiment that are shown in Table V. In the table, K represents the K -anonymity, δ is the time constraint of sliding window, ω is the time constraint for the pocket and re-usable K -anonymous clusters. In the experiment results graphs, for $X\text{-BAND}(x)$, x indicates the expiration-band parameter Γ .

TABLE V: Experiment parameters

Algorithm	Parameters
$K\text{-VARP}$	$K=50, \delta=1000, \omega=1000$
$X\text{-BAND}$	$K=50, \delta=1000, \omega=1000$

We performed comprehensive analysis on the impact of a increased Γ in anonymization quality and runtime for $X\text{-BAND}$ to select the Γ values for the experiment. In the experiment evaluation, Γ is set $\{4, 6, 8, 10\}$ and $\{2, 4, 6\}$ for *Adult* and *PM2.5* dataset respectively. To study the effect of the distance measurement for KNN (Definition 14) we executed each experiment on distance function with *four* varying weights of $\alpha\{1.0, 0.75, 0.5, 0.25\}$ and $\beta\{0.0, 0.25, 0.5, 0.75\}$ respectively.

A. Information loss

In Fig. 2(a) we illustrated the information loss of *FADS* and *X-BAND* having four different expiration-band sizes respectively *four*, *six*, *eight* and *ten* executed on *four* different distance function settings on *Adult* dataset. The average information loss is calculated using equation (4). *FADS* demonstrated higher information loss compared to any variations of *X-BAND*. Due to the nature of the expiration-band mechanism, information loss of *X-BAND* is reduced when Γ is increased. Moreover, we can see gradual increase of difference in information loss between *FADS* and *X-BAND* when the missingness of data is increasing. *X-BAND*(10) produced 5% less information loss compared to *FADS* when data have at most one missing value in their tuples, whereas *X-BAND*(10) showed 11% less information loss when the data have at most 10 missing values in their tuples. Moreover, increase of β leads to higher information loss of anonymization for both algorithms and this shows the effect of *QIDs* dissimilarity aspect of the distance function.

In Fig. 2(b) we demonstrated the information loss of *FADS* and *X-BAND* having three different expiration-band sizes respectively *two*, *four* and *six* executed on *PM2.5* data with four distance function settings on *PM2.5* dataset. Overall, *FADS* demonstrated slightly higher information loss compared to the three difference variations of *X-BAND*. The average difference of information loss between *FADS* and *X-BAND*(6) is around 3%. Also, similar to the experiment on *Adult* dataset, increase of α causes lesser information loss. Moreover, on *PM2.5* data, overall information loss is around 18% to 21% on four cities data: *Beijing*, *Chengdu*, *Shanghai* and *Shenyang* whereas on *Guangzhou* city's data, information loss is around 11% to 13%. The experiment on *Adult* and *PM2.5* show substantial difference in information loss, and this is caused by the *QID* composition, data distribution and amount of missingness in the data. *Adult* dataset contains eight categorical

attribute whereas *PM2.5* data has two categorical attribute, thus, generalization of the tuple creates more uncertainty due to the high number of categorical attribute on *Adult* dataset. In addition, *PM2.5* contains less missingness compared to the *Adult* which leads more precise distance measurement, and this creates clusters with less information loss.

B. Clustering

FADS and *X-BAND* are designed to create a cluster when buffer and pocket have enough tuples to produce K -anonymous clusters. In Fig. 3(a) and Fig. 3(b), the number of clusters for *FADS* and *X-BAND* are demonstrated by the varying expiration-band size and missingness for *Adult* and *PM2.5* dataset respectively. For *Adult* dataset, *FADS* creates more number of clusters compared to the most variations of *X-BAND*. On the other hand, for *PM2.5* dataset, *FADS* creates almost identical or less number of clusters. There are numerous factors that affected the number of clustering, these are pocketing strategy of expiration-band and time complexity of single iteration of anonymization on *X-BAND* and data distribution and missingness. The expiration-band mechanism is designed to publish clusters with less information loss by transferring few expired tuples to the pocket for effective anonymization. During one iteration of anonymization, *X-BAND* outputs one K -anonymous cluster and moves few tuples to the pocket that have high potential of suppression, in contrast, *FADS* only outputs one K -anonymous cluster or single tuple which is published by re-using or suppression. Also, less missingness leads to more compact anonymization, therefore, number of clustering is almost identical for *PM2.5* dataset.

The total number of suppression for *FADS* and *X-BAND* are demonstrated in Fig. 4(a) and Fig. 4(b) for *Adult* and *PM2.5* datasets. From Fig. 4(a) we can identify that *FADS* suppressed minimum number of tuples compared to the variations of *X-BAND* on both graphs. However, overall figure of the graphs show almost identical trend in Fig. 4. As we discussed, *FADS* does not suppress tuple unless the sliding window is not eligible to produce K -anonymous cluster, and this occurs only when data streams is interrupted or ended. In our experiment, data is received without any external interruption until the end, and most suppression for *FADS* occurred in the end of the anonymization. On the other hand, pocket tuples must be suppressed after staying for ω time, and this caused high number of suppression for *X-BAND*.

C. Clusters purity on Adult dataset

The *Adult* dataset contains significantly higher number of missingness compared to *PM2.5* dataset. Therefore, we evaluate cluster purity analysis on only *Adult* dataset. In Fig. 5(a) and Fig. 5(b) we illustrate the heatmap of cluster purity percentage on at most 6, 8 and 10 missing on *Adult* dataset for *FADS* and *X-BAND* respectively. *FADS* and *X-BAND* show almost identical result in terms of cluster purity in Fig. 5(a) and Fig. 5(b). It is expected that the less missingness of data leads to better data purity of clusters. The cluster purity for both *FADS* and *X-BAND* is 70% to 90% on dataset with

Fig. 2: Information loss for *FADS* and *X-BAND*

tuples having at most 6 missing values while having 40% to 70% on the dataset with at most 10 missing values in their tuples. Nevertheless, the increase of the β cause less amount of missingness in published clusters, and this proves the effect of the distance function (see Definition 14) to anonymize and publish varied data streams with minimized missingness. Also, there are few clusters having remarkably low purity when $\beta = 0.75$, and these clusters are created around the end of the clustering which are performed on the remaining tuples that have high percentage of missingness.

D. Re-Using of K -anonymous clusters

Fig. 6(a) and Fig. 6(b) demonstrates the number of tuples that are anonymized by re-using generalization information of recently published K -anonymous cluster on *Adult* and *PM2.5* datasets respectively. On *Adult* dataset, *FADS* show similar number of re-using compared to the several variations of the *X-BAND*. However, *FADS* re-use anonymized less number of tuples when α is decreased. On the other hand, *X-BAND* showed decrease of re-use when α is decreased from 1 to 0.5 and show substantial increase when $\alpha = 0.25$ and $\beta = 0.75$. This notable difference in re-use is expected when merging is more focused on the tuples similarity than the attributes distance. Also, for *PM2.5* dataset, *FADS* demonstrate substantial difference in re-use anonymization around 600 whereas three variations of *X-BAND* re-used average 50 recently anonymized cluster for re-use anonymization. The main reason of this substantial difference in re-use is that the information loss on *Adult* is relatively higher for all algorithms compared to the information loss occurred on *PM2.5* dataset. On *Adult* dataset

all algorithms show information loss between 0.5 to 0.72 depending on the distance function settings, in contrast, on *PM2.5* algorithms showed approximately 0.13 on *Guangzhou PM2.5* dataset and 0.2 on other four cities' *PM2.5* datasets. Therefore, clusters formed from *Adult* is not tightly clustered compared to *PM2.5* and this reduces the chance of re-use anonymization, thus, causes less number of re-using on *PM2.5* dataset.

E. Runtime

The runtime of the *FADS* and *X-BAND* are illustrated on Fig. 7(a) and Fig. 7(b). Overall, on *Adult* dataset, *FADS* shows very stable runtime for all the experiments, *X-BAND*(4) and *X-BAND*(6) spend approximately 2% less time compared to *FADS*. On the other hand, *FADS* uses approximately 4% less time compared to *X-BAND*(8) and *X-BAND*(10). Increase of the size of expiration-band Γ causes more number of runs of *KNN* to publish single K -anonymous cluster and this results in more runtime for *X-BAND*. On the other hand, *FADS* executes only one *KNN* to publish a K -anonymous cluster and this leads to less runtime for *FADS*. However, on *PM2.5* dataset, *FADS*, *X-BAND*(4) and *X-BAND*(6) show similar figure on runtime. In contrast, *X-BAND*(2) shows significant difference in runtime showing approximately 74000ms when $\alpha = 0.25, 0.5, 0.75$.

VI. CONCLUSION

In this paper, we presented an algorithm *X-BAND* with sliding window anonymization scheme expiration-band along with weighted distance function for anonymizing varied data

(a) Number of clusters created for varying missingness and distance settings on *Adult* (b) Number of clusters created for varying distance settings on *PM2.5*Fig. 3: Number of clusters created for *FADS* and *X-BAND* (a) Number of suppression for varying missingness and distance settings on *Adult* (b) Number of suppression for varying distance settings on *PM2.5*Fig. 4: Number of suppressions for *FADS* and *X-BAND*

streams with minimum missingness. *X-BAND* can be utilized to exploit varied data streams generated from IoT environment having uncontrollable missingness such as smart parking and smart healthcare. The mechanism of expiration-band helps

to find the best cluster to publish at any instant, and the effect of multiple expiration of tuple is compensated by a pocket strategy. *X-BAND* showed significant improvement, 5% to 11% less information loss on *Adult* and 1% to 3% on

(a) Effect of α and β on clusters' purity on *FADS* for varying missingness (b) Effect of α and β on clusters' purity on *X-BAND* for varying missingnessFig. 5: Heatmap of clusters purity on *FADS* and *X-BAND* on *Adult* dataset (a) Number of re-use anonymization for varying missingness and distance function on *Adult* (b) Number of re-use anonymization for varying distance function on *PM2.5*Fig. 6: Number of re-use anonymization for *FADS* and *X-BAND* on *Adult* and *PM2.5*

Fig. 7: Runtime for *FADS* and *X-BAND* on *Adult* and *PM2.5*

PM2.5 datasets, while resulting similar number of clustering and comparable suppression, re-using and runtime compared to *FADS*. However, the difference of suppression and re-use does not affect the overall performance of *X-BAND* as the information loss of anonymization showed substantial improvement over *FADS*. For future work, we will investigate on the optimization of distance function and expiration-band mechanism for anonymizing varied data streams. Finally, we will work on the development of self adaptive anonymization approaches for IoT data streams using artificial intelligence techniques. The main challenge will be the design of missingness prediction and adaptive distance metrics.

ACKNOWLEDGEMENT

The first author likes to acknowledge the support provided by the EU Erasmus Mundus project gLINK (552099-EM-1-2014-1-UK-ERA) to carry out this research at the University of the West of Scotland, UK.

REFERENCES

- [1] M. Oppitz and P. Tomsu, "Internet of things," in *Inventing the Cloud Century*. Springer, 2018, pp. 435–469.
- [2] K. Okada and H. Nishi, "Big data anonymization method for demand response services," in *Proceedings on the International Conference on Internet Computing (ICOMP)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2014, p. 1.
- [3] H.-C. Jo, S. Kim, and S.-K. Joo, "Smart heating and air conditioning scheduling method incorporating customer convenience for home energy management system," *IEEE transactions on consumer electronics*, vol. 59, no. 2, pp. 316–322, 2013.
- [4] P. Krishnan, "Design of collision detection system for smart car using li-fi and ultrasonic sensor," *IEEE Transactions on Vehicular Technology*, 2018.
- [5] W. Alsafery, B. Alturki, S. Reiff-Marganiec, and K. Jambi, "Smart car parking system solution for the internet of things in smart cities," in *2018 1st International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE, 2018, pp. 1–5.
- [6] X. Li, R. Lu, X. Liang, X. Shen, J. Chen, and X. Lin, "Smart community: an internet of things application," *IEEE Communications Magazine*, vol. 49, no. 11, 2011.
- [7] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013.
- [8] J. Tang, Y. Cui, Q. Li, K. Ren, J. Liu, and R. Buyya, "Ensuring security and privacy preservation for cloud data services," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, p. 13, 2016.
- [9] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.
- [10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005, pp. 49–60.
- [11] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian, "ell-diversity: Privacy beyond k-anonymity," in *null*. IEEE, 2006, p. 24.
- [12] T. M. Truta and B. Vinay, "Privacy protection: p-sensitive k-anonymity property," in *null*. IEEE, 2006, p. 94.
- [13] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 106–115.
- [14] X. Zhang, C. Liu, S. Nepal, and J. Chen, "An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud," *Journal of Computer and System Sciences*, vol. 79, no. 5, pp. 542–555, 2013.
- [15] A. Oganian and J. Domingo-Ferrer, "Local synthesis for disclosure limitation that satisfies probabilistic k-anonymity criterion," *Transactions on Data Privacy*, vol. 10, no. 1, pp. 61–81, 2017.
- [16] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 279–288.
- [17] J. Cao, B. Carminati, E. Ferrari, and K.-L. Tan, "Castle: Continuously

- anonymizing data streams,” *IEEE Transactions on Dependable and Secure Computing*, vol. 8, no. 3, pp. 337–352, 2011.
- [18] P. Wang, J. Lu, L. Zhao, and J. Yang, “B-castle: An efficient publishing algorithm for k-anonymizing data streams,” in *Intelligent Systems (GCIS), 2010 Second WRI Global Congress on*, vol. 2. IEEE, 2010, pp. 132–136.
- [19] K. Guo and Q. Zhang, “Fast clustering-based anonymization approaches with time constraints for data streams,” *Knowledge-Based Systems*, vol. 46, pp. 95–108, 2013.
- [20] H. Zakerzadeh and S. L. Osborn, “Faanst: fast anonymizing algorithm for numerical streaming data,” in *Data privacy management and autonomous spontaneous security*. Springer, 2011, pp. 36–50.
- [21] A. Otgonbayar, Z. Pervez, K. Dahal, and S. Eager, “K-varp: K-anonymity for varied data streams via partitioning,” *Information Sciences*, vol. 467, pp. 238–255, 2018.
- [22] J. Han, J. Yu, Y. Mo, J. Lu, and H. Liu, “Mage: A semantics retaining k-anonymization method for mixed data,” *Knowledge-Based Systems*, vol. 55, pp. 75–86, 2014.
- [23] S. Kim, M. K. Sung, and Y. D. Chung, “A framework to preserve the privacy of electronic health data streams,” *Journal of biomedical informatics*, vol. 50, pp. 95–106, 2014.
- [24] M. Al-Zobbi, S. Shahrestani, and C. Ruan, “Experimenting sensitivity-based anonymization framework in apache spark,” *Journal of Big Data*, vol. 5, no. 1, p. 38, 2018.
- [25] K. Al-Hussaini, B. C. Fung, and W. K. Cheung, “Privacy-preserving trajectory stream publishing,” *Data & Knowledge Engineering*, vol. 94, pp. 89–109, 2014.
- [26] A. Otgonbayar, Z. Pervez, and K. Dahal, “Toward anonymizing iot data streams via partitioning,” in *Mobile Ad Hoc and Sensor Systems (MASS), 2016 IEEE 13th International Conference on*. IEEE, 2016, pp. 331–336.
- [27] H. Zakerzadeh and S. L. Osborn, “Delay-sensitive approaches for anonymizing numerical streaming data,” *International journal of information security*, vol. 12, no. 5, pp. 423–437, 2013.
- [28] X. Yan, W. Xiong, L. Hu, F. Wang, and K. Zhao, “Missing value imputation based on gaussian mixture model for the internet of things,” *Mathematical Problems in Engineering*, vol. 2015, 2015.
- [29] J. W. Graham, “Multiple imputation with norm 2.03,” in *Missing data*. Springer, 2012, pp. 73–94.
- [30] B. Fekade, T. Maksymyuk, M. Kyryk, and M. Jo, “Probabilistic recovery of incomplete sensed data in iot,” *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2282–2292, 2018.
- [31] Z. Ding, G. Mei, S. Cuomo, Y. Li, and N. Xu, “Comparison of estimating missing values in iot time series data using different interpolation algorithms,” *International Journal of Parallel Programming*, pp. 1–15, 2018.
- [32] C.-F. Tsai, M.-L. Li, and W.-C. Lin, “A class center based approach for missing value imputation,” *Knowledge-Based Systems*, vol. 151, pp. 124–135, 2018.
- [33] Z.-g. Liu, Q. Pan, J. Dezert, and A. Martin, “Adaptive imputation of missing values for incomplete pattern classification,” *Pattern Recognition*, vol. 52, pp. 85–95, 2016.
- [34] K. Wagstaff, “Clustering with missing values: No imputation required,” in *Classification, Clustering, and Data Mining Applications*. Springer, 2004, pp. 649–658.
- [35] M. Ciglic, J. Eder, and C. Koncilia, “Anonymization of data sets with null values,” in *Transactions on Large-Scale Data and Knowledge-Centered Systems XXIV*. Springer, 2016, pp. 193–220.
- [36] B. Zhou, Y. Han, J. Pei, B. Jiang, Y. Tao, and Y. Jia, “Continuous privacy preserving publishing of data streams,” in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. ACM, 2009, pp. 648–659.
- [37] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, “Efficient knn classification algorithm for big data,” *Neurocomputing*, vol. 195, pp. 143–148, 2016.
- [38] W. Wang, J. Li, C. Ai, and Y. Li, “Privacy protection on sliding window of data streams,” in *Collaborative Computing: Networking, Applications and Worksharing, 2007. CollaborateCom 2007. International Conference on*. IEEE, 2007, pp. 213–221.
- [39] J. Li, B. C. Ooi, and W. Wang, “Anonymizing streaming data for privacy protection,” 2008.
- [40] A. B. Sakpere and A. V. Kayem, “On anonymizing streaming crime data: A solution approach for resource constrained environments,” in *IFIP International Summer School on Privacy and Identity Management*. Springer, 2017, pp. 170–186.
- [41] S. Li, “Poisson process with fuzzy rates,” *Fuzzy Optimization and Decision Making*, vol. 9, no. 3, pp. 289–305, 2010.
- [42] J. Barnard and X.-L. Meng, “Applications of multiple imputation in medical studies: from aids to nhanes,” *Statistical methods in medical research*, vol. 8, no. 1, pp. 17–36, 1999.
- [43] E.-L. Silva-Ramírez, R. Pino-Mejías, and M. López-Coello, “Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns,” *Applied Soft Computing*, vol. 29, pp. 65–74, 2015.
- [44] Z. Zhang, “Missing data imputation: focusing on single imputation,” *Annals of translational medicine*, vol. 4, no. 1, 2016.
- [45] P. Li, E. A. Stuart, and D. B. Allison, “Multiple imputation: a flexible tool for handling missing data,” *Jama*, vol. 314, no. 18, pp. 1966–1967, 2015.
- [46] S. Van Buuren, *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.
- [47] R. Dechter and J. Pearl, “Generalized best-first search strategies and the optimality of a,” *Journal of the ACM (JACM)*, vol. 32, no. 3, pp. 505–536, 1985.
- [48] J. Wang, C. Deng, and X. Li, “Two privacy-preserving approaches for publishing transactional data streams,” *IEEE Access*, vol. 6, pp. 23 648–23 658, 2018.
- [49] J. Cao, P. Karras, C. Raïssi, and K.-L. Tan, “ ρ -uncertainty: inference-proof transaction anonymization,” *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 1033–1044, 2010.
- [50] P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” technical report, SRI International, Tech. Rep., 1998.
- [51] S. Kosub, “A note on the triangle inequality for the jaccard distance,” *Pattern Recognition Letters*, vol. 120, pp. 36–38, 2019.

Ankhbayar Otgonbayar received his Master's degree in Computer Science and Engineering in Nanjing University of Information Science & Technology in 2013. Currently, he is studying his Ph.D at the University of the West of Scotland. His research interest are privacy and security aspects of internet-of-things, big data, cloud computing and artificial intelligence.

Zeeshan Pervez is with the University of the West of Scotland (UWS) as an Associate Professor (Reader in the UK). His research interests are security and privacy aspects of internet-of-things, cloud computing, and cybersecurity. He has published over 60 indexed journals, peer-reviewed conferences and book chapters. He has led and participated in several projects funded by the European Commission, UK and Korean Research Councils and industry.

Keshav Dahal is a Professor in Intelligent Systems and the leader of the Artificial Intelligence, Visual Communication and Network (AVCN) Research Centre at the University of the West of Scotland. Prior to this he was with the University of Bradford and University of Strathclyde UK. He obtained his PhD and MSc degrees from Strathclyde. His research interests lie in the areas of applied AI, trust and security modelling in distributed systems, and scheduling/optimization problems. He has published extensively in his research fields with award winning papers, and has sat on organizing/program committees of over 60 international conferences including general chair or program chair of five IEEE sponsored conferences. He is a senior member of the IEEE.