International Journal of Information Security and Privacy
Volume 14 • Issue 3 • July-September 2020

An Effective and Computationally Efficient Approach for Anonymizing Large-Scale Physical Activity Data:
Multi-Level Clustering-Based Anonymization

Pooja Parameshwarappa, University of Maryland, Baltimore County, USA
Zhiyuan Chen, University of Maryland, Baltimore County, USA
Gunes Koru, University of Maryland, Baltimore County, USA

ABSTRACT

Publishing physical activity data can facilitate reproducible health-care research in several areas such as population health management, behavioral health research, and management of chronic health problems. However, publishing such data also brings high privacy risks related to re-identification which makes anonymization necessary. One of the challenges in anonymizing physical activity data collected periodically is its sequential nature. The existing anonymization techniques work sufficiently for cross-sectional data but have high computational costs when applied directly to sequential data. This article presents an effective anonymization approach, multi-level clustering-based anonymization to anonymize physical activity data. Compared with the conventional methods, the proposed approach improves time complexity by reducing the clustering time drastically. While doing so, it preserves the utility as much as the conventional approaches.

KEYWORDS

De-identification, Differential Privacy, Health-Related Longitudinal Data, High-Dimensional Data, K-Anonymity, Microaggregation, Sequential Data

INTRODUCTION

There has been a rapid increase in the availability of physical activity data due to the increase in the use of wearable devices, smartphones, and smart environments. Publishing physical activity data can support reproducible research in personal and population health management, behavioral health research and management of chronic health problems. For example, data about vigorous activity and sedentary hours per day can help research studies investigating the types and amounts of physical activity necessary at the individual, cohort and population levels (Matthews et al., 2008; Pate et al., 1995). Physical activity is known to decrease the risk of various diseases such as cardiovascular diseases, diabetes and obesity (Dietz, Douglas, & Brownson, 2016; Thornton et al., 2016). Publishing activity data can support research in preventing such chronic diseases. Furthermore, it can facilitate research studies that aim to reduce health care costs and the costs related to social benefits and work absenteeism (CDC Foundation, 2015; Spenkelink, Hutten, Hermens, & Greitemann, 2002). Therefore, there is an important and increasing need for publishing physical activity data.

DOI: 10.4018/IJISP.2020070105

However, publishing physical activity data also brings high privacy risks related to re-identification. Although direct identifiers such as names, identification numbers, and other personally identifiable information (PII) are removed, many unique longitudinal patterns can easily reveal identities. For example, consider the publication of a data set which includes activity data of a group of people and their health status. Table 1 shows an example which contains activity data for four individuals collected every minute for a certain time duration. Additionally, the data contains health status of these individuals. Assume that, an adversary gets access to this data and knows that an individual whose record is in the data runs every Monday, Tuesday, and Wednesday at 6:00 am. Since there is only one person with this specific routine, his/her data is easily re-identifiable. As a result, the adversary gains access to sensitive information such as the health status. To reduce the probability of re-identification to acceptable levels, and ensure privacy, such activity data needs to be anonymized. Anonymization involves modifying the data, in order to protect the privacy of the individuals whose information is in the data, while preserving the utility of the data.

Table 1. Example showing physical activity data of four people and corresponding health status. S stands for Stationary, W stands for Walking and R stands for Running

	Physical Activity Data									Health Status
Day	**Mon**	**Mon**	**..**	**Tue**	**Tue**	**..**	**Wed**	**Wed**	**..**	
Time	6:00 am	6:01 am	..	6:00 am	6:01 am	..	6:00 am	6:01 am	..	
Person 1	S	S	..	S	W	..	S	S	..	Heart Disease
Person 2	R	R	..	R	R	..	R	R	..	Depression
Person 3	S	S	..	S	S	..	S	S	..	Cold
Person 4	S	S	..	S	S	..	W	W	..	Heart Disease

Unfortunately, most of the conventional anonymization techniques are suitable for cross-sectional data sets (El Emam et al., 2009; Gal, Chen, & Gangopadhyay, 2008; Loukides, Gkoulalas-Divanis, & Malin, 2010). An example for cross-sectional data is shown in Table 2. However, physical activity data is sequential in nature (shown in Table 1), which results in high dimensionality because every instance of time acts as a dimension. For example, if the data is collected every minute for a period of one month, then the number of dimensions would be 43,200.

Table 2. Sample cross-sectional data

	Age	**Sex**	**Zip Code**	**Disease**
Person 1	22	M	21220	Cold
Person 2	25	F	21222	Heart Disease
Person 3	33	F	21236	Cancer
Person 4	30	M	21239	Cancer

Furthermore, most of the existing techniques for anonymizing sequential data sets (Gkoulalas-Divanis & Loukides, 2012; He, Cormode, Machanavajjhala, Procopiuc, & Srivastava, 2015; Martinez-Bea & Torra, 2011; Pensa, Monreale, Pinelli, & Pedreschi, 2008) focus on preserving frequent

sequential patterns to maintain the utility of the data. On the other hand, for physical activity data, preserving aggregate statistics is more relevant than sequential pattern-preservation for meeting the utility requirements (Matthews et al., 2008; Spees, Scott, & Taylor, 2012). Additionally, most of the existing techniques do not consider the computational efficiency of the anonymization method, which is important for high dimensional data sets. Applying the existing techniques directly to high dimensional data results in very high computational costs.

In this paper, the authors propose a multi-level clustering (MC) based anonymization approach that improves time complexity by reducing the clustering time drastically. The approach focuses on preserving the aggregate statistics and correlations in the data in order to maintain the utility of the data. The approach includes two steps:

1. Clustering physical activity data using MC so that activity sequences that are similar to one another belong to the same cluster
2. Application of k-anonymity (Sweeney, 2002b) and differential privacy (Dwork, 2008; Dwork, McSherry, Nissim, & Smith, 2006) models on the clusters generated, thereby resulting in Multi-level Clustering based K-Anonymity (MCKA) and Multi-level Clustering based Differential Privacy (MCDP), respectively.

The rest of the paper is organized as follows. The next section presents the related work. It is followed by the methods section describing the proposed methodology. The next section describes the experiments and results, followed by the discussion section and the conclusion section.

RELATED WORK

Some of the commonly used privacy models include *k*-anonymity (Sweeney, 2002b), *l*-diversity (Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, 2007), *t*-closeness (Li, Li & Venkatasubramanian, 2007) and differential privacy (Dwork, 2008; Dwork et al., 2006). Some of the techniques for achieving these privacy models include generalization and suppression (Sweeney, 2002a), perturbative techniques such as microaggregation (Domingo-Ferrer, Mart´ınez-Ballest´e, Mateo-Sanz, & Seb´e, 2006a; Domingo-Ferrer & Torra, 2005; Templ, Meindl, & Kowarik, 2013) and Laplace Perturbation Algorithm (Dwork et al., 2006). The following sub-sections present the existing work for anonymizing cross-sectional and sequential data. At the end of the section, research gap addressed by the proposed approach is presented.

Anonymization of Cross-Sectional Data

Most of the conventional anonymization techniques are suitable for cross-sectional data sets. One of the approaches for anonymizing cross-sectional health data is Optimal Lattice Anonymization (OLA) (El Emam et al., 2009), which provides a globally optimal anonymization solution through generalization. El Emam (2008), and El Emam and Dankar (2008) present various heuristics for de-identifying cross-sectional health data, re-identification risks associated with using k-anonymization for health data, and improvements to control the risk. Another approach anonymizes data containing both clinical and genomic information by extracting linkable features from clinical data and generalizing these features such that it is no longer possible to link the genomic data with a small number of patients (Loukides et al., 2010). Another anonymization approach extends k-anonymity and l-diversity models for anonymizing patient data with multiple sensitive attributes (Gal et al., 2008). Gal et al. (2014) propose an anonymization framework which identifies a suitable de-identification method for cross-sectional data based on the requirements of the recipient of the data. Poulis et al. (2017) propose anonymization of health data that contains both demographics and diagnoses codes. The authors use (k, k^m) anonymity which assumes that the attacker knows the demographics and

up to *m* diagnoses codes of a patient. This approach takes into consideration the utility constraint set and provides low information loss. Zhang et al. (2014) propose anonymization of big data using MapReduce (Dean & Ghemawat, 2010). This approach combines Top–Down Specialization (TDS) and Bottom–Up Generalization (BUG) components for anonymization. It is a hybrid approach that automatically chooses the component suitable for anonymization and is implemented using the MapReduce paradigm. A two-phase clustering method using Map-Reduce was proposed in (Zhang, Dou, Pei, Nepal, Yang, Liu & Chen, 2014) where a k-means like clustering is used in the first phase to generate some initial clusters and then the clusters are recoded using agglomerative clustering.

Applying conventional techniques directly to sequential data is computationally expensive because of the high-dimensional nature of the sequential data. Although the approaches for big data using Map-Reduce address the issue of large-scale data, they focus only on cross-sectional data sets. In the following sub-sections, the existing techniques for anonymizing sequential data are discussed.

Anonymization of Trajectory Data

One of the techniques for anonymizing location information data uses microaggregation as the protection method (Martinez-Bea & Torra, 2011; Nin & Torra, 2009). In microaggregation, euclidean distance and short time series distance (Martinez-Bea & Torra, 2011) are used as distance measures and information loss is determined by using measures such as average, autocorrelation (Dunn & Davis, 2017), and the difference between the original time series and the protected time series. Nergiz et al. (2008) and Nergiz et al. (2007) proposed a technique for anonymizing trajectory data based on generalization. It has two components: 1) Ensuring k-anonymity i.e. every trajectory is made indistinguishable from k-1 other trajectories; 2) Reconstruction, in which atomic trajectories are randomly sampled from the area covered by the anonymized trajectories. Mohammed et al. (2009) proposed the LKC privacy model for anonymizing trajectory data. In this model, it is assumed that the adversary knows a sub-sequence of location and timestamp pairs that the victim visited (L) and a trajectory database is said to satisfy LKC privacy if and only if for any sub-sequence, k-anonymity (K) is preserved and the probability of obtaining a sensitive value associated with this sub-sequence is below a custom threshold (C). This is achieved through coarsening in which one or more trajectory points are removed. Differentially Private Trajectory Synthesis (DPT) (He et al., 2015) is another privacy protection model in which trajectory data is represented using reference systems of different resolutions. A prefix tree is constructed for each reference system. Laplace noise is added to a subset of prefix trees, which are then used to simulate trajectories that are differentially private. Dong and Pi (2018) propose a privacy-preserving algorithm for trajectories based on frequent path. Infrequent roads are removed, and trajectories are grouped based on similarity and k-anonymity is applied to these groups. Gao et al. (2014) propose a personalized anonymization model that takes trajectory angle into consideration while creating trajectory k-anonymity set. Hu et al. (2018) propose an anonymization technique for trajectory data in which equivalence classes of trajectories are created considering the fact that different users may have different privacy requirements at different time. Barack et al. (2016) propose a semantic cloaking-based anonymization framework in which exact locations are replaced by semantic categories such as home, work, and so on.

Anonymization of Transaction Data Sets

A variation of k-anonymity, k^m-anonymity (Terrovitis, Mamoulis, & Kalnis, 2008) is a privacy model for anonymizing transaction data. It is assumed that the maximum knowledge of an adversary is at most *m* items, and it is ensured that for any possible set of *m* items or less, there are at least *k* transactions. Gkoulalas-Divanis and Loukides (2012) propose two other algorithms for anonymizing transaction data sets: Privacy-Constrained Clustering-based Transaction Anonymization (PCTA) and Utility-guided Privacy-constrained Clustering-based Transaction Anonymization (UPCTA). These methods are based on agglomerative clustering. The clusters are merged at different levels, based on certain privacy (PCTA) and utility (UPCTA) constraints. Wang and Li (2018) propose anonymization of

transaction data that contain sensitive information in both relational and transactional attributes. They use graph-based approach for anonymization. In this approach, the associations between customers and products are represented using an uncertain graph. The method protects the multifold privacy of the data while maximizing its utility. In another work, Kakatkar and Spann (2019) highlight the importance of anonymized and fragmented data in retailing and present a methodology to analyze such data.

Anonymization of Other Sequential Data Sets

Randomization is one of the techniques for anonymizing time-series data (Moon, Kim, Kim, & Bertino, 2010). The original time-series is distorted by adding noise such that, the distance orders are preserved even after distortion. (k, P)-anonymity model (Shou, Shang, Chen, Chen, & Zhang, 2013) is another technique for anonymizing time-series data which ensures anonymity on two levels: k-anonymity is ensured for the entire data set, and P-anonymity is required for the pattern representations associated with the data points in the same group.

There has been some work on addressing privacy issues related to sensor data (Cavoukian, Mihailidis, & Boger, 2010; Chan & Perrig, 2003; He, Liu, Nguyen, Nahrstedt, & Abdelzaher, 2007; Li, Lou, & Ren, 2010; Sun, Fang, & Zhu, 2010). Most of these studies focus on security, access control, and encryption. A framework for anonymizing longitudinal Electronic Medical Record (EMR) data (Tamersoy, Loukides, Nergiz, Saygin, & Malin, 2012) ensures k-anonymity through generalization and suppression of International Classification of Diseases (ICD) codes and age details of patients. Longitudinal OLA (LOLA) (El Emam et al., 2012) is an extension of OLA which is used to anonymize longitudinal claims data. A system for anonymized public health data collection and intervention is proposed by Clarke and Steele (2014). In this system, statistics about the physical activity data and nutritional data from users' smartphones are computed at different resolutions based on disclosure risks and submitted to the central server. Another area of research uses Natural Language Processing (NLP) based techniques for de-identifying longitudinal clinical narratives (Dernoncourt, Lee, Uzuner, & Szolovits, 2017; Stubbs, Filannino, & Uzuner, 2017; Stubbs & Uzuner, 2015).

Research Gap

Table 3 summarizes the advantages and limitations of the existing methods. So far, no existing technique has dealt with anonymizing physical activity data. The existing techniques discussed may not be suitable when applied directly for anonymizing such data. This is because most of these techniques focus on retaining the sub-sequences for preserving the utility of the data. However, for physical activity data, aggregate statistics such as duration, frequency, and intensity, are more relevant from the utility perspective. Also, most of the existing techniques do not consider the efficiency of the anonymization method, which is important for high dimensional data sets. To address these gaps, the authors propose an efficient approach for anonymizing physical activity data using MC and ensuring k-anonymity and ε-differential privacy, while preserving the utility of the data.

MATERIALS AND METHODS

This section first presents the dataset used for the experiments. It is followed by the explanation of the proposed approach.

Data

Most publicly available activity datasets are small for various reasons such as unavailability of participants and need for data collection over long periods of time (Mendez-Vazquez, Helal, & Cook, 2009). However, evaluation of privacy-preserving approaches typically requires large data sets to achieve desired levels of accuracy (Kitamura, Chen, & Pendyala, 1997; Monekosso & Remagnino, 2009). To address this problem, a larger synthetic data set was generated from the *Student Life*

Table 3. Advantages and limitations of the existing anonymization approaches

Method	Advantages	Limitations
Anonymization of cross-sectional data	Suitable for cross-sectional data	Computationally expensive when applied to high dimensional sequential data
Anonymization of trajectory and transaction data	Suitable for sequential data	Focus on preserving frequent patterns. (For physical activity data, preserving aggregate statistics is more relevant)
Anonymization of other sequential data	Suitable for sequential data	Do not focus on the computational efficiency for high dimensional data

dataset (Wang et al., 2014) by paying attention to preserving the distributional characteristics of the original data.

The original dataset (*Student Life dataset*) consists of activity sequences of 49 students collected through a sensing application over a period of 10 weeks. Data for each student is a sequence of timestamp and activity inference pairs such as <1364356858; *stationary* >. This shows that, on March 27, 2013, at 04:00:58 AM, the student was stationary. Three types of activities are included: *running, walking* and *stationary*. For the same students, academic performance data, as well as behavioral and mental well-being data such as flourishing scale (i.e., self-perceived success in terms of relationships, self-esteem, purpose, and optimism) (Diener et al., 2010) are also available. Therefore, in this study, the correlations between (a) activity and Cumulative Grade Point Average (CGPA), and (b) activity and flourishing scale are used to represent utility-preservation capabilities of the proposed approach.

The synthetic data is generated by using a markov chain model (Gagniuc, 2017) in the R statistical environment (Spedicato, 2016). In the original data, activity inference is made every 2 seconds which results in very high dimensionality (approximately 450,000). In the synthetic data generation, this type of high dimensionality results in very large un-manageable state transition matrices. Also, the analyses of activity data are typically conducted at higher time levels such as daily and weekly (Matthews et al., 2008; Pate et al., 1995; Spees et al., 2012; Wang et al., 2014). Therefore, for this study, each activity sequence is aggregated to the level of one-minute intervals, and the most dominant activity during an interval is assigned as the activity for that interval. This results in 1,440 intervals a day.

In addition to the three activities, unknown values are also present. An unknown value is represented as an activity as well, and it is referred to as missing. This is because an adversary can make inferences based on the unknown data. For example, if the adversary knows that the victim turns the device off during class hours, the data set will contain unknown values consistently during the class hours. If there are only few records with this routine, the identity of the student is at stake.

A state transition matrix is constructed for every hour, for each student. The state transition matrices are used to simulate activity sequences for 336 hours at minute-level granularity. A total of 9,800 students' activity sequences was generated where each sequence is of length 20,160 (336 x 60). At each hour, a different student's state transition matrix is chosen at random for simulation with a probability of 0.01.

The difference between the probability distributions of the original data and the simulated data is measured using KL-divergence (Kullback & Leibler, 1951), and its value is 0.03. The low value of KL divergence indicates that the two distributions have similar behavior.

Proposed Approach

One of the most widely used methods for achieving k-anonymity is Microaggregation (Domingo-Ferrer et al., 2006a; Domingo-Ferrer & Torra, 2005). It has two steps:

1. Partitioning: the records are partitioned into clusters of size k.
2. Aggregation: the records in each cluster are replaced by the result of an aggregation operation leading to k-anonymity

The proposed approach is based on microaggregation and has the following two steps:

1. Multi-level Clustering (MC): sequences are partitioned into clusters of size k. Clusters are homogenous i.e. each cluster contains sequences that are similar to one another. This in turn helps preserve the utility of data.
2. Anonymization (similar to aggregation step): here two types of privacy models are applied to the clusters from step 1
   a. K-anonymity, resulting in MCKA (Multi-level Clustering based K-Anonymity)
   b. Differential Privacy, resulting in MCDP (Multi-level Clustering based Differential Privacy)

Multi-Level Clustering

One of the best-known heuristic methods for achieving the partitioning step of microaggregation is Maximum Distance to Average Vector (MDAV) (Domingo-Ferrer, Solanas, & Martinez-Balleste, 2006b; Solanas, Martinez-Balleste, & Domingo-Ferrer, 2006). MDAV involves the following steps:

1. Compute the centroid of the data set. Find a point *r* that is most distant to the centroid. Find a point *s* that is farthest to *r*
2. Find $k-1$ nearest data points around both *r* and *s*
3. If there are at least $2k$ data points remaining, then repeat steps 1 and 2 on the new set of data points, which is the previous set of data points minus the two clusters formed in step 2. Else, go to step 4
4. If there are between *k* and $2k-1$ points, form a new cluster with these points
5. If there are less than *k* points then, compute the centroids for all the clusters created so far and find the cluster whose centroid is closest to the centroid of the remaining *k* points and add the *k* points to that cluster

However, applying MDAV directly to high dimensional sequential data such as physical activity data results in very high computational cost. The proposed clustering method MC is based on MDAV. However, MC handles high dimensionality of the physical activity data by aggregating the sequences to different time intervals. This improves the efficiency of the anonymization process. In MC (shown in Algorithm 1) all the sequences are assigned to one cluster at the root level (line 1). The sequences are then aggregated to certain time intervals (line 6), for example daily intervals, and then clustered using MDAV (line 7). In the next level, the sequences are drilled down to smaller time intervals (for example, hourly intervals) and clustered using MDAV. These steps are repeated until each cluster at the leaf level has at least *k* sequences in it. *k* represents the desired level of anonymity. For clustering, weighted Euclidean distance is used for distance computation, with different weights for each activity.

Complexity of MDAV is $O\left(n^2 m\right)$ (Domingo-Ferrer et al., 2006b), where n is the number of data points and *m* is the number of dimensions. The complexity of MC at the root level is $O\left(n^2 m_L\right)$ since all the sequences are contained in one cluster. *L* represents the root level and m_L represents the number of dimensions at the root level $\left(m_L \ll m\right)$. At the i^{th} level, if n_i represents the size of each cluster, and m_i represents the number of dimensions, then, the complexity is $O\left(\frac{n}{n_i} {n_i}^2 m_i\right)$,

which is $O\left(nn_i m_i\right)$. There are L levels, so the complexity is given by $O\left(n\sum_{i=0}^{L} n_i m_i\right)$, where L represents the root level and 0 represents the leaf level ($n_L = n$).

MC is far more efficient than MDAV for two reasons. At the higher levels, data is aggregated to larger time intervals ($m_i << m$). At lower levels, data is aggregated ($m_i < m$), and additionally, the size of the clusters decreases ($n_i << n$). The quality of the clusters is preserved because, sequences that are similar to one another on a higher level are assigned to the same cluster early on in the clustering process.

The illustration shown in Figure 1 has eight sequences corresponding to physical activity data of eight students at 15-minute interval. For instance, let the k anonymity requirement be 2. At the root level, all the students are assigned to a single cluster. Sequences are aggregated to daily intervals and clustered using MDAV. The clustering results in U1, U2, U3 and U4 being assigned to one cluster, and U5, U6, U7 and U8 to another. In the next level, the sequences are aggregated to hourly intervals and further clustered. The clustering stops since each cluster has at least 2 sequences in it.

Algorithm 1: Multi-level Clustering

Data: Set of all sequences S, number of levels l, aggregation at each level $\{a_1,a_2..,a_l\}$, partition size at each level $\{s_1,s_2..,s_l\}$, required anonymity k

Result: Set of clusters of sequences C

```
C ←{c1}; /* c1 represents cluster of all sequences */
t←1; /* t is a loop variable */
R ← NULL; /* R stores generated clusters at a level */
while t<=l do
for c in C do
Aggregate all the sequences in c to level at;
Using MDAV and distEuc(X,Y) as the distance measure,
cluster the sequences into partitions of size st;
/* Let x be the number of clusters generated*/
R ←R ∪ {ct1, ct2,...,ctx} ;
C ←C −{c}; /* End of for loop*/
C ←R;
R ← NULL;
t←t+1; /* End of while loop*/
return C;
```

```
distEuc(X, Y)
/* X and Y represent two multivariate sequences with n dimensions
and m variables */
/* wi represents weight for each variable
(running, walking, stationary, and missing) */
```

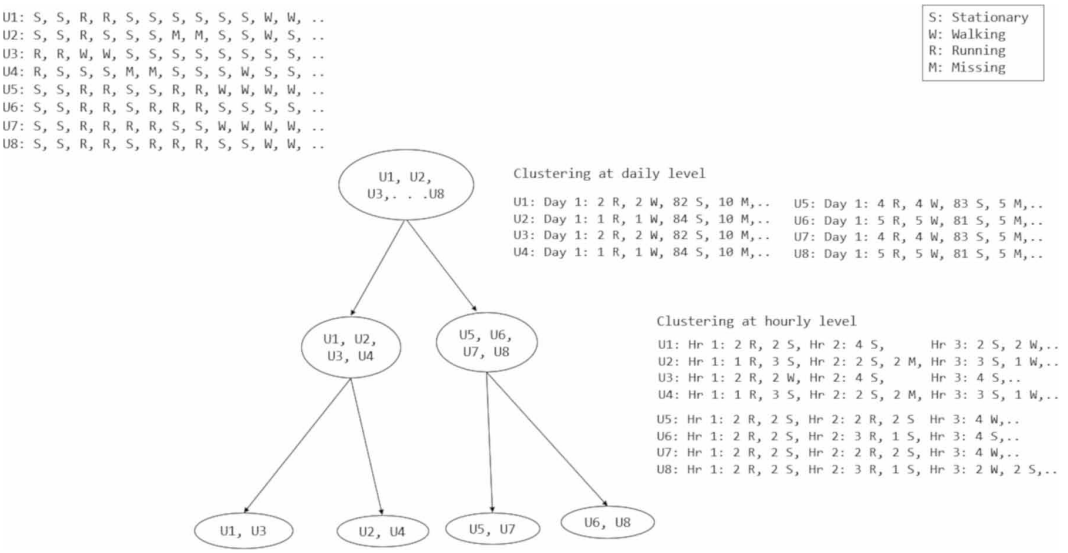
$$D \leftarrow \sum_{i=1}^{m} w_i \cdot \sqrt{\sum_{j=1}^{n} \left(x_{ij} - y_{ij}\right)^2}$$

```
/* m = 4 (running, walking, stationary,
and missing) */
return D;
```

Multi-Level Clustering Based K-Anonymity

K-anonymity is a privacy model which ensures that the quasi-identifiers identifying each person cannot be distinguished from at least $k-1$ other individuals (Sweeney, 2002b). In microaggregation,

Figure 1. Illustration of multi-level clustering



each cluster is replaced by its centroid to achieve k-anonymity. Similar step is performed in MCKA. However, instead of replacing the cluster with the centroid, the centroid is used to simulate sequences. The number of sequences simulated is equal to the size of the cluster. Simulation is done using probabilistic sampling. In Figure 2, the centroid for the cluster of two sequences is computed. Furthermore, using the centroid for probabilistic sampling, two sequences are generated.

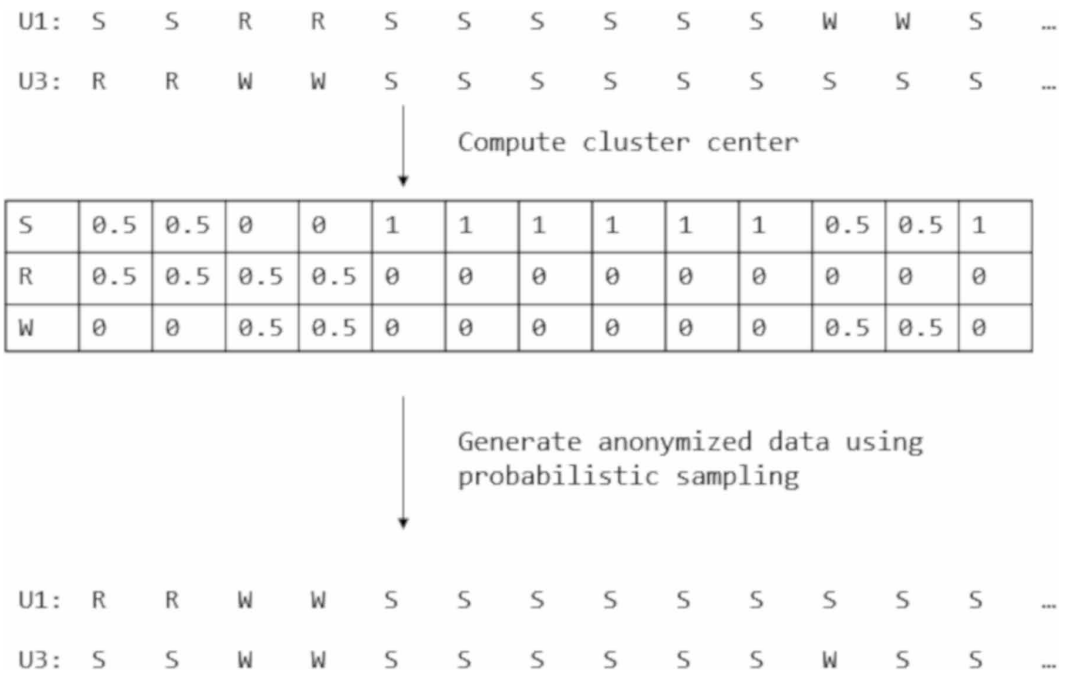
Multi-Level Clustering Based Differential Privacy

Differential privacy is a privacy model that ensures that the presence or absence of an individual's data in the dataset does not make an outcome more or less likely (Dwork, 2008). Differential privacy is stronger than k-anonymity because it can hide the impact of a single person even if the adversaries know the rest of the data set. A standard technique used to achieve differential privacy is the Laplace Perturbation Algorithm (LPA) (Dwork et al., 2006), which works by adding noise to the query results. Let Q be a sequence of queries $Q = \left\{Q_{1,} Q_2 ...\right\}$ and the answers to the queries be represented by $Q\left(I\right) = \left\{Q_1\left(I\right), Q_2\left(I\right) ...\right\}$ where *I* represents a dataset. In our case, Q_i is a query such as average time of an activity (stationary, walking, running), at an interval *t*, for a given cluster. If a set of queries request for the average time of each activity, at each time interval over the entire time period, then the answers to these queries together represent the centroid of the cluster. Each answer is perturbed by the addition of Laplace noise *Lap*(λ), therefore, with the increase in the number of queries, noise added also increases.

MCDP uses Fourier Perturbation Algorithm (FPA) (Rastogi & Nath, 2010), which greatly reduces noise to be added. FPA is based on compressing the answers of the query sequence by using orthonormal transformation such as Discrete Fourier Transform (DFT) (Ahmed, Natarajan, & Rao, 1974). The compressed sequence has fewer dimensions as compared to the original sequence, therefore, the overall amount of noise added is reduced. Steps involved in FPA are shown below (Rastogi & Nath, 2010).

1. Compute C, DFT of $Q\left(I\right)$
2. Consider first *l* co-efficients of C, C^l

Figure 2. Illustration of multi-level clustering-based k-anonymity

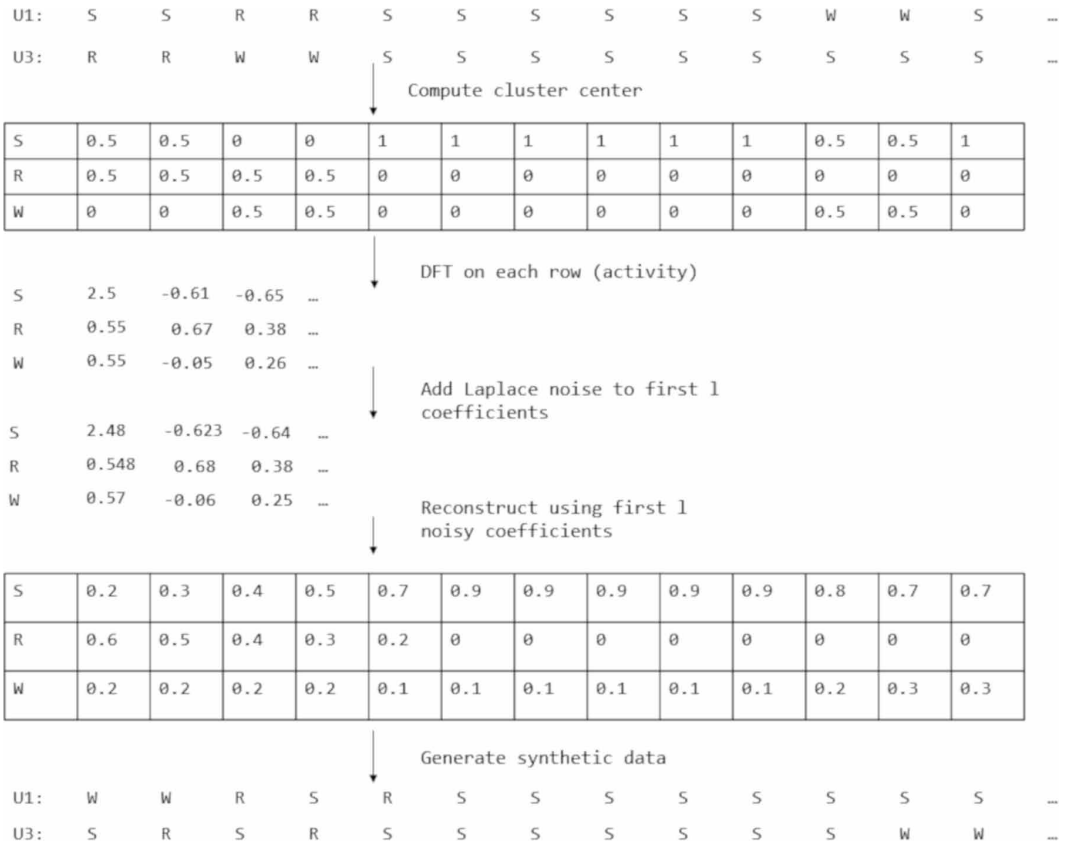


3. Add noise, $\overline{C^l} = C^l + Lap(\lambda)$
4. Pad $\overline{C^l}$ with *0s* ($C^{\bar{l}}_{pad}$), so the length is same as C
5. Compute Inverse DFT (IDFT) of $C^{\bar{l}}_{pad}$

Noise added is given by $\lambda = \frac{\sqrt{l}\Delta_2(Q)}{\epsilon}$, where $\Delta_2(Q)$ is the L_2 sensitivity of Q. Sensitivity is the maximum amount the query answers can change when data changes by a row i.e. maximum distance between vectors $Q(I)$ and $Q(I')$ given by $\max \left| Q(I) - Q(I') \right|_p$ where $p = \{1,2\}$. $p = 2$ represents L_2 distance metric i.e. Euclidean distance and the corresponding sensitivity is called L_2 sensitivity.

In the proposed approach, to apply FPA, DFT of the centroid is computed, and *Lap(λ)* is added to the first 14 (*l*=14) co-efficients. In this case, $\Delta_2(Q) = \sqrt{m} \cdot \frac{a}{s}$ and $\lambda = \frac{\sqrt{l}\Delta_2(Q)}{\epsilon}$, where a is the size of each interval ($a = 1$), s is size of the cluster, m is the maximal difference (in terms of intervals) between two sequences. Assuming that a sequence can differ at most by 24 hrs per day, and since we have two weeks of data, m =24∗60∗14=20160. ε is privacy budget and it is set to 1. After noise addition, the sequence is padded with 0s and IDFT is computed. This sequence can then be used to answer the queries. Similar to MCKA, based on the probability of each activity in each interval of the perturbed centroid, as many sequences as the size of cluster is generated and these sequences satisfy ε-differential privacy. Figure 3 shows the centroid for the two sequences U1 and U3, DFT of the centroid, addition of Laplace noise, reconstruction of the centroid and sequences simulated using the noisy centroid.

Figure 3. Illustration of multi-level clustering based differential privacy



EXPERIMENTS AND RESULTS

This section starts by describing the experimental set up. It is followed by the experiments for determining the optimal parameters for proposed methods (MC, MCKA and MCDP). Next, the proposed approach is compared with an existing approach as baseline for efficiency and data utility. This is followed by the description of the scaling capabilities of the proposed approach.

Experimental Setup

System Requirements

The experiments were conducted on a computer with 32 GB Random Access Memory (RAM) and 3.2 Gigahertz (GHz) processor running the Windows 10 operating system. R was used for programming and all the experiments were single thread implementations.

Data

For the experiments, synthetic data generated from student life data set was used (refer to the Data subsection in Methods section for a detailed explanation). The synthetic data used in the experiments consists of 9800 students' data (1.9 GB). This data contains activity information for every one-minute interval, for two weeks, for each student. The dimensions of the data set are 9800 rows and 20160 columns.

Algorithms

The algorithms compared in the experiments are as follows:

1. MCKA: this is the proposed approach which first clusters activity sequences using Multi-level Clustering (MC). This step is followed by applying k-anonymity model by computing centroid for each cluster and using the centroid to generate activity sequences.
2. MCDP: this is the proposed approach which clusters activity sequences using MC followed by implementation of differential privacy. Fourier Perturbation Algorithm (FPA) is used to implement differential privacy.
3. MDAV-KA: this is a baseline method that uses MDAV (Domingo-Ferrer et al., 2006b; Solanas, Martinez-Balleste, & Domingo-Ferrer, 2006) to cluster activity sequences and implements k-anonymity model afterwards.
4. MDAV-DP: this is also a baseline method that uses MDAV to cluster activity sequences and implements differential privacy model afterwards.

The MDAV method ran out of memory over the synthetic data set at minute-level, so the data was aggregated to daily level before running MDAV-KA and MDAV-DP.

Metrics

Execution time was used to compare the efficiency of the proposed approach with the baseline method. Each method was run five times and the average execution time is reported. The time complexity of the proposed and the baseline approach are presented in the Methods section.

In terms of privacy metrics, k=5 was used for MCKA and MDAV-KA where k measures the degree of privacy protection for k-anonymity model. For the two methods implementing differential privacy model (MCDP and MDAV-DP), privacy budget was set to ε=1.

Utility of anonymized data is also important for an anonymization method because anonymized data needs to be useful. Utility was measured by the following metrics:

1. Relative difference

Relative difference between un-anonymized data and anonymized data was used as the metric for evaluating the effectiveness of the proposed approach. Relative difference (Törnqvist, Vartia, & Vartia, 1985) is computed as:

$$d\left(x,y\right)=\frac{\left|x-y\right|}{\max\left(x,y\right)}$$

x and y represent un-anonymized and anonymized activity sequence, respectively. Entropy-based metrics (Bayardo & Agrawal, 2005; Sweeney, 2002a; Sweeney, 2001) for evaluating utility were not used since the proposed approach did not involve generalization and suppression-based anonymization. Relative error was not used because values of x or y are often zero.

2. t-test and Cohen's d (effect size)

Relative difference was computed for both proposed approach and the baseline method. These relative differences were then compared for any significant differences between their means using t-test. A t-test represents whether the difference between different methods is statistically significant.

However, it does not tell anything about the size of the effect. Therefore, in addition to the p-values, effect sizes are also reported. Effect size was calculated using Cohen's d value (Cohen, 2013).

3. Correlations

Pearson's correlation values between activity-flourishing scale, and activity-CGPA were computed for both un-anonymized and anonymized data. This shows whether anonymization preserves correlations in the data.

Parameter Setting

In this section, the experiments for empirically determining the optimal parameters for MC, MCKA and MCDP are presented. These parameters include 1) fan-out at intermediate level, 2) number of records in a leaf node, 3) number of levels, 4) aggregation of time intervals at each level, and 5) weights for Euclidean distance. The changes in the relative difference are used to determine the optimal values of the parameters. It should be noted that they can vary based on the domain and requirements.

Optimal Fan-Out at Intermediate Level

Let k be the number of sequences in a leaf node (these nodes represent the final clusters). The number of records in an intermediate node is $k*p^l$, where l is the level (leaf node has $l = 0$, penultimate node (one level above the leaf level) has $l = 1$ and so on) and p is the fan-out in the non-leaf level. Experiments were conducted for different values of p (2, 10, 50, and 250) with $k = 5$. Two levels were used. At the root level, sequences were aggregated to the entire period spanned by the data and at the penultimate level, sequences were aggregated to daily intervals. The relative differences between the un-anonymized and anonymized sequences for the running activity using MCKA are shown in Figure 4a. Figure 4b shows the time taken for clustering, for different values of p.

Relative difference decreases slightly as p increases. This could be because larger value of p (fan-out) leads to larger clusters at intermediate level, which may lead to better clusters in the subsequent levels. Time initially decreases with increase in p value. For a very small value of p such as $p=2$, a large number of clusters need to be created, which is time consuming. As p increases, number of clusters to be computed decreases but after $p = 50$, as the clusters get bigger, computing clusters for consecutive level results in increased time. From the graphs, the optimal value of p is 50 because, at this value of p, time taken for clustering is minimum and the relative difference is very close to the minimum value.

Optimal Number of Records in a Leaf Node (k)

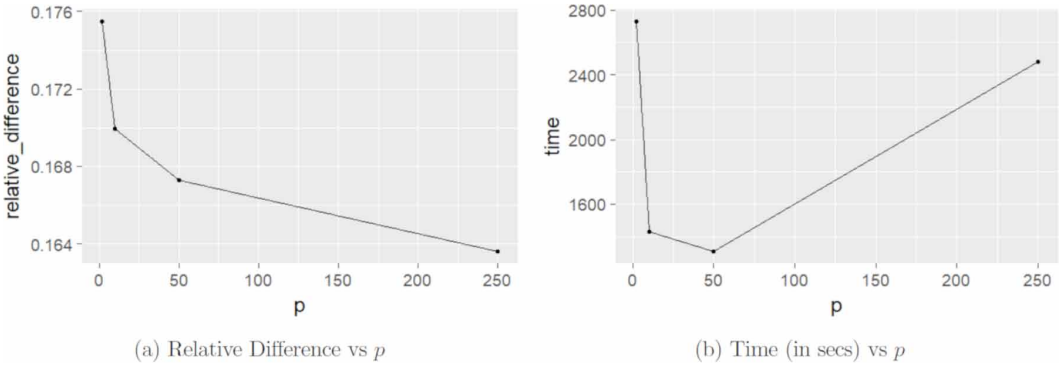
To determine the optimal value of k, experiments were conducted for different values of k (5, 10, 50, 100), with two levels of clustering. Number of records in an intermediate level node is the minimum of $k*50$ ($p = 50$ is the optimal setting for p) and $N/2$ (there should be at least two intermediate level nodes). Relative differences for the running activity using MCKA and MCDP are reported.

For MCKA (Figure 5a), the relative difference increases with the increase in k because, homogeneity of the clusters decreases with increase in k. For MCDP (Figure 5b), the relative difference decreases with the increase in k initially, because noise-to-be-added decreases with the increase in the size of cluster. After $k = 50$, relative difference slightly increases because of the decreased homogeneity of the clusters. The optimal value of k for MCKA is 5 and optimal value of k for MCDP is 50.

Optimal Number of Levels in MC

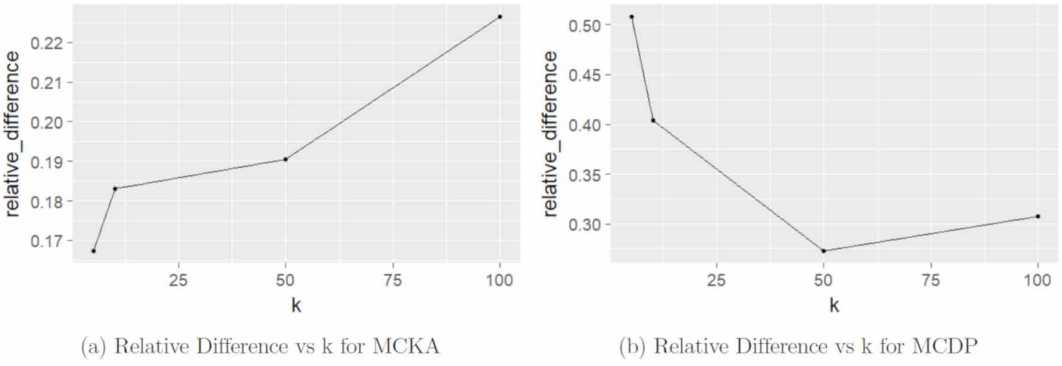
Experiments were conducted with two and three levels in MC. At the root level, the sequences were aggregated to the entire time duration and at all the other levels, they were aggregated to daily intervals. $k = 5$ for a leaf node and number of records at an intermediate level node is the minimum of $k*50^l$ and

Figure 4. Graphs for determining optimal fan-out at intermediate level



(a) Relative Difference vs p (b) Time (in secs) vs p

Figure 5. Graphs for determining optimal number of records in a leaf node



(a) Relative Difference vs k for MCKA (b) Relative Difference vs k for MCDP

N/2. Relative difference (for MCKA, for *running*) was very similar for two and three levels. However, time taken for clustering increased with the increase in the number of levels (1,280 seconds for two levels and 1,951 seconds for three levels). The optimal number of levels is 2.

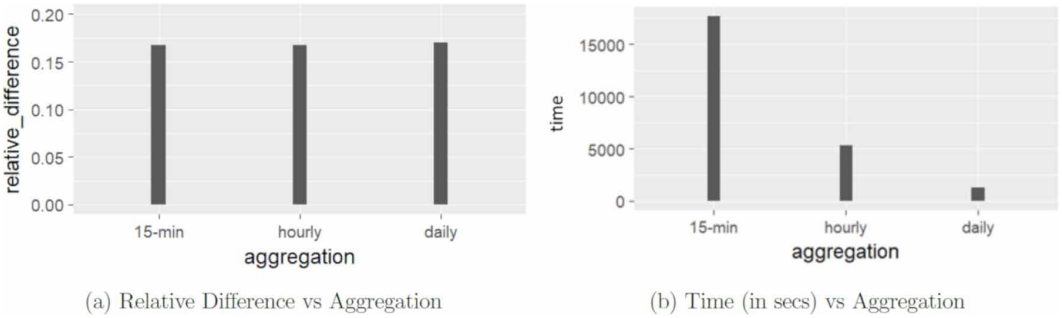
Optimal Aggregation of Time Intervals at Each Level

To determine optimal aggregation, three experiments with two levels were conducted. At the root level, sequences were aggregated to the entire time duration and at the penultimate level they were aggregated to daily, hourly and 15-minute intervals in each of the three experiments, respectively. Relative difference does not change much for different aggregations (Figure 6a). However, the time taken for clustering decreases drastically with the increase in the aggregation (17651 seconds for 15-min intervals and 1280 seconds for daily intervals) (Figure 6b). Therefore, optimal aggregation is entire-time-duration at the root level and daily interval at the penultimate level.

Weights for Euclidean Distance

Two experiments were conducted, one with equal weights for stationary, walking, running and missing, and another with weights in the ratio 1, 50, 150, 50. Running, walking and missing were given higher weights as compared to stationary because these states were infrequent in the data set. Relative difference (for running using MCKA) did not change much for equal weights and ratio-based weights. Therefore, equal weights were used for stationary, walking, running, and missing respectively.

Figure 6. Graphs for determining optimal aggregation



(a) Relative Difference vs Aggregation

(b) Time (in secs) vs Aggregation

Evaluation

Efficiency

The execution time of various anonymization algorithms is dominated by the time for clustering. So, using the optimal values for the parameters, experiments were conducted to compare the efficiency of MC (Multi-level clustering, which is used in MCKA and MCDP) and MDAV (clustering method used in MDAV-KA and MDAV-DP). MDAV ran out of memory when applied directly without aggregation (Table 4, column 4). Therefore, sequences were aggregated to daily level for MDAV (consistent with the aggregation in MC). Time taken for clustering (averaged over five runs) are shown in Table 4. The proposed approach took 21 minutes to complete, whereas the baseline method took more than two hours. In all subsequent experiments, MDAV-KA and MDAV-DP have their data aggregated to daily level.

Data Utility

Total duration of each activity on each day was computed for each un-anonymized activity sequence and its corresponding anonymized activity sequence. The relative difference between them was used to evaluate data utility. Lower value of relative difference means that they are close to each other. Hence, the lower the value, the higher the data utility of the anonymized data.

Table 5 reports the average relative difference for different activities. For example, using MCKA, anonymized data had on average 0.08 relative difference on daily duration of being stationary from un-anonymized data. Running and walking had higher relative difference as compared to stationary because they were infrequent (less than 3% of total time).

MCKA and MCDP had average relative difference comparable with that of MDAV-KA and MDAV-DP, respectively. Nevertheless, to check whether the difference is statistically significant between the proposed approach and the baseline, t-test was conducted. The t-test was conducted between the relative difference values from MCKA and MDAV-KA. It was also conducted between the relative difference values from MCDP and MDAV-DP. Cohen's d values were also computed to compare the effect size of the difference.

Table 6 shows the p-values and Cohen's d values. Under k-anonymity model (MCKA and MDAV-KA), p-values are significant. This means that, difference between MCKA and MDAV-KA

Table 4. Time for clustering

	MC (k = 5)	MDAV (k = 5) (with aggregation)	MDAV (k = 5) (No aggregation)
Time for clustering	21 mins	2.6 hrs.	>12.6 hrs. (memory issues)

Table 5. Average relative difference per day (Daily) (S stands for stationary, W stands for walking, R stands for running and M stands for Missing)

	MCKA	MDAV-KA	MCDP	MDAV-DP
	k=5	k=5	k=50	k=50
Daily (S)	0.08	0.08	0.14	0.14
Daily (W)	0.23	0.22	0.35	0.35
Daily (R)	0.17	0.16	0.27	0.27
Daily (M)	0.22	0.21	0.42	0.42

is statistically significant. However, since Cohen's d is less than 0.1 the difference is considered small (Cohen, 2013). Under differential privacy model, the difference between MCDP and MDAV-DP is mostly not statistically significant except for missing values. The results mean that our proposed methods (MCKA and MCDP) lead to similar or slightly worse data utility than the existing methods (MDAV-KA and MDAV-DP), but they are far more efficient than existing methods (see Table 4 where our methods achieved over 5 times speedup even if MDAV aggregates data). Given that physical activity data has very high dimensions, our methods achieve better performance-utility tradeoff than existing methods.

In addition, as an indicator of preserving utility, experiments were conducted to compare activity-flourishing scale (self-perceived success) correlation and activity-CGPA correlation. Pearson's correlation co-efficient for the un-anonymized and anonymized data are shown in Table 7. This shows that the proposed approach preserves both direction and magnitude of the correlations after anonymization.

Scalability

To demonstrate the scaling capabilities, MC was conducted for different number of students (2450, 4900, 7350, 9800) and different duration of the data (one-day, one-week, two-weeks, one-month). The results are shown in Figure 7a and Figure 7b. It can be seen that the algorithm scales linearly. This is because MC spends most time on clustering in lower levels due to increased dimensionality and at lower level n_i (size of cluster) is very small and the cost of clustering is almost linear to n (number of rows) and m_i (number of dimensions).

Table 6. Result of the t-test and Cohen's d between relative difference values from the proposed and the baseline approach

	MCKA & MDAV-KA		MCDP & MDAV-DP	
	p-value	Cohen's d	p-value	Cohen's d
Daily (S)	1.535e-06	0.07	0.35	0.01
Daily (W)	5.309e-07	0.07	0.16	0.02
Daily (R)	0.00014	0.05	0.38	0.01
Daily (M)	1.396e-08	0.08	0.0001	0.05

Table 7. Pearson's correlation co-efficient for activity-flourishing scale and activity-CGPA

	Activity-Flourishing Scale		Activity-CGPA	
	Correlation (*r*)	**p-value**	**Correlation (*r*)**	**p-value**
Un-anonymized data	0.146	<2.20E-16	-0.289	<2.20E-16
MCKA (k=5)	0.146	<2.20E-16	-0.290	<2.20E-16
MCDP (k=50)	0.129	<2.20E-16	-0.293	<2.20E-16

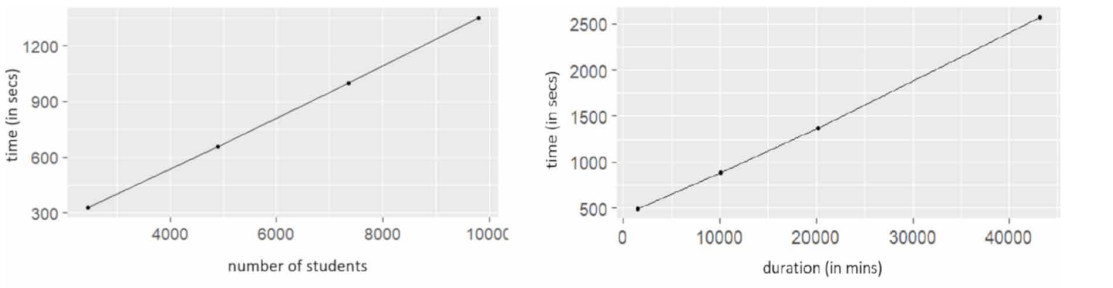
DISCUSSION

The usefulness of any anonymization technique relies greatly on its capability of preserving data utility. This is important because, if neglected, research studies performed using anonymized data might result in inaccurate and unreliable results. This would in turn defeat the purpose of data publishing. In this study, relative difference between un-anonymized and anonymized data is used to demonstrate the utility preserving capability of the proposed approach. In addition, correlation analysis conducted demonstrates the utility preserving capabilities of the approach. Table 5 shows that MCKA and MCDP have relative difference comparable with MDAV-KA and MDAV-DP. Correlation between activity and flourishing scale for un-anonymized data shows significant positive correlation, and the correlation between activity and CGPA for un-anonymized data shows significant negative correlation (Table 7). The proposed approach preserves both direction and magnitude of the correlations after anonymization. The results show that the proposed approach keeps the utility of the data set intact.

With the increase in the use of wearable sensors, the magnitude of the data collected is increasing dramatically (Banaee, Ahmed, & Loutfi, 2013). This requires the anonymization algorithms to be highly efficient. There are existing approaches that deal with large-scale data anonymization (Zhang et al., 2014). However, they do not take into consideration sequential data sets. Table 4 shows that MC reduces computation time from hours to minutes when compared to the conventional clustering technique such as MDAV (with the proposed dimensionality reduction). Without the dimensionality reduction, MDAV runs into memory issues.

Theoretically we show that our approach has complexity significantly lower than MDAV due to two reasons: 1) at higher level our method aggregates data to very low dimensions; 2) at lower level our method only needs to further divide small clusters rather than the whole data set. In practice, our study shows that it is possible to achieve drastic speed up with the cost of slight drop in data utility when anonymizing large scale physical activity data. These findings agree with existing studies on large scale data anonymization (Zhang, Liu, Nepal, Yang, Dou, & Chen, 2014; Zhang, Dou, Pei, Nepal, Yang, Liu, & Chen, 2014), where they also achieve significant speed up with little loss of data

Figure 7. Scalability of MC



(a) Scaling with Number of Students

(b) Scaling with Duration of the Data

utility. The main difference though is that their approach considers cross-sectional data while our approach focuses on sequential physical activity data. In addition, their approach uses Map-Reduce, which can be an interesting future research direction for our approach as at lower level clusters can be divided into smaller clusters in parallel.

The proposed approach protects privacy while preserving utility, and it speeds up the process drastically as compared to the conventional methods. Therefore, the authors recommend the use of the proposed approach for publishing physical activity data. When published data is used for large number of data points, MCDP can be used because, in MCDP, as the number of data points increases the noise added decreases. MCKA is suitable for both small and large data sets. It preserves utility better than MCDP but provides weaker privacy protection because *k*-anonymity is a weaker privacy model than differential privacy. Both MCKA and MCDP involve the same steps for clustering and only differ in the final step, which is the centroid generation. Therefore, data curators can reuse most of the code for MCKA to compute MCDP and vice-versa.

Like any study, there are certain limitations: 1) Due to unavailability of demographic information in the data set, the study lacks the analysis at the sub-population levels. 2) Synthetic data is used for demonstrating the approach and the original data set contains information for only three types of activities. However, in real-life there are more day-to-day activities such as cycling, climbing and sleeping, that could be considered. 3) Parameter values proposed are suitable for the data set used and might vary depending on the domain and requirements. The last two limitations can be overcome by applying the technique to different data sets. The authors leave this challenge as a part of the future work. 4) Dataset used in this study focuses on student population, however, health related issues are generally associated with elderly population (over the age of 65) (Yanco & Haigh, 2002). Nevertheless, it is worth noting that mental health problems associated with students itself is a highly significant problem in the United States, in terms of severity and number (Hunt & Eisenberg, 2010). The proposed approach is generic enough to be extended to other similar data sets, for example, those related to the activities of the elderly.

In addition, NIH (National Institutes of Health) data sharing policy (National Institutes of Health, 2003) supports and endorses data publication in order to enable free flow of information. However, organizations might refrain from publishing data in order to avoid HIPAA (Health Insurance Portability and Accountability Act of 1996) non-compliance penalties. The proposed approach can enable organizations to follow the encouragements stated in the NIH data sharing policy.

CONCLUSION

This paper presents a multi-level clustering approach that addresses the issue of high dimensionality in anonymizing sequential data sets such as physical activity data. K-anonymity and differential privacy are then applied to the resulting clusters in order to ensure privacy protection. Compared with the conventional methods, MC based anonymization improves time complexity by reducing the clustering time drastically (from hours to minutes) as compared to MDAV. Both MCKA and MCDP show results (in terms of relative difference) comparable with privacy models applied to conventional clustering technique such as MDAV. The proposed approach also preserves correlation between activity and flourishing scale, and activity and CGPA. In addition, as a part of the future work, the authors intend to apply the same approach to different health related sequential data sets to verify the results and make potentially generalizable recommendations of optimal parameters for the approach.

REFERENCES

Ahmed, N., Natarajan, T., & Rao, K. R. (1974). Discrete cosine transform. *IEEE Transactions on Computers*, *100*(1), 90–93. doi:10.1109/T-C.1974.223784

Banaee, H., Ahmed, M. U., & Loutfi, A. (2013). Data mining for wearable sensors in health monitoring systems: A review of recent trends and challenges. *Sensors (Basel)*, *13*(12), 17472–17500. doi:10.3390/s131217472 PMID:24351646

Barak, O., Cohen, G., & Toch, E. (2016). Anonymizing mobility data using semantic cloaking. *Pervasive and Mobile Computing*, *28*, 102–112. doi:10.1016/j.pmcj.2015.10.013

Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on*, *Data Engineering ICDE 2005* (pp. 217-228). IEEE. doi:10.1109/ICDE.2005.42

Cavoukian, A., Mihailidis, A., & Boger, J. (2010). Sensors and in-home collection of health data: A privacy by design approach. *Information and Privacy Commissioner*.

Chan, H., & Perrig, A. (2003). Security and privacy in sensor networks. *Computer*, *36*(10), 103–105. doi:10.1109/MC.2003.1236475

CDC Foundation. (2015, January 28). *Worker Illness And Injury Costs U.S. Employers $225.8 Billion Annually*. Retrieved from www.cdcfoundation.org: https://www.cdcfoundation.org/pr/2015/worker-illness-and-injury-costs-us-employers-225-billion-annually

Clarke, A., & Steele, R. (2014). A smartphone-based system for population-scale anonymized public health data collection and intervention. In *Proceedings of the 2014 47th Hawaii International Conference on*, *System Sciences (HICSS)* (pp. 2908-2917). IEEE. doi:10.1109/HICSS.2014.363

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge. doi:10.4324/9780203771587

Dean, J., & Ghemawat, S. (2010). MapReduce: A flexible data processing tool. *Communications of the ACM*, *53*(1), 72–77. doi:10.1145/1629175.1629198

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, *24*(3), 596–606. PMID:28040687

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*(2), 143–156. doi:10.1007/s11205-009-9493-y

Dietz, W. H., Douglas, C. E., & Brownson, R. C. (2016). Chronic disease prevention: Tobacco avoidance, physical activity, and nutrition for a healthy start. *Journal of the American Medical Association*, *316*(16), 1645–1646. doi:10.1001/jama.2016.14370 PMID:27668419

Domingo-Ferrer, J., Mart'ınez-Ballest'e, A., Mateo-Sanz, J. M., & Seb'e, F. (2006a). Efficient multivariate data-oriented microaggregation. *The VLDB Journal*, *15*(4), 355–369. doi:10.1007/s00778-006-0007-0

Domingo-Ferrer, J., Solanas, A., & Martinez-Balleste, A. (2006b). Privacy in statistical databases: k-anonymity through microaggregation. In *Proceedings of the 2006 IEEE International Conference on Granular Computing* (pp. 774–777). IEEE.

Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, *11*(2), 195–212. doi:10.1007/s10618-005-0007-5

Dong, Y., & Pi, D. (2018). Novel privacy-preserving algorithm based on frequent path for trajectory data publishing. *Knowledge-Based Systems*, *148*, 55–65. doi:10.1016/j.knosys.2018.01.007

Dunn, P. F., & Davis, M. P. (2017). *Measurement and data analysis for engineering and science*. CRC press.

Dwork, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference* (pp. 265–284). Springer. doi:10.1007/11681878_14

El Emam, K. (2008). Heuristics for de-identifying health data. *IEEE Security and Privacy*, *6*(4).

El Emam, K., Arbuckle, L., Koru, G., Eze, B., Gaudette, L., Neri, E., & Gluck, J. et al. (2012). De-identification methods for open health data: The case of the heritage health prize claims dataset. *Journal of Medical Internet Research*, *14*(1), e33. doi:10.2196/jmir.2001 PMID:22370452

El Emam, K., & Dankar, F. K. (2008). Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, *15*(5), 627–637. doi:10.1197/jamia.M2716 PMID:18579830

El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., & Bottomley, J. et al. (2009). A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, *16*(5), 670–682. doi:10.1197/jamia.M3144 PMID:19567795

Gagniuc, P. A. (2017). *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons. doi:10.1002/9781119387596

Gal, T. S., Chen, Z., & Gangopadhyay, A. (2008). A privacy protection model for patient data with multiple sensitive attributes. [IJISP]. *International Journal of Information Security and Privacy*, *2*(3), 28–44.

Gal, T. S., Tucker, T. C., Gangopadhyay, A., & Chen, Z. (2014). A data recipient centered de-identification method to retain statistical attributes. *Journal of Biomedical Informatics*, *50*, 32–45. doi:10.1016/j.jbi.2014.01.001 PMID:24412834

Gao, S., Ma, J., Sun, C., & Li, X. (2014). Balancing trajectory privacy and data utility using a personalized anonymization model. *Journal of Network and Computer Applications*, *38*, 125–134. doi:10.1016/j.jnca.2013.03.010

Gkoulalas-Divanis, A., & Loukides, G. (2012). Utility-guided clustering-based transaction data anonymization. *Transactions on Data Privacy*, *5*(1), 223–251.

He, W., Liu, X., Nguyen, H., Nahrstedt, K., & Abdelzaher, T. (2007). Pda: Privacy-preserving data aggregation in wireless sensor networks. In *Proceedings of the 26th IEEE International Conference on Computer Communications INFOCOM 2007* (pp. 2045–2053). IEEE Press.

He, X., Cormode, G., Machanavajjhala, A., Procopiuc, C. M., & Srivastava, D. (2015). Dpt: Differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, *8*(11), 1154–1165. doi:10.14778/2809974.2809978

Hu, Z., Yang, J., & Zhang, J. (2018). Trajectory privacy protection method based on the time interval divided. *Computers & Security*, *77*, 488–499. doi:10.1016/j.cose.2018.05.001

Hunt, J., & Eisenberg, D. (2010). Mental health problems and help-seeking behavior among college students. *The Journal of Adolescent Health*, *46*(1), 3–10. doi:10.1016/j.jadohealth.2009.08.008 PMID:20123251

Kakatkar, C., & Spann, M. (2019). Marketing analytics using anonymized and fragmented tracking data. *International Journal of Research in Marketing*, *36*(1), 117–136. doi:10.1016/j.ijresmar.2018.10.001

Kitamura, R., Chen, C., & Pendyala, R. (1997). Generation of synthetic daily activity-travel patterns. *Transportation Research Record: Journal of the Transportation Research Board*, *1607*(1), 154–162. doi:10.3141/1607-21

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79–86. doi:10.1214/aoms/1177729694

Li, M., Lou, W., & Ren, K. (2010). Data security and privacy in wireless body area networks. *Wireless Communications, IEEE*, *17*(1), 51–58. doi:10.1109/MWC.2010.5416350

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the IEEE 23rd International Conference on Data Engineering ICDE 2007* (pp. 106-115). IEEE.

Loukides, G., Gkoulalas-Divanis, A., & Malin, B. (2010). Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(17), 7898–7903. doi:10.1073/pnas.0911686107 PMID:20385806

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 3, es. doi:10.1145/1217299.1217302

Martinez-Bea, S., & Torra, V. (2011). Trajectory anonymization from a time series perspective. In *Proceedings of the 2011 IEEE International Conference on Fuzzy Systems (FUZZ)* (pp. 401–408). IEEE. doi:10.1109/FUZZY.2011.6007405

Matthews, C. E., Chen, K. Y., Freedson, P. S., Buchowski, M. S., Beech, B. M., Pate, R. R., & Troiano, R. P. (2008). Amount of time spent in sedentary behaviors in the united states, 2003–2004. *American Journal of Epidemiology*, *167*(7), 875–881. doi:10.1093/aje/kwm390 PMID:18303006

Mendez-Vazquez, A., Helal, A., & Cook, D. (2009). Simulating events to generate synthetic data for pervasive spaces. In *Workshop on Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing Research*. Academic Press.

Mohammed, N., Fung, B., & Debbabi, M. (2009). Walking in the crowd: anonymizing trajectory data for pattern analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1441–1444). ACM. doi:10.1145/1645953.1646140

Monekosso, D., & Remagnino, P. (2009). Synthetic training data generation for activity monitoring and behavior analysis. In *Proceedings of the European Conference on Ambient Intelligence* (pp. 267–275). Springer. doi:10.1007/978-3-642-05408-2_31

Moon, Y. S., Kim, H. S., Kim, S. P., & Bertino, E. (2010, August). Publishing time-series data under preservation of privacy and distance orders. In *Proceedings of the International Conference on Database and Expert Systems Applications* (pp. 17-31). Springer. doi:10.1007/978-3-642-15251-1_2

National Institutes of Health. (2003, February 26). *FINAL NIH STATEMENT ON SHARING RESEARCH DATA*. Retrieved from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

Nergiz, M. E., Atzori, M., & Saygin, Y. (2008). Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS* (pp. 52–61). ACM. doi:10.1145/1503402.1503413

Nergiz, M. E., Atzori, M., Saygin, Y., & Prog, C. (2007). Perturbation-driven anonymization of trajectories.

Nin, J., & Torra, V. (2009). Towards the evaluation of time series protection methods. *Information Sciences*, *179*(11), 1663–1677. doi:10.1016/j.ins.2009.01.024

Pate, R. R., Pratt, M., Blair, S. N., Haskell, W. L., Macera, C. A., & Bouchard, C. et al.. (1995). Physical activity and public health: A recommendation from the centers for disease control and prevention and the American college of sports medicine. *Journal of the American Medical Association*, *273*(5), 402–407. doi:10.1001/jama.1995.03520290054029 PMID:7823386

Pensa, R. G., Monreale, A., Pinelli, F., and Pedreschi, D. (2008). Pattern-preserving k-anonymization of sequences and its application to mobility data mining. In *International Workshop on Privacy in Location-Based Applications PiLBA'08* (pp. 44-60). Academic Press.

Poulis, G., Loukides, G., Skiadopoulos, S., & Gkoulalas-Divanis, A. (2017). Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints. *Journal of Biomedical Informatics*, *65*, 76–96. doi:10.1016/j.jbi.2016.11.001 PMID:27832965

Rastogi, V., & Nath, S. (2010). Differentially private aggregation of distributed time-series with transformation and encryption. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 735–746). ACM. doi:10.1145/1807167.1807247

Shou, L., Shang, X., Chen, K., Chen, G., & Zhang, C. (2013). Supporting pattern-preserving anonymization for time-series data. *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 877–892. doi:10.1109/TKDE.2011.249

Solanas, A., Martinez-Balleste, A., & Domingo-Ferrer, J. (2006). V-mdav: a multivariate microaggregation with variable group size. In *Proceedings of the 17th COMPSTAT Symposium of the IASC* (pp. 917–925). Academic Press.

Spedicato, G. A. (2016). *markovchain: Discrete Time Markov chains made easy*. R package version 0.6.

Spees, C. K., Scott, J. M., & Taylor, C. A. (2012). Differences in amounts and types of physical activity by obesity status in us adults. *American Journal of Health Behavior*, *36*(1), 56–65. doi:10.5993/AJHB.36.1.6 PMID:22251783

Spenkelink, C., Hutten, M. M., Hermens, H., & Greitemann, B. O. (2002). Assessment of activities of daily living with an ambulatory monitoring system: A comparative study in patients with chronic low back pain and nonsymptomatic controls. *Clinical Rehabilitation*, *16*(1), 16–26. doi:10.1191/0269215502cr463oa PMID:11841065

Stubbs, A., Filannino, M., & Uzuner, O¨. (2017). De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of Biomedical Informatics*, *75*, S4–S18. doi:10.1016/j.jbi.2017.06.011 PMID:28614702

Stubbs, A., & Uzuner, Ö. (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, *58*, S20–S29. doi:10.1016/j.jbi.2015.07.020 PMID:26319540

Sun, J., Fang, Y., & Zhu, X. (2010). Privacy and emergency response in e-healthcare leveraging wireless body sensor networks. *Wireless Communications, IEEE*, *17*(1), 66–73. doi:10.1109/MWC.2010.5416352

Sweeney, L. (2001). *Computational disclosure control: a primer on data privacy protection* [PhD thesis]. Massachusetts Institute of Technology.

Sweeney, L. (2002a). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, *10*(5), 571–588. doi:10.1142/S021848850200165X

Sweeney, L. (2002b). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, *10*(05), 557–570. doi:10.1142/S0218488502001648

Tamersoy, A., Loukides, G., Nergiz, M. E., Saygin, Y., & Malin, B. (2012). Anonymization of longitudinal electronic medical records. *IEEE Transactions on Information Technology in Biomedicine*, *16*(3), 413–423. doi:10.1109/TITB.2012.2185850 PMID:22287248

Templ, M., Meindl, B., and Kowarik, A. (2013). Introduction to statistical disclosure control.

Terrovitis, M., Mamoulis, N., & Kalnis, P. (2008). Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, *1*(1), 115–125. doi:10.14778/1453856.1453874

Thornton, J. S., Frémont, P., Khan, K., Poirier, P., Fowles, J., Wells, G. D., & Frankovich, R. J. (2016). Physical activity prescription: a critical opportunity to address a modifiable risk factor for the prevention and management of chronic disease: a position statement by the Canadian Academy of Sport and Exercise Medicine. *British Journal of Sports Medicine*, *50*(18), 1109–1114. doi:10.1136/bjsports-2016-096291 PMID:27335208

Törnqvist, L., Vartia, P., & Vartia, Y. O. (1985). How should relative changes be measured? *The American Statistician*, *39*(1), 43–46.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., & Campbell, A. T. et al. (2014). Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. doi:10.1145/2632048.2632054

Wang, L. E., & Li, X. (2018). A graph-based multifold model for anonymizing data with attributes of multiple types. *Computers & Security*, *72*, 122–135. doi:10.1016/j.cose.2017.09.003

Yanco, H. A., & Haigh, K. Z. (2002). Automation as caregiver: A survey of issues and technologies. *Am. Assoc. Artif. Intell*, *2*, 39–53.

Zhang, X., Liu, C., Nepal, S., Yang, C., Dou, W., & Chen, J. (2014). A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud. *Journal of Computer and System Sciences*, *80*(5), 1008–1020. doi:10.1016/j.jcss.2014.02.007

Zhang, X., Dou, W., Pei, J., Nepal, S., Yang, C., Liu, C., & Chen, J. (2014). Proximity-aware local-recoding anonymization with mapreduce for scalable big data privacy preservation in cloud. *IEEE Transactions on Computers*, *64*(8), 2293–2307. doi:10.1109/TC.2014.2360516

Zhiyuan Chen is an Associate Professor in the Department of Information Systems at the University of Maryland Baltimore County. He received a PhD degree in Computer Science from Cornell University in August 2002. He has more than 10 years of extensive research experience in data privacy, privacy preserving data mining, database management, data science, and cyber security. His main research focus is in algorithms for preserving privacy of data and at the same time allows accurate analysis of the data. He has published over 40 papers in peer reviewed journals and publications and over 20 of them are in the area of privacy and security. More information can be found at https://userpages.umbc.edu/~zhchen/

Gunes Koru is a Professor in the Department of Information Systems at the University of Maryland, Baltimore County. He conducts research on the intersection of health services and information technology and teaches related undergraduate and graduate courses. As the Principal Investigator of Federal and State sponsored research projects, Dr. Koru has employed many graduate students in his lab over the years and served as their academic advisor. His student--co-authored publications and presentations appeared in quality journals and conferences on nursing, health informatics, software engineering, home care, gerontology, and health services. Dr. Koru is among the founders of the Home Care and Hospice Information Technology (H3IT) Conference and continues to serve as its organizing chair. He served in the Program Committee of the Annual Symposium of the American Medical Informatics Association (AMIA), Editorial Board of the American Medical Directors Association (AMDA), and Research Work Group for the Alliance for Home Health Quality and Innovation (AHHQI). He also served his institution as a member of the Academic Planning Budget Committee, as the Chair of the Research Council, as a Faculty Senator, and recently as the President of the Faculty Senate in addition to performing various other service activities in his department, university, and in the University System of Maryland (USM) over the years.