Anonymization of Daily Activity Data by Using ℓ-diversity Privacy Model

POOJA PARAMESHWARAPPA, ZHIYUAN CHEN, and GÜNEŞ KORU, University of Maryland, Baltimore County, USA

In the age of IoT, collection of activity data has become ubiquitous. Publishing activity data can be quite useful for various purposes such as estimating the level of assistance required by older adults and facilitating early diagnosis and treatment of certain diseases. However, publishing activity data comes with privacy risks: Each dimension, i.e., the activity of a person at any given point in time can be used to identify a person as well as to reveal sensitive information about the person such as not being at home at that time. Unfortunately, conventional anonymization methods have shortcomings when it comes to anonymizing activity data. Activity datasets considered for publication are often flat with many dimensions but typically not many rows, which makes the existing anonymization techniques either inapplicable due to very few rows, or else either inefficient or ineffective in preserving utility. This article proposes novel multi-level clustering-based approaches using a non-metric weighted distance measure that enforce ℓ-diversity model. Experimental results show that the proposed methods preserve data utility and are orders more efficient than the existing methods.

CCS Concepts: • **Security and privacy → Security services**; **Pseudonymity, anonymity and untraceability**; • **Information systems → Information systems applications**; **Data mining**; **Clustering**; • **Computing methodologies → Machine learning**; **Learning paradigms**; **Unsupervised learning**; **Cluster analysis**;

Additional Key Words and Phrases: k-anonymity, ℓ-diversity, clustering, privacy, anonymization

ACM Reference format:
Pooja Parameshwarappa, Zhiyuan Chen, and Güneş Koru. 2021. Anonymization of Daily Activity Data by Using ℓ-diversity Privacy Model. *ACM Trans. Manage. Inf. Syst.* 12, 3, Article 23 (May 2021), 21 pages.
https://doi.org/10.1145/3456876

1 INTRODUCTION

Daily activity data, referred to as activity data henceforth, typically belong to the activities representing daily routines such as bathing, dressing, feeding, walking, driving, shopping, and so on [28, 33]. Table 1 shows an example of activity data. With the increase in the availability of reliable sensors, collection of activity data has become a commonplace in many application domains [6, 44]. Publishing such data can be useful in a number of ways. For example, analyzing the activities performed by older adults can be useful in estimating the level of assistance they will require. In addition, it can also support various quality improvement and research activities in healthcare

Authors' addresses: P. Parameshwarappa, Z. Chen, and G. Koru, Department of Information Systems, Healthcare Informatics and Technologies Lab, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD, 21250; emails: {poojap1, zhchen, gkoru}@umbc.edu.

2158-656X/2021/05-ART23 $15.00
https://doi.org/10.1145/3456876

Table 1. Daily Activity Data (B stands for Bathing, E stands for Eating, S stands for Sleeping, and V stands for Vacation)

	Mon		..	Tue		..	..	Fri		..	Sat		..	Sun		..
	6AM	7AM		6AM	7AM			6AM	7AM		6AM	7AM		6AM	7AM	
Person 1	S	E		S	S			V	V		V	V		V	V	
Person 2	S	E		S	S			V	V		V	V		V	V	
Person 3	B	E		S	S			S	S		V	V		V	V	
Person 4	B	E		S	S			V	V		V	V		V	V	

such as predicting hospital admissions, utilization of home care services, planning of insurance services, and so on [21, 41].

However, publishing activity data come with certain privacy risks. Although personally identifiable information may be removed from the data, there can be other sensitive information available in the data that might lead to privacy breach. For example, if most of the people whose information is in the dataset take a vacation during the same time of a year, then this information might help an adversary to plan a break-in. There have been many incidents where burglars used publicly available information to plan the thefts [46, 56]. Statistics show that over 75% of the convicted burglars believe that other burglars make use of the activity data available in social media sites to identify their targets [55]. One burglar in California used social media for this purpose and stole over $250,000 worth of items from 33 women [39]. Such incidents demonstrate the existence of motivations to use publicly available data to commit crime and the importance and need for anonymizing datasets before making them publicly available.

One of the most widely used privacy models is k-anonymity [51]. k-anonymity ensures that for every record in the data, there are at least $k-1$ other records that have exactly same values for the quasi-identifiers (attributes that when combined can identify a person). This reduces the probability of correct re-identification when attackers link the quasi-identifiers with external datasets such as voters' databases to find out the identity of a person. However, k-anonymity is susceptible to homogeneity and background-knowledge attacks [35] when people in the same equivalence class (i.e., those with the same values for quasi-identifier attributes) have similar sensitive attribute values (say a certain type of disease). To overcome such attacks several privacy models such as ℓ-diversity [35] and t-closeness [34] have been proposed. ℓ-diversity ensures that the values for sensitive attribute are well represented in each equivalence class. t-closeness [34] further requires that the distribution of sensitive attribute in the equivalence class is close to its distribution in the whole dataset. In practice, though, ℓ-diversity model is used more widely.

The above-mentioned traditional approaches were developed for a limited number of quasi-identifiers. However, for activity data, the quasi-identifiers include all time dimensions. The quasi-identifiers themselves may also contain sensitive information. For example, people going on vacation for a certain time-interval is the sensitive information in Table 1. If these four people are in the same equivalence class, then burglars may figure out that they can break into any or all of the four houses over the weekend. In addition, activity datasets are typically very flat with many dimensions but typically not many rows. Most available activity datasets contain fewer than 100 people. This poses challenges to existing data anonymization techniques. For example, differential privacy [16] is a stronger privacy model than k-anonymity and ℓ-diversity, but it cannot be used for datasets with very few rows as it works well for a large population. k-anonymity and ℓ-diversity models are applicable to the datasets with few rows. However, most of such techniques are suitable for data with relatively few dimensions. Activity data are collected over a period of time making it longitudinal in nature with very high dimensionality. For example, in a dataset

used in this article each record has over 10,000 dimensions. Applying these models directly to activity data becomes computationally expensive [43]. Furthermore, the techniques available for anonymizing longitudinal data mostly focus on preserving frequent patterns; however, for daily activity data, preserving statistics such as duration and frequency might be more relevant [4, 5, 14].

There has been relatively little work on applying ℓ-diversity for longitudinal data. In Reference [43], a **Multi-level Clustering-based *K*-Anonymity (MCKA)** approach was proposed that is orders faster than existing methods when anonymizing high-dimensional longitudinal data. However, MCKA does not enforce ℓ-diversity. In this article, we propose several approaches that implement ℓ-diversity for daily activity data.

This article makes the following contributions:

- We propose several **Multi-level Clustering- (MC)** based ℓ-diversity approaches for anonymizing longitudinal data such as daily activity data. Compared to existing methods, MC-based methods are more efficient and require a smaller memory footprint.
- We propose a non-metric weighted distance measure that can be used in any clustering method. This distance measure can help create better equivalence classes and better preserve utility of data.

This article is organized as follows. In Section 2 we describe the related work. Section 3 presents the proposed approaches. Experimental results are presented in Section 4. Section 5 discusses the results. Section 6 concludes the article.

2 RELATED WORK

ℓ-diversity [35] is one of the most common anonymization techniques along with k-anonymity [51] and differential privacy [15]. Although differential privacy is a stronger privacy model, we do not use it in this article because for differential privacy to work the dataset needs to have a large population, but most daily activity datasets contain a very small population (typically fewer than 100), making it inappropriate to use differential privacy.

A number of techniques for achieving k-anonymity for cross-sectional data have been proposed [17, 19, 20, 58, 59]. In References [13, 18, 25, 26, 36, 40, 43, 50], techniques for applying k-anonymity to longitudinal data are presented. In this section, we discuss some of the existing techniques to achieve ℓ-diversity privacy model for different types of data including cross-sectional data and longitudinal data.

In Reference [35], the authors mention that ℓ-diversity can be achieved by using lattice search algorithms similar to k-anonymity. Gal et al. [23] propose a method that extends k-anonymity and ℓ-diversity for datasets with multiple sensitive attributes. Rao et al. [45] propose a scalable anonymization technique for k-anonymity and ℓ-diversity privacy models using MapReduce technique. In this technique equivalence classes for scalable k-anonymity and **scalable ℓ-diversity (SLD)** are generated and given as input to the anonymization component. The authors proposed an improved algorithm ImSLD in Reference [37]. In this approach, columns that are quasi-identifiers are arranged in ascending order of number of unique values. Data are then grouped with respect to these columns to generate equivalence classes. Equivalence classes are merged until k-anonymity and l-diversity conditions are satisfied.

Typically, for ℓ-diversity, the dataset is assumed to have non-overlapping quasi-identifiers and sensitive attributes. In Reference [47], the authors propose a privacy model $l_1 \ldots l_q$ diversity and $t_1 \ldots t_q$ closeness that assumes sensitive quasi-identifiers. In this article, we assume that all attributes (activities at each time interval) are quasi-identifiers and one type of activity is sensitive.

In Reference [57], the authors propose a method called Data Privacy Preservation with Perturbation for anonymizing trajectory data by enhancing the ℓ-diversity model. This method first

identifies critical trajectories that can identify individuals and then perturbs them. In Reference [54], authors propose a technique for anonymizing trajectories to prevent re-identification and semantic attacks. Semantic attacks happen when adversary is able to attach semantics to the points of interest. However they do not consider data with very high dimensions as daily activity data.

Most of the existing techniques for anonymizing longitudinal data apply k-anonymity privacy model to the data. Not a lot of work has been done in applying ℓ-diversity privacy model, and they usually do not consider data with very high dimensions. So far, to the best of our knowledge, no work has been done is anonymizing daily activity data using ℓ-diversity model. In this article, we present approaches for anonymizing activity data using ℓ-diversity privacy model.

3 METHODS

Section 3.1 presents a novel (δ, ϵ)-diversity model that is suitable for activity data. Section 3.2 describes several existing methods to implement l-diversity. Section 3.3 describes the proposed approaches.

3.1 Privacy Model

We first introduce some notations. Let D be a dataset with n rows and each row represents activities of a person. Each row consists of m columns where m is the number of time intervals (can be minute, second, etc.), and q is the number of possible activities. Let D_{ij} be the value at row i and column j. D_{ij} represents the activity at time interval j. Since we often need to compute distance between two records, we use a size q bitmap to represent the activity at each time interval. For example, if at interval j at record i the activity is ath activity, then the ath bit of the vector is 1 and all other bits are 0. Each data record can now be represented as a size mq bit vector that is concatenation of m q-bit vectors. Distance computation can be done directly on these bit vectors. Note that, if data are aggregated at larger intervals (e.g., hours instead of minutes), then we can simply replace each bit with duration of that activity in a time interval.

We also assume that an activity s is sensitive. For example, being on vacation is a sensitive activity because burglars can use that to find targets.

ℓ-diversity: ℓ-diversity requires that for every equivalence class, the values for the sensitive-attribute be well-represented. The term "well represented" is defined in three different ways [35]:

- Distinct ℓ-diversity: Every equivalence class has at least l distinct values for the sensitive attribute
- Entropy ℓ-diversity: For every equivalence class E, $Entropy(E) >= log\ell$
- Recursive (c, ℓ)-diversity: Let m' be the number of distinct sensitive attribute values in an equivalence class, and r_i be the frequency of the ith most frequent value, $r_1 < r_\ell + r_{\ell+1} + \ldots r_{m'}$. This ensures that for every equivalence class, the most frequent value in the sensitive attribute is not too frequent and the infrequent values are not too infrequent.

Existing definitions focus on the frequency of sensitive values across rows. For activity data, we are also concerned with the duration of the sensitive activity. For example, someone who is away from home for a few days will have higher risk than someone who is away for just a few hours. So we propose a (δ, ϵ)-diversity model for activity data.

Definition 1. (δ, ϵ)-diversity: Let D be the dataset and E represent an equivalence class with all m dimensions as quasi-identifiers. Let N_{Sj} represent the number of records that have the sensitive value for at least δ consecutive time intervals from time unit j to time unit $\delta + j - 1$, $1 \leq j \leq m - \delta + 1$. E is said to satisfy (δ, ϵ)-diversity if $\frac{N_{Sj}}{|E|} \leq \epsilon, \forall j$. D satisfies (δ, ϵ)-diversity if every equivalence class in D satisfies (δ, ϵ)-diversity.

For example, assuming that the data are at hourly level granularity. The dataset shown in Table 1 satisfies $(72, 0.75)$-diversity as at most three of four people take 3 day (72 hours) vacation at the same time.

Definition 1 can be also generalized to multiple sensitive activities by requiring for each sensitive activity the (δ, ϵ)-diversity is satisfied.

3.2 Background

This section briefly reviews two existing algorithms that implement ℓ-diversity on cross sectional data. We will modify these two algorithms for activity data in Section 3.3.

3.2.1 ℓ-MDAV. Microaggregation [9] is one of the widely used methods for achieving k-anonymity. It has two steps: (1) *partitioning*, in which the records are partitioned into clusters of at least size k, and (2) *replacement*, where the records in each cluster are replaced by the result of an aggregation operation such as average. In this article, once the clusters are formed, all methods replace each cluster with its centroid.

One of the best-known algorithm for achieving the partitioning step of microaggregation is **Maximum Distance to Average Vector (MDAV)** [8, 11, 12]. Domingo et al., proposed a method called p-sensitive k-anonymity [10], which extends MDAV to ensure that each equivalence class contains p distinct sensitive attribute values. We refer to this as ℓ-MDAV. Below is the sketch of the algorithm.

(1) Compute the centroid of the dataset. 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 r and form a cluster with these points and r, form a similar cluster around s.
(3) If there are at least $2k$ data points remaining, then repeat steps (1) and (2) on the remaining points. Else, go to step (4).
(4) If there are between k and $2k-1$ points remaining, then form a new cluster with these points.
(5) If there are fewer than k points remaining, 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 points and add them to that cluster.
(6) If any of the resulting equivalence class (cluster) does not satisfy ℓ-diversity requirement, then increment k ($k = k + 1$) and repeat steps (1) to (5).

Steps (1) to (5) partition data into clusters with sizes at least k using MDAV. Step (6) is basically a backtrack step that increases k if ℓ-diversity is not satisfied.

The complexity of MDAV is $O(n^2mq)$ where n is number of records, m is number of time intervals, and q is number of activities. Checking for ℓ-diversity can be done in $O(km\delta)$ time for a cluster with size k as we just need to check for each record in the cluster whether it has sensitive activity at every δ consecutive time interval. There are around n/k clusters so total time for checking ℓ-diversity is $O(nm\delta)$. Suppose ℓ-MDAV backtracks b times. Its cost is thus $O(b(n^2mq + nm\delta))$. The most expensive step of this algorithm is to find nearest neighbors at step (2). So it is possible to speed up the algorithm by using a fast nearest neighbor search based on k-d tree or R-tree [24, 30].

3.2.2 ℓ-VMDAV. Han et al. [27] proposed a method to implement ℓ-diversity by extending a variant of MDAV called VMDAV [49], which generates variable sized clusters . This often generates more homogeneous clusters than MDAV and avoids backtrack. The sketch of the algorithm [27] is as follows:

(1) Compute a $n \times n$ distance matrix M for the dataset to save future distance computation (M_{ij} represents distance between record i and j).
(2) Compute the centroid c of the dataset.
(3) Repeat steps (4) to (9) until there are no more than $k-1$ unassigned records.
(4) Compute most distant record r from the centroid.
(5) Form a group $g = \{r\}$.
(6) Build a priority queue Q for the unassigned records in ascending order of their distance to r.
(7) Repeat steps (a) to (c) until g satisfies ℓ-diversity or there is no more points to add
  (a) Compute $D(g)$, which is the ℓ-diversity value of g
  (b) Pop the head v from Q and compute $D(g \cup \{v\})$
  (c) If $D(g \cup \{v\}) > D(g)$, then add v to g.
(8) If group g does not satisfy ℓ-diversity, then suppress records in g.
(9) Extend g with details in Reference [27].
(10) Assign the remaining records to the closest group.

This method does not backtrack. Instead, it tries to add unassigned points directly to a group g as long as the group's ℓ-diversity keeps improving. If the group never becomes ℓ-diverse, then all the records in the group will be suppressed.

After group g becomes ℓ-diverse, step (9) further extends g by adding unassigned points that are closer to g than to remaining unassigned points and the new group has same or better degree of ℓ-diversity. The extension ends when g contains $2k-1$ data points or all unassigned points have been checked.

In our implementation, we modify this algorithm by first computing clusters of size k and then checking for ℓ-diversity. Clusters are extended using priority queue (similar to step (6)) until l-diversity is satisfied or all the unassigned points have been checked. We do this to reduce the clustering time because extending the group using step (9) would be very inefficient due to the high-dimensional nature of the data. We refer to this as ℓ-VMDAV.

There are two limitations associated with this method: (1) There is no guarantee that each group (cluster) will become ℓ-diverse, so many records could be suppressed, and (2) this method is still quite expensive for high-dimensional data such as activity data as they require us to compute distances between all pairs of records on all dimensions.

The cost of computing distance matrix (step (1)) is $O(n^2mq)$. The cost of checking ℓ-diversity of a group g is $O(|g|m\delta)$, and each group can be checked at most $O(n)$ times (the number of possible records). The size of g is k to $2k-1$. So the cost of generating a group is $O(nkm\delta)$. There are up to n/k groups so the total cost is $O(n^2m(q+\delta))$.

3.3 Proposed Approaches

The existing methods including ℓ-MDAV and ℓ-VMDAV are quite expensive over high-dimensional data. An MC method was proposed to address this issue [43].

MC handles high dimensionality of the activity data by aggregating the records to different time intervals. This improves the efficiency of the anonymization process. In MC (shown in Algorithm 1) all the records are assigned to one cluster at the root level (line 1). The records are then aggregated to certain time intervals (line 5), for example daily intervals, and then clustered using MDAV (line 6). In the next level t, the records are aggregated to smaller time intervals (for example, hourly intervals) and each cluster c at the previous level is further divided into smaller clusters of size s_t

using MDAV. These steps are repeated until each cluster at the leaf level has at least k records in it.

At the level t, the complexity of this algorithm is $O(ns_t m_t q)$, where s_t is cluster size at level t and m_t is number of intervals at level t. There are l levels, so the complexity of MC is $O(nq \sum_{t=1}^{l} s_t m_t)$.

MC is more efficient than MDAV for two reasons. At the higher levels, data are aggregated to larger time intervals so $m_t << m$. At lower levels, data are aggregated so $m_t < m$, and additionally, the size of the clusters decreases ($s_t << n$).

ALGORITHM 1: Multi-level Clustering

Data: Set of all records 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 C

```
C ← {S};                                          /* at root level all records are in one cluster */
t ← 1; R ← NULL
while t <= l do
    for c in C do
        Agg(c, a_t);                              /* Aggregate all the records in c to level a_t*/
        {c_t1, c_t2, ...c_tx} ← MDAV(c, s_t);     /* Using MDAV to cluster records in c into partitions of size s_t*/
        R ← R ∪ {c_t1, c_t2, ...c_tx};            /* x is the number of clusters generated */
        C ← C − {c}
    C ← R; R ← NULL; t ← t + 1
return C
```

In this article, we first propose a weighted distance measure that improves the clustering quality. We then propose three new methods. In the first two, we modify ℓ-MDAV and ℓ-VMDAV to combine them with MC method and use the weighted distance. We refer to the two methods as MC-ℓ-MDAV-WD and MC-ℓ-VMDAV-WD, respectively. The third one is called as MC-RoundRobin (MC-RR).

3.3.1 Weighted Distance Measure. Euclidean or other existing distance measures treat the sensitive activities the same way as non-sensitive activities. This is inappropriate for privacy protection. For non-sensitive activities, we want records in the same cluster to have such activities happening at similar time because this will preserve the utility of the data. However, we do *not* want records with sensitive activities happening at the same time to be in the same cluster because this will increase N_{Sj} in Definition 1, and thus reduce ℓ-diversity. The weighted distance measure is motivated by this observation.

Suppose for two activity sequences X and Y, we used size q bit vector to represent activities at each time interval. Without loss of generalizability, we assume that the sensitive activity is represented by the last bit. Let X_{NS} and Y_{NS} represent the concatenation of bit vectors containing only non-sensitive activity (i.e., the first $q-1$ bits for every time interval). There are m time intervals in each record so X_{NS} (or Y_{NS}) contains $(q-1)m$ bits. Let X_S and Y_S represent the concatenation of the bits representing the sensitive activity. X_S (or Y_S) contains m bits.

Let $d_1 = EuclideanDist(X_{NS}, Y_{NS})$, $d_2 = EuclideanDist(X_S, Y_S)$. $d(X, Y)$ is a weighted distance defined as follows:

$$d(X, Y) = d_1 - w_d * d_2. \tag{1}$$

If X and Y are aggregated to larger time intervals (e.g., daily or hourly), then we can replace each bit with the duration of that activity at that time interval and the computation is still the same.

This distance measure is not a metric distance measure as it could be negative and as a result it may not satisfy triangular inequality. For example, suppose sequence s_1 has non-sensitive activity a at time 1 and sensitive activity s at time 2, and s_2 has s at time 1 and a at time 2. It is easy to verify that $d_2 = d_1$ and if $w_d = 1.1$, then the distance is negative. In practice, we may need to set w_d greater than one if sensitive activity is rare.

The weighted distance can be computed in $O(qm)$ time. However, since it is not metric, it will be difficult to use an index structure such as R-tree or k-d tree to speed up finding the nearest neighbors (step (2) of MDAV). However, experimental results in Section 4 will later show that the weighted distance often leads to better clusters and higher data utility. We have also pre-computed the distances in all proposed algorithms that significantly reduced execution time.

3.3.2 MC-ℓ-MDAV-WD. MC-ℓ-MDAV-WD is shown in Algorithm 2. It is similar to MC at non leaf levels. All records are initially assigned to one cluster (line 4). For intermediate levels, all records are aggregated to the required time interval (line 8) and clustered using MDAV with weighted distance (line 9). The size of the intermediate level clusters at level t is given by $s_t = k * p^{l-t}$, where k is the anonymity requirement and p is the fan-out and l is total number of levels. At leaf level $s_l = k$. The algorithm keeps newly generated clusters in C (line 21), increments t (line 23), and goes to next level.

The main difference between Algorithm 2 and MC is that at the leaf level. After generation of new clusters, the resulting clusters are checked for ℓ-diversity (line 10). If all the clusters satisfy ℓ-diversity at the leaf level, then the algorithm returns the resulting clusters (line 25). If the any of the clusters fail to satisfy ℓ-diversity, then backtracking is needed. At line 14, the cluster size s_l at the leaf level is incremented by 1 and Line 15 checks whether the new s_l is greater than half the size of the clusters in the previous level. If this happens, then it is impossible to split these clusters further at the leaf level. So Algorithm 2 backtracks to the root level and increments cluster size at all the levels (line 18). Otherwise, backtrack only happens at the leaf level with the incremented cluster size s_l.

After the clustering step, each cluster is replaced with the centroid of the cluster. This step remains same for all the proposed approaches.

ℓ-diversity is only checked at leaf level. At leaf level each cluster has size around k. So checking for ℓ-diversity for a cluster costs $O(km_l\delta)$ where m_l is the number of aggregated intervals at the leaf level. Since there are $O(n/k)$ clusters, the cost of checking ℓ-diversity is $O(nm_l\delta)$. Let b be the number of backtracks, then the cost is b times of the cost of MC plus the time to check ℓ-diversity, i.e., $O(b(nq\sum_{t=1}^{l} s_t m_t + nm_l\delta))$.

3.3.3 MC-ℓ-VMDAV-WD. Algorithm 3 combines MC and ℓ-VMDAV (described in Section 3.2.2) and uses weighted distance. In the intermediate levels, the records are clustered using MDAV with weighted distance and only in the leaf level the records are clustered using ℓ-VMDAV with weighted distance. MC-ℓ-VMDAV-WD is more efficient than ℓ-VMDAV as it uses multi-level clustering to reduce the cost of clustering high-dimensional data. Using weighted distance also improves the quality of clusters.

The cost of MC-ℓ-VMDAV-WD equals cost of MC at all non-leaf levels plus the cost of running ℓ-VMDAV at leaf level. Since at leaf level each cluster has size around k and m_l intervals, the cost of running ℓ-VMDAV at each leaf level cluster is $O(k^2 m_l(q+\delta))$. There are about n/k clusters so the cost at leaf level is $O(nkm_l(q+\delta))$. The cost of MC-ℓ-VMDAV-WD is $O(nq\sum_{t=1}^{l-1} s_t m_t + nkm_l(q+\delta))$.

ALGORITHM 2: MC-ℓ-MDAV-WD

Data: S set of all records, l number of levels, $agg = [a_1, a_2.., a_l]$ aggregation at each level , $size = [s_1, s_2.., s_l]$ partition size at each level where $s_t = s_l * p^{l-t}$ where p is fanout

Result: C Set of clusters of records

Check for ℓ-diversity for entire data at a_l aggregation. If yes, then continue. Otherwise ℓ-diversity cannot be satisfied.

while *anonymization is not done* **do**
  $brk \leftarrow 0$; /*brk is used to indicate need for backtrack*/
  $C \leftarrow \{S\}; t \leftarrow 1$; /*t is current level*/
  while $t <= l$ **do**
    $R \leftarrow NULL$; /*R is generated clusters*/
    for c *in* C **do**
      $c_{agg} \leftarrow Agg(c, a_t)$; /*Aggregate all the records in c to level a_t*/
      $C_{res} \leftarrow$ MDAV-WD(c_{agg}, s_t); /*C_{res} is the set of clusters resulting from MDAV*/
      if $t == l$ & C_{res} *does not satisfy* ℓ*-diversity* **then**
        $brk \leftarrow 1$; break; /*backtrack is needed*/
      $R \leftarrow R \cup C_{res}$
    if $brk == 1$ **then**
      $s_l \leftarrow s_l + 1$; /*increment size at leaf level by 1*/
      if $(l! = 1)$ & $(size[l] > 0.5 * size[l-1])$; /*If the leaf is bigger than half of the previous level then entire size array needs to be changed*/
      then
        for $i = 1$ *to* $l - 1$ **do**
          $s_i \leftarrow s_l * p^{l-i}$
        break; /*Need to backtrack from root and break out of inner loop*/
    else
      $C \leftarrow R$
      if $t < l$ **then**
        $t \leftarrow t + 1$; /*Non-leaf level, just continue to next level*/
      else
        return C; /*Leaf level, all clusters satisfy ℓ-diversity so return result*/

ALGORITHM 3: MC-ℓ-VMDAV-WD

Data: S set of all records, l number of levels, $agg = [a_1, a_2.., a_l]$ aggregation at each level, $size = [s_1, s_2.., s_l]$ partition size at each level

Result: C Set of clusters of records

1 Check for ℓ-diversity for entire data at a_l aggregation. If yes, then continue. Otherwise ℓ-diversity cannot be satisfied
2 $C \leftarrow \{S\}$;
3 $t \leftarrow 1$;
4 $R \leftarrow NULL$;
5 **while** $t <= l$ **do**
6 **for** c *in* C **do**
7 $c_{agg} \leftarrow Agg(c, a_t)$; /*Aggregate all the records in c to level a_t*/
8 **if** $t! = l$ **then**
9 $C_{res} \leftarrow$ MDAV-WD(c_{agg}, s_t); /*C_{res} is the set of clusters resulting from MDAV*/
10 **else**
11 $C_{res} \leftarrow \ell$-VMDAV-WD(c_{agg}, s_t)
12 $R \leftarrow R \cup C_{res}$
13 $C \leftarrow R$
14 $R \leftarrow NULL$
15 $t \leftarrow t + 1$
16 **return** C

3.3.4 MC-RR. One problem of the above two methods is that they use the new non-metric distance so faster k-nearest neighbor search cannot be used. MC-RR (Algorithm 4) is a more efficient method that does not rely on this new distance measure.

In line 3, records containing sensitive activities are added to a set V. The remaining records are added to a set N and are clustered using MC (algorithm 1) using a cluster size $k' < k$. We can set $k' = k|N|/n$, where n is total number of records and $|N|$ is number of records in N. Since none of the records in N has sensitive activity, it just uses Euclidean distance.

The intuition of MC-RR is that if we strengthen the (δ-ϵ)-diversity definition by requiring that in each equivalence class, at most ϵ fraction of records have sensitive activity, then we just need to ensure that each class does not have over $\epsilon|E|$ records from V ($|E|$ is size of cluster). One way to achieve this is simply evenly distributing records in V to the existing clusters.

After the clusters are formed, the records in V are added to these clusters in a round-robin fashion. For each record $v \in V$, the algorithm checks whether there is a cluster c created from N that contains less than k records. And it also checks if after adding v to c, cluster c still satisfies ℓ-diversity (lines 8 to 13). If so, then v is removed from V and added to that cluster (lines 11 to 13). At the end of this process all clusters still have at most k records.

There may be still records left in V. So the algorithm assigns the remaining records to existing clusters as long as ℓ-diversity is not violated. Note that, at this point, some clusters may have size over k. Those records that cannot be assigned to any cluster are suppressed.

At the end, if any cluster c has size less than k, then the algorithm checks whether c can be merged with any existing cluster without violating ℓ-diversity (lines 23 to 30). The existing clusters are checked in the order of their centroids' distance to the centroid of c. If c cannot be merged with any existing cluster, then c is suppressed.

Suppose there are six records: r_1, r_2, r_3, r_4, r_5, and r_6. $k = 2$ and $k' = 1$. $V = \{r_4, r_5, r_6\}$, $N = \{r_1, r_2, r_3\}$. At step (5), N is divided into three clusters: $c_1 = \{r_1\}$, $c_2 = \{r_2\}$, and $c_3 = \{r_3\}$. At steps (6) to (13), suppose r_4 is added to c_1, r_5 is added to c_2, and r_6 is suppressed because it cannot be added to any cluster without violating ℓ-diversity. At steps (22)–(29), suppose c_3 can be merged with c_2. So finally, two clusters are returned: $c_1 = \{r_1, r_4\}$, $c_2 = \{r_2, r_3, r_5\}$.

MC-RR will generate about n/k clusters after running MC on N. After that each record in V will be added to an existing cluster if ℓ-diversity is satisfied. It costs $O(km_l\delta)$ to check ℓ-diversity for one cluster with size around k. Let n_c be the average number of cluster that needs to be checked before the record can be added to a cluster. So the total cost of MC-RR is $O((n - |V|)q \sum_{t=1}^{l} s_t m_t + |V|n_c k m_l \delta))$, where $|V|$ is number of records with sensitive activities and the first term is the cost of running MC on N and the second term is the cost of assigning records in V. In practice the algorithm often does not need to check many clusters so n_c is quite small. In our experiments MC-RR is faster than the other two proposed methods (with weighted distance) as it uses Euclidean distance in clustering. However, this comes at the cost of utility as records in V are assigned in a round-robin fashion, without considering their distance to existing clusters.

4 EXPERIMENTS AND RESULTS

4.1 Experimental Setup

Experiments were conducted on a computer with 32 GB RAM and 3.2 GHz processor running the Windows 10 operating system. All the algorithms were implemented in R.

4.2 Data

Most of the activity related datasets that are publicly available have very few records. However, to demonstrate the effectiveness of any anonymization approach a larger dataset is necessary [29, 38]. Therefore, we used a publicly available human activity dataset known as **Activity Recognition with Ambient Sensing (ARAS)** [2] as seed to generate synthetic data. ARAS data contain activity information of four people collected from two real homes. The data are collected in seconds for a duration of 1 month. It has 27 different activities including preparing breakfast, having breakfast, sleeping, having shower, toileting, going out, and so on.

With ARAS data as the seed, we generated synthetic data using Markov chain model [22]. We first aggregated the data to minute level. For each person's data (seed), a state transition matrix was constructed for every hour. The matrices were then used to generate synthetic data for a certain number of people. Noise was also introduced by selecting a different person's state transition matrix at random with a probability of 0.01 at each hour. Synthetic data were generated for 100 people (so each seed was used to generate 25 people's data) at minute-level granularity for a period of two weeks.

To check the quality of the simulated data, we measured KL divergence [32] between the original data and the simulated synthetic data. The KL divergence value was 0.014. Low KL-divergence value shows that these two datasets have similar distributions.

ALGORITHM 4: MC-RR

Data: S set of all records, l number of levels, $agg = [a_1, a_2.., a_l]$ aggregation at each level, $size = [s_1, s_2.., s_l]$ partition size at each level, k'
Result: C Set of clusters of records

Check for ℓ-diversity for entire data at a_l aggregation. If yes, then continue.
$C \leftarrow \{S\}$
Compute V as the set of records that contain sensitive activity
$N \leftarrow C - V$
$C_{res} \leftarrow MC(N, \ k')$; /* let k' be the partition size*/
for $v \in V$ **do**
  $v \leftarrow Agg(v, a_l)$; /* Aggregate v to the same level as the leaf level*/
for $v \in V$ **do**
  for $c \in C_{res}$ **do**
    if *size*$(c) < k$ *and* $c \cup \{v\}$ *satisfies* ℓ*-diversity* **then**
      add v to c
      remove v from V
      break
for $v \in V$ **do**
  for $c \in C_{res}$ **do**
    if *adding* v *to* c *still satisfies* ℓ*-diversity* **then**
      remove v from V
      add v to c
      break;
  if v *is not added to any* c **then**
    suppress v
for $c \in C_{res}$ **do**
  if *size*$(c) < k$ **then**
    Create a priority queue$Q$ based on the distance from c's centroid to other clusters' centroids
    for *each* $c' \in Q$ **do**
      if $c \cup c'$ *satisfies* ℓ*-diversity* **then**
        merge c with c'
        break
    suppress c if c cannot be merged with any existing cluster
Return(C_{res})

The original data do not have sensitive activity. So, to demonstrate the ℓ-diversity requirement another activity called *Vacation* was added. Two different patterns of vacation data were added resulting in two datasets:

- Based on statistics published by AAA [1], about one third of Americans travelled in December holiday season in 2018. For the first dataset, we consider one week of data. Around 35% of the people are randomly selected to go on a vacation around a long weekend. The length of the vacation is drawn from a normal distribution with $\mu = 2$ (days) and $\sigma = 0.5$. All vacation will end on Sunday. We refer to this dataset as the *long weekend data.*

Table 2. Properties of the Datasets

Data	No. of records	Duration	No. of activities	No. of dimensions	Fraction of people going on vacation
Long weekend data	100	1 week	28	10080	35%
Long vacation data	100	2 weeks	28	20160	35%

- For the second dataset, we consider two weeks of data. Similarly to the previous data, 35% of people are randomly selected to go on vacation around Christmas, or around the New Year's day, or go on a long vacation including both Christmas and New Year's day. The length of the vacation is generated from 3 normal distributions. For Christmas and New Year's day, $\mu = 2$ (days) and $\sigma = 0.5$. For long vacation, $\mu = 6$ (days) and $\sigma = 1.5$. We refer to this dataset as the *long vacation data.*

Table 2 shows properties of the synthetic datasets generated for experiments. '*Vacation*' activity is considered as the sensitive activity.

4.3 Algorithms Compared in the Experiments

- MC-ℓ-MDAV-WD: This is proposed Algorithm 2, which uses multi-level clustering and MDAV with backtrack to cluster the activity sequences. The proposed weighted distance measure is used during clustering.
- MC-ℓ-VMDAV-WD: This is proposed Algorithm 3, which uses multi-level clustering and ℓ-VMDAV to cluster the activity sequences. The proposed weighted distance measure is used during clustering.
- MC-ℓ-MDAV: This is same as MC-ℓ-MDAV-WD except that regular Euclidean distance is used in clustering.
- MC-ℓ-VMDAV: This is same as MC-ℓ-VMDAV-WD except that Euclidean distance is used in clustering.
- MC-RR: This is proposed Algorithm 4, which assigns records with sensitive activity in a round robin fashion to clusters of records without sensitive activity.
- ℓ-MDAV: This is a baseline approach (refer to Section 3.2.1). Compared to our approach it does not use multi-level clustering and does not use new weighted distance. Data are not aggregated, since this method does not use multi-level clustering and aggregation is a part of multi-level clustering.
- ℓ-VMDAV: This is a baseline approach (refer to Section 3.2.2). Compared to our approach it does not use multi-level clustering and does not use new weighted distance. Similarly to the previous one, data are not aggregated, since this method does not use multi-level clustering and aggregation is a part of multi-level clustering.
- MSLD: ImSLD is an existing approach presented in Reference [37]. In this approach, the authors use MapReduce technique for achieving l-diversity for Big Data. However, activity datasets are typically small (few hundred records) therefore, MapReduce is not necessary. Additionally, ImSLD enforces a simpler version of ℓ-diversity that is not applicable to activity data. We modified ImSLD to make it applicable to activity dataset and used it as a baseline approach for comparing the proposed approaches. We refer to this baseline approach as MSLD (Modified ImSLD).

Table 3. Parameter Settings for Our Algorithms

No. of levels	No. of records at each level	Fanout	Aggregation level	w_d in Equation (1)
2	top: 50, leaf: 10	5	top: weekly, leaf: hourly	1

All methods other than MC-ℓ-MDAV-WD and MC-ℓ-VMDAV-WD use Euclidean distance. So these methods used a fast k-nearest neighbor search based on k-d tree [24] to speed up their execution.

The use of the existing standard techniques as benchmark approaches is not only important for comparing our approach with them but also important for obtaining and presenting evidence about whether longitudinal data are in fact ill suited to be de-identified using those standard techniques or not.

4.4 Parameter Settings

We used methods in Reference [43] to determine the optimal values of parameters in our algorithms such as number of levels in multi-level clustering. Table 3 shows the values. For privacy parameters, we set $k = 10$, $\epsilon = 0.75$, $\delta = 48$ (hours).

4.5 Metrics for Comparison

To measure the efficiency, we focus on the time to cluster data for two reasons: (1) this is the most time-consuming step, and (2) once clusters (equivalence classes) are generated, all algorithms follow the same steps to replace original data with anonymized data.

To measure the utility of anonymized data, we use the following metrics:

- Relative difference [53] between un-anonymized data D and anonymized data D'. We define relative difference $r(x, y)$ between two values as $\frac{|x-y|}{max(x,y)}$ if at least one of x or y is not zero and 0 otherwise.
  $r(D, D')$ is the relative difference between D and D' and equals $= \frac{1}{nmq} \sum_{1 \le i \le n, 1 \le j \le m, 1 \le v \le q} r(D_{ijv}, D'_{ijv})$
  Where, D_{ijv} is the duration of activity v in record i at time interval j in original data and D'_{ijv} is the duration in anonymized data. We use relative difference rather than relative error because many values in the dataset are zero.
  We present relative difference at daily level, i.e., for example, a relative difference of 0.1 for activity a means that anonymized data on an average had 0.1 relative difference with respect to unanonymized data for daily duration of activity a. Also, the relative difference values shown in the tables are average relative difference over all the activities.
- The number of points suppressed for each approach are also presented that can be an important factor for choosing the right method.
- We also provide an empirical evaluation of the approaches by comparing the average duration of the following activities before and after anonymization: *Sleep*, *Talking on the Phone*, and *Having Conversation*. Pearson's correlation between *Sleep* and *Having Conversation* before and after anonymization is also presented. Correlation between *Sleep* and *Talking on the Phone* is not shown because it is not statistically significant in the original data.

Table 4. Results for Long Weekend Data for Different Sized Datasets

Method	100 records			200 records			300 records		
	rel	ns	time	rel	ns	time	rel	ns	time
MC-ℓ-MDAV	0.33	0	20.2	0.42	0	302.95	0.33	0	186.79
MC-ℓ-VMDAV	0.25	10	9.16	0.2	40	17.44	0.18	60	27
MC-ℓ-MDAV-WD	0.27	0	21.14	0.23	0	39.87	0.28	0	539.39
MC-ℓ-VMDAV-WD	0.27	0	21.94	0.24	0	42.48	0.21	20	67.19
MC-RR	0.36	0	13.83	0.28	0	27.5	0.28	0	38.07
ℓ-MDAV†	0.24	0	426.94	0.26	0	8600.38	0.24	0	13690.89
ℓ-VMDAV†	0.23	10	569.13	0.19	20	1620.95	0.19	40	2503.08
MSLD†	0.27	0	644.98	0.23	0	1316.78	0.23	0	2063.33

rel is relative difference, *ns* is number of suppressed records, *time* is time for clustering. † represents baseline approach.

Table 5. Results for Long Vacation Data for Different Sized Datasets

Method	100 records			200 records			300 records		
	rel	ns	time	rel	ns	time	rel	ns	time
MC-ℓ-MDAV	0.32	0	40.31	0.37	0	123.34	0.37	0	329.53
MC-ℓ-VMDAV	0.25	10	15.77	0.25	20	31.95	0.21	30	49.65
MC-ℓ-MDAV-WD	0.22	0	39.38	0.24	0	76.78	0.22	0	115.69
MC-ℓ-VMDAV-WD	0.22	0	42.89	0.24	0	82.41	0.22	0	122.48
MC-RR	0.35	0	32.67	0.26	0	57.67	0.28	0	80.53
ℓ-MDAV†	0.34	0	4337.66	—	—	—	—	—	—
ℓ-VMDAV†	0.25	10	1149.23	—	—	—	—	—	—
MSLD†	0.26	0	1354.04	0.22	0	2797.47	—	—	—

rel is relative difference, *ns* is number of suppressed records, *time* is time for clustering. † represents baseline approach.

4.6 Experimental Results

Results when varying number of records: Typically, activity datasets are very flat, and therefore we consider up to 300 data points. Table 4 and 5 show results when the size of the data was varied from 100 to 300.

Table 4 shows the results for varying number of records for long weekend data. Time taken for clustering increases with number of records. Time taken by the proposed approaches is orders less than the time taken by the three baseline approaches. Baseline approach *l*-MDAV has slightly lower relative difference compared to proposed approaches for 100 and 300 records; however, it takes at least 20 times longer than the proposed approaches. Proposed approaches with weighted distance, i.e., MC-ℓ-MDAV-WD and MC-ℓ-VMDAV-WD show better balance in terms of utility and time. They have low relative difference and do not suppress records. Compared to these two approaches all other approaches have higher relative loss, suppress records, or have very high execution time.

For long vacation data (Table 5), for 200 and 300 records the baseline approaches ran into memory error. For 200 records, baseline approach MSLD has low relative difference but takes orders longer than proposed approaches with weighted distance that have similar relative difference. MC-ℓ-MDAV-WD and MC-ℓ-VMDAV-WD have low relative difference without any data suppression and execution time much lower than the baseline approaches and comparable to that of the other proposed approaches. For 100 records, MC-ℓ-MDAV-WD and MC-ℓ-VMDAV-WD have lower

Table 6. Results When Varying ℓ-diversity Constraints for Long Weekend Data

Method	Increased vacation			$\delta = 72$			$\epsilon = 0.5$		
	rel	ns	time	rel	ns	time	rel	ns	time
MC-ℓ-MDAV	0.5	0	122.95	0.24	0	7.44	0.46	0	93.59
MC-ℓ-VMDAV	0.24	30	8.46	0.24	0	8.53	0.23	20	8.55
MC-ℓ-MDAV-WD	0.27	0	20.64	0.27	0	20.17	0.24	0	174.24
MC-ℓ-VMDAV-WD	0.27	0	21.3	0.27	0	21.18	0.26	20	23.48
MC-RR	0.4	0	15.72	0.36	0	13.83	0.36	0	14.04
ℓ-MDAV†	0.23	0	794.23	0.24	0	432.13	0.39	0	5988.64
ℓ-VMDAV†	0.2	20	559.68	0.24	0	564.57	0.21	20	545.08
MSLD†	0.27	0	649.64	0.27	0	649.04	0.27	0	655.16

Table 7. Results When Varying ℓ-diversity Constraints for Long Vacation Data

Method	Increased vacation			$\delta = 72$			$\epsilon = 0.5$		
	rel	ns	time	rel	ns	time	rel	ns	time
MC-ℓ-MDAV	0.41	0	98.3	0.32	0	42.08	0.35	0	53.06
MC-ℓ-VMDAV	0.29	30	15.87	0.25	10	16.94	0.25	10	16.45
MC-ℓ-MDAV-WD	0.26	0	40.9	0.22	0	39.63	0.22	0	39.18
MC-ℓ-VMDAV-WD	0.26	0	42.97	0.22	0	42.92	0.22	0	42.58
MC-RR	0.43	0	37.26	0.35	0	32.77	0.35	0	32.47
ℓ-MDAV†	0.46	0	11612.3	0.34	0	4328.62	0.32	0	8885.4
ℓ-VMDAV†	0.25	20	1176.36	0.25	10	1190.15	0.25	10	1134.08
MSLD†	0.26	0	1401.08	0.26	0	1390.14	0.26	0	1428.4

relative difference than baseline methods possibly because the weighted distance is more effective for diverse vacation patterns.

The relative difference for all methods in the long vacation dataset is slightly lower than the loss in the long weekend dataset. One possible reason is that in long vacation dataset the vacation patterns are more diverse (from three normal distributions). As our privacy model restricts the fraction of people going on vacation during the same time period, more diverse vacation patterns result in better clusters.

It can be seen MC-ℓ-VMDAV and ℓ-VMDAV suppress data points. Proposed approach MC-ℓ-VMDAV is the fastest method in all the settings but suppresses data points in a few cases. MC-ℓ-MDAV takes longer as compared to the other proposed approaches because of the backtracking and also has higher relative difference because of large clusters.

Results when varying privacy constraints: Tables 6 and 7 show the results of various algorithms when we change the percentage of people going on vacation as well as ϵ and δ parameters in our ℓ-diversity model. The first set of experiments were run on a dataset with the default settings except that the percentage of people going on vacation is increased to 50%. In the second set δ is increased to 72 (i.e., we look at 3-day window) but ϵ is still 0.75. This means the privacy requirement is relaxed. In the third set of experiments ϵ is changed to 0.5 while δ is still 48. In this case, privacy requirement is more restrictive.

As the number of people going on vacation increases to 50%, it becomes more difficult to satisfy the (δ, ϵ)-diversity. For long weekend data (Table 6), baseline ℓ-MDAV has low relative difference but takes about 40 times more time as compared to proposed approaches with weighted distance

MC-ℓ-MDAV-WD and MC-ℓ-VMDAV-WD, which have slightly higher relative difference but take much lesser time. All other approaches either take longer or have higher relative difference or result in data suppression. For long vacation data (Table 7), proposed approaches with weighted distance MC-ℓ-MDAV-WD and MC-ℓ-VMDAV-WD have lowest relative difference without any data suppression. Baseline MSLD has similar relative difference as these approaches for both the datasets but takes much longer time to complete.

Similarly, when ϵ is reduced to 0.5, the privacy constraint is more restrictive. In this case, the proposed approach MC-ℓ-MDAV-WD achieves lower loss (low relative difference and no suppression) compared to all other methods. This method does take longer to cluster the data (especially on long weekend data) due to using weighted distance and the need for backtracks. MC-ℓ-VMDAV-WD has the same relative difference in the long vacation data but suffers data suppression in long weekend data. The base line methods are more expensive and have comparable loss as MC-ℓ-MDAV-WD.

When δ is increased to 72, the privacy constraint is relaxed. Note that for long weekend data, 72 hours translates to mean duration of vacation (which is 48 hours) plus 2 times standard deviation (which is 12 hours), meaning only a very small fraction of people go on vacation that long. Therefore, almost all the methods have similar loss for long weekend data except MC-RR. MC-ℓ-MDAV and MC-ℓ-VMDAV (the two variants of proposed methods without weighted distance) achieve the same loss as existing methods (ℓ-MDAV and ℓ-VMDAV) but have much shorter execution time.

For the long vacation data, two proposed methods MC-ℓ-MDAV-WD and MC-ℓ-VMDAV-WD have lower loss compared to others when $\delta = 72$ in Table 7 possibly because the weighted distance is more effective when there are diverse vacation patterns.

4.7 Empirical Evaluation

Some of the daily activities that are of high value to the medical field, especially mental health and behavioral health communities are *sleep* [3, 42, 52] and *social engagement* [7, 31, 48]. Studies have found that insomnia is related to depression, anxiety and other health conditions [3, 42, 52]. Use of technological devices can also impact the amount of sleep [7]. Additionally, studies have found that social engagement can also influence the quality of sleep [31, 48]. Therefore, it becomes important to preserve the characteristics of such activities after anonymization.

ARAS dataset used for the experiments includes *Sleep, Having Conversation, Talking on the Phone*. Table 8 presents the average duration (per day) of these three activities before and after anonymization. Table 8 also shows Pearson's correlation between *Sleep* and *Having Conversation* before and after anonymization. Correlation between *Sleep* and *Talking on the Phone* is not shown as it was not statistically significant in the original data. For 100 records, since the anonymized data are represented by cluster centers and there can be 10 clusters at the most (with $k = 10$), it is not possible to see statistical significance. Therefore, we present the results for 200 records and default parameters.

Results show that the proposed approaches with weighted distance (MC-ℓ-MDAV-WD and MC-ℓ-VMDAV-WD) preserve the averages after anonymization. Baseline approaches ℓ-MDAV and MSLD also preserve the averages. However, the baseline approaches have much higher execution time as compared to the proposed approaches as seen earlier. The proposed weighted distance-based approaches also preserve the direction and magnitude of the correlation more effectively as compared to all the other proposed and baseline approaches. All methods have higher p values than original data, which is expected as records in each cluster is replaced with cluster mean so there are fewer distinctive records after anonymization.

Table 8. Average Duration of Sleep (in hrs) per Day, Average Duration of Having Conversation (in mins) per Day, Average Duration of Talking on the Phone (in mins) per Day, Correlation between Sleep and Having Conversation

	Sleep	Having Conversation	Talking on the Phone	Sleep - Having Conversation	
	(in hrs per day)	(in mins per day)	(in mins per day)	Pearson's corr	p-value
Before Anonymization					
	6.46	3.5	14.32	0.42	8.91e-10
After Anonymization					
MC-ℓ-MDAV	6.47	3.46	13.54	−0.19	0.81
MC-ℓ-VMDAV	6.87	3.86	16.15	0.58	0.01
MC-ℓ-MDAV-WD	6.46	3.5	14.32	0.46	0.03
MC-ℓ-VMDAV-WD	6.46	3.5	14.32	0.48	0.03
MC-RR	6.39	3.68	15.9	0.51	0.03
ℓ-MDAV†	6.46	3.5	14.32	0.67	0.03
ℓ-VMDAV†	6.67	3.67	13.68	0.57	0.02
MSLD†	6.46	3.5	14.32	0.51	0.02

5 DISCUSSION

For any anonymization method to be effective, it should be able to preserve the utility of the dataset and it should be efficient. Activity data are sequential in nature and applying existing approaches directly is computationally expensive. This is also verified by the experiments (see the last three rows of Tables 4, 6, 5, and 7). Proposed multi-level clustering-based approaches are more efficient than existing methods. Approaches using backtrack (MC-ℓ-MDAV, MC-ℓ-MDAV-WD, ℓ-MDAV) take longer time in some cases than those approaches not using backtrack because the backtracking might cause the clustering to start again from the root level, which is time-consuming. MC-ℓ-VMDAV takes the least time as it uses multi-level clustering to reduce execution time and does not need backtrack.

In terms of data utility, the proposed MC-ℓ-MDAV-WD method has relative difference either comparable or better than that of the baseline approaches. It avoids data suppression, which in turn improves the data utility. Additionally, the execution time is orders lower than the baseline approaches. MC-ℓ-VMDAV-WD has results quite similar to that of MC-ℓ-MDAV-WD except when privacy requirement is more restrictive (when $\epsilon = 0.5$ for long weekend data) and larger data (300 records for long weekend data).

Although our proposed weighted distance is non-metric, its execution time is much lower than the baseline approaches and comparable to other proposed approaches. It also leads to high quality clusters and improves data utility. In the empirical evaluation, the two proposed methods MC-ℓ-VMDAV-WD and MC-ℓ-MDAV-WD also better preserve the mean of sleep and social engagement activities as well as the correlation between them.

Overall, for all datasets, the two proposed methods MC-ℓ-WMDAV-WD and MC-ℓ-MDAV-WD perform well both in terms of data utility and efficiency.

6 CONCLUSION

This article studies the problem of anonymizing longitudinal daily activity data based on the ℓ-diversity privacy model. Several different methods have been proposed to combine multi-level clustering with methods to enforce ℓ-diversity. A non-metric distance measure is also proposed to improve the quality of clustering.

Experiments were conducted on two datasets and results show that longitudinal data is in fact ill-suited to be anonymized using standard techniques. Two of our proposed methods MC-ℓ-MDAV-WD and MC-ℓ-VMDAV-WD lead to similar data utility as existing methods but are orders more efficient. They also have better utility than other proposed methods.

REFERENCES

[1] AAA. 2018. AAA: One-in-Three Americans Will Travel This Holiday Season, the Most on Record. Retrieved from https://newsroom.aaa.com/2018/12/2018-busiest-holiday-travel-season-on-record/.

[2] Hande Alemdar, Halil Ertan, Ozlem Durmaz Incel, and Cem Ersoy. 2013. ARAS human activity datasets in multiple homes with multiple residents. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*. ICST, 232–235.

[3] Najib T. Ayas, David P. White, JoAnn E. Manson, Meir J. Stampfer, Frank E. Speizer, Atul Malhotra, and Frank B. Hu. 2003. A prospective study of sleep duration and coronary heart disease in women. *Arch. Internal Med.* 163, 2 (2003), 205–209.

[4] Eduardo Casilari, José-Antonio Santoyo-Ramón, and José-Manuel Cano-García. 2017. Analysis of public datasets for wearable fall detection systems. *Sensors* 17, 7 (2017), 1513.

[5] Prafulla Nath Dawadi, Diane Joyce Cook, and Maureen Schmitter-Edgecombe. 2015. Automated cognitive health assessment from smart home-based behavior data. *IEEE J. Biomed. Health Inf.* 20, 4 (2015), 1188–1194.

[6] Christian Debes, Andreas Merentitis, Sergey Sukhanov, Maria Niessen, Nikolaos Frangiadakis, and Alexander Bauer. 2016. Monitoring activities of daily living in smart homes: Understanding human behavior. *IEEE Sign. Process. Mag.* 33, 2 (2016), 81–94.

[7] Kadir Demirci, Mehmet Akgönül, and Abdullah Akpinar. 2015. Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students. *J. Behav. Addict.* 4, 2 (2015), 85–92.

[8] Josep Domingo-Ferrer, Antoni Martínez-Ballesté, Josep Maria Mateo-Sanz, and Francesc Sebé. 2006. Efficient multivariate data-oriented microaggregation. *VLDB J.* 15, 4 (2006), 355–369.

[9] Josep Domingo-Ferrer and Josep Maria Mateo-Sanz. 2002. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.* 14, 1 (2002), 189–201.

[10] Josep Domingo-Ferrer, Francesc Sebé, and Agusti Solanas. 2007. Microaggregation heuristics for p-sensitive k-anonymity. In *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality* (2007).

[11] Josep Domingo-Ferrer, Agusti Solanas, and Antoni Martinez-Balleste. 2006. Privacy in statistical databases: k-anonymity through microaggregation. In *Proceedings of the 2006 IEEE International Conference on Granular Computing*. IEEE, 774–777.

[12] Josep Domingo-Ferrer and Vicenç Torra. 2005. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Min. Knowl. Discov.* 11, 2 (2005), 195–212.

[13] Yulan Dong and Dechang Pi. 2018. Novel privacy-preserving algorithm based on frequent path for trajectory data publishing. *Knowl.-Bas. Syst.* 148 (2018), 55–65.

[14] Dorothy D. Dunlop, Jing Song, Emily K. Arntson, Pamela A. Semanik, Jungwha Lee, Rowland W. Chang, and Jennifer M. Hootman. 2015. Sedentary time in US older adults associated with disability in activities of daily living independent of physical activity. *J. Phys. Activ. Health* 12, 1 (2015), 93–101.

[15] Cynthia Dwork. 2011. Differential privacy. *Encyclopedia of Cryptography and Security* (2011), 338–340.

[16] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 3–4 (2014), 211–407.

[17] Khaled El Emam. 2008. Heuristics for de-identifying health data. *IEEE Secur. Priv.* 6, 4 (2008), 58–61.

[18] Khaled El Emam, Luk Arbuckle, Gunes Koru, Benjamin Eze, Lisa Gaudette, Emilio Neri, Sean Rose, Jeremy Howard, and Jonathan Gluck. 2012. De-identification methods for open health data: The case of the Heritage Health Prize claims dataset. *J. Med. Internet Res.* 14, 1 (2012), e33.

[19] Khaled El Emam and Fida Kamal Dankar. 2008. Protecting privacy using k-anonymity. *J. Am. Med. Inf. Assoc.* 15, 5 (2008), 627–637.

[20] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, et al. 2009. A globally optimal k-anonymity method for the de-identification of health data. *J. Am. Med. Inf. Assoc.* 16, 5 (2009), 670–682.

[21] Paying for Senior Care. 2019. Activities & Instrumental Activities of Daily Living—Definitions, Importance and Assessments. Retrieved August 7, 2019 from https://www.payingforseniorcare.com/longtermcare/activities-of-daily-living.html.

[22] Paul A. Gagniuc. 2017. *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons.

[23] Tamas S. Gal, Zhiyuan Chen, and Aryya Gangopadhyay. 2008. A privacy protection model for patient data with multiple sensitive attributes. *Int. J. Inf. Secur. Priv.* 2, 3 (2008), 28–44.

[24] Patrick J. Grother, Gerald T. Candela, and James L. Blue. 1997. Fast implementations of nearest neighbor classifiers. *Pattern Recogn.* 30, 3 (1997), 459–465.

[25] Xi He, Graham Cormode, Ashwin Machanavajjhala, Cecilia M. Procopiuc, and Divesh Srivastava. 2015. DPT: Differentially private trajectory synthesis using hierarchical reference systems. *Proc. VLDB Endow.* 8, 11 (2015), 1154–1165.

[26] Zhaowei Hu, Jing Yang, and Jianpei Zhang. 2018. Trajectory privacy protection method based on the time interval divided. *Comput. Secur.* 77 (2018), 488–499.

[27] Han Jian-min, Cen Ting-Ting, and Yu Hui-Qun. 2008. An improved V-MDAV algorithm for l-diversity. In *Proceedings of the 2008 International Symposiums on Information Processing*. IEEE, 733–739.

[28] Sidney Katz. 1983. Assessing self-maintenance: Activities of daily living, mobility, and instrumental activities of daily living. *J. Am. Geriatr. Soc.* 31, 12 (1983), 721–727.

[29] Ryuichi Kitamura, Cynthia Chen, and Ram M. Pendyala. 1997. Generation of synthetic daily activity-travel patterns. *Transport. Res. Rec.* 1607, 1 (1997), 154–162.

[30] Joseph Kuan and Paul Lewis. 1997. Fast k nearest neighbour search for R-tree family. In *Proceedings of the 1997 International Conference on Information, Communications and Signal Processing (ICICS'97)*, Vol. 2. IEEE, 924–928.

[31] Daniel Kuhn, Perry Edelman, and Bradley R. Fulton. 2005. Daytime sleep and the threat to well-being of persons with dementia. *Dementia* 4, 2 (2005), 233–247.

[32] Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[33] M. Powell Lawton and Elaine M. Brody. 1969. Assessment of older people: Self-maintaining and instrumental activities of daily living. *Gerontologist* 9, 3_Part_1 (1969), 179–186.

[34] Ninghui Li, Tiancheng Li, and Suresh Enkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 106–115.

[35] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. 2006. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*. IEEE, 24–24.

[36] Sergi Martínez-Bea and Vicenç Torra. 2011. Trajectory anonymization from a time series perspective. In *Proceedings of the 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'11)*. IEEE, 401–408.

[37] Brijesh B. Mehta and Udai Pratap Rao. 2019. Improved l-Diversity: Scalable anonymization approach for privacy preserving big data publishing. *J. King Saud Univ. Comput. Inf. Sci.* (2019).

[38] Dorothy Monekosso and Paolo Remagnino. 2009. Synthetic training data generation for activity monitoring and behavior analysis. In *Proceedings of the European Conference on Ambient Intelligence*. Springer, 267–275.

[39] NBC. 2016. Some Burglars Using Social Media to Find Targets, I-Team Survey Shows. Retrieved August 12, 2019 from https://www.nbcnewyork.com/news/local/Investigations-I-Team-Social-Media-Use-Survey-New-York-New-Jersey-390938211.html.

[40] Mehmet Ercan Nergiz, Maurizio Atzori, and Yucel Saygin. 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*. ACM, 52–61.

[41] U.S. Department of Health and Human Services. 1990. Measuring the Activities of Daily Living: Comparisons across National Surveys. Retrieved August 7, 2019 from https://aspe.hhs.gov/basic-report/measuring-activities-daily-living-comparisons-across-national-surveys.

[42] Consensus Conference Panel, Nathaniel F. Watson, M. Safwan Badr, Gregory Belenky, Donald L. Bliwise, Orfeu M. Buxton, Daniel Buysse, David F. Dinges, James Gangwisch, Michael A. Grandner, et al. 2015. Joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society on the recommended amount of sleep for a healthy adult: Methodology and discussion. *Sleep* 38, 8 (2015), 1161–1183.

[43] Pooja Parameshwarappa, Zhiyuan Chen, and Gunes Koru. 2020. An effective and computationally efficient approach for anonymizing large-Scale physical activity data: Multi-Level clustering-Based anonymization. *Int. J. Inf. Secur. Priv.* 14, 3 (2020), 72–94.

[44] Ivan Pires, Nuno Garcia, Nuno Pombo, and Francisco Flórez-Revuelta. 2016. From data acquisition to data fusion: A comprehensive review and a roadmap for the identification of activities of daily living using mobile devices. *Sensors* 16, 2 (2016), 184.

[45] Udai Pratap Rao, Brijesh B. Mehta, and Nikhil Kumar. 2019. Scalable l-Diversity: An extension to scalable k-Anonymity for privacy preserving big data publishing. *Int. J. Inf. Technol. Web Eng.* 14, 2 (2019), 27–40.

[46] Chris Rose. 2011. The security implications of ubiquitous social media. *International Journal of Management & Information Systems (IJMIS)* 15, 1 (2011).

[47] Yuichi Sei, Hiroshi Okumura, Takao Takenouchi, and Akihiko Ohsuga. 2017. Anonymization of sensitive Quasi-Identifiers for l-diversity and t-closeness. *IEEE Trans. Depend. Sec. Comput.* (2017).

[48] Eti Ben Simon and Matthew P. Walker. 2018. Sleep loss causes social withdrawal and loneliness. *Nat. Commun.* 9, 1 (2018), 1–9.

[49] Agusti Solanas, Antoni Martinez-Balleste, and J. Domingo-Ferrer. 2006. V-MDAV: A multivariate microaggregation with variable group size. In *Proceedings of the 17th COMPSTAT Symposium of the IASC*. 917–925.

[50] Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks track 1. *J. Biomed. Inf.* 75 (2017), S4–S18.

[51] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *Int. J. Uncertainty Fuzz. Knowl.-Bas. Syst.* 10, 05 (2002), 557–570.

[52] Daniel J. Taylor, Kenneth L. Lichstein, H. Heith Durrence, Brant W. Reidel, and Andrew J. Bush. 2005. Epidemiology of insomnia, depression, and anxiety. *Sleep* 28, 11 (2005), 1457–1464.

[53] Leo Törnqvist, Pentti Vartia, and Yrjö O Vartia. 1985. How should relative changes be measured? *Am. Stat.* 39, 1 (1985), 43–46.

[54] Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, and Depeng Jin. 2018. Protecting trajectory From semantic attack considering k-Anonymity, l-Diversity, and t-Closeness. *IEEE Trans. Netw. Serv. Manage.* 16, 1 (2018), 264–278.

[55] Prince William County Virginai. [n.d.]. How Burglars Use Social Media. Retrieved August 12, 2019 from https://www.pwcgov.org/government/dept/police/Pages/How-Burglars-Use-Social-Media.aspx.

[56] George R. S. Weir, Fergus Toolan, and Duncan Smeed. 2011. The threats of social networking: Old wine in new bottles? *Inf. Secur. Techn. Rep.* 16, 2 (2011), 38–43.

[57] Lin Yao, Xinyu Wang, Xin Wang, Haibo Hu, and Guowei Wu. 2019. Publishing sensitive trajectory data under enhanced l-Diversity model. In *Proceedings of the 2019 20th IEEE International Conference on Mobile Data Management (MDM'19)*. IEEE, 160–169.

[58] Xuyun Zhang, Wanchun Dou, Jian Pei, Surya Nepal, Chi Yang, Chang Liu, and Jinjun Chen. 2014. Proximity-aware local-recoding anonymization with mapreduce for scalable big data privacy preservation in cloud. *IEEE Trans. Comput.* 64, 8 (2014), 2293–2307.

[59] Xuyun Zhang, Chang Liu, Surya Nepal, Chi Yang, Wanchun Dou, and Jinjun Chen. 2014. A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud. *J. Comput. Syst. Sci.* 80, 5 (2014), 1008–1020.

Received December 2019; revised November 2020; accepted March 2021