Chapter 32
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.

DOI: 10.4018/978-1-7998-8954-0.ch032

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.

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