Chapter 5

Protecting Big Data Through Microaggregation Technique for Secured Cyber-Physical Systems

Shakila Mahjabin Tonni
East West University, Bangladesh

Sazia Parvin
Melbourne Polytechnic, Australia

Amjad Gawanmeh
Khalifa University, UAE

Joanna Jackson
Melbourne Polytechnic, Australia

ABSTRACT

Secured cyber-physical systems (CPS) requires reliable handling of a high volume of sensitive data, which is in many cases integrated from several distributed sources. This data can usually be interconnected with physical applications, such as power grids or SCADA systems. As most of these datasets store records using numerical values, many of the microaggregation techniques are developed and tested on numerical data. These algorithms are not suitable when the data is stored as it is containing both numerical and categorical data are stored. In this chapter, the available microaggregation techniques are explored and assessed with a new microaggregation technique which can provide data anonymity regardless of its type. In this method, records are clustered into several groups using an evolutionary attribute grouping algorithm and groups are aggregated using a new operator.

DOI: 10.4018/978-1-5225-5510-0.ch005
INTRODUCTION

With the growing use of cloud computing technology, a new generation of SCADA systems (Wikipedia.org, SCADA, 2017) is emerging that is incorporating new technologies. Many CPS require continuous communications accessing big data. In addition, several CPS applications, such as power grid, and SCADA systems impose high reliability, security and availability. Moreover, possibilities of physical or cyberspace intrusions may lead to exposure of sensitive data. Therefore, protecting data in critical CPS applications should be considered a priority.

Increasing utilization of cyber-physical systems is opening new possibilities of enormous impact on the society and economy. Many CPSs are now trying to enhance its capability using evolving cloud (Zhang, Qiu, Tsai, Hassan, & Alamri, 2017) and distributed technology (Jaskolka & Villasenor, 2017), where different types of sensitive data are collected. Although, it can be said from the past experiences, most of the attacks on CPSs are physical or environmental in nature (Frey, Rashid, Zanutto, Busby, & Follis, 2016), there are evidences of cyber-attacks on ICS (Industrial Control System). For instance, Stuxnet (Wikipedia.org, Stuxnet, 2017) is a worm that caused considerable damage to Iran’s nuclear program by intercepting both sensor and actuator data, and tampering the centrifuge. According to (Martínez, Sánchez, & Valls, 2012), the wide use of CPSs from medical devices to smart cars, CPS security is crucial due to the vulnerabilities of the legacy systems that are used to build them. So, in such contexts, securing control data produced by the systems is highly needed. Secured controlled data can be protected by using widely used data microaggregation techniques. Microaggregation is a technique massively used as a mean of statistical disclosure control (Domingo-Ferrer J. &.-S., 2002). The main idea is to group and then apply an aggregation operator on microdata (Kabir & Wang, 2011), (Martínez, Sánchez, & Valls, 2012) to produce a confined record. In this chapter, we discuss different microaggregation techniques and assess the usefulness of our new microaggregation algorithm proposed in (Tonni, Rahman, Parvin, & Gawanmeh, 2017) for CPSs.

To serve the purpose of protecting data privacy, numerous techniques are introduced for privacy in statistical databases that are collectively known as Statistical Disclosure Control (SDC). Data protection methods can be divided into two categories:

- **Perturbative:** In these methods, the actual data set is altered using some technique like noise addition, and the deformed new data set might have some fake information. Naturally, it introduces new sets of attribute values while losing a few. The generated data set cannot be matched with the real data set. Therefore, it protects the dataset by preventing external intrusion. Among all the other methods k-anonymity and microaggregation techniques are highly studied.

- **Non-Perturbative:** Protection is achieved through replacing an original value by another one that is not incorrect but less specific. For example, we replace a real number by an interval. In general, non-perturbative methods reduce the level of detail of the data set. This detail reduction causes different records to have the same combinations of values, which makes disclosure difficult to intruders.

Either way, the target of SDC is to produce a dataset in such a way so that the risk of re-identification of some sensitive data is low and produces same or close results when statistical techniques are applied on the newly generated data set.