Jens Hjort Schwee, jehs@mmmi.sdu.dk
SDU Centre for Energy Informatics
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark

Software-Based Methods for Fusing and Anonymizing Building-Related Data for Cross-Organizational Data Analytics

Background

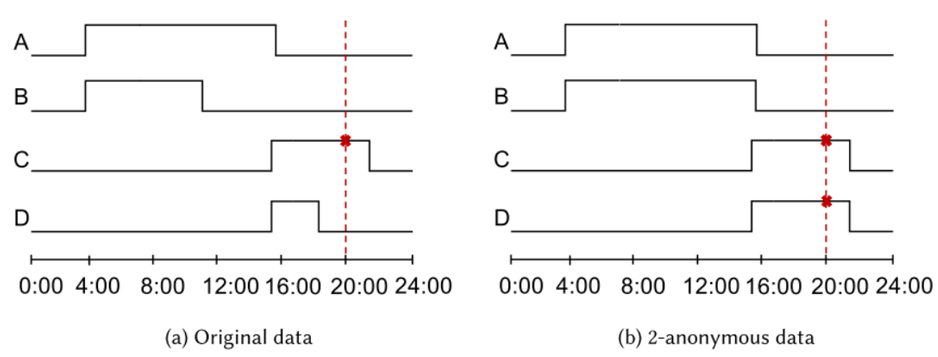
The large sensor-networks of current and future smart buildings can capture a large amount of data about building usage, resources consumption, and occupant comfort. This data enables new possibilities for both monitoring and optimization of the building. For example, some applications have been developed which are using the gathered data as input for the building management systems about the conditions of the building.
In a modern smart building, there can be many different types of sensors, which can measure the level of light, air quality, the humidity, the level of noise, the proximity of people on workstations, temperature, occupant counters and CO2. The readings are typically stored in the form of time-series, which is a sensor reading with a time-stamp. The sensor readings are then used to change the conditions of the building. Other systems are using the gathered data for monitoring how the facilities of the building are being used by the occupants, both for facility management and for preventive maintenance of the inventory. Likewise, in a commercial building, there can be a lot of contractors which have to perform a job on a regular basis, e.g., cleaning the areas in the building. Meaning that the sensor readings are capturing data about a lot of different occupants and that external contractors might want to have access to the sensor readings for analyzing how they are to perform their tasks inside the building.

Sharing the time-series, with external contractors, could be done in the form of open data, data which have been shared by the one owning or using the building. Sharing these data could be problematic as the data can be used for different types of analysis, e.g., investigating what the occupants are doing in the building, how much of the time an occupant is at a workstation or how well a company is doing. By investigating how much traffic there is a given area of the building.

Anonymity is a concept which covers that we are not able to trace a piece of data in the released data back to an individual. This can be done using several methods, we can take the inputs and add a random among of noise to each individual stream, this would create data which would be hard to infer an individual in the stream afterward. The usefulness of these streams is however limited. We can also, use a method like PAD: Protecting Anonymity in Publishing Building Related Datasets [1], which consider the use of the output data before anonymization of the data. Which likely gives a more useful dataset. The method of PAD considers the privacy model of k-anonymity which have some limitations, e.g. it does not support multiple or incremental releases. Therefore, the Ph.D. project will address strong privacy models like (X, Y)-privacy [2], l-Diversity [3], or differential privacy.

Sangogboye et al. [1] have made an example where it is highlighted why k-anonymous can be useful between data records: If we consider an example where we have data records which consist of offices occupied status, labeled A, B, C, and D. If the original data was released an attack might be able to infer when a person left the office; suppose that the attacker knows that C stays in the office until 20:00, then by linking the information with the data, the attacker can get the complete occupied status of C in the published data. If we provided 2-anonymous on the data it would not be possible to do this, see Figure 1 for a visual representation of the example.

Figure 1 Office occupancy Linage attack, provide by Sangogboye et al. [1]



(a) Original data (b) 2-anonymous data

When sharing and publishing data, there is a need for considering what can be inferred in the data before sharing it and ensure the occupant's privacy. We also must consider relevant laws and regulations, e.g., if monitoring on EU citizens or in the EU we must, among others, comply with the General Data Protection Regulation (GDPR).

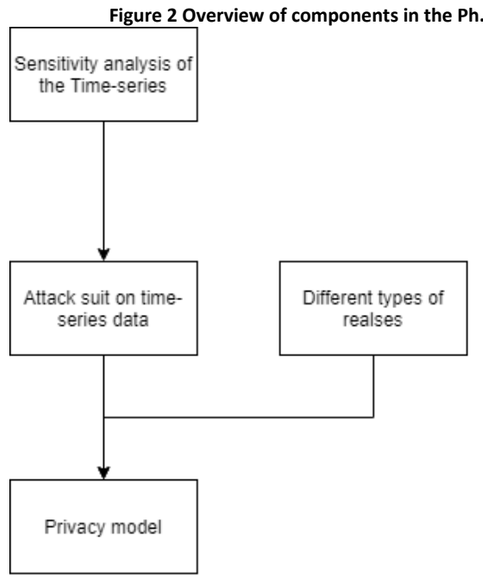
Problem Statement

When working with time-series data, we can have privacy sensitive data, depending on the sensor deployment. If an attacker had some prior knowledge of a target victim in the released dataset, the series can be used to identify a victim, e.g. consider the one occupant office rooms example earlier. The various privacy models each prevents some types of attacks. The Ph.D. project will investigate to what degree sensor data can be used to perform attacks on the data, this is to be used for developing methods for protection the privacy of the individuals and still have some usability of the data.

Overall Project

The Ph.D. work will cover contributions to two projects: HBODEx and the IEA EBC Annex 79. It will cover four components as shown in Figure 2. First a sensitivity analyses of variances sensor data, this is to be used for creating an attack suit. Then exploring how best to handle the various type of resales, the level of privacy which the methods provide is to be tested using the attack suit. Finally, an investigation of some difference privacy models for finding, the most appreciate balancing different objectives. The work will be done iteratively over the various cases.

Figure 2 Overview of components in the Ph.D.



Project Outcomes

The outcome of the project will include methods for sanitizing datasets, both using a single release, multiple releases and even live releases of data. Furthermore, will there be developed some tools using the data for showcasing the usefulness of the data.

Project Period

1 July 2018 to 30 June 2021

Ph.D. Student

Jens Hjort Schwee

Supervisor

Professor, Ph.D. Mikkel Baun Kjærgaard
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark

References

[1] R. Jia, F. C. Sangogboye, and M. B. Kjærgaard, “PAD : Protecting Anonymity in Publishing Building Related Datasets,” *4th ACM Int. Conf. Syst. Energy-Efficient Built Environ.*, 2017.
[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing,” *ACM Comput. Surv.*, vol. 42, no. 4, pp. 1–53, Jun. 2010.
[3] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, “Secure Anonymization for Incremental Datasets,” pp. 48–63, 2006.