Demo: Privacy-Preserving Building-Related Data Publication Using PAD

Ruoxi Jia*
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
ruoxijia@berkeley.edu

Fisayo Caleb Sangogboye*
Center for Energy Informatics
Mærsk McKinney Møller Institute
University of Southern Denmark
fsan@mmmi.sdu.dk

Tianzhen Hong
Building Technology Department
Lawrence Berkeley National Laboratory
thong@lbl.gov

Costas Spanos
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
spanos@berkeley.edu

Mikkel Baun Kjærgaard
Center for Energy Informatics
Mærsk McKinney Møller Institute
University of Southern Denmark
mbkj@mmmi.sdu.dk

ABSTRACT

The massive data collected from buildings provide opportunities for data- and information-based building management. Furthermore, to benefit from collective efforts in research communities, there arises a need for methods to share building-related data in a privacy-preserving manner while being able to ensure the utility of published datasets. In this demo abstract, we present PAD, an open-sourced data publication system that offers k-anonymity guarantee. The novelty of this system is to incorporate data recipients' feedbacks into the publication process in order to improve data utility. We demonstrate the interface of PAD and highlight how participants (as data publishers) can generate sanitized datasets using this interface. Also, we demonstrate how participants (as data users) can provide feedback to PAD for improving data quality.

ACM Reference format:
Ruoxi Jia, Fisayo Caleb Sangogboye[1], Tianzhen Hong, Costas Spanos, and Mikkel Baun Kjærgaard. 2017. Demo: Privacy-Preserving Building-Related Data Publication Using PAD. In *Proceedings of BuildSys '17, Delft, Netherlands, November 8–9, 2017,* 2 pages.
DOI: 10.1145/3137133.3141436

1 INTRODUCTION

The deployment of large-scale sensor networks in smart buildings gives rise to massive data, which is useful for informing building management to achieve better energy efficiency and comfort. Driven by the benefits mutual to occupants, building managers and research communities, there is a continually rising demand for publication of datasets collected in buildings. However, the data published in the original form arouses occupants' concern about privacy. Current practice in publishing building-related datasets mainly relies on policy and agreement to regulate data use, sharing and retention. However, this prescriptive approach does not prevent privacy breaches from happening in the first place. At the same time, simple anonymization might be applied to data records by suppressing the identity of record owners. It has been shown that such anonymization is prone to linkage attacks, i.e., the record owners can be re-identified if adversaries have access to some easily obtained auxiliary knowledge [2].

We designed the PAD system, a privacy-preserving data publication system that is specialized for high-dimensional datasets collected in buildings. PAD can protect privacy even against adversaries with prior knowledge about some snippets of targeted individuals' data. In our demo, we demonstrate the efficacy and usability of PAD using various data mining applications in buildings. Our demo will show that PAD can protect privacy without introducing any significant data fidelity penalties. Here we present a brief summary of PAD, and provide details about the data mining applications that will be presented in our demo. More details about the PAD system can be found in the conference paper in [2].

2 PAD SUMMARY

The attack model considered in the PAD is the so-called "trusted model", in which the data publisher is trustworthy yet the trust is not transitive to the data recipient who conducts data mining tasks on the published data. PAD ensures that the published datasets satisfy k-anonymity, i.e., any row in the published dataset cannot be differentiated from $k-1$ other rows. K-anonymity is a widely used privacy notion for publishing numerous public datasets, e.g., location data, webpage visits, energy data, etc.

K-anonymity is achieved in PAD by applying *microaggregation*-a popular data perturbation technique-prior to publication [1]. Microaggregation works as follows: First, all rows in the database are partitioned into small groups of k or more rows; Then, each individual row is replaced with the centroid of the group it belongs to. Due to the distortion introduced in the second step, the main problem in microaggregation is to retain as much information as possible while preserving k-anonymity. In order to minimize the information loss caused by microaggregation, rows that are similar to each other should be grouped together. The similarity of data is

*Both authors contributed equally to the paper.

BuildSys '17, Delft, Netherlands
 978-1-4503-5544-5/17/11…$15.00
DOI: 10.1145/3137133.3141436

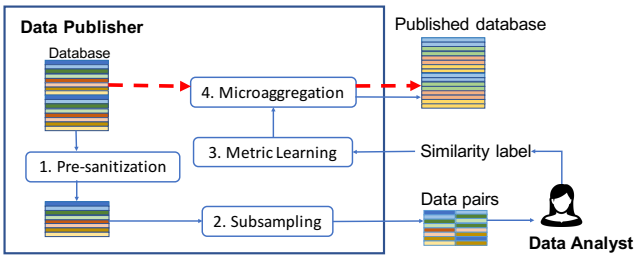

Figure 1: Data publication workflow in PAD. The red arrows illustrate general-purpose publication, and the blue ones demonstrate application-specific publication.

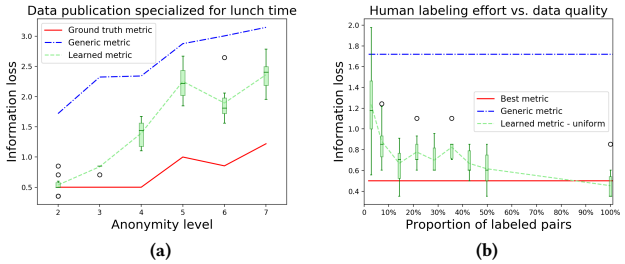

Figure 2: Tradeoff between (a) privacy and data quality, (b) labeling effort and information loss.

inherently task-dependent. Consider two data analysts who want to analyze the same occupancy dataset. The first one is interested in the occupancy patterns during electricity peak demand hours in order to estimate the demand response potential, whereas the second one is interested in the aggregate occupancy over the day for energy modeling purposes. Given the nature of their respective tasks, both should use very different distance metrics to measure the similarity of data points.

Inspired by the above observations, PAD offers two working modes: *application-specific publication* and *general-purpose publication* (Figure 1). They both produces k-anonymous datasets, while the former one can achieve improved data utility by incorporating data recipients' feedbacks on which two data points are considered "similar" into the data publication process.

Application-specific publication. If the intended use of the dataset is known at the time publication, then PAD processes the database in the following steps: (1) The data is first pre-sanitized and formed into pairs; (2) The data pairs are subsampled and returned to the data analyst to solicit their labels on which pair of rows are considered similar in accordance with the data purpose; (3) A distance metric is learned from the data pairs and corresponding similarity labels; (4) The learned metric is used by microaggregation to generate the sanitized dataset for final publication. In the case where the desired distance metric can be explicitly defined, labeling effort can be greatly alleviated by using computer programs to automatically label similarity of data points based on the desired metric.

General-purpose publication. If the purpose of the published data cannot be determined at the time of publication or the data analyst does not want to devote labeling efforts for improved data utility, then PAD directly applies microaggregation with an uninformed distance metric, such as Euclidean distance, to sanitize the dataset.

3 DATA MINING APPLICATIONS

Our demo will showcase the two working modes of PAD using various data mining applications (Table 1) and real-world datasets on occupancy and plug load consumption. We will also demonstrate the development of automatic similarity labeling algorithms for reducing labeling efforts in the application-specific publication mode. The datasets published by PAD retain useful information for data mining tasks. The moderate degree of anonymization is even helpful for improving models' robustness. Readers are referred to [2] for more details.

PAD working mode	Data mining applications
General-purpose	Occupancy prediction
	Retrieve the distribution of occupancy statistics (e.g., arrival and departure time, total occupation time)
Application-specific	Occupancy pattern during lunch time
	Total energy usage during the peak hours

Table 1: Data mining applications with PAD.

If the data purpose is known *a priori*, PAD can achieve better data quality by learning a proper metric for microaggregation. For example, if we consider a data analyst who wants to study the occupancy patterns during lunch time, i.e., 11 : 00-14 : 00. A script can be developed to label the similarity of data points. This can be achieved by clustering the data points and assigning the similarity label according to whether the pair of points belong to the same cluster. The distance metric used in clustering should be able to measure the difference of two time series during the lunch time. Figure 2 illustrates the tradeoff between privacy and data quality as well as the one between labeling efforts and information loss. Note that using the learned metric can significantly reduce the information loss compared with using the generic metric.

4 ACKNOWLEDGEMENTS

This work is supported by the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program, and the Innovation Fund Denmark for the project COORDICY (4106-00003B) at the Center for Energy informatics, University of Southern Denmark.

REFERENCES

[1] Josep Domingo-Ferrer and Josep Maria Mateo-Sanz. 2002. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and data Engineering* 14, 1 (2002), 189–201.

[2] Ruoxi Jia, Fisayo Caleb Sangogboye, Tianzhen Hong, Costas Spanos, and Mikkel Baun Kjærgaard. 2017. PAD: Protecting Anonymity in Publishing Building Related Datasets. In *Proceedings of the 4th ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM.