A Framework for Privacy-Preserving Data Publishing with Enhanced Utility for Cyber-Physical Systems

FISAYO CALEB SANGOGBOYE, University of Southern Denmark, Denmark
RUOXI JIA, University of California, Berkeley
TIANZHEN HONG, Lawrence Berkeley National Laboratory
COSTAS SPANOS, University of California, Berkeley
MIKKEL BAUN KJÆRGAARD, University of Southern Denmark, Denmark

Cyber-physical systems have enabled the collection of massive amounts of data in an unprecedented level of spatial and temporal granularity. Publishing these data can prosper big data research, which, in turn, helps improve overall system efficiency and resiliency. The main challenge in data publishing is to ensure the usefulness of published data while providing necessary privacy protection. In our previous work (Jia et al. 2017a), we presented a privacy-preserving data publishing framework (referred to as PAD hereinafter), which can guarantee *k*-anonymity while achieving better data utility than traditional anonymization techniques. PAD learns the information of interest to data users or features from their interactions with the data publishing system and then customizes data publishing processes to the intended use of data. However, our previous work is only applicable to the case where the desired features are linear in the original data record. In this article, we extend PAD to nonlinear features. Our experiments demonstrate that for various data-driven applications, PAD can achieve enhanced utility while remaining highly resilient to privacy threats.

CCS Concepts: • **Security and privacy → Pseudonymity, anonymity and untraceability**; **Privacy-preserving protocols**; **Data anonymization and sanitization**; • **Theory of computation → Unsupervised learning and clustering**;

Additional Key Words and Phrases: Privacy preservation, *k*-anonymity, smart buildings, deep learning, cyber-physical systems

Sangogboye and Jia contributed equally to this work.

This work is supported in part 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. BEARS has been established by the University of California, Berkeley as a center for intellectual excellence in research and education in Singapore. This work is also supported by the Innovation Fund Denmark for the project COORDICY (4106-00003B) at the Center for Energy informatics, University of Southern Denmark. This work was also supported by the Office of Energy Efficiency and Renewable Energy, the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Authors' addresses: F. C. Sangogboye, University of Southern Denmark, Mærsk McKinney Møller Institute, Odense M, Odense Denmark; email: fsan@mmmi.sdu.dk; R. Jia, University of California, Berkeley, Department of Electrical Engineering and Computer Sciences, Berkeley, California; email: ruoxijia@berkeley.edu; T. Hong, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California; email: thong@lbl.gov; C. Spanos, University of California, Berkeley, Department of Electrical Engineering and Computer Sciences, Berkeley, California; email: spanos@berkeley.edu; M. B. Kjærgaard, University of Southern Denmark, Mærsk McKinney Møller Institute, Odense M, Odense Denmark; email: mbkj@mmmi.sdu.dk.

1550-4859/2018/11-ART30 $15.00
https://doi.org/10.1145/3275520

ACM Reference format:
Fisayo Caleb Sangogboye, Ruoxi Jia, Tianzhen Hong, Costas Spanos, and Mikkel Baun Kjærgaard. 2018. A Framework for Privacy-Preserving Data Publishing with Enhanced Utility for Cyber-Physical Systems. *ACM Trans. Sensor Netw.* 14, 3–4, Article 30 (November 2018), 22 pages.
https://doi.org/10.1145/3275520

1 INTRODUCTION

A seamless integration of computation, networking, and the physical world is being featured in a multitude of engineering systems such as civil infrastructure, energy grid, transportation, and health care, among others. In these systems, embedded computers and networks are used to monitor and control physical processes with feedback loops where these processes affect computation and vice versa. In light of the tight coupling between cyber and physical processes, these systems are commonly termed cyber-physical systems (CPSs). CPSs have enabled various applications where decisions are driven by the sensory information. For instance, the deployment of large-scale sensing and actuation networks in buildings has driven the evolution to "smart" buildings that can collect fine-grained information about indoor environments, energy usage, and occupants. This information is further leveraged to control lighting; heating, ventilation, and air conditioning (HVAC); and other building equipment in an energy-efficient and occupant-responsive manner. Smart buildings, as a salient example of CPSs, will be considered throughout this article.

Due to the distributed nature and fast increase of system complexities, the operation of CPSs involves sensing, processing, and storage of massive amounts of data. Driven by benefits mutual to the stakeholders, there is a continually rising demand for publishing datasets collected in CPSs. In particular, publishing datasets collected in smart buildings is beneficial to occupants, building managers, and research communities. Large-scale and high-quality datasets are often enablers of robust and sophisticated models. Promoting research on advanced data analytics will eventually give rise to building operations that provide more cost savings for building managers and better adapt to occupants' needs. Occupancy modeling and energy profiling are two good examples of building applications with a significant reliance on data-driven analytics. Occupancy modeling derives occupancy schedules from data and further enables on-demand control over lighting and HVAC systems (Jia and Spanos 2017; Sangogboye et al. 2017). Energy profiling refers to the characterization of occupants' energy use, which can help gain insights into buildings' operational conditions (Gul and Patidar 2015).

However, data published in the original form can come with the risk of privacy breach, especially when the CPSs involve humans in the loop. Pristine databases may reveal detailed information about occupants' behaviors. Previous studies (D'Oca and Hong 2015; Jin et al. 2014) have shown that occupants' schedules and activities can be easily retrieved from occupancy and energy datasets. Tech-savvy criminals are already exploiting unintentional occupancy leaks to select victims for burglaries (Bloxham 2011). In addition, electricity data also indirectly reveal private information that is of interest to insurance companies, marketers, potential employers, or the government for setting premium rates, directing ads, vetting an applicant's background, or monitoring its citizens (McKenna et al. 2012). In light of the risks of privacy violation, the European Commission has proposed a comprehensive reform of data protection rules in the European Union (EU) to protect personal data from misuse, and the regulation will apply from May 25, 2018 (European Commission 2012).

Current practice in publishing CPSs' datasets mainly relies on policy and agreements to regulate data use, sharing, and retention (Bharathan 2015). However, this prescriptive approach does not prevent privacy breaches from happening. Before publication, privacy-sensitive datasets are

often anonymized by suppressing direct identifiers such as the identity of record owners. However, datasets resulting from applying simple suppression operations are vulnerable to adversaries with auxiliary knowledge. Given that an adversary possesses some prior knowledge of a person's data, the record of this person can be easily re-identified from the anonymized database by matching the records with the auxiliary information. This prior knowledge can often be easily obtained via external observations or interactions with the target.

k-anonymity (Sweeney 2002) is a stronger notion for "being anonymous" than just suppressing direct identifiers. It can mitigate the risks of re-identification by allowing data owners to "hide in the crowd." To be specific, k-anonymity ensures that each record in a database is indistinguishable from at least $k-1$ other records in the database. Since k-anonymity is conceptually simple and can be easily implemented, it has been extensively used in publishing various datasets including location data collected from mobile devices (Gruteser and Grunwald 2003). Some states in the U.S., such as California, Colorado, and Illinois, have enacted a privacy standard, often referred to as "15/15" rule, for utility companies in order to help ensure customer anonymity when energy data is released to third parties without customer consent (ElevateEnergy 2013). The privacy standard is based on the k-anonymity concept, requiring that aggregated data include a minimum of 15 customers with no one customer's load exceeding 15% of the group's energy consumption.

The main challenge in applying k-anonymity to data publishing is the information loss introduced inevitably by the anonymization process, which is also remarked in the Article 29, "Opinion 05/2014 on Anonymization Techniques" (El Emam and Álvarez 2014), composed by representatives from all EU Data Protection Authorities, the European Data Protection Supervisor and the European Commission:

> *It is clear from case studies and research publications that the creation of a truly anonymous dataset from a rich set of personal data, whilst retaining as much of the underlying information as required for the task, is not a simple proposition.*

The challenge becomes even acuter for publishing CPSs' data, as decision making and control in CPSs are highly sensitive to data quality. In the aforementioned occupancy modeling example, operating lighting and HVAC according to inaccurate occupancy schedules would affect the comfort and well-being of occupants. For the energy profiling example, without a truthful profiling grid operators can hardly preempt disturbances and ensure a stable and resilient energy supply.

In our previous work, we presented PAD (Jia et al. 2017b)—an open-sourced system to publish data collected from CPSs with k-anonymity and enhanced utility. The underlying idea of PAD for improving data utility is to customize the data privatization process to the subsequent usage of the data. To illustrate the idea, we can consider two researchers who are interested in performing different analysis on the same dataset. Suppose that one is interested in the occupancy patterns during lunchtime while the other is interested in people's arrival time. It is evident that if we want to publish a dataset that is more useful for the first researcher, the occupancy records with similar patterns during lunch time should form a size-k group so that replacing the original record with any of the records in this group would not cause severe information loss for the lunchtime occupancy patterns. In contrast, to publish a privatized dataset that is more valuable for the second researcher, the occupancy records with similar arrival time should be grouped in order to retain more arrival time information.

Although customizing k-anonymization to the interest of data users is promising to increase data utility, due to the diversity of potential data uses, it will be cumbersome to enumerate and hard-code every possible data use and design the corresponding anonymization process. In PAD, we proposed a unified protocol to comprehend users' diverse interests by learning from their interactions with the data publishing system. More specifically, PAD will first provide data users

with some data that does not involve privacy risks such as publicly available datasets, and the data users will label the similarity of these data points according to the features of particular interest to them. PAD will then learn these features from the similarity labels provided by the data users and optimize the anonymization processes accordingly. However, the current implementation of PAD is only applicable when the features are linear functions of original data records.

In this article, we resolve the constraint of linear features present in our previous work and extend the PAD framework to enhance data utility even if the features of interest to data users are nonlinear. We introduce a new learning approach for accommodating diverse interests of data users based on deep neural networks (DNNs). An example of the new abilities enabled by this extension is the accurate estimation of arrival and departure time from a database containing daily occupancy profiles. We demonstrate the value of the extension through extensive experiments on real-world smart building data of occupant presence and plug-load energy consumption.

This article is structured as follows. Section 2 presents the related work. Section 3 introduces the concept of k-anonymity, its privacy implications, and a simple technique for achieving k-anonymity. Section 4 presents an overview of the architecture of PAD. Section 5 provides the details of the algorithm to learn potential data uses. Section 6 elaborates on how to optimize the anonymization process according to the learned data use. In Section 7, we evaluate PAD using real-world smart building datasets and present the results. Section 8 presents the future work. Section 9 concludes the article.

2 RELATED WORK

Privacy-preserving data publication has been extensively studied in various contexts, including social networks (Hsu et al. 2014), smart meter data (Sankar et al. 2013), and so on. Depending on the underlying definition of privacy, data publication procedures can be categorized into three types: (1) differentially-private, (2) information-theoretically private, and (3) k-anonymous.

Differential privacy (Dwork 2008) is one of the most popular metrics for privacy, which enjoys mathematical rigorousness and often acts as a worst-case privacy measure against any possible adversaries. It is typically assured by adding appropriately chosen random noise to database outputs. One known challenge for differentially-private publication is that for high-dimensional streaming data, it often adds too much noise, which may lead to unsatisfactory data utility. Hence, it is not applicable for releasing CPSs' datasets, which are typically in the form of time series. Differential private systems have been successfully deployed to collect data on the Chrome Web browser. RAPPOR (Erlingsson et al. 2014) is a data collection and publication system that provides differential privacy guarantee. RAPPOR extends the idea of the randomized response technique where true data is perturbed to a random value with some probability depending on the strength of privacy protection. RAPPOR is only applicable to one- or two-dimensional crowdsourced data for estimating data distribution. Plausible deniability (Bindschaedler et al. 2017) is a privacy notion that has recently been used for generating synthetic datasets for publication. It ensures at least k input records that could have generated the observed output with similar probability. Plausible deniability is closely related to differential privacy. The authors in Bindschaedler et al. (2017) show that a differentially private mechanism can be obtained by slightly modifying a plausibly deniable mechanism. The difference between k-anonymity and plausible deniability is that the former is a syntactic condition on the published dataset, whereas the latter is a condition on the synthetic data generation algorithm.

Information-theoretic privacy ensures that limited knowledge can be learned about individuals from a published database, and the amount of information leakage is characterized via information theory (du Pin Calmon and Fawaz 2012; Jia et al. 2017b). Calmon et al. (du Pin Calmon and Fawaz 2012) pioneered research on applying information theory and statistical decision frameworks to

study privacy leakage. The framework models privacy using a probabilistic argument and data utility to be the distance between the true value of a data record and the perturbed value. Under this framework, the problem of solving the optimal perturbation can be converted to the rate-distortion problem, which has been extensively studied in information theory. Rajagopalan et al. (2011) apply the framework to smart meter data publication. This framework facilitates the analysis of privacy-utility tradeoff for data publication. However, the caveat is that it requires a model of the joint distribution of private information and sensor measurements, which is nevertheless difficult to be obtained in practice.

K-anonymity has received a great deal of attention during the last decade, and has been successfully implemented in various areas among which the most prominent one is location-based services (Gkoulalas-Divanis et al. 2010). Gruteser and Grunwald (2003) present a location data collection system that adjusts the resolution of location information along spatial or temporal dimensions to meet anonymity constraints. Location data takes the form of time series and often has strong time correlation. Our work is partially inspired by the wide adoption of *k*-anonymity in location-based services.

3 K-ANONYMITY

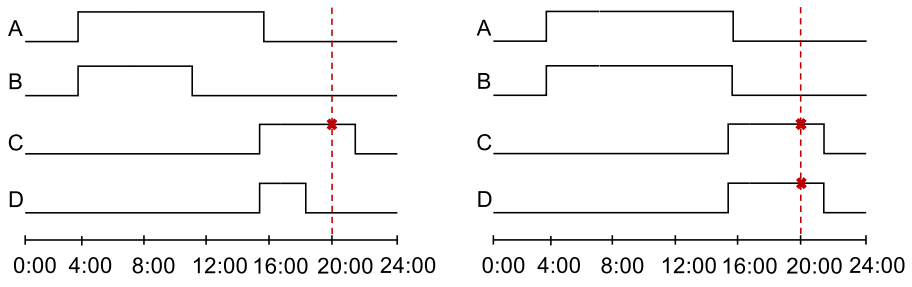
In this section, we will discuss the privacy value of *k*-anonymity and attacker models, followed by a brief introduction of basic techniques for achieving *k*-anonymity. We will close the section by discussing the intrinsic tradeoff between privacy and data utility and some limitation of basic techniques to motivate the design of the proposed system.

3.1 Privacy Value

The concept of *k*-anonymity (Sweeney 2002) was originally introduced in the context of relational data privacy. The idea behind *k*-anonymity can be described as "hiding in the crowd," as it requires that each individual cannot be identified within a set of k individuals in the released data. In this article, we deal with a slightly more general definition of *k*-anonymity, i.e., we consider a *row* in a database as *k-anonymous* if and only if it is indistinguishable from at least $k-1$ other rows. Depending on the contents of a row, this definition can incorporate the privacy guarantee at different levels. For instance, if each row is a daily energy or occupancy profile of a person, then this definition ensures that the profile of each day cannot be differentiated from $k-1$ other profiles. If we consider that each row in the database contains information of a person, then we recover user-level privacy, which guarantees the indistinguishability of k persons and, therefore, offers a stronger privacy notion.

We illustrate the privacy value of the *k*-anonymity model by comparing it with the strategy that only masks the identifier of each row in a database. Assuming a data analyst requests data publishing and the database is sanitized solely by suppressing names of the data owners, we want to show that the information retained in this database can still create a threat against data privacy when combined with external observations or knowledge.

As an example, consider the scenario depicted in Figure 1 where the database contains four rows corresponding to the office occupancy status of four persons labeled as A, B, C, and D. If no *k*-anonymization is performed by the data curator, then the following linkage attack can be conducted: Suppose the adversary knows that C stays in this office at 20:00; then, by linking this information with the data trajectories it has at hand, it can find the complete occupancy status of C in the time horizon of the published data. However, such linkage attack is not effective if proper data perturbation is performed by the data curator to maintain *k*-anonymity. Consider the 2-anonymized version of the original dataset illustrated by Figure 1(b). Now, even if the adversary can have access to the knowledge of occupancy status of C via external observations, it cannot



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

Fig. 1. Linkage attack.

recover the complete data trajectories with certainty, as 2-anonymity guarantees that at least two rows in the database have the same values.

In this article, we wish to achieve data protection against the adversaries with the following capabilities: (1) Having access to the published data; (2) Knowing short snippets of truthful private data by external observations.

3.2 Microaggregation

Microaggregation is a popular perturbation technique to achieve k-anonymity for databases with quantitative records. It processes the data in the following two steps prior to publication:

- **Step 1 (k-partition):** All rows in the database are partitioned into small aggregates of k or more rows.
- **Step 2 (substitution):** Each row is replaced with the centroid of the group it belongs to.

Following this procedure ensures that every record in the released database corresponds to at least k individual records; hence, k-anonymity is guaranteed.

Due to the data distortion introduced in the substitution step, the main problem in microaggregation is to retain as much information as possible while offering sufficient privacy protection. In order to minimize the information loss caused by microaggregation, groups should be formed by maximizing their within-group homogeneity. The more homogeneous the records in a group are, the lower information loss is incurred when replacing the true value of a record by the group average. The sum of squared distances (SSD) criterion is a common measure to estimate group heterogeneity and this is defined as

$$SSD = \sum_{i=1}^{g} \sum_{j=1}^{n_i} d(x_{ij}, \bar{x}_i), \tag{1}$$

where x_{ij} denotes the j-th row of the i-th group, $\bar{x}_i$ represents the centroid of the group i, n_i is the number of elements in the i-th group, and g stands for the number of groups.

The distance metric $d(\cdot, \cdot)$ in Equation (1) is often chosen to be a uninformed norm, such as Euclidean distance. Although Euclidean distance is simple and intuitive, it ignores the fact that the semantic meaning of "information loss" is inherently task- and data-dependent (Weinberger et al. 2006). To illustrate this point, imagine two researchers 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 information loss.

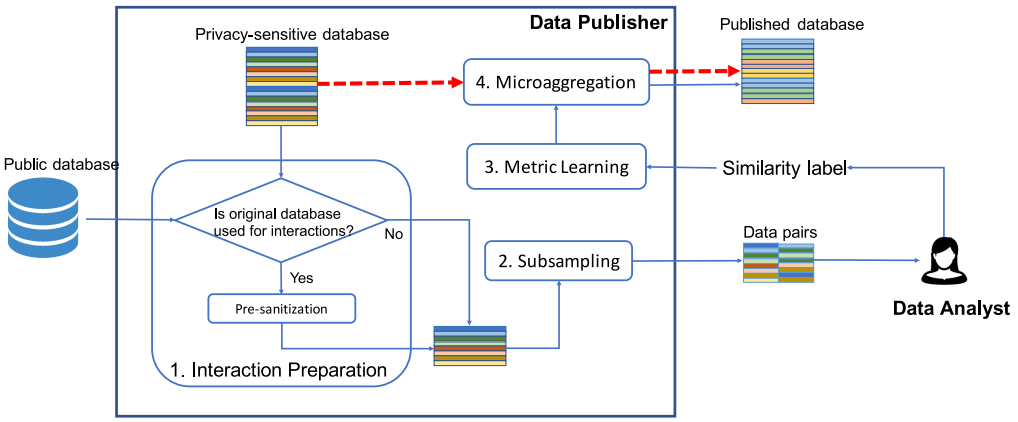


Fig. 2. PAD diagram: If the purpose of the dataset to be published is not known prior to publication, then PAD directly applies microaggregation with an uninformed distance metric to sanitize the dataset (shown by the red dashed arrow). Otherwise, PAD processes the data in the following steps: (1) Prepare the training data used for learning potential data uses. The training data can either come from the original database or a similar dataset that is already public. Pre-sanitize the data if the original database is used. (2) The data pairs are subsampled from the prepared training data and returned to the data analyst to solicit their labels on which data pairs are considered similar (The labels can be assigned manually or automatically using custom programs); (3) PAD learns a metric from the similarity labels; (4) The learned metric is used by microaggregation to generate the sanitized dataset for final publication.

If the purpose of the data is known at the time of publication, it can be taken into account during microaggregation to better retain information. But clearly, building a system to parse data users' interest is not the most robust and scalable approach due to the diversity of different data analysts' interest. It is, therefore, more desirable to have a standard protocol for different users to express their respective data purposes. Our approach implemented in PAD is to learn the distance metric explicitly for each specific application from data points' similarity labeled by the user.

4 OVERVIEW OF PAD

We assume that the *data publisher* collects data records and releases the collected data to the *data recipient*, who will then conduct data mining on the published data. We will use "*data recipient*," "*data analyst*," and "*data user*" interchangeably in this article. Further, we assume that the data publisher is trustworthy yet the data recipients are not. This assumption is also referred to as the *trusted* model (Fung et al. 2010). Since, in our framework, data analysts can interact with the data publication system to improve the usefulness of the published data, it is important to ensure that data analysts do not have access to the original database during any part of the data publication process.

Figure 2 illustrates the design of PAD. The objective of the system is to publish the dataset with k-anonymity guarantee as well as high quality in support of the required data analysis. The core idea of the system is to improve the data fidelity by learning how the data is intended to be used and then adjusting the data perturbation algorithm accordingly.

If the data is not used for specialized purposes, then PAD directly applies microaggregation and publishes the database. Otherwise, PAD processes the original database in the following four steps.

(1) Interaction preparation. The objective of this step is to provide a dataset for the data analyst to label data points' similarity, which will be later used to learn the purpose of data analysis. The dataset can either come from the original database or a dataset that is already public. Since

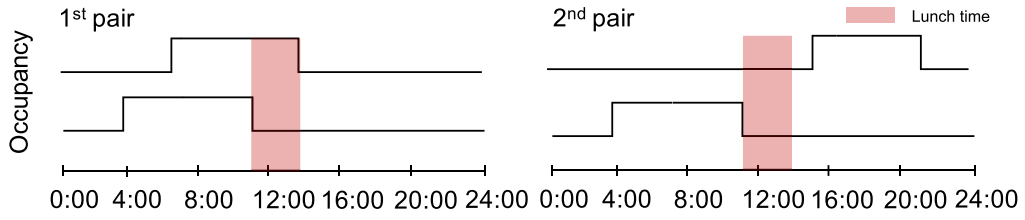


Fig. 3. Illustration of determining similarity labels.

this dataset should not cause additional privacy concerns, it must be pre-sanitized if the original databased is used for interacting with the data analyst. At this step, the system has not received any inputs from the data analyst yet. Pre-sanitization is, therefore, performed via microaggregation with a simple generic distance metric, e.g., Euclidean distance.

(2) Subsampling. As the second step, PAD processes the rows in the prepared database into pairs and randomly selects some pairs to be returned to the data analyst, who will then assign a binary label indicating if the two rows are similar or not in accordance with the particular data purpose to each returned data pair. Consider, for example, the two pairs of occupancy records depicted in Figure 3. If the data analyst wants the published dataset to maximally retain the information regarding the occupancy patterns during lunchtime, then he will assign "dissimilar" to the first pair and "similar" to the second one; however, if the data analyst is interested in the occupancy patterns during the entire day, then the first pair will be labeled as "similar" and the second one as "dissimilar." In the case where the desired metric for comparing similarity can be explicitly defined, the labeling effort can be greatly alleviated by using computer programs to automatically label similarity of data points based on the desired metric.

(3) Metric learning. In this step, a distance metric over the data record is automatically learned from data pairs and the corresponding similarity relationships specified by the data analyst.

(4) Microaggregation. This step uses the distance metric learned from the previous step for microaggregation so that the database can be sanitized in a way that the information of interest to the data analyst is maximally retained.

The detailed algorithms for (3) metric learning and (4) microaggregation will be presented in Sections 5 and 6, respectively. Before closing the section, we want to point out that the existence, amount and quality of similarity labels provided by the data analyst affect the usefulness of the published data; however, the privacy level remains the same regardless because the dataset is always microaggregated before publication.

5 DISTANCE METRIC LEARNING

We will firstly summarize the linear distance metric learning method presented in our previous study. Next, we will introduce a more flexible metric learning method based on deep neural networks and it can learn distance metrics for both linear and nonlinear features.

5.1 Linear Metric Learning

Let the original and finally published dataset be denoted by X and $\tilde{X}$, respectively. In addition, we denote by $\hat{X}$ the dataset prepared for learning a distance metric. In the metric learning step, the data analyst is provided with some data pairs $(\hat{x}_i, \hat{x}_j)$ $(i, j = 1, \ldots, |\hat{X}|)$ from $\hat{X}$, and assigns a similarity label to each of the data pairs. The objective is to learn a distance metric $d(x_i, x_j)$ between points x_i and x_j so that "similar" points end up close to each other. In our previous study, we have adopted the Mahalanobis distance metric as the underlying metric for the learning mechanism.

The Mahalanobis distance metric is given by:

$$d(x_i, x_j) = d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A(x_i - x_j)} \tag{2}$$

The Mahalanobis distance is a generalization of Euclidean distance by admitting linear scalings and rotations of the original data space. The metric learning goal is to learn the matrix A such that it reflects the similarity relationship labeled by the data analyst. A is often termed as inverse covariance (IC) matrix. Setting A to be the identity matrix I gives the Euclidean distance; Restricting A to be diagonal corresponds to learning a metric where the different axes are weighted differently. Note that $d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A(x_i - x_j)} = \|A^{\frac{1}{2}}x_i - A^{\frac{1}{2}}x_j\|_2$, and, therefore, learning a full matrix A is equivalent to finding a scaling and rotation of data that replaces each point x with $A^{\frac{1}{2}}x$ and applying the Euclidean distance to the transformed data. Similar to Xing et al. (2003), we formulated the linear distance metric learning as an optimization problem, which solves for A such that the distance between the data pairs labeled as "similar" $\mathcal{S}$ are minimized and pushes the "dissimilar" $\mathcal{D}$ pairs far away. We also applied proper regularization to address the over-fitting problem. The interested readers are referred to our previous study (Jia et al. 2017b) for more details.

5.2 Deep Metric Learning

One challenge with the Mahalanobis distance metric is that it performs well only when the feature is linear in the original data record. This is because learning the Mahalanobis distance metric from labeled data pairs is equivalent to seeking a linear transformation $A^{\frac{1}{2}}$. This implies that our previous approach cannot adequately capture the nonlinearities presented in a number of scenarios such as arrival and departure time analysis of an occupancy dataset.

To overcome this limitation, several approaches have been proposed to learn a nonlinear distance function, including the use of kernels (Tsang et al. 2003; Yeung and Chang 2007) and deep neural networks (Hadsell et al. 2006; Hu et al. 2014). Kernel methods map each data instance to a high-dimensional feature space and then learn a distance metric in the high-dimensional space. The challenge with kernel methods is that they require a user-specified kernel function. Conversely, deep metric learning can learn a nonlinear representation of data and enjoys more flexibility. To the best of our knowledge, no previous approach has applied DNNs for improving data utility with regard to k-anonymization.

DNNs pass the dataset through several layers of nonlinear transformations achieved by compositing linear transformations and nonlinear activation functions, as illustrated in Figure 4. The neural network we used comprises two identical branches, representing the nonlinear feature function applied to both points in a data pair. Let us first consider a single branch. Suppose that there are N layers in a deep network and that for each layer such as the nth layer, there are $k^{(n)}$ activation units, where $n = 1, 2, 3, \ldots, N$. The first layer takes one of the points in a data pair $x \in \mathbb{R}^d$ as input and outputs $h^{(1)} = g(W^{(1)}x + b^{(1)}) \in \mathbb{R}^{p^{(1)}}$. $W^{(1)} \in \mathbb{R}^{p^{(1)} \times d}$ is a projection matrix, $b^{(1)} \in \mathbb{R}^{p^{(1)}}$ is a bias vector, and $g : \mathbb{R} \mapsto \mathbb{R}$ is the non-linear activation function. Examples of commonly used activation functions include `sigmoid`, `tanh`, and `rectified` functions. The output $h^{(1)}$ from the first layer becomes the input for the second layer, and the output of the second layer is given by $h^{(2)} = g(W^{(2)}h^{(1)} + b^{(2)}) \in \mathbb{R}^{p^{(2)}}$, where $W^{(2)} \in \mathbb{R}^{p^{(2)} \times p^{(1)}}$, $b^{(2)} \in \mathbb{R}^{p^{(2)}}$. We can compute the outputs of other layers in a similar fashion and the output of the topmost layer, i.e., the Nth layer, is as follows:

$$f(x) = h^{(N)} = g\left(W^{(N)}h^{(N-1)} + b^{(N)}\right) \in \mathbb{R}^{p^{(N)}}, \tag{3}$$

where the $f : \mathbb{R}^d \mapsto \mathbb{R}^{p^{(N)}}$ is a non-linear function determined by the foregoing parameters $W^{(n)}$ and $b^{(n)}, n = 1, 2, 3, \ldots, N$, as well as the nonlinear active function. Hence, we compute the distance

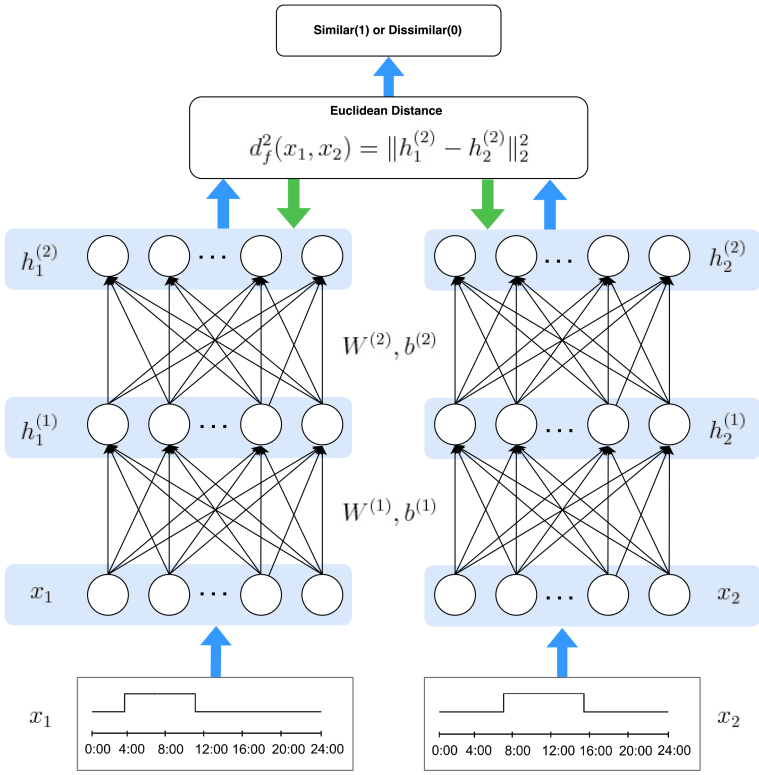


Fig. 4. Deep Metric Learning with a two-layer neural network: A pair of data samples x_1 and x_2 are transformed to $h_1^{(2)}$ and $h_2^{(2)}$ through the same hierarchical non-linear transformation specified by the neural network. The Euclidean distance between $h_1^{(2)}$ and $h_2^{(2)}$ are computed to determine if x_1 and x_2 are similar.

between any pair of data samples x_i and x_j by first performing the transformation $f(x_i) = h_i^{(N)}$ and $f(x_j) = h_j^{(N)}$ and then calculating the Euclidean distance between $f(x_i)$ and $f(x_j)$:

$$d_f^2(x_i, x_j) = \|f(x_i) - f(x_j)\|_2^2 \tag{4}$$

The objective of our deep network is to find a non-linear mapping f such that for similar pairs $\mathcal{S}$ with the label $Y = 0$, $d_f^2(x_i, x_j)$ is smaller than for dissimilar pairs $\mathcal{D}$ with the label $Y = 1$. Given Equation (4), Hadsell et al. (2006) proposed a contrastive loss function that learns the parameters of f such that data pairs in $\mathcal{S}$ are pulled closer and those in $\mathcal{D}$ are pushed apart. This contrastive loss function is defined as:

$$L(f, Y, x_i, x_j) = (1 - Y)\frac{1}{2}\Big(d_f^2(x_i, x_j)\Big) + (Y)\frac{1}{2}\Big(max\{0, m - d_f^2(x_i, x_j)\}\Big), \tag{5}$$

where $m > 0$ is a margin that separates $\mathcal{S}$ and $\mathcal{D}$. $\mathcal{D}$ only contributes to the loss function if their distance is within the margin (Hadsell et al. 2006).

It is worth noting that this network can also be adapted to learning linear distance metrics by simply replacing the activation functions in each hidden layer with identity functions.

6 EFFICIENT ALGORITHM FOR MICROAGGREGATION

As discussed previously, microaggregation includes two steps, namely, k-partition that clusters the data into group sizes of at least k records and a substitution step that perturbs the data by replacing the true values by the group centroid. It is possible that the data type of group centroid is not consistent with the original data. For instance, the centroid of multiple occupancy time series is not necessary to be in an integer form. In such cases, proper post-processing, like rounding, should be conducted to make the published database meaningful.

The information loss in the published dataset is mainly determined by the k-partition step. An optimal k-partition is defined to be the one that minimizes the heterogeneity of group members characterized by Equation (1). Note that k-partition is different from the classical clustering problem where the goal is to split the dataset into a fixed number of groups irrespective of the group size. In the case of k-partition, the constraints are on the group size instead of the number of groups. Nevertheless, we can modify the classical agglomerative clustering to make it serve for the k-partition purposes by terminating the agglomeration process at the proper level where the size of each group formed satisfies the constraints desired by the optimal k-partition.

The following proposition states the properties of the sizes of groups formed by optimal k-partition.

Proposition 1. *An optimal solution to the k-partition problem of a set of data exists such that its groups have size greater than or equal to k and less than $2k$.*

The proof can be found in Domingo-Ferrer and Mateo-Sanz (2002a). Proposition 1 indicates that the search space of the optimal k-partition can be reduced to the partition where all groups have size between k and $2k$. Therefore, we modify a widely used agglomerative clustering algorithm, Ward's method (Domingo-Ferrer and Mateo-Sanz 2002b), to provide a heuristic and efficient solution that fulfills the group size requirements. The detailed algorithm is presented in Algorithm 1.

ALGORITHM 1: k-ward algorithm

Input: Database $X_i, i = 1, \ldots, n$

Group initialization
Define the extreme data points as the two that are most distant
For each of the extreme data points, take $k - 1$ data closest to it and form the first two groups
The rest of the data points in the dataset constitute single-element groups
Agglomerative clustering via Ward's method
while there exists some group of the size less than k **do**
    Find the nearest pair of distinct groups, at least one of which must have size less than k
    Merge the two groups and decrement the number of groups by one
end while
if there exists some group containing $2k$ or more data **then**
    Apply k-ward algorithm recursively on those groups
end if

7 EVALUATION

We evaluate the performance of PAD using various datasets collected in real-world buildings. The questions we would like to answer from the experiments are:

—How useful are the sanitized datasets for typical data mining tasks?
—If the use purpose of a dataset is predetermined, can a dataset sanitized with the learned metric retain more relevant information than the one sanitized with an uninformed metric?

To answer these questions, we differentiate between three evaluation cases, namely:

(1) The utility of PAD with a generic distance metric
(2) The utility of PAD with a customized distance metric when the feature of interest to the data analyst is linear in the original data record
(3) The utility of PAD with a customized distance metric when the feature of interest to the data analyst is nonlinear in the original data record

7.1 Experimental Setup

7.1.1 Datasets. Our datasets include occupancy and plug load power consumption, which represent typical building data types that may arouse occupants' privacy concerns. Two different occupancy datasets are employed in this study. One occupancy dataset, lasting about half a year, was collected at a resolution of 1 minute in four classrooms of the OU44 building at the University of Southern Denmark. In the following, this dataset will be referred to as *OU44 occupancy dataset*. Another occupancy dataset, which we call *smart home occupancy dataset*, was collected from thermostat motion sensors in 49 users' houses. Each user's data has a resolution of 5 minutes and lasts 1 day. Both datasets contain a binary occupancy time series, indicating whether or not the room is occupied. Occupancy data can potentially reveal privacy-sensitive information such as daily routines and detailed schedules of the inhabitants. The *plug load dataset* consists of 15-minute-resolution power consumption data over 3 months. This dataset was collected at the individual desks of five occupants in Cory Hall on the UC Berkeley campus. Plug load data also raises privacy concerns. As shown in the previous studies (Jin et al. 2017; Molina-Markham et al. 2010), occupants' presence or even more detailed activities can be easily identified from power data.

Since OU 44 occupancy dataset and plug load dataset contain a relatively small population of individuals, we will consider anonymity protection at the daily profile level instead of the user level. That is, k-anonymity ensures that k day profiles, rather than k users, are indistinguishable. In this regard, we process these two datasets into the form where each row corresponds to a person's daily occupancy or energy profile. We would like to stress that the framework can also protect the anonymity at the user level by feeding a dataset where each row corresponds to the data of a different user, such as the smart home occupancy dataset.

7.1.2 Implementation. The deep metric learning algorithm was implemented using *Keras* (Chollet et al. 2015) and *Tensorflow* (Abadi et al. 2015). We used `rectified` as the activation function for the hidden layers. The `Adam` algorithm was adopted for learning the weights of the network. The implementation code of the framework is open-sourced at https://github.com/PAD-Protecting-Anonymity/PAD.

7.1.3 Evaluation Procedure. We evaluate PAD in two training scenarios. One is where public datasets are available to train a distance metric. In that case, we divide the data into two parts. The first part is assumed to be the privacy-sensitive database that will be sanitized by PAD. The second part plays the role of a publicly available dataset and is used for training distance metric functions. Another training case considered in our experiments is where publicly available datasets are difficult to find and, thus, the pre-sanitized version of the original database is used for training the distance metric. We demonstrate the results of both cases for all experiments performed in this article. We further split training data into two portions: one for fitting the distance function

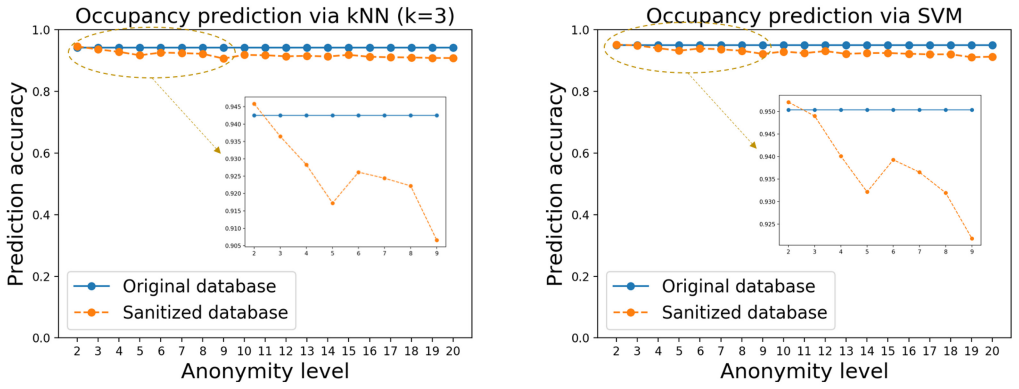


Fig. 5. Comparison of prediction performance of occupancy models constructed by using the original vs. sanitized database.

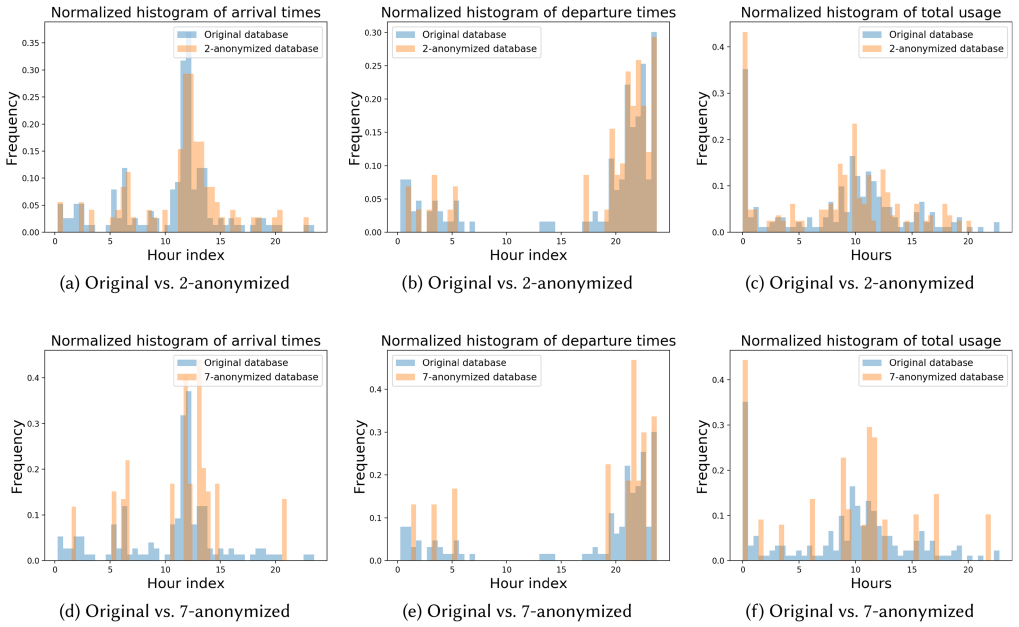
and another for testing the fitted function and preventing overfitting. We do not implement hyper-parameter tuning due to the lack of training data. In addition, to examine the performance variation of the learned metrics caused by changes in the training dataset, we conduct five Monte Carlo (MC) simulations, and in each MC simulation, 80% of the training samples are randomly drawn to learn the distance metrics.

7.2 Utility of PAD with Generic Distance Metric

We first focus on a general scenario where the system does not have access to similarity labels. This can happen either when the purpose of the data is now known before publication, or when the data analyst does not want to interact with PAD. In that case, a generic metric, i.e., Euclidean distance, is used for performing micro-aggregation. We validate the usefulness of the k-anonymized dataset through several typical data mining tasks, including occupancy prediction and occupancy statistics extraction.

7.2.1 Prediction. K-nearest neighbor (KNN) based occupancy prediction models are built using the original and sanitized database, respectively, with varying anonymity levels. To make a prediction at time t, we compute the distance between the testing profile and all profiles in the training set during the interval $[t - \Delta t, t - 1]$ where Δt is the length of the window used for prediction, and then pick the most common occupancy value at t among the K nearest training profiles. Cross-validation is performed to compute the average prediction accuracy across all time steps in the day. The results are shown in Figure 5(a), where the prediction accuracies with the original and sanitized dataset are both above 90%. There is a tradeoff between anonymity protection level and data utility. We can see that the prediction accuracy drops as the anonymity level of the published dataset is increased.

It is important to note that a moderate degree of anonymization is helpful for improving model's robustness and better fitting unseen data. Particularly, the KNN model constructed with a 2-anonymized dataset achieves higher prediction accuracy than that built with the original dataset. We also implement an occupancy prediction model based on Support Vector Machine (SVM) and the corresponding results are shown in Figure 5(b) where we can observe the similar patterns. This is because the training data points usually contain both the useful information that can be used to predict unseen cases, as well as the useless noise that can degrade the model. Essentially,



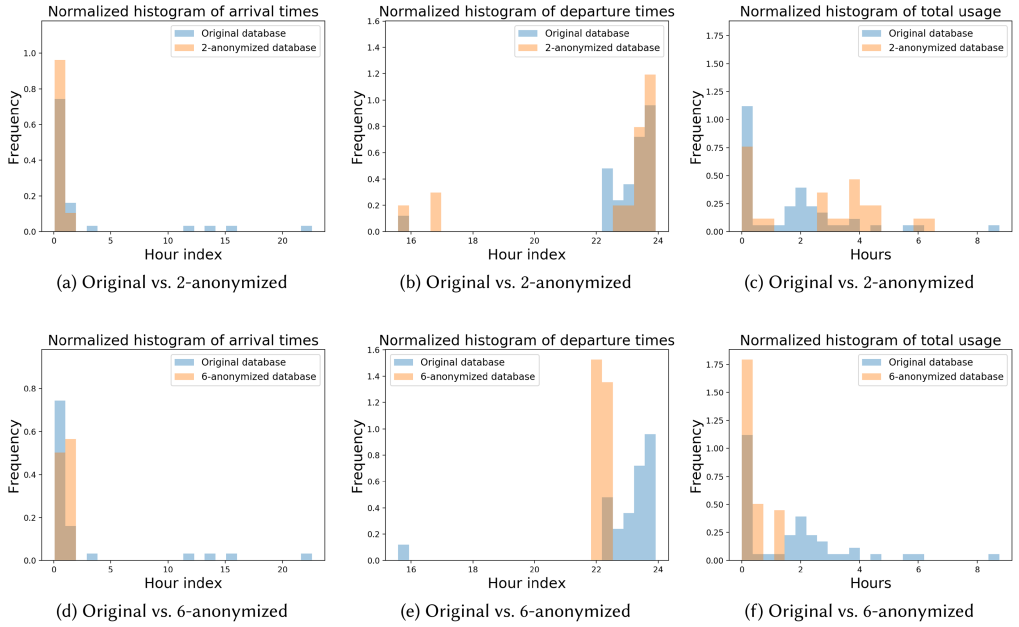
(a) Original vs. 2-anonymized (b) Original vs. 2-anonymized (c) Original vs. 2-anonymized

(d) Original vs. 7-anonymized (e) Original vs. 7-anonymized (f) Original vs. 7-anonymized

Fig. 6. Comparison of occupancy statistics extracted from the OU44 occupancy dataset and the corresponding sanitized dataset.

k-anonymization reduces the "harmful" noise by aggregating similar data points and avoids overfitting. This suggests that for a data publication with moderate anonymity requirement the sanitized dataset is more advantageous than the original dataset since the sanitized one can achieve privacy protection as well as an improved model quality.

7.2.2 Statistics. The raw time series collected in buildings are often processed into some key information that is directly useful for informing various control applications. For instance, occupancy statistics, such as arrival time, are particularly useful for designing occupant-responsive HVAC control algorithms. In light of this, we want to test if the sanitized database can retain these useful statistics. We compare the histograms of the useful occupancy statistics including arrival time, departure time, and total occupation time extracted from the original and sanitized database, respectively. Figures 6 and 7 illustrate the results on the OU44 occupancy dataset and the smart home occupancy dataset, respectively. We can see that the anonymized datasets can preserve the distribution of these statistics, especially the mean and modes of the distribution. Take the OU44 occupancy dataset for example: the relative errors of using the 2-anonymized datasets to estimate the mean of arrival time, departure time, and total occupation time are 8.13%, 8.37%, and 6.21%, respectively; for 7-anonymized datasets, the relative errors are 6.80%, 5.34%, and 0.47%, respectively. In other words, we can still retrieve accurate information about typical behaviors of occupants from the sanitized database. However, it is worth noting that data sanitization reduces the variability of the dataset, which is getting more pronounced when the anonymity level is increased to 7 as shown in Figure 6(d), (e), and (f). For instance, the departures at noon cannot be detected with the anonymized dataset. This is a direct consequence of "hide in the crowd" philosophy of k-anonymity. Therefore, it will be easier to mine population properties than atypical patterns from the sanitized data.



(a) Original vs. 2-anonymized (b) Original vs. 2-anonymized (c) Original vs. 2-anonymized

(d) Original vs. 6-anonymized (e) Original vs. 6-anonymized (f) Original vs. 6-anonymized

Fig. 7. Comparison of occupancy statistics extracted from the smart home occupancy dataset and the corresponding sanitized dataset.

7.3 Utility of PAD with Customized Distance Metric for Linear Features

In this part, we investigate scenarios where the purpose of the data is known at the time of publication and there exists a "best" distance metric for microaggregation, which retains the maximum amount of information pertaining to the data analyst's interest. For instance, if the data is used for studying occupancy patterns of a building during lunchtime, then the best metric will be the Euclidean distance over the lunch period. The data records with similar lunch patterns will be grouped by the "best" metric; as a result, the information loss with respect to lunchtime occupancy patterns incurred by the substitution step will be minimized.

First, we consider that the feature that interests the data analyst is a linear function of the original data record. The aforementioned lunchtime occupancy pattern is an example of the linear feature because the lunchtime occupancy is equivalent to multiplying the whole-day occupancy data by a diagonal matrix that has non-zero entries only at the coordinates corresponding to the lunchtime. In the sequel, we will use two use cases, namely, occupancy data segments and peak-hour energy usage, to demonstrate the utility of PAD for linear features.

Note that although there has been fruitful previous research on data publishing, different approaches may not be directly comparable because they may have different viewpoints on what is considered "private." Existing work on k-anonymization always relies on a generic metric in the microaggregation step. Therefore, PAD with the generic distance metric is used as the baseline approach for comparison here. Also, as described in Section 5, DNNs can represent features in linear and nonlinear forms. In the following, we will also compare these two representations, referred to as *linear metric* and *nonlinear metric*.

7.3.1 Segment. Consider that the data analyst wants to study the occupancy patterns during lunchtime, i.e., $11:00-14:00$. In Figure 8(a) and (b), we compare sanitization procedures that use a generic metric, the metric learned by a linear neural network (i.e., linear metric), the metric

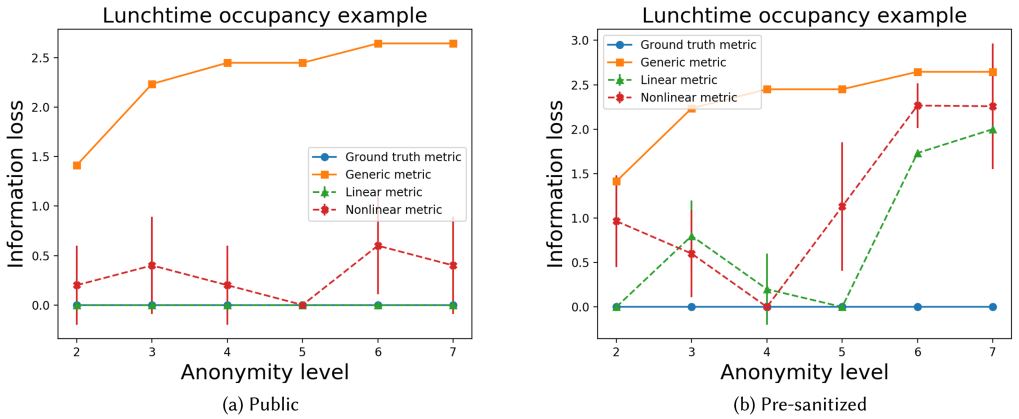
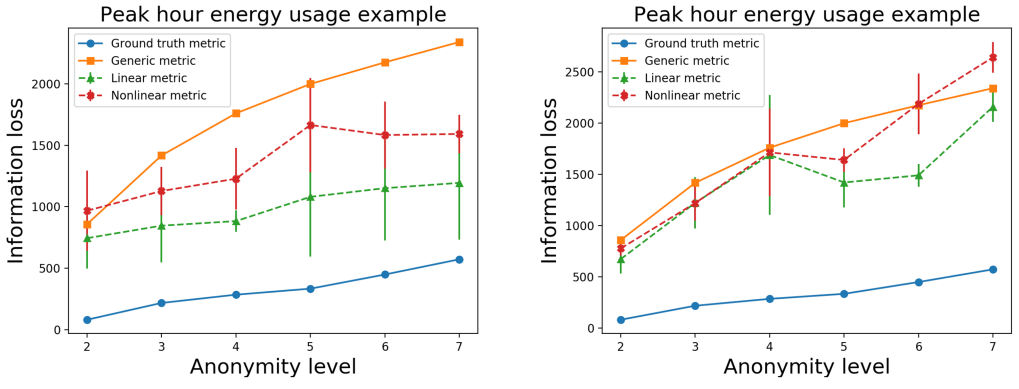


Fig. 8. The tradeoff between anonymity level and information quality for the customized publication for preserving lunch-time occupancy patterns.

learned by a nonlinear neural network (i.e., nonlinear metric), and the ground truth metric, respectively. The performance of the metrics learned from a separate public dataset are shown in Figure 8(a), and the ones learned from the pre-sanitized dataset are shown in Figure 8(b). The information loss for special-purpose publication measures the difference between the interesting information in the original data record and that in the sanitized record. Here, the information loss refers to the Euclidean distance of the lunch periods between the record in the original database and its sanitized version in the published database. The errorbars indicate ±1 standard deviation of the information losses for different MC simulations. We can see that the information loss can be significantly reduced by learning a proper metric for microaggregation. As discussed before, the lunchtime occupancy pattern is a linear feature and, therefore, using a linear distance metric can indeed well-preserve the lunchtime pattern. The information loss increases by a large amount at high anonymity levels when the metric is learned from the pre-sanitized dataset. This is because the number of unique data points decreases in the pre-sanitized dataset as the anonymity level goes higher and fewer data points can be used for learning the distance metrics.

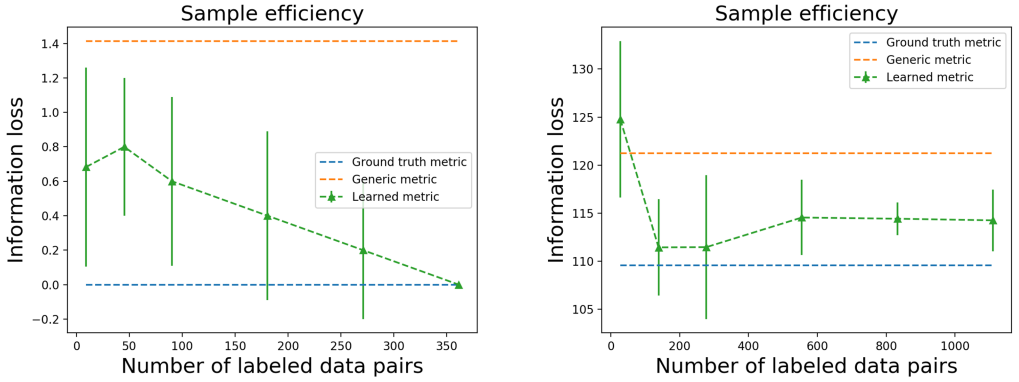
7.3.2 Peak-hour Energy Usage. We consider an energy data use case that mines occupants' peak-hour energy use patterns. More specifically, the data analyst is interested in acquiring accurate information on total energy consumption during the peak hours, i.e., $17:00-20:00$. The ground truth metric associated with this example can be defined as $d_p(x,x') = \|f(x) - f(x')\|_2$ where f calculates the sum of the coordinates during peak hours for x and x' and, therefore, is also a linear feature. Figure 9(a) shows the information loss of peak-time usage in the published datasets using the generic metric, the linear metric, the nonlinear metric, and the ground truth metric, respectively, under different anonymity guarantees. The information loss is measured by the difference between the peak-hour total usage of the original record and that of the sanitized version in the published database. We can observe a similar tradeoff between privacy and data utility to what we have seen in the use case of the lunchtime segment. The information loss can be reduced by replacing a generic metric with the learned metrics. Since the feature, per se, is linear, using a linear neural network for metric learning can, in effect, outperform the nonlinear ones.

7.3.3 Sample Efficiency. Figure 10 demonstrates the variation of published data quality with respect to the change in the number of samples used for metric learning. Figure 10(a) and (b) show that with more labeled data pairs, PAD can generally achieve better data utility.



(a) Public (b) Pre-sanitized

Fig. 9. The tradeoff between anonymity level and information quality for the customized publication for preserving peak-time energy usage patterns.



(a) Lunchtime occupancy example (b) Peak-hour energy usage example

Fig. 10. The tradeoff between labeling effort and information loss. The total number of data pairs is 903 for the lunchtime occupancy example, and 2,775 for the peak-hour energy usage example.

7.4 Utility of PAD with Customized Distance Metric for Nonlinear Features

In this part, we will switch our focus to nonlinear features, which are quite common in mining smart building datasets. For instance, the data analyst is interested in modeling the arrival and departure time of a building from the occupancy datasets. Assume that each row in the database contains occupancy measurements throughout the day, denoted by a vector x. Let f be the function that calculate the arrival time of x. Then, we have $f(x)$ is equal to the first non-zero element in x, which is apparently a nonlinear function of x.

Figure 11(a) and (b) compare the ability of different metrics to retain arrival time information. We can see that learning a proper non-linear metric for microaggregation is beneficial to the preservation of nonlinear features. Linear metrics require fewer examples to train because they have fewer parameters. Since the number of unique training samples decreases as anonymity level increases, we observed in Figure 11(b) that the linear metric is more performant than the nonlinear one when a high level of anonymity is desired. The results corresponding to departure time are illustrated in Figure 11(c) and (d), where the advantage of non-linear metrics can be observed for both training scenarios.

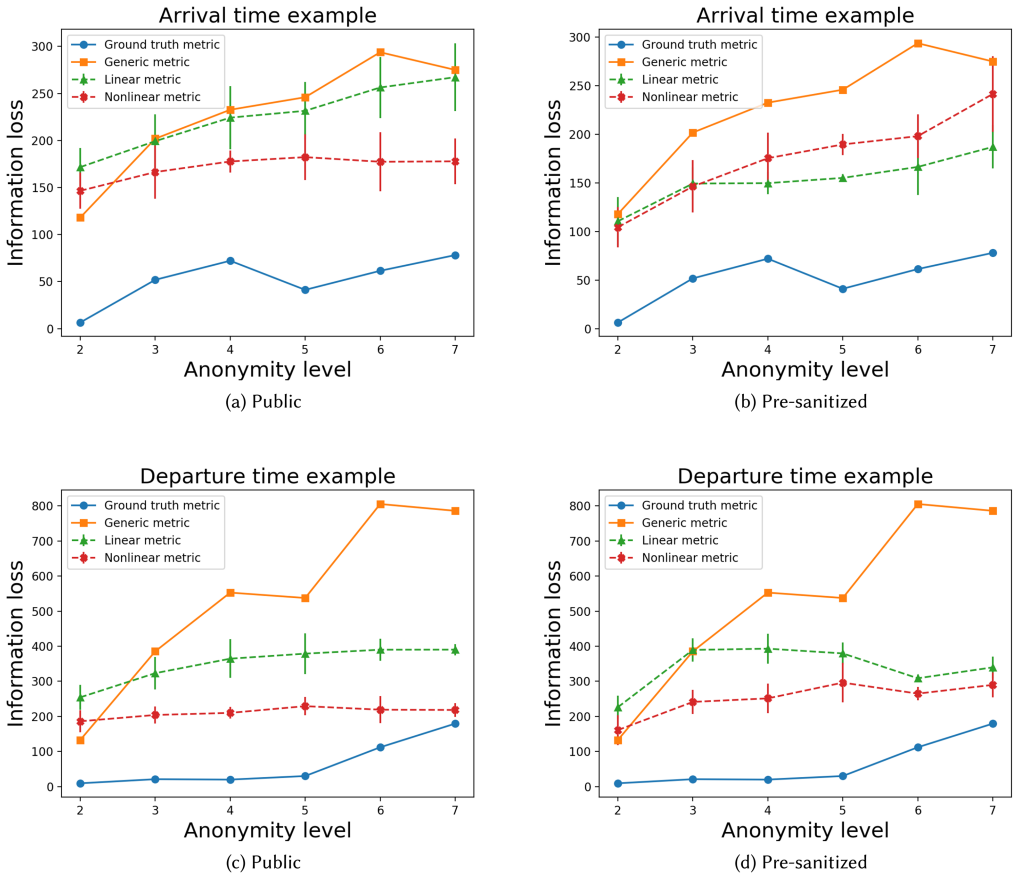


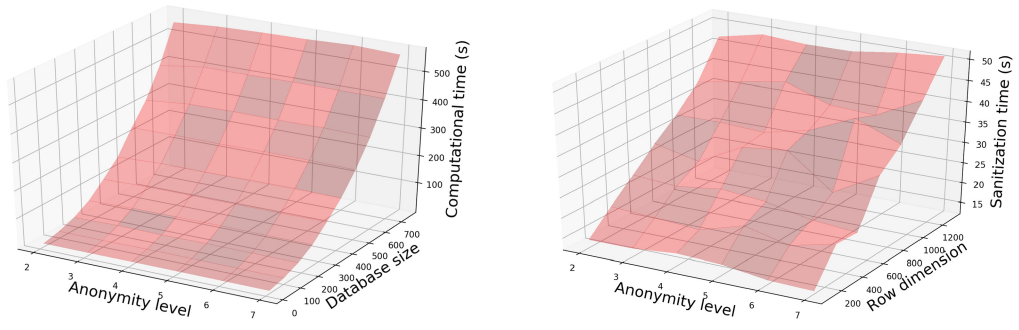
Fig. 11. The tradeoff between anonymity and information loss for data publication specialized for arrival and departure times.

In order to better understand the reason for the performance discrepancy between different distance metrics, we calculate the correlation between the learned distances and the ground truth distances for the four aforementioned use cases, namely, lunch time occupancy pattern, peak hour energy usage, arrival time, and departure time. The correlation is measured in terms of the Pearson correlation coefficient, which has a value between +1 and −1, where 1 is the total positive linear correlation, 0 is no linear correlation, and −1 is the total negative linear correlation. The results are listed in Table 1. The correlation between the generic distances (i.e., Euclidean distance) and the ground truth distances is also listed as a baseline. We can see that when the ground truth metric is nonlinear, the nonlinear metrics can produce distances that have the highest correlation with the ground truth distances. On the other hand, if the ground truth metric is linear, linear metrics have the highest correlation with the ground truth distances. In addition, we can observe that both linear and nonlinear metrics are more indicative of the ground truth, compared to the Euclidean distance. In practice, the data publishing system does not have access to data analysts' interests *a priori*; instead, only a set of data pairs with similarity labels are provided for the system. Therefore, the data publishing system can implement a nonlinear metric (e.g., via neural networks) since it can work sufficiently well for both linear and nonlinear features.

Table 1. Correlation Between the Learned Distances and the Ground Truth Distances for Different Use Cases

Distance metrics	**Use cases**			
	Lunchtime occupancy	Peak hour energy usage	Arrival time	Departure time
Euclidean	0.32	0.64	0.41	0.39
Linear	**0.95**	**0.93**	0.34	0.36
Nonlinear	0.43	0.72	**0.68**	**0.96**

Correlation is measured in terms of Pearson correlation coefficients. The correlation between the generic distance (i.e., euclidean distance) and the ground truth distance is also listed as a baseline. The Pearson correlation coefficients are calculated at anonymity level 4 and averaged over 5 MC simulations.



(a) (b)

Fig. 12. Computational complexity of microaggregation.

7.5 Computational Overhead

We study the computational overhead associated with PAD. We first look into the complexity of the microaggregation part. Let the size of the database be n, the dimension of the row be m, and the anonymity level be k. The microaggregation complexity mainly comprises $O(n^2m)$ computations of distance values and the complexity of the clustering process, which is shown to be $n(1-1/k)$ in the best case and $(n/k-1)(n/2+k-2)$ in the worst case (Domingo-Ferrer 2006). Figure 12 demonstrates the computation time of microaggregation as a function of n, m, and k. We can see that the overhead is approximately quadratic in the database size and linear in the dimension of the row. In addition, changing the anonymity level requirement does not affect the computational time significantly.

The complexity of the deep metric learning step depends on the actual algorithm used for optimization and the convergence criterion. Figure 13 illustrates the relationship between computational time of metric learning and database dimension. `Adam` is used for solving the optimization involved in the deep metric learning. Given a fixed number of epochs (the number of times that the learning algorithm goes through the training data pairs), learning rate, and batch size, computational time associated with the metric learning part increases with the number of labeled data pairs and the dimension of the data records. Moreover, the number of labeled pairs dominates the computational overhead of the metric learning step.

8 FUTURE WORK

For future work, we aim to analyze the tradeoff between the multiple release of private datasets and the inherent privacy vulnerabilities with regard to possible linkage and correlation attacks. This tradeoff is not particular to k-anonymity; indeed, it is a universal challenge for various

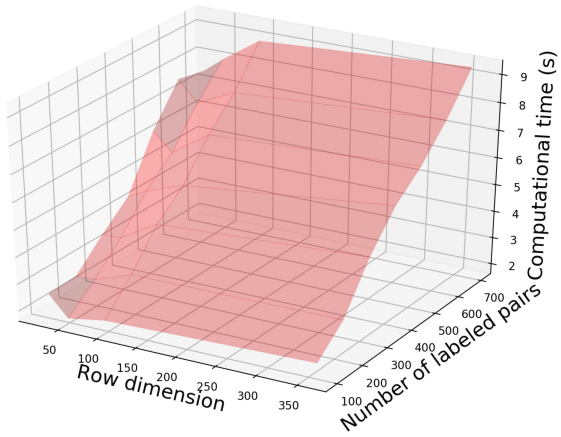


Fig. 13. Computational overhead of the deep metric learning step.

privacy metrics. For instance, differential privacy also suffers from the similar issue. The more queries the dataset answers, the worse the privacy guarantee is for a given utility tolerance (or the worse the data utility is for a given privacy guarantee). Prior work (Dwork 2008) in differential privacy domain has proposed to employ a privacy budget management module that constrains the number of queries allowed for a specific user. Inspired by this, we can also control the number of sanitized versions that a data analyst can process in order to prevent re-identification. This approach requires comprehensive modeling and analysis of the privacy leakage for every sanitized version of a single database. Another potential solution is to introduce uncertainties into the published data (Domingo-Ferrer and Soria-Comas 2016; Holohan et al. 2017). For instance, Facebook has adopted a similar method to answer the queries on its advertisement platform (Venkatadri et al. 2018). Typical methods to introduce uncertainties include rounding the entries in the dataset, adding noise, and the like.

9 CONCLUSION

In this article, we present an open-sourced data publication system, PAD, for protecting k-anonymity of time series data collected in buildings. Particularly, PAD can achieve better data utility than traditional anonymization techniques. This feat is achieved by customizing the data privatizing process to the potential data use. In order to tackle the scalability issues with hard-coding different data uses and their corresponding optimized anonymization procedures, we propose a simple protocol for data users to convey their diverse interests, i.e., the system provides a batch of data pairs and the analyst labels the similarity of each data pair according to their interests. PAD can then learn a more context-aware distance metric from the labeled data. We show through extensive experiments on real-world datasets that PAD can better preserve the usefulness of the published data while proving privacy protection under a variety of use cases. By proposing PAD, we hope to revolutionize the way that CPSs' datasets are published.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg,

Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Retrieved from https://www.tensorflow.org/ Software available from tensorflow.org.

Bharathan Balaji. 2015. Zodiac dataset publication agreement. Retrieved June 15, 2017 from http://www.synergylabs.org/bharath/datasets.html.

Vincent Bindschaedler, Reza Shokri, and Carl A. Gunter. 2017. Plausible deniability for privacy-preserving data synthesis. *Proceedings of the VLDB Endowment* 10, 5 (2017), 481–492.

Andy Bloxham. 2011. Most burglars using Facebook and Twitter to target victims, survey suggests. Retrieved September 26, 2011 from http://www.telegraph.co.uk/technology/news/8789538/Most-burglars-using-Facebook-and-Twitter-to-target-victims-survey-suggests.html.

François Chollet and others, 2015. Keras. Retrieved from https://github.com/keras-team/keras.

Josep Domingo-Ferrer. 2006. Microaggregation for database and location privacy. In *International Workshop on Next Generation Information Technologies and Systems*. Springer, 106–116.

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

J. Domingo-Ferrer and J. M. Mateo-Sanz. 2002b. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering* 14, 1 (Jan. 2002), 189–201. DOI: http://dx.doi.org/10.1109/69.979982

Josep Domingo-Ferrer and Jordi Soria-Comas. 2016. Anonymization in the time of big data. In *Privacy in Statistical Databases*, Josep Domingo-Ferrer and Mirjana Pejić-Bach (Eds.). Springer International Publishing, Cham, 57–68.

Simona D'Oca and Tianzhen Hong. 2015. Occupancy schedules learning process through a data mining framework. *Energy and Buildings* 88 (2015), 395–408.

Flávio du Pin Calmon and Nadia Fawaz. 2012. Privacy against statistical inference. In *Proceedings of the 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 1401–1408.

Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*. Springer, 1–19.

Elevate Energy. 2013. Aggregated Data Access: The 15/15 Rule in Illinois and Beyond. Retrieved June 15, 2017 from http://www.elevateenergy.org/wp/wp-content/uploads/1515-Rule-Factsheet-FINAL.pdf.

Khaled El Emam and Cecilia Álvarez. 2014. A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques. *International Data Privacy Law* 5, 1 (2014), 73–87.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1054–1067.

European Commission. 2012. Protection of personal data. Retrieved January 13, 2017 from http://ec.europa.eu/justice/data-protection/.

Benjamin Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)* 42, 4 (2010), 14.

Aris Gkoulalas-Divanis, Panos Kalnis, and Vassilios S. Verykios. 2010. Providing k-anonymity in location based services. *ACM SIGKDD Explorations Newsletter* 12, 1 (2010), 3–10.

Marco Gruteser and Dirk Grunwald. 2003. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services*. ACM, 31–42.

Mehreen S. Gul and Sandhya Patidar. 2015. Understanding the energy consumption and occupancy of a multi-purpose academic building. *Energy and Buildings* 87 (2015), 155–165.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2. IEEE, 1735–1742.

Naoise Holohan, Spiros Antonatos, Stefano Braghin, and Pól Mac Aonghusa. 2017. (k, ϵ)-Anonymity: k-Anonymity with ϵ-differential privacy. *arXiv Preprint arXiv:1710.01615* (2017).

Tsan-sheng Hsu, Churn-Jung Liau, and Da-Wei Wang. 2014. A logical framework for privacy-preserving social network publication. *Journal of Applied Logic* 12, 2 (2014), 151–174.

Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2014. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1875–1882.

Ruoxi Jia, Fisayo Caleb Sangogboye, Tianzhen Hong, Costas Spanos, and Mikkel Baun Kjærgaard. 2017a. PAD: Protecting anonymity in publishing building related datasets. In *Proceedings of the 4th ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM.

Ruoxi Jia, Roy Dong, S. Shankar Sastry, and Costas J. Spanos. 2017b. Privacy-enhanced architecture for occupancy-based HVAC control. In *Proceedings of the 8th International Conference on Cyber-Physical Systems*. ACM, 177–186.

Ruoxi Jia and Costas Spanos. 2017. Occupancy modelling in shared spaces of buildings: A queueing approach. *Journal of Building Performance Simulation* 10, 4 (2017), 406–421.

Ming Jin, Ruoxi Jia, Zhaoyi Kang, Ioannis C. Konstantakopoulos, and Costas J. Spanos. 2014. Presencesense: Zero-training algorithm for individual presence detection based on power monitoring. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM, 1–10.

Ming Jin, Ruoxi Jia, and Costas Spanos. 2017. Virtual occupancy sensing: Using smart meters to indicate your presence. *IEEE Transactions on Mobile Computing* 16, 11 (2017), 3264–3277.

Eoghan McKenna, Ian Richardson, and Murray Thomson. 2012. Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy* 41 (2012), 807–814.

Andrés Molina-Markham, Prashant Shenoy, Kevin Fu, Emmanuel Cecchet, and David Irwin. 2010. Private memoirs of a smart meter. In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*. ACM, 61–66.

S. Raj Rajagopalan, Lalitha Sankar, Soheil Mohajer, and H. Vincent Poor. 2011. Smart meter privacy: A utility-privacy framework. In *Proceedings of the 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 190–195.

Fisayo Caleb Sangogboye, Krzysztof Arendt, Ashok Singh, Christian T. Veje, Mikkel Baun Kjærgaard, and Bo Nørregaard Jørgensen. 2017. Performance comparison of occupancy count estimation and prediction with common versus dedicated sensors for building model predictive control. *Building Simulation* 10, 6 (Dec. 2017), 829–843. DOI: http://dx.doi.org/10.1007/s12273-017-0397-5

Lalitha Sankar, S. Raj Rajagopalan, and H. Vincent Poor. 2013. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security* 8, 6 (2013), 838–852.

Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.

Ivor W. Tsang, James T. Kwok, C. Bay, and H. Kong. 2003. Distance metric learning with kernels. In *Proceedings of the International Conference on Artificial Neural Networks*. 126–129.

Giridhari Venkatadri, Athanasios Andreou, Yabing Liu, Alan Mislove, Krishna P. Gummadi, Patrick Loiseau, and Oana Goga. 2018. Privacy Risks with Facebooks PII-based Targeting: Auditing a Data Brokers Advertising Interface.

Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. 2006. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*. 1473–1480.

Eric P. Xing, Michael I. Jordan, Stuart J. Russell, and Andrew Y. Ng. 2003. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*. 521–528.

Dit-Yan Yeung and Hong Chang. 2007. A kernel approach for semisupervised metric learning. *IEEE Transactions on Neural Networks* 18, 1 (2007), 141–149.

Received January 2018; revised April 2018; accepted September 2018