Title

A review of preserving privacy in data collected from buildings with differential privacy

Authors

Janghyun, K

Barry, H

Tianzhen, H

et al.

Publication Date

2022-09-01

DOI

10.1016/j.jobe.2022.104724

A Review of Preserving Privacy in Data Collected from Buildings with Differential Privacy

Janghyun K.*[a], Barry H.[b], Tianzhen H.[c], and Marc A.P.[d]

[a]National Renewable Energy Laboratory, Golden, Colorado, USA, janghyun.kim@nrel.gov
[b]San Francisco Department of the Environment, San Francisco, California, USA, barry.e.hooper@sfgov.org
[c]Lawrence Berkeley National Laboratory, Berkeley, California, USA, thong@lbl.gov
[d]Recurve Analytics, Mill Valley, California, USA, marc@recurve.com

February 2022

Abstract

Significant amounts of data are collected in buildings. While these data have great potential for maximizing the energy efficiency of buildings in general, only a small portion of the data are accessible to researchers, government, and industry for analyses. Concerns about privacy are one of the major barriers prohibiting access to these data. Privacy preservation techniques are generally applied to this problem not only to preserve underlying privacy but also to improve the usefulness of data. Among various privacy preserving techniques, differential privacy has become one of the more popular solutions since its introduction in 2006. Differential privacy is a mathematical measure for protecting privacy so that one's privacy cannot be incurred by participating in a database. Although significant research improvements have been made for more than a decade, applying differential privacy to data collected in buildings is still an immature field of study. This literature review aims to introduce what has been done to implement differential privacy in data collected in buildings, and to discuss associated challenges and potential future research opportunities.

Keywords: differential privacy, building, meter, data
Word count: 13,418

1 Introduction

The residential and commercial buildings sector accounted for 20% of global energy consumption in 2018 [1], and a much higher 39% in the United States in 2019 [2]. Many research efforts are focused on reducing the energy consumption, increasing the energy efficiency of buildings, and reducing carbon emission. Buildings can also provide services to utilities to enable deeper penetration of renewable energy on the grid. One of the major pathways to achieve these goals is to first understand the operational performance of buildings with collected data. The performance reflected in these data either informs the reality of existing buildings, reflects the effects of short- or long-term events and improvements in buildings, or aids the development of innovative approaches for maximizing building energy efficiency. The focus of this literature review aligns with the last use case, where extracting insights from the data greatly benefits further research.

The big data era saw a massive increase in data collection as well as unprecedented privacy threats. Various approaches (e.g., encryption, anonymization, differential privacy) have been studied and applied to achieve a trade-off between preserving privacy and extracting accurate insights from the data. Lane [3], Hayashi [4], and Fang et al. [5] describe the importance of ensuring privacy while handling and extracting insights from large data sets (data that is broader than data collected in buildings). While Lane [3] emphasized the new demand created by the big data era to both disseminate data and protect privacy and confidentiality of data, Hayashi [4] presented and compared actual privacy approaches between the United States and the United Kingdom, focusing on possible policy considerations, an area that needs significant attention to realize the implementation of privacy preserving techniques into the marketplace. Fang et al. [5] conducted a survey on privacy preserving techniques for big data and also considered legal measures and industry specifications.

Regarding data specifically collected for buildings, Chau and Little [6], Pillitteri and Brewer [7], Finster and Baumgart [8], Begum and Nausheen [9], Asghar et al. [10], Desai et al. [11], and Sookhak et al. [12] discussed privacy in different levels of smart applications, such as smart spaces, smart buildings, and smart cities. Chau and Little [6] analyzed limitations of existing privacy preserving approaches for smart spaces and proposed adaptations to ensure strong privacy preservation. Pillitteri and Brewer [7] developed guidelines for smart grid cybersecurity to enable relevant organizations to effectively construct their cybersecurity strategies around the smart grid. Their study also covered the privacy aspect of the data in smart grids by assessing the privacy impact, discussing mitigating factors, and identifying potential privacy issues. Finster and Baumgart [8] conducted a survey on privacy preserving approaches and solutions in smart grids and classified the problem into two areas: problems in energy metering for billing and metering for operations. Asghar et al. [10] also reviewed privacy preserving approaches for smart meter data in three application

*corresponding author

areas (billing, operation, value-added service) and covered various use cases of smart meter data and related privacy legislation highlighting shortcomings, recommendations, and future research directions. Desai et al. [11] reviewed privacy preserving approaches for smart meter data and presented solutions for detailed privacy problems. Sookhak et al. [12] surveyed security and privacy issues in smart cities and presented a thematic taxonomy of the issues to support the security design of smart cities.

While these studies have clearly touched upon important aspects of privacy as it relates to smart meter data, other studies take different approaches. Ruddell et al. [13] emphasized the usefulness of smart meter data (collected by utility companies) and how these data can result in public benefit by analyzing the trade-off between usability and privacy. The study also noted that the strict privacy rule applied to California utilities (in the United States) does not meet the needs of compelling public interest and suggested a relaxed privacy rule based on their statistical analysis. Schwee et al. [14] developed a tool for assessing privacy risk as part of the data sharing process. The tool creates a report that presents potential risks associated with each data type. The tool was tested on real-world building data sets that include state measurements of indoor air temperature, humidity, CO_2, illuminance, weather, noise, pressure, etc. at various time intervals (10 seconds to 20 minutes).

It is also worth noting what privacy means in this field of study because the definition of privacy has evolved over the years. An early definition of privacy started with "the right to be left alone" [15]; however, social situations have evolved since then, as mentioned by Hayashi [4] who notes "self-determination such as sexual orientation or contraception is admitted as one of the privacy elements." Similar evolution also occurred and continues to occur in data collected from buildings. Granular data, such as smart meter data collected from advanced metering infrastructure (AMI) in sub-hourly (e.g., down to 15 minutes) intervals, has become available (owned by many utility companies around the world), but research has also shown that these data include various types of private information given the strong correlation of occupants (or building operators) with building energy consumption. Thus, there is a need to define privacy, especially for data collected in buildings. While Warren and Brandeis [15], Prosser [16], and Clark [17] defined privacy in general, Hayashi [4], Begum and Nausheen [9], and Jain et al. [18] considered privacy in terms of big data, and Pillitteri and Brewer [7] specifically focused on the context of smart grids. Pillitteri and Brewer [7] translated the privacy classifications defined by Clark [17] in the context of the smart grid application as shown in Table 1. This literature review follows the definition of privacy depicted by Pillitteri and Brewer [7] and reviews previous studies to better understand how private information can be obfuscated with differential privacy.

Table 1: Definition of privacy by Pillitteri and Brewer [7]

"There is no one universal, internationally accepted definition of privacy, it can mean many things to different individuals. ... Privacy is not a plainly delineated concept and is not simply the specifications provided within laws and regulations. Furthermore, privacy should not be confused, as it often is, with being the same as confidentiality; and personal information is not the same as confidential information. Confidential information is information for which access should be limited to only those with a business need to know and that could result in compromise to a system, data, application, or other business function if inappropriately shared. It is important to understand that privacy considerations with respect to the Smart Grid include examining the rights, values, and interests of individuals; it involves the related characteristics, descriptive information and labels, activities, and opinions of individuals, to name just a few applicable considerations."

The main objective of this paper is to open the door for building researchers to introduce differential privacy and understand how the applications are implemented around the data that building researchers are interested in. To achieve this goal, this paper reviews previous studies that focus on preserving privacy in data collected in buildings with differential privacy. The remainder of this article is organized as follows: Section 2 describes the specific scope considered in this review, Section 3 introduces privacy risks associated with building data, Section 4 provides use cases that leverage insights extracted from the building data, Section 5 reviews all studies in terms of various configurations of differential privacy, Section 6 discusses research gaps summarized from the review, and Section 7 provides concluding remarks.

2 Scope of the literature review

This section presents specific scope and definitions used in the literature review. The detailed methodology for our literature review and metadata analysis based on the gathered literature are provided in the Supplementary Material.

What type of data is this article interested in? While the data is a key starting element of applying differential privacy, differential privacy can be applied to any type of data (e.g., numeric, categorical, timeseries). While various data can be collected in buildings, the phrase "building data" as used in this literature review covers data listed in Table 2. Note that this is a relatively extensive list and not all data types were found in reviewed studies.

Table 2: Types of building data considered in the literature review

Data Type	Data Description	Typical Data Owner
smart meter	electricity, gas, water, on-site power generation (e.g., photovoltaic [PV]) data	utility companies

Table 2 continued from previous page

Data Type	Data Description	Typical Data Owner
submeter data	data measuring building sub-system totals: HVAC, lighting, plug load, or water heating, etc.	homeowner or building operator (typically tracked in BAS)
state measurement data	data measuring the state of building systems: temperature, pressure, airflow, occupancy, control signal, etc.	homeowner or building operator (typically tracked in BAS)
smart home device data	data measured with smart Internet of Things (IoT) devices: smart thermostat, occupant health monitor, etc.	homeowner

What type of privacy preserving technique is this article interested in? While there are different approaches available, such as anonymization, encryption, or differential privacy, this review focuses on the approach of differential privacy. Differential privacy was first introduced by Dwork [19]. To provide additional context, Figure 1 provides mathematical and narrative definitions of differential privacy. The contextual and formal definition of differential privacy reflected in Figure 1 can also be described as follows: "Differential privacy describes a promise, made by a data holder, or curator, to a data subject: You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available. At their best, differentially private database mechanisms can make confidential data widely available for accurate data analysis, without resorting to data clean rooms, data usage agreements, data protection plans, or restricted views [20]." The ϵ in the equation represents the privacy loss, privacy parameter, or privacy budget. Theoretically, it is the maximum distance between the same query on database D_1 and D_2. It is the key parameter in differential privacy implementation, where the selection of ϵ will determine the trade-off between privacy protection (e.g., increased protection with smaller ϵ) and usability of the data (e.g., less usability with smaller ϵ).

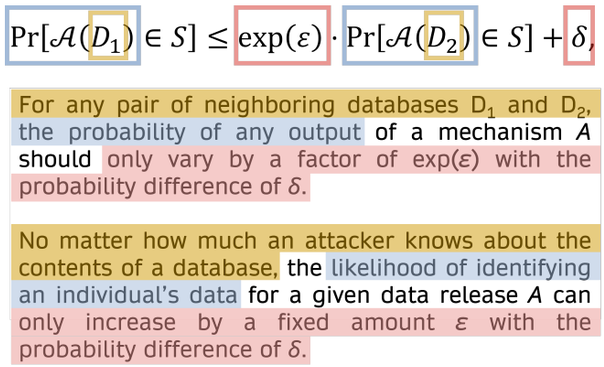

Figure 1: Mathematical and narrative definitions of differential privacy

Unlike other anonymization methods, differential privacy provides a mathematically rigorous definition of privacy. This definition of privacy is implemented in a wide variety of "mechanisms" suitable for different analytical workloads. The guarantee of differential privacy is composable, allowing data owners to bound privacy risk across multiple data releases, even for very different statistics. Appealingly, this guarantee holds no matter how much additional information is released about individuals in a data set. Finally, differential privacy can be realized in a number of different architectures depending on the threat model for a given data set.

There are additional review studies that focus on differential privacy but with relatively broader scope compared to this literature review. These review studies provide useful theories, concepts, configurations, and general perspectives of differential privacy regardless of the type of data. Desfontaines and Pejó [21] emphasized that approximately 200 different notions that are either extensions or variants of differential privacy have been introduced since 2006. The evolution of differential privacy definitions necessitated a proper classification, and the study proposed a systematic taxonomy that is unified, comprehensive, and flexible enough to categorize numerous types of differential privacy definitions. This study provides key insights on what type of variations can happen in differential privacy implementations in general; however, it can be difficult for an early practitioner to digest all dimensions and variants in the context of the building data. Hassan et al. [22] relatively distilled the scope and surveyed differential privacy applications around cyber physical systems (CPS). The term CPS represents any computer system that includes a mechanism of monitoring and/or controlling by a computer-based algorithm and is connected to an internet where data can be transferred to other CPSs. Application areas such as energy systems (e.g., smart grid), transportation systems, healthcare and medical systems, and the industrial internet of things are covered in the study. The study provides a very comprehensive review of many previous studies, classifying them into different applications and various configurations of differential privacy. While there is some overlap in scope between Hassan et al. [22] (via applications in smart grid) and this literature review, the review needs more distilling for a building researcher to understand all possible variations of differential privacy around building data. For this reason, the purpose of this literature review is to provide a gateway for differential privacy applications and to inform researchers who are interested in data collected in buildings.

3 Review of Privacy Risks

Before examining individual studies that have applied differential privacy in building data, readers should be aware of the private information that these data might contain. Privacy risks are summarized from the reviewed literature categorizing risks into 1) research studies that revealed private information from real data, 2) theoretical privacy risks mentioned in studies, and 3) privacy risks in real life.

3.1 Risks revealed from studies (from data not related to buildings)

The first examples include general (and popular) context that revealed private information based on the public data that are not related to buildings. In the 1990s, the Massachusetts Group Insurance Commission in the United States decided to release anonymized health information of state employees to help researchers. While the governor assured the public that the data would be anonymized enough to protect underlying privacy, Sweeney [23] combined the anonymized health data with additional information such as proximity of the governor's residence (i.e., Cambridge, Massachusetts) and a $20 database including the city of Cambridge voters' information (e.g., name, address, ZIP code) and was able to narrow down the anonymized health information specific to the governor. In 2006, Netflix released an anonymized data set including movie rankings of 500,000 customers. The purpose of this release was to hold an open competition where competitors could use this data to create better movie recommendation algorithms. Knowing from past experience that anonymization does not always guarantee the protection of underlying privacy in data, Narayanan and Shmatikov [24] leveraged public data from the Internet Movie Database (IMDb) to reidentify the anonymized customer and his or her movie rankings. This reidentification of anonymized data not only revealed movie rankings of a certain customer but can also reveal a customer's political preferences, which can easily be deemed private. The last example is Homer et al. [25], which showed that a person's participation in a genome study can be inferred using the publicly released genome-wide association study statistics. Once it is verified that a person has participated in the genome-wide association study, the genotype of the person including certain risk factors related to their health can also be revealed. These examples not only provide real-life privacy risks in other data types but also demonstrate the weakness of anonymized data, where differential privacy can play a role to better preserve privacy.

3.2 Risks revealed from studies (from data collected in buildings)

The second examples include similar examples but for risks related to data collected in buildings. These examples mainly consider private information extracted from electricity consumption data (i.e., meter data) and expressing concerns of privacy threats. The non-intrusive load monitoring (NILM) method was first introduced by Hart [26] in 1992, where the method infers usage of various home appliances by analyzing and disaggregating household-level current and voltage measurements. The original purpose of this method was to understand a household's home appliance usage and to inform utility companies about a customer managing use cases. However, the usage profiles of home appliances can reveal private lifestyles of the occupants, such as whether an occupant is in the shower or not. For this reason, the NILM technique has also been used as an adversary (attack) model for evaluating the performance of privacy preserving techniques.

The emergence of smart grids associated with smart meters in AMI led researchers to focus more on privacy in smart meter data because data sharing via smart meters in the smart grid is not only required for efficiently managing the smart grid, but also exposes data relatively more to the outer world than the conventional automatic meter reading (AMR) infrastructure. Lisovich et al. [27], Berenguer et al. [28], and Molina-Markam et al. [29] specifically focused on this concern by conducting NILM attacks on fine-grained real smart meter data. Without a priori knowledge and leveraging off-the-shelf (or newly developed) statistical methods, the studies revealed either 1) the number of occupants at home, 2) sleeping routines of occupants, 3) eating routines of occupants, and other characteristics based on home appliance usage patterns.

Rouf et al. [30] also emphasized the privacy risks in the existing AMR infrastructure, where there are 40 million meters being used in the United States. Because the existing AMR infrastructure was receiving less attention compared to AMI, the study performed reverse engineering to infer private information from AMR data and proved that the existing infrastructure is also vulnerable to privacy threats. The study conducted an eavesdropping experiment in a neighborhood with AMR infrastructure, and experiments were conducted in a nearby location with a laptop installed with a commercially available low noise amplifier (LNA) to increase the physical range of eavesdropping. Most of the meters in the neighborhood were decoded, and based on the energy consumption readings revealed from these meters, private information such as daily routines of occupants (e.g., household occupancy) and appliance (e.g., water heater, washing machine, stove) usage patterns were also revealed.

Greveler et al. [31] focused on even more granular disaggregation by using relatively more granular metered data (compared to typical smart meter readings) to infer not just TV usage patterns but also display type (e.g., CRT, plasma, or LCD) and TV channels that occupants are watching. The developed algorithm first requires digital movie data as an input and to use as a basis for comparison. The data is then split into 5-minute intervals, and the brightness of each frame in each interval is calculated. Because the brightness (i.e., backlighting) is mainly correlated with the power consumption of the TV, the timeseries TV power consumption can also be calculated. The calculated power consumption can be compared against the TV power consumption inferred from NILM. Once the best match is verified based on the movie data, the 5-minute window can specifically decipher which scene the occupant was watching with his or her TV.

While the examples related to buildings described above mostly focus on meter data as a basis to infer, there are other measurements typically being collected in buildings. Wang and Tague [32] focused on occupancy sensor measurements that can be stored in the building automation system (BAS, or Building

Management System [BMS]), which provides timeseries occupancy status (occupied/unoccupied or even counts of occupants) in each room installed with an occupancy sensor. And these data are considered valuable as an input for optimally controlling the heating, ventilation, and air conditioning (HVAC) system in a building. The study begins with a concern that these data can reveal as much information as looking into a surveillance camera installed in buildings, revealing the whereabouts of occupants. Although typical data measured in BAS stay within BAS and generally never leave the building operator's hands (who are considered trustworthy), another assumption justifying the study's concern is the scenario for smart buildings where data exchange against the outside world (e.g., cloud) is necessary for maximizing the efficiency of buildings in a community. Other than occupancy sensor measurements, the developed algorithm also requires additional contextual information, such as the floor plan of the building and office directories, which are often publicly available. The factorial hidden Markov model (FHMM), which is suitable for transforming the characteristics of occupancy sensor measurements in different locations in a building, was applied in the study to reconstruct and reveal location traces of individuals. The proposed algorithm was evaluated with real and synthetic occupancy sensor data, proving the accuracy of the reconstruction.

3.3 Risks in theory

The third examples include theoretical privacy risks mentioned in studies reviewed in this article. Most of them can be grouped into daily (or even more granular) activities and lifestyles of occupants, not only in residential homes but also in commercial buildings. For example, specific locations of an occupant in a building, general occupancy status (either occupied or unoccupied) of the building, number of occupants in the building, and other characteristics are described as risks that can be revealed from the data collected in buildings.

In addition to the lifestyle of an occupant, two studies (Barbosa et al. [33, 34] and Hassan et al. [22]) also noted that the information of equipment in buildings not operating at the desired efficiency level can be inferred and used for targeted business advertisement. Anecdotally, it is not uncommon to see homeowners being surprised and questioning "How did they know?" in response to advertisement material delivered to their homes. Other private information includes the timings of on-site electricity generation with renewable energy sources (e.g., PV), mentioned mostly by Hassan et al [22, 35]. The information of when and how long these on-site generations occur can be used by a utility to change their dynamic pricing scheme to one unfavorable to the customer; this information can also be sold to commercial parties for targeted advertisements.

Liu et al. [36] and Xu et al. [37], which focused on IoT device data, described a privacy risk in individual's health data monitored in smart home devices. Liu et al. [36] specifically provided examples of Intel's IoT solutions targeting the healthcare industry and Intel and General Electric's QuietCare, where the purpose of these solutions is creating a new paradigm that reduces unnecessary societal costs by leveraging IoT. For example, sensors developed in these solutions can monitor the occupant's health status almost in real time. These data can also be shared with the healthcare providers, where timely instructions and feedback can be provided to minimize emergency visits and healthcare office visits. However, if these data are leaked, they would seriously threaten the privacy of the occupant (or patient) because health information is one of the most highly regarded types of private data in real life.

3.4 Risks in real life

To the best of the authors' knowledge, the building-related privacy risk examples presented previously have never reached the point where the harm resulted in public shaming, monetary expense, or legal proceedings. Thus, it is worth knowing if there was any real-life harm due to the privacy invasion inferred by data collected from buildings. Amador [38] analyzed domestic and international terrorism events that occurred in Germany between 1960 and the 1990s. It described the transformation of radical students to become terrorists and described how the government responded to these terrorism events. The thesis specifically mentioned an example where a member of the Red Army Faction was arrested by utilizing meter data. At the time, it was thought that the terrorists preferred high-rise apartments close to highways and with underground garages. Officials reached out to the local electricity company to verify apartment units with extremely low electricity consumption (because safe houses are only used at certain times) and where payments were made in cash. By narrowing down apartment units that fit these criteria, officials were able to detect the safe house and arrested the member of the Red Army Faction. However, this was later deemed unconstitutional by German courts giving the inference from private billing information.

Several news articles (Guest [39] and Smith [40]) in 2007 reported that the Austin Police Department in Texas in the United States utilized electricity consumption data acquired from Austin Energy (a local utility company) to verify illegal marijuana growing operations. Similar to the high energy use intensity of data centers, marijuana growing facilities are also a huge energy consumer per unit floor area. On one side, targeting narrowed-down addresses for an investigation can save a large amount of police enforcement effort. This case resulted in a legal court debate, and revealed that the Austin Police Department used thousands of Austin Energy customers' data without their consent. Thus, on the other side, thousands of customers' underlying privacy in metered data was at risk. At the time, Austin Energy's decision on sharing the data with the law enforcement was based on a previous ruling from 1994. However, because data collected in buildings is becoming more granular (e.g., smaller measurement intervals) and high-quality (e.g., without erroneous measurements), these data are increasingly being considered private. A similar court case was Naperville Smart Meter Awareness v. City of Naperville [41] in 2018. The city of Naperville in Illinois in the United States received funding from the Department of Energy to upgrade AMR infrastructure into AMI. While other cities had an option for residents to opt out from the smart meter transition, Naperville's residents did not have that option. And because intimate personal information can be inferred from the smart

meter data if the data falls into the hands of adversaries, concerned residents sued the city of Naperville. However, the court decided the following: "Because of the significant government interests in the program, and the diminished privacy interests at stake, the search is reasonable. We therefore AFFIRM the district court's denial of leave to amend."

The last examples are from news articles by Glionna [42] (Las Vegas, Nevada, United States) and Horwath [43] (Santa Fe, New Mexico, United States). These articles reported a list of top residential owners and commercial users who have consumed the most water in the city based on the data provided by the local water utility companies. These articles not only included commercial users but also included residential owners with their names clearly shown. The intention of these articles was not to publicly shame these users because it is also revealed that those large bills can be due to leaky pipes, incorrect billing, or not accounting for the usage of recycled water. However, releasing the exact names of businesses and individuals can easily cause readers to shame the users.

Some readers might argue that many of these examples are based on theory, with only a handful of real-life examples. However, it is worth noting that data collection is still evolving in terms of its quality and quantity and, at the same time, skills of researchers or data scientists for analyzing these data (e.g., machine learning) are also quickly evolving. Thus, before these threats become more definite, the topic of preserving privacy in data collected in buildings should be carefully scrutinized.

4 Review of Use Cases

Applying differential privacy often means implementing additional mathematical equations in an existing query that supports the use case of the analyst. Here, the term query is an algorithm developed by the analyst (or data curator) to extract insights by running the algorithm on data. An analyst is the person or entity who needs insights for their use case. For example, considering the use case of a utility company's monthly billing scheme, the algorithm for calculating the monthly electricity consumption from granular (e.g., in 15 minutes interval) timeseries data can easily be formulated without differential privacy implementation. But if the threat that customers may not trust the data curator (i.e., internal attack) working for the utility is perceived as significant, then the timeseries electricity consumption of individual customers should be obfuscated by adding another layer of mathematical equations with differential privacy before the data gets transmitted to the utility company's database. If the data curator can be trusted, but releasing monthly energy consumption information to the public is the concern (i.e., external attack), then the results of monthly energy consumption should also be obfuscated. For this reason, the implementation of differential privacy is highly dependent on the use case of the analyst or the underlying queries used for various use cases. This section summarizes all use cases identified in the literature (which is also summarized as Table S2 in the Supplementary Material) and also provides other potential use cases.

4.1 Building portfolio level use cases

Energy consumption data of buildings joined or related to other data (e.g., metadata such as building footprint, primary business type) is a very underutilized resource. Utility companies that provide energy (e.g., electricity, natural gas, and/or water) to customers have a relatively strong track record in energy management and energy data analysis because utilities have much less friction limiting their ability to obtain data to conduct such analyses. And as a result of having the means, analyses that serve a utility's business interests are more likely to be well developed. Thus, the description and conceptual framing of many use cases (shown in Table S2) for data privacy centers on (a) applying energy data to meet an objective or serve a business interest of a utility company, (b) meeting moral and/or legal standards for providing privacy to utility customers, and (c) focusing on numerous customers' (portfolio-level) energy consumption rather than focusing on individual building performance.

Aside from use cases, which mostly focus on a utility company's interest, there are also problems and potential solutions from the point of view of other stakeholders such as local governments, owners and their representatives from individual properties to portfolios, real estate investors, academics, and other parties performing research in the public interest. The interests of these non-utility parties may rely on similar or identical analytic methods to those developed to address a utility company's use cases. For example, local governments have interest in planning for reduction and elimination of building operational greenhouse gas (GHG) emissions. To understand which buildings (within the jurisdiction or portfolio) contribute the most toward GHG emissions, an analysis can be performed by joining building energy use by fuel type with spatial and demographic data. This can enable proper planning and targeting of customers for building electrification.

It is not only local governments but also state governments, utility companies, and electrical engineers that have interest in right-sizing electrical infrastructure with thinner tolerance due to more precise safety factors and diversity factors in load estimation. An analysis of this use case can be conducted by identifying and quantifying distribution statistics for different end uses that present identifiable signals by occupancy in usage data. These statistics and a transparent methodology can be provided to officials for drafting National Electrical Code and local policy. This will provide a supplemental and alternative source of data to inform load diversity factors or related calculation methods used by electrical engineers to resize electrical infrastructure. Another type of interest from similar stakeholders is estimating the potential and impact of electrification or implementing best practices for selecting minimum efficiency appliances. This use case can be constrained by the electrical capacity of the local distribution network. Additionally, the use case considers grid harmonization from the point of view of utility companies, but it also considers transparency about potential scale, methods, priorities, and timing for energy system transformation for the other parties. Local governments are also interested in facilitating community input as part of a master plan for developing new

or replacement utility infrastructure and decommissioning existing utility infrastructure (i.e., decarbonization entailing electric system upgrades and "pruning" of gas infrastructure, or development of community microgrids enabled by concentrated deployment of utility-controllable distributed energy resources). The analysis should be based on where (specific location) local government, stakeholders, and/or utility should prioritize investment to advance public safety, racial equity, and public health by facilitating concentrated transition of energy systems.

4.2 Individual building level use cases

As shown in Table S2, many studies have also discussed use cases that focus on performance within a building. Two data types are covered in this scope: 1) IoT devices (and their measurements) in smart-home, -city, -community, and -grid scenarios where these devices have capability of communicating with the cloud and 2) BAS data that collects various state measurements (e.g., temperature, flow, control signal) as well as energy consumption information for the purpose of automating and efficiently controlling sub-systems (e.g., HVAC, lighting) in buildings. Hassan et al. [22], Jelasity and Birman [44], Liu et al. [36], and Pappachan et al. [45] described a use case of monitoring the quality and structural health of each smart meter device and their target to maintain the quality service of IoT devices. Pappachan et al. [45] also mentioned that "information captured about the building and its inhabitants will aid in development of services that improve productivity, comfort, social interactions, safety, energy savings and more." However, because the useful "information captured about the building and its habitants" can also be private, these studies are making an effort to develop privacy preserving techniques while providing valuable services.

According to the Commercial Buildings Energy Consumption Survey (CBECS, [**?**]), which provides results of a bottom up approach (via statistical sampling) for estimating nationwide energy consumption characteristics of commercial buildings in the United States, it is estimated that 36% (780,662 buildings covering 3.4 billion m^2 of floor area) of commercial buildings in the United States have BAS integrated in their buildings. The comprehensiveness level of BAS can vary widely by product, but it typically measures various state measurements such as temperature in rooms (i.e., thermal zones) and temperature, pressure, airflow, and control signals around the HVAC system that are necessary for automating and controlling the HVAC system to deliver comfort to occupants while maximizing the efficiency of the system. Some extended capabilities of BAS can include measurements around the lighting system. For example, by measuring the indoor/outdoor brightness levels and calculating the solar angle based on the location of the building and time, it is possible to automate the control of motorized roller shades on the facade and lighting power output (i.e., dimming) depending on the available daylight that penetrates into the rooms throughout the day. Hassan et al. [22], Sookhak et al. [12], Jia et al. [46], Chau and Little [6], Xiao et al. [47], and Ny and Mohammady [48] described or focused on the use case of improving HVAC and/or lighting systems control by using the data measured in BAS. Similar to value-added services that the utility company can offer to customers based on portfolio-level measurements, energy efficiency providers can also take advantage of the BAS data for optimizing service offerings with data-driven applications for building owners and operators.

There are certainly many more use cases that are not highlighted in the literature that may need a proper obfuscation because of the underlying privacy included in the data. One of the popular research topics around buildings is a data-driven analysis leveraging machine learning techniques. Today, this field of study is an enormous research area in which studies are expanding in both quantity and quality every day. While reviewing all use cases in this field of study is not the scope of this paper, some of the recent review articles that focus on several major use cases can represent the current trends and reviews. Similar to the use case of load forecasting in the grid level, individual building level load prediction with machine learning techniques is also a popular research topic as it was reviewed by Zhang et al. [49]. Mirnaghi and Haghighat [50] conducted a review of how data-driven methods can be used for the automated fault detection and diagnostics (AFDD). Data-driven AFDD is, in short, analyzing various state and energy consumption measurements in buildings to detect or classify the signatures in data that correspond to certain faults (e.g., condenser fouling, duct leaking) in buildings. Wang and Hong [51] focused on the review of improved building control realized by machine learning techniques. While the previous three studies focused on the performance of buildings, Hong et al. [52] expanded the review scope of machine learning applications to the building's entire life cycle, including design, construction, commissioning, operation, maintenance, control, and retrofit.

5 Review of Differential Privacy Implementation Approaches

This section reviews differential privacy implementation approaches around building data. Figure 2 presents various simplified implementations of differential privacy summarized from the literature, including approach classification, required key components/actors from collecting data to extracting insights from the data, the location of the data obfuscation, and relative characteristics between approaches.

As shown in Figure 2, a local model adds noise on end use devices (e.g., at the meter) before the individual data is combined (or processed) with any other data in a database. Conversely, a global model is distinctive compared to the local model because the differential privacy filter is applied after all data are collected in the database. Compared to the global model, local models have larger noise once they are aggregated into various grouped levels (e.g., utility customers by different building types) for a given level of accuracy. This also means that in order to achieve a certain level of accuracy for extracting insights from very noisy aggregated consumption data, the number of samples (e.g., meters, buildings, or homes) should also be large enough to compensate the noise. The global model has merits in this perspective where the noise is added once for a certain query after numerous meter data is gathered in the database, and the configuration of differential privacy can be tailored for each and every query for maximizing the accuracy of insights while preserving privacy. However, when (and if) catastrophic failure happens with differential privacy, the global

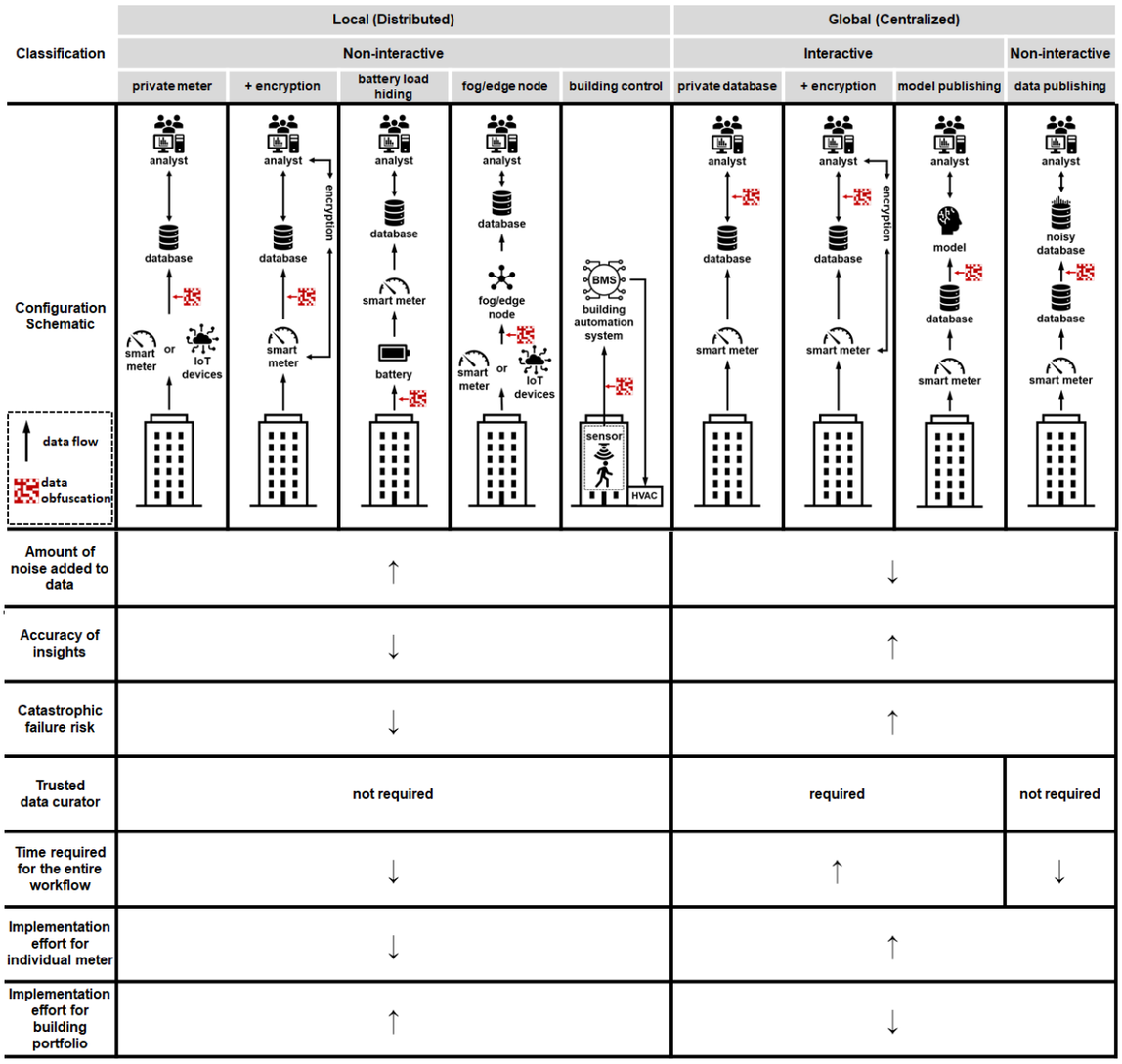

Figure 2: Different configurations of differential privacy implemented for building data

models have more risk compared to the local model because the failure can reveal lots of data that are placed in one database while several privacy preserving failures in local models will still ensure safety for the other majority of data.

From the data query perspective for extracting insights from the data, the local model can provide a privacy preserved database that includes noisy meter data while the global model provides a database with raw meter data that contains the private information of the building or occupants. For this reason, the global model requires a trusted data curator who has access to the raw data but also has responsibility for preserving privacy while extracting insights from the raw data. The global models in the literature are either interactive or non-interactive. The interactive global model requires the data curator to review and respond to individual queries, requiring independent interaction for each query type. This can result in additional processing time (if the new query has never been reviewed before) for the entire querying process compared to the non-interactive models where any query can be made on the noisy database. Additionally, the data curator can also be a risk over time. For a local model, because the data is obfuscated before any data curators look into the raw data, this configuration does not require an additional human resource. For the same aspect of not involving the data curator and specific interaction for individual queries, the local model is also represented as a non-interactive model.

5.1 Local (distributed) and non-interactive models

The literature review found a total of 35 studies applying "local" differential privacy on data collected in buildings, including five major variants within the local (and non-interactive) model in terms of the application as shown in Figure 2: simple private meter, private meter with differential privacy combined with encryption, battery load hiding, fog/edge node, and building control applications. These applications differ slightly in terms of components that constitute the entire workflow between measuring the data, obfuscating the data, and querying the data (by the analyst).

Configurations and mechanisms varied across these studies, as well as the type of queries analysts were interested in. While ϵ-differential privacy was adopted in 19 studies, (ϵ, δ)-differential privacy was adopted in 6 studies. In order to evaluate the performance of privacy preserving techniques proposed from these studies, many studies reviewed in this literature review shared a similar study component by structuring attack (or adversary or threat) models and considering internal (e.g., honest-but-curious) and/or external (e.g., malicious eavesdropper) attacks. The honest-but-curious adversary in this context means the actors

(i.e., internal attackers including data curator and analyst) involved in the algorithm are "honest" enough not to tamper with the algorithm of differential privacy protocols but are "curious" enough to seek and infer insights from data that they are handling. Most of the studies selected one of the distributions (Laplace, Gaussian, Uniform, and Geometric) for the noise addition mechanism while only one of the studies explored and quantified the performance of different differential privacy mechanisms testing all distribution types [53].

The most common query (or use case) type was calculating aggregated consumption of numerous meter data, which is important in terms of understanding the magnitude and profile of energy consumption on the grid level. Meter data aggregation (across the time horizon and/or across different meters) is one of the popular queries that the utility company can mainly use for customer billing, creating strategies for mitigating peak demand, and/or creating cost-effective plans for balancing the supply and demand. Other queries included 1) perturbating individual timeseries metered profile [54–58], 2) spectral analysis [59, 60], and 3) state estimation in distribution network [61].

5.1.1 Private meter including differential privacy combined with encryption

The applications of private meter and differentially private meter combined with encryption shown in Figure 2 include a meter (or IoT device) connected to a building, database that includes data collected from multiple meters, and analyst that conducts a query on the database to extract insights. The data obfuscation with the differential privacy happens before the meter data are transferred to the database; thus, the database includes noisy (or differentially private) data compared to the actual meter readings. The 27 studies that focused on these applications are the most common type of studies among those reviewed in this study [34, 35, 44, 53–76].

Determining the value of ϵ is one of the big questions in differential privacy implementation, and studies in this review mostly tested the performance with either a fixed (and arbitrary) value of ϵ or tested the variation of performance across a range of ϵ values (where the ϵ value varied from 0.001 to 2.5 within 25 studies). Because the implementation of the differential privacy algorithm is highly dependent on a specific use case, analyzing the implementation against a use case can provide relatively more physical meanings of ϵ [34, 65, 68, 70]. For example, Yang et al. [68] focused on the use case of a utility company's optimal power flow to minimize the total generation cost by applying (ϵ, δ)-differential privacy. The use case of optimal power flow aims to minimize the total generation cost of the utility company and constructs the optimal schedules for power generators. While the optimal power flow is a complicated problem that is constrained by various parameters (e.g., to meet the generation capacity and also to not violate lower/upper bounds of generation capacity), it not only serves to minimize the generation cost but also supports other use cases such as real-time pricing for customers. The study constructed a workflow 1) to create a differential privacy algorithm for obfuscating local smart meter data (and for a group of meters), 2) inform an optimal power flow algorithm to create generation planning, and 3) calculate the additional generation cost associated with differentially private input. The study also considered the impact where the change in the power supply on the utility side can affect the cost of electricity on the customer side by considering a locational marginal price scenario. In this context, the study actually connects the relationship between how much customers want in terms of privacy level (via ϵ) to how much they actually have to pay additionally (in their electric bill) because of the elevated privacy preservation guarantee.

Eibl and Engel [67] pointed out that the global sensitivity calculated from the actual data can be misinformed by the nature of malfunctioning smart meters (e.g., transmitting unreasonably high value) in the real world, resulting in too much noise in aggregated results (i.e., decreased usability). Thus, combining encryption techniques with differential privacy has been studied [62–64, 69], which supports the identification of malfunctioning smart meters and removes the necessity of trusted data curator. Barthe et al. [63] considered a realistic scenario by counting the number of meters that are malfunctioning or non-responding to disregard those meters and to provide the correct customer base for the total aggregation of energy consumption. The study combined differential privacy with an encryption technique to share encrypted keys between an analyst and smart meter that supports identifying malfunctioning meters. Because the encryption technique requires relatively more computational and communication overhead (compared to differential privacy), most of the other studies that adopted encryption techniques quantified the overhead with experiments to measure the weight of the proposed algorithm and to see if the proposed algorithm is feasible for practical implementation. Ács and Castelluccia [62] also assumed the smart meters to be trusted devices (i.e., tamper resistant) that can store encryption keys and perform encryption and differential privacy computations. Based on this assumption, encryption techniques are introduced and combined with differential privacy to include the capability of identifying malfunctioning smart meters as well as to obfuscate the individual smart meter readings before they are transmitted to the database. Unlike most other studies, this study considers internal attackers to be dishonest-but-non-intrusive, where the adversary may not follow the differential privacy protocol correctly and is allowed to provide false information to manipulate the collected data. Bao and Lu [64] also emphasized a feature of handling faults (e.g., malfunctioning) in smart meters while differentially privatizing the meter data. The study combined differential privacy with an encryption technique by proposing a novel key management technique that is used for handling the fault in smart meters.

Some of the studies [35, 77] highlighted specific scenarios between actors (e.g., data owner, data analyst, and smart meter customer) showing how the data access framework policy can be constructed around the differential privacy implementation. Lou et al. [77] focused on the use case of a utility company's economic dispatch control that plans the active power outputs of generators to meet demand at the lowest total generation cost subject to various transmission and operational constraints. Compared to the conventional method where the economic dispatch control analysis starts on a group of customer readings in the group connection point (i.e., bus) level, the study suggested a differentially private aggregated demand forecasting and reporting algorithm at the smart meter level (or building level) as a solution to the use case. Within this context, the study quantified the cost impact of differential privacy (by adding noise) of different smart

meter customer groups while assuming different groups of customers will choose their own level of privacy. The study also considered how to allocate fair shares of the total cost of differential privacy for those groups that differ in privacy levels. Simulations were evaluated by comparing electricity generation costs between a conventional method and several different configurations of the differential privacy algorithm. The study suggested designing incentive programs that will not only motivate customers to participate in demand reporting but will also let the customers pay less compared to when they are not reporting at all.

When and where to add noise can also happen either on each time stamp (several studies referred this application as point-wise differential privacy) on time dependent energy consumption data or after some level (e.g., monthly, annual) of aggregation is performed (this scenario needs a trusted data curator). The former implementation mostly focuses on protecting the individual lifestyle embedded in the time dependent energy consumption data where the privacy can be threatened by NILM attack. Some of the previous studies [54–60] focused specifically on perturbing the timeseries data by adding noise in each time stamp to hide the underlying privacy in the time dependent data stream. For example, Pöhls and Karwe [54] proposed a ϵ-differential privacy algorithm to obfuscate the energy usage pattern of a household. While this algorithm requires the data curator to be involved in the smart meter data obfuscation, a redactable signature scheme that removes parts of the data when it is not allowed by the smart meter user is also considered to remove the need for trusted data curator. In this algorithm, the utility company (or analyst) should be defining the accuracy of the query, and the smart meter user should also define the privacy level. And whenever there is a conflict between two, the smart meter reading is not reported to the utility company by the redactable signature scheme. The study leans against the importance of a utility company's use case (i.e., accuracy), thus emphasizing that the decision about the accuracy should first be determined (by the utility company), and the noise (or privacy level ϵ) should be driven by the accuracy limit. But the study also acknowledges the criticism that this scenario leads to weak privacy protection.

Only 1 of the 25 studies [74] focused on data that were collected through an IoT device (e.g., any device connected to Wi-Fi); the other studies focused on the smart meter data. Chen et al. [74] considered several representative queries in the study such as (but not limited to) 1) how many users are connected to a certain access point at a certain point in time, 2) how many users are connected to a certain access point at a certain time window, 3) whether the number of users who have accesse to the access point is larger than a certain threshold at a certain previous point in time, 4) whether the number of users with access to the access point larger than a certain threshold at a certain time window, and 5) how many sensors in a building were successfully connected to the access point in terms of being installed in the same room, same floor, and/or the same building. The proposed algorithm included three main modules: 1) "perturber" differentially privatizes the real-time data (increasing privacy guarantee), 2) "grouper" performs differentially privatized grouping of incoming timeseries data, which later informs smoother, and 3) "smoother" utilizes two previous outputs and performs smoothing of noisy timeseries data to increase data usability.

5.1.2 Battery load hiding

The analysis of local energy storage (e.g., rechargeable battery) attached to a building has gained increased attention in past decades, and there have also been studies [78–81] that concentrated on implementing differential privacy by hiding the raw energy consumption characteristics with a battery. The use of the battery (shown in Figure 2) in a building has the ability to shift the peak of end users' power consumption, which can result in reduced electricity prices when dynamic electricity pricing is adopted. The differences among the studies focused on battery load hiding are the mechanisms when the battery charge level reaches the bottom or top, for example, controlling the charge/discharge rate differently in those circumstances. While these studies included features of these mechanisms to mitigate the limit on batteries, the physical limits of the battery were still expressed as a limitation.

Backes and Meiser [78] developed a differential privacy algorithm leveraging a rechargeable battery that is connected to a household's power supply to obfuscate the electricity consumption profile of the household. In the proposed algorithm, the rechargeable battery is used to add noise on top of the original consumption data to achieve differential privacy. There are two major constraints when leveraging a rechargeable battery: 1) the power (denoted as throughput in the study) limit at each time step for either drawing or charging and 2) the energy (denoted as capacity in the study) limit for either drawing or charging during a certain time window. Zhao et al. [79] developed a differential privacy algorithm for the battery load hiding application by applying (ϵ, δ)-differential privacy. Attackers who are interested in revealing the timeseries profile of home appliances are assumed as an adversary model, and it is also assumed they have no common sense about typical consumption characteristics (e.g., expecting higher consumption in the evenings compared to after midnight). The multiple armed bandit problem that is a sequential decision problem defined by a set of actions was adopted in this study to mitigate the constraints (i.e., rate and capacity) stemming from leveraging a battery in the algorithm. More specifically, two separate noise generating mechanisms were proposed in this study, and the multiple armed bandit problem was used to decide which noise mechanism would be optimal when facing the battery's limit. The study noted that while there are various studies evaluating the differential privacy algorithms with certain metrics, there is no clear definition to connect these metrics to real privacy threats, thus emphasizing the need to properly define privacy around these use cases. Zhang et al. [81] also developed a differential privacy algorithm for the battery load hiding application. The study also incorporated the multiple armed bandit problem into the algorithm as well as a switching mechanism to deny the reporting of the smart meter reading to the utility company if necessary. The study also attempted to convert the impact of privacy preservation to cost impact by adopting the time of use pricing policy with three pricing models (square, triangle, and sinusoidal). An honest-but-curious internal attacker is assumed for the adversary model who can apply NILM on smart meter readings. Because of the constraints in the battery (i.e., rate and capacity), not all noise drawn from the Laplace distribution can be drawn from the battery. For this reason, a technique for adjusting and scaling the noise generation is

applied to resolve these constraints. Zellner et al. [80] developed a model predictive controller leveraging the local energy storage (e.g., battery) for differentially privatizing the load profile of smart meters while minimizing the cost of the electricity they use. An optimization problem for distributed (local) smart meters was considered, and two main objectives were targeted: 1) minimization of the energy costs for smart meter users under a dynamic electricity pricing scheme and 2) smoothing (and obfuscating) the load profile to lower the operation cost of a utility company. A dynamic pricing of the electricity is considered for calculating the total electricity cost for a smart meter user, and the net consumption used for the cost calculation is calculated by subtracting the electricity generation (from the renewable energy sources, such as PV) from the demand and adding electricity drawn from the battery to the demand. In the proposed algorithm, a data curator is required to perform analytics to calculate the cost of electricity based on a group of smart meter readings in order to achieve a proper smoothing. For this reason, both trusted and not-to-be trusted data curators are considered as two separate scenarios where the noise is either added in the local smart meter (not-to-be trusted data curator scenario) or added when the data curator performs cost minimization analytics (trusted data curator scenario).

5.1.3 Fog/Edge node

Most local differential privacy applications are in nature representing edge or fog computing environments because the obfuscation analytics are performed either in or close to the end use devices (e.g., smart meter). Edge and fog computing have the same concept in terms of the capability they provide by moving the analytics down to the end use devices rather than computing everything at the highest cloud (or database) level. The difference between the two environments is primarily the physical distance between end use devices and where the analytics happen: edge computing occurs at the device or in the gateway close to the sensor, and fog computing occurs in the LAN, which can be relatively distant from the sensor. Fog and edge computing depicted in Figure 2 essentially provide improved response time and relaxed bandwidth in the network, resulting in efficient communication of information. Because IoT devices and even smart meters can also be installed under these environments, studies [37, 82] have also focused on implementing differential privacy in these environments.

Xu et al. [37] focused on developing a local differential privacy algorithm for IoT device measurements in an edge computing environment considering both the computation limits of the edge devices and underlying privacy threats in those devices. The ability to share information and perform lightweight tasks among edge devices (e.g., laptop, Wi-Fi router) means that the edge computing environment can avoid significant response delays that affect the operations of each edge device if computations are all conducted on the cloud server. However, because lightweight tasks are only feasible in these devices, the mechanism of differential privacy should also be lightweight. For this reason, the study implemented two layers of data obfuscation by applying a feature distillation method to minimize the type and size of the data and by applying differential privacy against the output of the first layer to add noise and preserve privacy. An autoencoder model that automatically extracts features based on the useful inference objective function is used in the feature distillation method. Cao et al. [82] proposed a differential privacy algorithm in a fog computing environment leveraging a Factorial Hidden Markov model as a basis for smart meter reading obfuscation. Instead of adding noise directly on the smart meter readings, this study developed a mechanism where the noise is added in a different state (applying Factorial Hidden Markov model) of the user consumption information. The proposed algorithm provided better trade-offs (compared to existing algorithms) between usability and privacy by creating another state based on the smart meter data and by adding noise in that additional state before the obfuscated consumption data is sent to the fog node. More specifically, smart meter data representing the energy consumption of a single household was used to 1) disaggregate appliance level consumption profiles, 2) convert appliance level consumption into on/off switching sequences for each appliance (using Factorial Hidden Markov model), 3) add noise into switching sequences, and 4) regenerate obfuscated smart meter readings based on differentially private switching sequences.

5.1.4 Building control

While studies introduced previously mostly consider the smart meter data (measuring the whole building level performance) as the main source for obfuscation, some of the previous studies [46–48, 83] focused on differentially privatizing buildings' sub-system measurements to improve the operation efficiency of the building (or buildings) with privacy protection. Figure 2 also includes the schematic of actors and components of this architecture.

Ny and Mohammady [48] developed a differential privacy algorithm to obfuscate the multiple-input multiple-output (MIMO) system by considering an example of numerous occupancy status measurements with motion sensors in a building. The study also considered short- to medium-term occupancy forecasting based on differentially private motion sensor measurements, and the goal of the privacy protection was to protect an individual's location in the building. The sensitivity was analytically calculated for both single-input multiple-output (SIMO) and multiple-input multiple-output (MIMO) systems to apply (ϵ, δ)-differential privacy with Gaussian noise. A data set including 200 motion sensor measurements in a two-story building over several months was used to demonstrate the proposed algorithm in terms of how well the occupancy can be predicted with the differentially private motion sensor measurements with ϵ of 1 and δ of 0.5. Jia et al. [46] focused on the use case of occupancy-based HVAC control while preserving the location traces of individuals in a building that are measured by occupancy sensors. While measurements in occupancy sensors only provide information about whether a space is occupied, without revealing who those individuals are, the study noted that if the sensor measurements across various spaces (or rooms) are combined with additional information such as an office directory (showing who usually occupies a certain space) of the building, location traces of individuals can be inferred with high accuracy.

An HVAC system's performance and the thermal comfort in a building can be optimized with advanced control techniques using highly accurate sensor measurements. Because the noise added to the sensor measurement will impact the HVAC system, Jia et al. [46] analyzed the trade-off between privacy preservation versus the performance (e.g., comfort and expenditure) of the HVAC system controller. Technically, this study implemented a different occupancy distortion mechanism compared to differential privacy to enhance optimizing the noise distribution. However, it is introduced in this literature review given that the mechanism of adding noise and setting the framework is still viable for applying differential privacy. Several mathematical models are formulated to develop a model predictive controller: 1) a state model representing the comfort (expressed with zone temperature) of thermal zones depending on supply air flow rate and temperature set by the controller, 2) a cost function calculating the cost of using HVAC components (e.g., fan, cooling coil, reheat coil) to provide comfort in thermal zones, and 3) constraints for maintaining certain HVAC requirement levels (e.g., thermal comfort, indoor air quality, HVAC system capacity). Additionally, uncertainty limits of zone temperature and cost of energy consumption are bounded as input parameters to limit the amount of noise. Instead of applying differential privacy, the addition of noise was implemented by adopting the mutual information as a privacy loss metric, thus solving the optimization problem for ensuring the performance of the HVAC system while quantifying the privacy loss with the mutual information provided the appropriate amount of noise for the application. While the use case focus in previous studies was to improve a single building's operation, Xiao et al. [47] considered a use case for optimizing the control of a group (or cluster) of buildings in a privatized manner. This use case includes more risk in terms of revealing the privacy of occupants (or building operation) because the data collected in buildings needs to be shared with the data curator to optimize the control of the building cluster. A differential privacy algorithm was developed and applied on a model predictive control method, and three main targets were considered: 1) minimize energy consumption, 2) minimize energy cost, and 3) maintain privacy. An optimization problem for the building cluster control was formulated to minimize the energy cost of the building cluster (while satisfying comfort and capacity constraints). The proposed algorithm requires a data curator who collects locally optimized data from individual buildings to perform cluster level analysis. To avoid a trusted data curator scenario, a local (ϵ, δ)-differential privacy with Gaussian noise was implemented in each building to obfuscate the energy demand of individual buildings.

Both modern smart buildings and buildings equipped with BAS collect vast amounts of data around a building to understand the status of various systems (e.g., HVAC, lighting, plug load, occupancy) and to control the building in real time to optimize operation (in terms of comfort and energy efficiency). Some readers might argue whether the submetered data stored within the building needs to be privatized if it is just used for HVAC control while not sharing with the outside world. The argument is valid at the moment, however, there is a significant potential for these submetered data (including IoT device measurements in buildings), which are typically (and privately) stored within the building once they can be shared with the outside world and combined with data analytics. While there has been significant research advancement in the past couple of decades, buildings are complex, dynamic, and evolving in nature, which makes them very difficult to understand in terms of every aspect. There can be many different approaches for understanding buildings, however, opening up these data in a differentially private manner will boost the analytics on building research.

5.2 Global (centralized) and interactive/non-interactive models

As shown in Figure 2, the global (centralized) model includes both interactive and non-interactive models, which are then classified into variants such as private database, differential privacy combined with encryption, model publishing, and data publishing. A total of 11 studies [84–94] that are related to the implementation of global differential privacy are identified, and their implementation approaches are reviewed in this section.

The obfuscation in the global model happens after data is gathered in the database, which is represented as centralized storage. While the specific configurations between local and global models can vary significantly depending on the variants, the type of a query (or the use case) of a differential privacy implementation can still be similar for studies focusing on smart meter data. For example, the query of data aggregation that was introduced as the most frequent query in the local model implementation studies was also one of the query types that was studied in the global model implementations. However, compared to the queries analyzed in the local model implementations, the global model implementations included more queries that are suitable for extracting insights from large data. These included machine learning techniques such as data clustering. The variation of configurations between studies was less diverse compared to the local model implementations. For example, nine studies adopted ϵ-differential privacy while one study adopted (ϵ, δ)-differential privacy, and eight studies used the Laplacian distribution for the noise addition while one study considered both Laplacian and Geometric distributions.

5.2.1 Private database including differential privacy combined with encryption

The two implementation approaches covered in this section include the 1) private database and 2) differentially private database combined with encryption, shown in Figure 2 under the global model. Studies [84–91, 93] that focused on these implementations covered queries such as data aggregation, data clustering, and other machine learning techniques.

Hassan et al. [91] focused on the use case of dynamic pricing for enhancing demand response programs with global differential privacy. One of the major goals of a modern utility company is to reshape (or shift peak) the grid level load profile by utilizing various programs, such as a demand response program, and adopting dynamic pricing to efficiently match the supply and demand. However, because dynamic pricing directly leverages timeseries energy consumption of individual customers, the privacy of these customers can be threatened. Smart meter readings of individual households are gathered in the local database handled

by a trusted data curator. The data curator applies the differential privacy mechanism on these data to obfuscate underlying privacy before the timeseries data gets transmitted to the utility company. Then, the utility checks the noisy data for every month to generate billing for each customer based on the dynamic pricing scheme. The study applied ϵ-differential privacy with a Laplacian distribution for adding noise, and both external and internal (honest-but-curious) adversary models are considered a threat. The study also developed a dynamic pricing strategy by understanding usage at certain times, where the customers get charged for the peak load only if they have contributed to the peak load.

Jonsson [84] proposed a global ϵ-differential privacy algorithm considering the use case of data aggregation to support a utility company's efficient load forecasting. The study focused on improving the query mechanism by adopting simpler queries with low sensitivity, thus increasing the accuracy of the query while preserving privacy (in terms of ϵ). For example, a simple aggregation query asking "how many kW is being consumed by a group of smart meter users?" that typically includes a very large noise is reconstructed to a histogram type of query asking "how many smart meter users' consumptions are between x and y kW?" The basis of this reconstruction is from a counting query, which is theoretically proven to have very low sensitivity. An additional contribution was made in the study for setting the bins (how to set ranges of multiple x an y) of the histogram with four different partitioning strategies. These four strategies are either creating equally sized bins for the entire range, creating equally sized bins starting from the mean value, creating equally sized bins starting from each end of the range, or creating adaptively sized bins for the entire range. Once the complete histogram is generated with multiple counting queries, the original query is answered by mathematically integrating the histogram with average kWhs of each bin and corresponding counts.

Another common query that has been studied [87, 89, 90, 93] in global model implementations is data clustering. Data clustering analysis is a frequently used unsupervised data mining technique for efficiently extracting high-level insights or classifications from a massive amount of data. There are different methods available even within the data clustering analysis, and the k-means clustering method is the most popular method, where the mathematical centroid of each cluster is determined and used to cluster and classify the data into subgroups. Data clustering analysis has been widely used in many applications (e.g., market research, image processing) in machine learning, and it is also used by utility companies for various use cases (e.g., efficient power distribution operation, load forecasting). For example, Xiong et al. [89] focused on the query of data clustering with smart meter data that supports various use cases for the utility company. The goal of the study is to maximize the accuracy of the data clustering results while preserving underlying privacy in the smart meter data. The k-means clustering method was considered for the clustering algorithm, and improvements in selecting the appropriate (but noisy) centroids were made while combining with the differential privacy algorithm. An outlier detection method was also proposed in the study to remove the outlier data from the data set, thus reducing the sensitivity of the data and reducing the noise (but still retaining the same privacy level). Two adversary models were considered in the study, where one attack happens based on the center point and the other attack happens based on background knowledge. During the iterative process of k-means clustering, the calculations of the distance between the original data and the centroid in each iteration can reveal the underlying data, meaning successfully attacking the data based on the centroid information. The second attack occurs when the attacker has even more background information that can be used to deduce the private information more easily with the same information available during the clustering analysis. ϵ-differential privacy was applied in the study and Laplacian distribution was used to add noise to the centroid in each iteration during the data clustering analysis.

Applying machine learning techniques other than data clustering is another emerging area for implementing differential privacy. While a black box model trained with machined learned techniques already includes uncertainty in terms of the output that the model is trying to predict, the accuracy of the model can also be in the level of threatening the privacy of certain data if the quality of the training data is enough and the configurations of the machine learning processes are optimized. Differential privacy can also play a role in this area if the data needs to be obfuscated before being released to other parties. There are several review studies [95, 96] that discuss the deep learning or machine learning techniques around differential privacy implementation; for example, Zhao et al. [96] reviewed differential privacy implementation in deep learning by introducing privacy attacks (e.g., membership inference, training data extraction, and model extracting) related to deep learning models, classifying differential privacy mechanisms based on layers (e.g., input, hidden, and output layers) of deep learning. While this field of study is receiving growing attention, only one study related to applications in buildings was found. Soykan et al. [88] focused on the use case of load forecasting and implemented differential privacy leveraging open source libraries (for both machine learning and differential privacy) to preserve underlying privacy in smart meter data. The study adopted the forecasting model of Long-Short Term Memory (LSTM), which is one of the models of the Recurrent Neural Network (RNN) method for predicting the load in an hourly interval. A membership inference attack where the attacker tries to verify if a record they own is included in the training data set was considered as the adversary model. (ϵ, δ)-differential privacy was directly integrated in the LSTM model by adding random noise in one of the internal functions (e.g., gradient descent method optimizer) during the model training process. Because adding noise (to preserve privacy) to the training process will decrease the accuracy of load forecasting, the trade-off between accuracy of load prediction against the level of privacy preservation was studied.

5.2.2 Model publishing

There is certainly less literature focusing on global differential privacy implemented around building data compared to local differential privacy. However, it was also noticeable during the literature search process that some of these variants of differential privacy have potential for application to building data. The last two columns in Figure 2 show these variants. Zhu et al. [92] proposed a differentially private model publishing

method that can be applied in CPSs. In order to increase the usability compared to the conventional method (e.g., private database in Figure 2), the existing framework of releasing the obfuscated query results is transformed into releasing prediction models trained by machine learning techniques, as illustrated in Figure 2. More specifically, the models being released to the analyst are trained based on the raw data and raw results of the query, thus providing the relation between the data and the query with black box models. In this implementation, 1) the raw data is used to generate raw outputs for a set of queries, 2) differential privacy is implemented in these outputs of queries by inducing noise, 3) the relations between the raw data and noisy outputs of queries are used as training data for training the black box models, and 4) models trained with the noisy training data are released to the analyst. Various learning algorithms, such as linear regression, neural network, support vector machine, and others, are considered for the evaluations.

5.2.3 Data publishing

The last variant in the global model is data publishing. Rather than publishing the noisy query results or publishing black box models trained with queries, this approach involves publishing the anonymized or noisy data. This approach is actually the most common practice today with an anonymization technique (e.g., anonymous Netflix Prize data set[1]). While no literature was found where the differential privacy implementation was targeting building data, Fung et al. [97] reviewed existing privacy preserving data publishing methods not only for differential privacy but also for other privacy preserving techniques. The study considers various possible attacks (e.g., record linkage, attribute linkage, table linkage, probabilistic attack) and classifies different privacy preserving techniques that are immune to some of these attacks. The review includes theoretical aspects between different privacy preserving techniques that can be used for data publishing; however, no actual implementation or experiments were conducted. The study concluded with a future direction emphasizing needed efforts beyond technical advancements: "Privacy protection is a complex social issue, which involves policy-making, technology, psychology, and politics. Privacy protection research in computer science can provide only technical solutions to the problem. Successful application of privacy preserving technology will rely on the cooperation of policy makers in governments and decision makers in companies and organizations. Unfortunately, while the deployment of privacy-threatening technology, such as RFID and social networks, grows quickly, the implementation of privacy preserving technology in real-life applications is very limited. As the gap becomes larger, we foresee that the number of incidents and the scope of privacy breach will increase in the near future."

6 Discussion

This review provides a landscape view of data privacy in the buildings field, and an in depth discussion on differential privacy as an essential means to protect data privacy while unlocking the value from the widely collected and growing amount of data in buildings. It is clear from the literature that this research area, although attractive, is still in its early stage, and there are still a lot of topics to be studied. Thus, this section attempts to draw key findings gathered from the literature along with implications, limitations, and future work related to these key findings.

Understanding the current state of data access frameworks: Access to AMI smart meter data is important for energy efficiency providers to cost-effectively inform energy efficiency programs. State commissions, such as those in California and Illinois in the United States, established rules where customers can authorize distributed energy resources (DER) providers to access their smart meter usage data via platforms such as Green Button[2]. However, the narrow focus on permission-based access to AMI data became inadequate because energy efficiency service providers, building operators, and building researchers want granular data (not only usage, but also metadata such as demographics and building characteristics) without individual consent. Still, utility companies (as the custodians of smart meter data) control the information that each requester seeks, even when the customer approves the sharing of their data with an authorized third party by establishing constraints on the release of information that may or may not have an explicit basis in law. It is inevitable that practices for accessing private data will continue to evolve. Keeping data access rules and protocols up to date will be essential in a future in which more and more actors will either be enabled or thwarted by data access rules. Thus, the implementation of the differential privacy requires not only technical solutions about how to effectively add noise into the data but also comprehensive communications between actors to come to a mutual agreement to properly balance privacy and usability.

Determining the trade-off between privacy and usability: On the technical side, differential privacy has been increasingly adopted as an essential technique to protect data privacy at various levels, from individual buildings to portfolios of buildings. However, the determination of ϵ, which sets the privacy level of differentially private data, is still an open question. The closest attempt for selecting the appropriate ϵ from reviewed studies was reflecting the physical meanings of ϵ in a realistic scenario. For example, if a utility company is using private data of an individual customer for load forecasting, then the elevated privacy guarantee can result in additional generation cost because of the noise induced in the data. So, if the utility can decide on the upper bound of uncertainty of a load forecasting use case (e.g., how much error in load forecasting the utility company can accept), then the maximum limit of noise can also be determined mathematically. However, this does not consider the appropriate level of privacy from the customer's (person whose private information can be reflected in data) perspective. Some other studies leave this as an open question or an option where, in the future, the implementer will provide a set of privacy options to the customers and the customers will select their own privacy level from "completely okay to share everything" to "will not share anything." Other than studies that were reviewed in this article, there are studies (Lee

[1] https://en.wikipedia.org/wiki/Netflix$prize$
[2] https://www.greenbuttondata.org/

and Clifton [98], Hsu et al. [99], Yao et al. [100], Nissim et al. [101]) that discussed guidelines or theoretical aspects for selecting an appropriate ϵ. Based on the understanding gathered from the literature, while ϵ is implemented via a mathematical expression of differential privacy, the determination of ϵ should involve all stakeholders: differential privacy implementer, person or entity whose privacy is embedded in the data, policymaker, and end user (or analyst) of the data for extracting insights.

Differential privacy as a technical solution: As illustrated in Figure 2, this literature review found at least nine different variants of differential privacy implementation approaches from the gathered literature, where each of the variants was targeting specific query types or use cases. Furthermore, even within the same approach, the detailed differential privacy algorithms were different by making advancements from conventional approaches. As described by Desfontaines and Pejó [21], these variants can grow significantly more along with the advancement in research because the literature reviewed in this article only represents the early phase of the research. Some of the reviewed studies specifically expressed that the formulation of differential privacy has to be constructed for each query type. However, the authors also believe a systematic classification (e.g., leveraging differential privacy taxonomy from Desfontaines and Pejó [21]) and some level of generalization by grouping similar queries will be necessary and beneficial for researchers to align their contributions and for end users to leverage differential privacy as a tool with fewer hurdles. There are various existing open source libraries[3] that contribute to this path; however, libraries that can be used for applying differential privacy around building data are still minimal. Thus, a collaborative effort to develop open source tools (e.g., eeprivacy[4]) for implementing differential privacy in building data will be essential.

Building data coverage: Smart meter data is primarily considered in reviewed studies because of the emerging focus on smart grid scenarios where information sharing is one of the key aspects (but which increases privacy risk). However, buildings collect much more than just meter data. As it was mentioned previously, 36% of commercial buildings in the United States are supposedly integrated with BAS, measuring more specific measurements around the building. Furthermore, the deployment of IoT devices in buildings will also increase the amount of data collected in buildings significantly in the near future. While several studies focused on implementing differential privacy with regards to this aspect, further research efforts are required to examine privacy risks in various forms of data collected in buildings, especially occupant-related data such as internet connection (Wi-Fi signals), occupancy, and activities (movement, presence, human-building interactions).

Data and adversary models for testing differential privacy implementations: The reviewed studies tend to use their own or ad hoc publicly available data sets with limited coverage or data quality for testing and evaluating the performance of differential privacy implementations. For FAIR principles[5], there is a strong need for an open access high-quality data set as well as standardized performance metrics to support the evaluation and benchmarking of differential privacy techniques.

7 Conclusions

This article reviewed previous literature to provide building researchers with the basics of differential privacy implementation around data collected in buildings. Although the specific scope of this article is still in an early research stage, fast-growing data collection in buildings along with the breadth and scale of data available necessitated an examination of privacy preserving research for differential privacy. Because the topic is also not common in the representative building research journals, this study aimed to provide a greater level of detail for readers who might never have heard of differential privacy. To provide relevant context around privacy preservation in data, this literature review presented 1) privacy risks associated with data collected in buildings, 2) use cases that could be supported by analyses from these data, and 3) reviews of differential privacy implementation. The findings from the literature emphasize not only technical development but also engagement from stakeholders and policymakers in properly configuring differential privacy and protecting underlying privacy in data collected in buildings.

8 Acknowledgments

The authors thank Harry Bergmann of the U.S. Department of Energy's Building Technologies Office (BTO) for the support of this work. The authors also thank Carmen Best, McGee Young, Phil Ngo, Mariano Teehan, Lin Ainsworth, Deanna Cook, and Emily Laidlaw for their efforts towards the project. This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. Lawrence Berkeley National Laboratory's contribution to this work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Office, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

[3] https://github.com/google/differential-privacy
[4] https://github.com/recurve-inc/eeprivacy
[5] https://www.go-fair.org/fair-principles/

References

[1] Global energy consumption driven by more electricity in residential, commercial buildings - Today in Energy - U.S. Energy Information Administration (EIA), . URL https://www.eia.gov/todayinenergy/detail.php?id=41753.

[2] Frequently Asked Questions (FAQs) - U.S. Energy Information Administration (EIA), . URL https://www.eia.gov/tools/faqs/faq.php.

[3] Julia Lane. O privacy, where art thou?: Protecting privacy and confidentiality in an era of big data access. *Chance*, 25(4):39–41, 2012.

[4] Koichiro Hayashi. Social issues of big data and Cloud: privacy, confidentiality, and public utility. In *2013 International Conference on Availability, Reliability and Security*, pages 506–511. IEEE, 2013.

[5] Wei Fang, Xue Zhi Wen, Yu Zheng, and Ming Zhou. A survey of big data security and privacy preserving. *IETE Technical Review*, 34(5):544–560, 2017.

[6] Jimmy C. Chau and Thomas D. C. Little. Challenges in Retaining Privacy in Smart Spaces. *Procedia Computer Science*, 19:556–564, January 2013. ISSN 1877-0509. doi: 10.1016/j.procs.2013.06.074. URL https://www.sciencedirect.com/science/article/pii/S1877050913006820.

[7] Victoria Y. Pillitteri and Tanya L. Brewer. Guidelines for Smart Grid Cybersecurity. September 2014. URL https://www.nist.gov/publications/guidelines-smart-grid-cybersecurity.

[8] Sören Finster and Ingmar Baumgart. Privacy-aware smart metering: A survey. *IEEE communications surveys & tutorials*, 17(2):1088–1101, 2015.

[9] Sayyada Hajera Begum and Farha Nausheen. A comparative analysis of differential privacy vs other privacy mechanisms for big data. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pages 512–516. IEEE, 2018.

[10] Muhammad Rizwan Asghar, György Dán, Daniele Miorandi, and Imrich Chlamtac. Smart meter data privacy: A survey. *IEEE Communications Surveys & Tutorials*, 19(4):2820–2835, 2017.

[11] Sanket Desai, Rabei Alhadad, Naveen Chilamkurti, and Abdun Mahmood. A survey of privacy preserving schemes in IoE enabled smart grid advanced metering infrastructure. *Cluster Computing*, 22 (1):43–69, 2019.

[12] Mehdi Sookhak, Helen Tang, Ying He, and F. Richard Yu. Security and Privacy of Smart Cities: A Survey, Research Issues and Challenges. *IEEE Communications Surveys Tutorials*, 21(2):1718–1743, 2019. ISSN 1553-877X. doi: 10.1109/COMST.2018.2867288.

[13] Benjamin L. Ruddell, Dan Cheng, Eric Daniel Fournier, Stephanie Pincetl, Caryn Potter, and Richard Rushforth. Guidance on the usability-privacy tradeoff for utility customer data aggregation. *Utilities Policy*, 67:101106, December 2020. ISSN 0957-1787. doi: 10.1016/j.jup.2020.101106. URL https://www.sciencedirect.com/science/article/pii/S0957178720301004.

[14] Jens Hjort Schwee, Fisayo Caleb Sangogboye, Flora D Salim, and Mikkel Baun Kjærgaard. Tool-chain for supporting Privacy Risk Assessments. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 140–149, 2020.

[15] Samuel Warren and Louis Brandeis. The Right to Privacy-Harvard Law Review. In *Ethical issues in the use of computers*, volume 4, pages 172–183. Wadsworth Publ. Co, 1890.

[16] William L. Prosser. Libel Per Quod. *Virginia Law Review*, 46(5):839–855, 1960. ISSN 0042-6601. doi: 10.2307/1070563. URL https://www.jstor.org/stable/1070563.

[17] Roger Clarke's 'What's Privacy?', . URL http://www.rogerclarke.com/DV/Privacy.html.

[18] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. Differential privacy: its technological prescriptive using big data. *Journal of Big Data*, 5(1):15, April 2018. ISSN 2196-1115. doi: 10.1186/s40537-018-0124-9. URL https://doi.org/10.1186/s40537-018-0124-9.

[19] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[20] DworkCynthia and RothAaron. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, August 2014. doi: 10.1561/0400000042. URL https://dl.acm.org/doi/abs/10.1561/0400000042.

[21] Damien Desfontaines and Balázs Pejó. Sok: differential privacies. *arXiv preprint arXiv:1906.01337*, 2019.

[22] Muneeb Ul Hassan, Mubashir Husain Rehmani, and Jinjun Chen. Differential privacy techniques for cyber physical systems: a survey. *IEEE Communications Surveys & Tutorials*, 22(1):746–789, 2019.

[23] Latanya Sweeney. Weaving Technology and Policy Together to Maintain Confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997. ISSN 1748-720X. doi: https://doi.org/10.1111/j.1748-720X.1997.tb01885.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1748-720X.1997.tb01885.x.

[24] Arvind Narayanan and Vitaly Shmatikov. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, May 2008. doi: 10.1109/SP.2008.33.

[25] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLOS Genetics*, 4 (8):e1000167, 2008. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000167. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000167.

[26] George W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, December 1992. ISSN 1558-2256. doi: 10.1109/5.192069.

[27] Mikhail A. Lisovich, Deirdre K. Mulligan, and Stephen B. Wicker. Inferring Personal Information from Demand-Response Systems. *IEEE Security Privacy*, 8(1):11–20, January 2010. ISSN 1558-4046. doi: 10.1109/MSP.2010.40.

[28] M. Berenguer, M. Giordani, F. Giraud-By, and N. Noury. Automatic detection of activities of daily living from detecting and classifying electrical events on the residential power line. In *HealthCom 2008 - 10th International Conference on e-health Networking, Applications and Services*, pages 29–32, July 2008. doi: 10.1109/HEALTH.2008.4600104.

[29] Andrés Molina-Markham, Prashant Shenoy, Kevin Fu, Emmanuel Cecchet, and David Irwin. Private memoirs of a smart meter. In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building - BuildSys '10*, page 61, Zurich, Switzerland, 2010. ACM Press. ISBN 978-1-4503-0458-0. doi: 10.1145/1878431.1878446. URL http://portal.acm.org/citation.cfm?doid=1878431.1878446.

[30] Ishtiaq Rouf, Hossen Mustafa, Miao Xu, Wenyuan Xu, Rob Miller, and Marco Gruteser. Neighborhood watch: security and privacy analysis of automatic meter reading systems. In *Proceedings of the 2012 ACM conference on Computer and communications security - CCS '12*, page 462, Raleigh, North Carolina, USA, 2012. ACM Press. ISBN 978-1-4503-1651-4. doi: 10.1145/2382196.2382246. URL http://dl.acm.org/citation.cfm?doid=2382196.2382246.

[31] Ulrich Greveler, Peter Glosekotter, Benjamin Justus, and Dennis Loehr. Multimedia Content Identification Through Smart Meter Power Usage Profiles. page 8.

[32] Xiao Wang and Patrick Tague. Non-invasive user tracking via passive sensing: Privacy risks of time-series occupancy measurement. In *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*, pages 113–124, 2014.

[33] Pedro Barbosa, Andrey Brito, Hyggo Almeida, and Sebastian Clau\s s. Lightweight privacy for smart metering data by adding noise. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 531–538, 2014.

[34] Pedro Barbosa, Andrey Brito, and Hyggo Almeida. A Technique to provide differential privacy for appliance usage in smart metering. *Information Sciences*, 370-371: 355–367, November 2016. ISSN 0020-0255. doi: 10.1016/j.ins.2016.08.011. URL https://www.sciencedirect.com/science/article/pii/S0020025516305862.

[35] Muneeb Hassan, Mubashir Husain Rehmani, Ramamohanarao Kotagiri, Jiekui Zhang, and Jinjun Chen. Differential privacy for renewable energy resources based smart metering. *Journal of Parallel and Distributed Computing*, 131:69–80, September 2019. ISSN 0743-7315. doi: 10.1016/j.jpdc.2019.04.012. URL https://www.sciencedirect.com/science/article/pii/S0743731518309201.

[36] Jianqing Liu, Chi Zhang, and Yuguang Fang. EPIC: A Differential Privacy Framework to Defend Smart Homes Against Internet Traffic Analysis. *IEEE Internet of Things Journal*, 5(2):1206–1217, April 2018. ISSN 2327-4662. doi: 10.1109/JIOT.2018.2799820.

[37] Chugui Xu, Ju Ren, Deyu Zhang, and Yaoxue Zhang. Distilling at the Edge: A Local Differential Privacy Obfuscation Framework for IoT Data Analytics. *IEEE Communications Magazine*, 56(8): 20–25, August 2018. ISSN 1558-1896. doi: 10.1109/MCOM.2018.1701080.

[38] Brian S Amador. The federal republic of germany and left wing terrorism. Technical report, NAVAL POSTGRADUATE SCHOOL MONTEREY CA, 2003.

[39] Robert Guest. Austin PD Lawyers Up Over Warrantless Surveillance Program, November 2007. URL https://www.dallascriminaldefenselawyerblog.com/austin-pd-lawyers-up-over-warr/.

[40] Jordan Smith, Fri., Nov. 16, and 2007. APD Pot-Hunters Are Data-Mining at AE. URL https://www.austinchronicle.com/news/2007-11-16/561535/.

[41] Naperville Smart Meter Awareness v. City of Naperville, No. 16-3766 (7th Cir. 2018), . URL https://law.justia.com/cases/federal/appellate-courts/ca7/16-3766/16-3766-2018-08-16.html.

[42] John Glionna. Las Vegas outs its water hogs – at least when asked - Los Angeles Times, 2015. URL https://www.latimes.com/nation/la-na-vegas-water-hogs-20151030-story.html.

[43] Justin Horwath. Top 10 WATER GUZZLERS, 2015. URL http://www.sfreporter.com/news/coverstories/2015/03/31/top-10-water-guzzlers/.

[44] Márk Jelasity and Kenneth P. Birman. Distributional differential privacy for large-scale smart metering. In *Proceedings of the 2nd ACM workshop on Information hiding and multimedia security - IH&MMSec '14*, pages 141–146, Salzburg, Austria, 2014. ACM Press. ISBN 978-1-4503-2647-6. doi: 10.1145/2600918.2600919. URL http://dl.acm.org/citation.cfm?doid=2600918.2600919.

[45] Primal Pappachan, Martin Degeling, Roberto Yus, Anupam Das, Sruti Bhagavatula, William Melicher, Pardis E. Naeini, Shikun Zhang, Lujo Bauer, Alfred Kobsa, Sharad Mehrotra, Norman Sadeh, and Nalini Venkatasubramanian. Towards Privacy-Aware Smart Buildings: Capturing, Communicating, and Enforcing Privacy Policies and Preferences. In *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 193–198, June 2017. doi: 10.1109/ICDCSW.2017.52.

[46] Ruoxi Jia, Roy Dong, S. Shankar Sastry, and Costas J. Sapnos. Privacy-Enhanced Architecture for Occupancy-Based HVAC Control. In *2017 ACM/IEEE 8th International Conference on Cyber-Physical Systems (ICCPS)*, pages 177–186, April 2017.

[47] Yingying Xiao, Xiaodong Hou, Jie Cai, and Jianghai Hu. A Differentially Private Distributed Solution Approach to the Model Predictive Control of Building Clusters. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 7289–7295, December 2018. doi: 10.1109/CDC.2018.8619017.

[48] J. Le Ny and M. Mohammady. Differentially private MIMO filtering for event streams and spatio-temporal monitoring. In *53rd IEEE Conference on Decision and Control*, pages 2148–2153, December 2014. doi: 10.1109/CDC.2014.7039716.

[49] Liang Zhang, Jin Wen, Yanfei Li, Jianli Chen, Yunyang Ye, Yangyang Fu, and William Livingood. A review of machine learning in building load prediction. *Applied Energy*, 285:116452, March 2021. ISSN 0306-2619. doi: 10.1016/j.apenergy.2021.116452. URL https://www.sciencedirect.com/science/article/pii/S0306261921000209.

[50] Maryam Sadat Mirnaghi and Fariborz Haghighat. Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review. *Energy and Buildings*, 229:110492, December 2020. ISSN 0378-7788. doi: 10.1016/j.enbuild.2020.110492. URL https://www.sciencedirect.com/science/article/pii/S037877882031327X.

[51] Zhe Wang and Tianzhen Hong. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy*, 269:115036, July 2020. ISSN 0306-2619. doi: 10.1016/j.apenergy.2020.115036. URL https://www.sciencedirect.com/science/article/pii/S0306261920305481.

[52] Tianzhen Hong, Zhe Wang, Xuan Luo, and Wanni Zhang. State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, 212:109831, April 2020. ISSN 0378-7788. doi: 10.1016/j.enbuild.2020.109831. URL https://www.sciencedirect.com/science/article/pii/S0378778819337879.

[53] Muneeb Ul Hassan, Mubashir Husain Rehmani, and Jinjun Chen. Performance Evaluation of Differential Privacy Mechanisms in Blockchain based Smart Metering. *arXiv:2007.09802 [cs]*, July 2020. URL http://arxiv.org/abs/2007.09802. arXiv: 2007.09802.

[54] Henrich C Pöhls and Markus Karwe. Redactable signatures to control the maximum noise for differential privacy in the smart grid. In *International Workshop on Smart Grid Security*, pages 79–93. Springer, 2014.

[55] Fabian Laforet, Erik Buchmann, and Klemens Böhm. Individual privacy constraints on time-series data. *Information Systems*, 54:74–91, 2015.

[56] Xiaojing Liao, Preethi Srinivasan, David Formby, and Raheem A Beyah. Di-PriDA: Differentially private distributed load balancing control for the smart grid. *IEEE Transactions on Dependable and Secure Computing*, 16(6):1026–1039, 2017.

[57] Alaa Gohar, Farida Shafik, Frank Duerr, Kurt Rothermel, and Amr ElMougy. Privacy-preservation mechanisms for smart energy metering devices based on differential privacy. In *2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW)*, pages 1–6. IEEE, 2019.

[58] Haoxiang Wang and Chenyu Wu. Understanding Differential Privacy in Non-Intrusive Load Monitoring. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*, pages 401–403, 2020.

[59] Lu Ou, Zheng Qin, Shaolin Liao, Tao Li, and Dafang Zhang. Singular spectrum analysis for local differential privacy of classifications in the smart grid. *IEEE Internet of Things Journal*, 7(6):5246–5255, 2020.

[60] Kendall Parker, Prabir Barooah, and Matthew Hale. Spectral Differential Privacy: Application to Smart Meter Data. 2021.

[61] Henrik Sandberg, György Dán, and Ragnar Thobaben. Differentially Private State Estimation in Distribution Networks with Smart Meters. *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 4492–4498, December 2015. doi: 10.1109/CDC.2015.7402921. URL `http://arxiv.org/abs/1503.08490`. arXiv: 1503.08490.

[62] Gergely Ács and Claude Castelluccia. I Have a DREAM! (DiffeRentially privatE smArt Metering). In Tomáš Filler, Tomáš Pevný, Scott Craver, and Andrew Ker, editors, *Information Hiding*, Lecture Notes in Computer Science, pages 118–132. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-24178-9.

[63] Gilles Barthe, George Danezis, Benjamin Grégoire, César Kunz, and Santiago Zanella-Béguelin. Verified Computational Differential Privacy with Applications to Smart Metering. In *2013 IEEE 26th Computer Security Foundations Symposium*, pages 287–301, June 2013. doi: 10.1109/CSF.2013.26.

[64] Haiyong Bao and Rongxing Lu. A New Differentially Private Data Aggregation With Fault Tolerance for Smart Grid Communications. *IEEE Internet of Things Journal*, 2(3):248–258, June 2015. ISSN 2327-4662. doi: 10.1109/JIOT.2015.2412552.

[65] Marco Savi, Cristina Rottondi, and Giacomo Verticale. Evaluation of the Precision-Privacy Tradeoff of Data Perturbation for Smart Metering. *IEEE Transactions on Smart Grid*, 6(5):2409–2416, September 2015. ISSN 1949-3061. doi: 10.1109/TSG.2014.2387848.

[66] Vincenzo Gulisano, Valentin Tudor, Magnus Almgren, and Marina Papatriantafilou. Bes: Differentially private and distributed event aggregation in advanced metering infrastructures. In *Proceedings of the 2nd ACM International Workshop on Cyber-Physical System Security*, pages 59–69, 2016.

[67] Günther Eibl and Dominik Engel. Differential privacy for real smart metering data. *Computer Science - Research and Development*, 32(1):173–182, March 2017. ISSN 1865-2042. doi: 10.1007/s00450-016-0310-y. URL `https://doi.org/10.1007/s00450-016-0310-y`.

[68] Zequ Yang, Peng Cheng, and Jiming Chen. Differential-privacy preserving optimal power flow in smart grid. *IET Generation, Transmission & Distribution*, 11(15):3853–3861, 2017.

[69] Jianbing Ni, Kuan Zhang, Khalid Alharbi, Xiaodong Lin, Ning Zhang, and Xuemin Sherman Shen. Differentially Private Smart Metering With Fault Tolerance and Range-Based Filtering. *IEEE Transactions on Smart Grid*, 8(5):2483–2493, September 2017. ISSN 1949-3061. doi: 10.1109/TSG.2017.2673843.

[70] Jingyi Wang, Xinyue Zhang, Haijun Zhang, Hai Lin, Hideki Tode, Miao Pan, and Zhu Han. Data-Driven Optimization for Utility Providers with Differential Privacy of Users' Energy Profile. In *2018 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, December 2018. doi: 10.1109/GLOCOM.2018.8647839.

[71] Matthew Hale, Prabir Barooah, Kendall Parker, and Kasra Yazdani. Differentially private smart metering: Implementation, analytics, and billing. In *Proceedings of the 1st ACM International Workshop on Urban Building Energy Sensing, Controls, Big Data Analysis, and Visualization*, pages 33–42, 2019.

[72] Xin Lou, David K. Y. Yau, Rui Tan, and Peng Cheng. Cost and Pricing of Differential Privacy in Demand Reporting for Smart Grids. *IEEE Transactions on Network Science and Engineering*, 7(3): 2037–2051, July 2020. ISSN 2327-4697. doi: 10.1109/TNSE.2020.2971723.

[73] Junfang Wu, Weizhong Qiang, Tianqing Zhu, Hai Jin, Peng Xu, and Sheng Shen. Differential Privacy Preservation for Smart Meter Systems. In Sheng Wen, Albert Zomaya, and Laurence T. Yang, editors, *Algorithms and Architectures for Parallel Processing*, Lecture Notes in Computer Science, pages 669–685. Springer International Publishing, 2020. ISBN 978-3-030-38991-8.

[74] Yan Chen, Ashwin Machanavajjhala, Michael Hay, and Gerome Miklau. Pegasus: Data-adaptive differentially private stream processing. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1375–1388, 2017.

[75] Zhigao Zheng, Tao Wang, Ali Kashif Bashir, Mamoun Alazab, Shahid Mumtaz, and Xiaoyan Wang. A Decentralized Mechanism Based on Differential Privacy for Privacy-Preserving Computation in Smart Grid. *IEEE Transactions on Computers*, pages 1–1, 2021. ISSN 1557-9956. doi: 10.1109/TC.2021.3130402.

[76] Khadija Hafeez, Mubashir Husain Rehmani, and Donna O'Shea. DPNCT: A Differential Private Noise Cancellation Scheme for Load Monitoring and Billing for Smart Meters. In *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6, June 2021. doi: 10.1109/ICCWorkshops50388.2021.9473837.

[77] Xin Lou, Rui Tan, David K. Y. Yau, and Peng Cheng. Cost of differential privacy in demand reporting for smart grid economic dispatch. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pages 1–9, May 2017. doi: 10.1109/INFOCOM.2017.8057062.

[78] Michael Backes and Sebastian Meiser. Differentially private smart metering with battery recharging. In *Data Privacy Management and Autonomous Spontaneous Security*, pages 194–212. Springer, 2013.

[79] Jing Zhao, Taeho Jung, Yu Wang, and Xiangyang Li. Achieving differential privacy of data disclosure in the smart grid. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 504–512. IEEE, 2014.

[80] Martin Zellner, T. Tinoco De Rubira, Gabriela Hug, and Melanie Nicole Zeilinger. Distributed Differentially Private Model Predictive Control for Energy Storage. *IFAC-PapersOnLine*, 50(1):12464–12470, July 2017. ISSN 2405-8963. doi: 10.1016/j.ifacol.2017.08.1922. URL https://www.sciencedirect.com/science/article/pii/S2405896317325508.

[81] Zijian Zhang, Wenqiang Cao, Zhan Qin, Liehuang Zhu, Zhengtao Yu, and Kui Ren. When privacy meets economics: Enabling differentially-private battery-supported meter reporting in smart grid. In *2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS)*, pages 1–9, June 2017. doi: 10.1109/IWQoS.2017.7969167.

[82] Hui Cao, Shubo Liu, Longfei Wu, Zhitao Guan, and Xiaojiang Du. Achieving Differential Privacy against Non-Intrusive Load Monitoring in Smart Grid: a Fog Computing approach. *arXiv:1804.01817 [cs]*, April 2018. URL http://arxiv.org/abs/1804.01817. arXiv: 1804.01817.

[83] Sameera Ghayyur, Yan Chen, Roberto Yus, Ashwin Machanavajjhala, Michael Hay, Gerome Miklau, and Sharad Mehrotra. IoT-Detective: Analyzing IoT Data Under Differential Privacy. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pages 1725–1728, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-4703-7. doi: 10.1145/3183713.3193571. URL https://doi.org/10.1145/3183713.3193571. event-place: Houston, TX, USA.

[84] Hedvig Jonsson and Boel Nelson. Applied Differential Privacy in the Smart Grid. 2015. URL https://odr.chalmers.se/handle/20.500.12380/218681.

[85] Valentin Tudor, Magnus Almgren, and Marina Papatriantafilou. Employing Private Data in AMI Applications: Short Term Load Forecasting Using Differentially Private Aggregated Data. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, pages 404–413, July 2016. doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0076.

[86] Xiuxia Tian, Qian Song, and Fuliang Tian. Multidimensional Data Aggregation Scheme For Smart Grid with Differential Privacy. *IJ Network Security*, 20(6):1137–1148, 2018.

[87] Zefang Lv, Lirong Wang, Zhitao Guan, Jun Wu, Xiaojiang Du, Hongtao Zhao, and Mohsen Guizani. An Optimizing and Differentially Private Clustering Algorithm for Mixed Data in SDN-Based Smart Grid. *IEEE Access*, 7:45773–45782, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2909048.

[88] Elif Ustundag Soykan, Zeki Bilgin, Mehmet Akif Ersoy, and Emrah Tomur. Differentially Private Deep Learning for Load Forecasting on Smart Grid. In *2019 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6, December 2019. doi: 10.1109/GCWkshps45667.2019.9024520.

[89] Jinbo Xiong, Jun Ren, Lei Chen, Zhiqiang Yao, Mingwei Lin, Dapeng Wu, and Ben Niu. Enhancing Privacy and Availability for Data Clustering in Intelligent Electrical Service of IoT. *IEEE Internet of Things Journal*, 6(2):1530–1540, April 2019. ISSN 2327-4662. doi: 10.1109/JIOT.2018.2842773.

[90] Zhitao Guan, Zefang Lv, Xianwen Sun, Longfei Wu, Jun Wu, Xiaojiang Du, and Mohsen Guizani. A Differentially Private Big Data Nonparametric Bayesian Clustering Algorithm in Smart Grid. *IEEE Transactions on Network Science and Engineering*, 7(4):2631–2641, October 2020. ISSN 2327-4697. doi: 10.1109/TNSE.2020.2985096.

[91] Muneeb Ul Hassan, Mubashir Husain Rehmani, and Jinjun Chen. Differentially Private Dynamic Pricing for Efficient Demand Response in Smart Grid. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pages 1–6, June 2020. doi: 10.1109/ICC40277.2020.9149131.

[92] Tianqing Zhu, Ping Xiong, Gang Li, Wanlei Zhou, and Philip S. Yu. Differentially private model publishing in cyber physical systems. *Future Generation Computer Systems*, 108:1297–1306, July 2020. ISSN 0167-739X. doi: 10.1016/j.future.2018.04.016. URL https://www.sciencedirect.com/science/article/pii/S0167739X17325554.

[93] Shuai Guo, Mi Wen, and Xiaohui Liang. A Differentially Private K-means Clustering Scheme for Smart Grid. page 9.

[94] Saurab Chhachhi and Fei Teng. Market Value of Differentially-Private Smart Meter Data. In *2021 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, February 2021. doi: 10.1109/ISGT49243.2021.9372228.

[95] Zhanglong Ji, Zachary C Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*, 2014.

[96] Jingwen Zhao, Yunfang Chen, and Wei Zhang. Differential privacy preservation in deep learning: Challenges, opportunities and solutions. *IEEE Access*, 7:48901–48911, 2019.

[97] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)*, 42(4):1–53, 2010.

[98] Jaewoo Lee and Chris Clifton. How much is enough? choosing for differential privacy. In *International Conference on Information Security*, pages 325–340. Springer, 2011.

[99] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C Pierce, and Aaron Roth. Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410. IEEE, 2014.

[100] Xiaoming Yao, Xiaoyi Zhou, and Jixin Ma. Differential privacy of big data: An overview. In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pages 7–12. IEEE, 2016.

[101] Kobbi Nissim, Aaron Bembenek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David R O'Brien, Thomas Steinke, and Salil Vadhan. Bridging the gap between computer science and legal approaches to privacy. *Harv. JL & Tech.*, 31:687, 2017.