Analysis of Preprocessing Steps for z-Anonymity in IoT Architectures

Mateusz Kaczynski
TU Dresden
mateusz.kaczynski@mailbox.tu-dresden.de

January 9, 2026

Abstract

This thesis investigates the impact of preprocessing steps on the effectiveness of z-anonymity in multi-layered IoT architectures. (To be continued)

1 Introduction

In the era where Internet of Things (IoT) continuously gathers importance [1], the collection of high-frequency data from smart meters offers unprecedented opportunities for operational efficiency [2]. Smart meters are a fundamental component of demand response systems, supplying them with the consumption data that is needed to operate effectively. These systems leverage this data to detect changes in the energy usage in response to electricity price fluctuations or incentive payments, which are designed to encourage lower electricity use at times of high wholesale market prices [3]. This knowledge can be used by the energy provider to balance the energy production levels, ultimately improving the system efficiency [4].

However, the collection of such detailed consumption data from smart meters poses significant privacy risks. Smart meter reading can reveal sensitive characteristics about marital status, employment status, income, or the number of residents [5]. Moreover, behavioral patterns, including residents' presence, sleep schedule or meal times can be inferred from this consumption data. When combined with information about specific appliances, this information may be exploited by adversaries for malicious purposes [6].

Consequently, protecting this sensitive information is not merely a technical necessity for user safety, but a strict legal requirement. Within the European Union, General Data Protection Regulation (GDPR), enacted in 2018, establishes strict rules for processing, storing, managing personal data from individuals. Its primary goal is to give individuals more responsibility over their own data and ensure that companies handle data responsibly [7].

A naive approach might suggest removing user's private information in the published data to avoid those threats. Yet, it is well-established that simply removing direct identifiers is insufficient for protecting user privacy. Infamous real-world examples, such as the de-anonymization of the Netflix Prize dataset, and foundational academic research have proven that individuals can be easily re-identified by linking seemingly innocuous data points known as quasi-identifiers [8].

To counter this threat, formal privacy models like k-anonymity were developed for static datasets [9]. However, these traditional models are often too computationally intensive for the real-time, zero-delay requirements of modern data streams. This has led to the development of stream-specific algorithms, such as z-anonymity, which are designed to provide lightweight, continuous privacy protection [10].

While z-anonymity provides the core privacy, we argue that its effectiveness and the resulting data utility can be significantly influenced by data preprocessing techniques. These steps, applied before the main anonymization algorithm, can modify the data to influence the privacy-utility trade-off. For instance, generalization could possibly increase the amount of data that is ultimately published, while aggregation could drastically reduce communication overhead and system load.

The availability of these powerful tools raises a critical and under-explored architectural question: at which layer of the given architecture should the privacy-preserving preprocessing steps be deployed? The conventional Centralized model requires placing a high degree of trust in the central entity that collects all raw, sensitive data [11]. However, modern Internet of Things (IoT) networks are multi-layered [1], consisting of smart meters, gateways, and a central entity. This motivates our investigation into how shifting preprocessing steps, such as prefiltering, generalizing, and aggregating, affect the final privacy-utility trade-off when performed closer to the source data itself.

This research will systematically evaluate the impact of deploying these data preprocessing steps at different layers within a centralized smart meter architecture. Specifically, this work will establish a baseline scenario where raw data is processed only by the z-anonymity algorithm at the Central Entity. This baseline will then be systematically compared against variations where data preprocessing steps are first applied at the edge (on Smart Meters) or at the gateway level, allowing for a precise measurement of their impact on the final privacy-utility trade-off. By measuring the effects on data utility, information loss, and system performance, this research will provide a clear framework for understanding the trade-offs of moving computation closer to the user. All experiments will be grounded in a realistic context by using the SmartMeter Energy Consumption Data in London Households [12] dataset.

This thesis is structured as follows. Chapter 2 introduces the background knowledge required to understand this work. Chapter 3 reviews the related literature. Chapter 4 describes the research methodology, while Chapter 5 presents the experimental results. Chapter 6 discusses these results, and Chapter 7 concludes the thesis.

2 Background

2.1 The Smart Meters and IoT Architecture

Smart meters are digital energy measurement devices that automatically record electricity consumption and transmit these measurements to the utility provider at regular intervals [2]. Depending on the deployment and configuration, the reporting interval usually ranges from seconds to hours. This continuous data collection enables detailed monitoring of energy usage. While this data is essential for applications such as demand response and load forecasting [3, 4], it also introduces significant challenges related to data volume, communication overhead, and user privacy [5, 6].

Smart meters operate within the Internet of Things (IoT) architectures [1]. While early implementations often relied on direct device-to-cloud communication, modern systems are increasingly organized as multi-layered architectures consisting of the Edge Layer, the Fog Layer, and the Central Layer [13, 14]. The split into different layers brings its advantages.

The Edge Layer consists of the smart meters themselves. Those devices are deployed in close proximity to end users and generate raw consumption data. Due to constraints in computational power, memory, and energy consumption, edge devices are typically limited in the complexity of calculcations they can execute [15]. However, performing lightweight preprocessing at the edge offers important advantages. By executing initial tasks such as data filtering or local aggregation, the system can scale effectively. This is particularly critical given that the amount of data produced by smart meters will continue to grow in the upcoming years, potentially exceeding the capacity of conventional cloud computing to handle raw streams [13]. Morevoer, it also brings the advantage with respect to privacy, as sensitive data can be transformed before leaving the user's premises, thereby reducing the trust needed in the central entity, which is currently seen as a potential risk [11, 13].

The Gateway of Fog Layer represents an intermediate layer between the edge devices and the central entity. They work by analyzing the data sent by edge nodes and reducing it significantly by removing unimportant data or performing preprocessing tasks such as filtering, aggregation, or generalization. By reducing the volume of the data transmitted, this layer can decrease communication overhead and address the inherent cloud problems such as "unreliable latency, lack of mobility support and location-awareness" [14], while still retaining more computational resources than individual smart meters.

The Central Entity, often described as a cloud-based system, is responsible for long-term storage, large-scale analysis, and decision-making processes. While centralized processing simplifies system mananegement and enables complex analytics [2], it requires the transmission of raw or minimally processed data, which increases both privacy risks and the level of trust placed in the data collector [6].

This layered architecture reflects a fundamental trade-off between efficiency, privacy, and system complexity. Moving computation closer to the data source can reduce latency and bandwidth requirements while limiting the exposure of sensitive information. At the same time, it constrains the available computational resources. These trade-offs are central to this thesis, which investigates how the placement of privacy-preserving preprocessing steps within this architecture affects the final privacy-utility balance.

2.2 Data Privacy Fundamentals

The collection of smart meter data raises significant privacy concerns, as energy consumption patterns can reveal sensitive information about individuals and households [6]. In this matter, it is essential to distinguish between explicit identifiers and quasi-identifiers.

Explicit identifiers are attributes that uniquely identify an individual on their own, such as a name, address, or customer identification number. In the context of smart metering, examples include the physical address of a household or a unique meter identifier. These attributes are typically removed before data is shared or published [9].

Quasi-identifiers are attributes that are not identifying in isolation but can be used to re-identify individuals when combined with other data sources. In smart meter datasets, quasi-identifiers may include timestamps, energy consumption values, and location-related information. Although such attributes appear innocuous, their combination can form distinctive usage patterns that uniquely characterize households [9].

The risk posed by quasi-identifiers is demonstrated by linkage attacks, in which anonymized datasets are combined with external information, often publicly available, from an auxiliary dataset that shares common attributes to re-identify individuals [9].

A foundational example of this vulnerability is the work of Latanya Sweeney, who demonstrated that 87% of the U.S. population could be uniquely identified using only a combination of ZIP code, gender, and date of birth [16]. Sweeney proved the validity of this threat by successfully re-identifying the medical records of the Governor of Massachusetts, linking the "anonymized" hospital data (which contained diagnosis and demographics) with the public Voter Registration List (which contained names and demographics). Similarly, Narayanan and Shmatikov de-anonymized the Netflix Prize dataset [8]. Although the dataset had been stripped of user IDs, the researchers were able to re-identify specific users by linking the timestamped movie ratings in the dataset with public reviews posted on IMDb.

In the smart meter context, linkage attacks may exploit publicly available information, such as occupancy schedules or appliance usage statistics, to infer private attributes including presence patterns, daily routines, or socio-economic characteristics [5]. These risks illustrate that simply removing explicit identifiers is insufficient to guarantee privacy and motivate the use of formal privacy models that provide quantifiable protection against re-identification [9].

2.3 Formal Privacy Models

To mitigate the re-identification risks described in the previous section, several formal privacy models have been developed. These models transition privacy from a heuristic process to a mathematically provable state. This section outlines the foundational model of k-anonymity, the alternative paradigm of Differential Privacy, and the stream-specific derivation known as z-anonymity.

2.3.1 K-Anonymity and Equivalence Classes

The foundational concept in privacy-preserving data publishing is k-anonymity, proposed by Sweeney. Formally, a dataset satisfies k-anonymity if every record is indistinguishable from at least $k - 1$ other records with respect to the set of quasi-identifiers [9].

The core mechanism relies on partitioning the dataset into equivalence classes. An equivalence class is a set of records that share identical values for their quasi-identifiers. To achieve a specified threshold k, algorithms employ generalization (reducing granularity, e.g., converting a specific timestamp to a 1-hour interval) or suppression (removing the value entirely) until the cardinality of every equivalence class is $\geq k$ [9].

However, k-anonymity is inherently ill-suited for the IoT data streams described in Section 2.1. Standard algorithms require a global view of a static dataset to calculate optimal generalization hierarchies. In a streaming context, data arrives sequentially and indefinitely [17]. To form an equivalence class of size k, the gateway would need to buffer incoming tuples until enough matching records arrive. This introduces "blocking," creating significant latency that contradicts the real-time requirements of smart grids [10]. Furthermore, the computational cost of recalculating equivalence classes as new data arrives is often prohibitive for resource-constrained edge devices [17].

2.3.2 Differential Privacy

Differential Privacy (DP) is widely regarded as the "gold standard" in modern privacy research. Unlike k-anonymity, which focuses on the structure of the output data, DP focuses on the algorithm itself. It guarantees that the output of a query is substantially similar whether or not any single individual's data is included in the input [18].

Mechanically, DP is achieved by injecting calibrated noise (typically via Laplace or Gaussian mechanisms) into the data or query results [18]. While robust against linkage attacks, this noise injection presents a fundamental conflict with the operational requirements of smart metering. As emphasized in [3], energy data must be reliable and precise to avoid substantial financial costs, with billing meters subject to rigorous accuracy standards (e.g., IEC 61036). Consequently, the probabilistic distortion introduced by DP is often considered unacceptable for raw energy data streams where precise measurements are mandatory.

Furthermore, in a continuous streaming context, the Basic Composition Theorem [18] dictates that privacy loss accumulates with every query ($\sum \epsilon_i$). As established in Section 3.5 of the foundational literature, this cumulative property makes it mathematically impossible to maintain a fixed privacy guarantee over an indefinite period without eventually stopping transmission or introducing infinite noise.

2.3.3 Z-Anonymity

To address the latency constraints of k-anonymity while avoiding the noise injection of DP, z-anonymity was proposed as a lightweight, stream-centric alternative. Unlike k-anonymity, which enforces distinct groups of k current records, z-anonymity operates on the principle of historical frequency. It assumes that privacy risks are highest for "outliers"—values that appear rarely in the data stream [10].

The algorithm processes data in a single pass with O(1) complexity, making it ideal for the Edge and Fog layers. It functions through the following steps, based on the definition in [10]:

1. **History Maintenance**: The system maintains a sliding window or historical table of observed values for a specific quasi-identifier.

2. **Lookup**: For every incoming data tuple, the algorithm queries the historical count of that attribute value.

3. **Threshold Check**: If the count exceeds the threshold z, the value is deemed "safe" (common enough to not be identifying) and is transmitted immediately.

4. **Suppression**: If the count is below z, the value is considered a rare outlier and is suppressed or generalized.

This approach eliminates the need for buffering, allowing for zero-delay transmission [10]. Crucially, z-anonymity is probabilistically linked to k-anonymity. The authors demonstrate that if a value appears z times in a sufficiently large historical window, the probability that the record belongs to a crowd of size k approaches 1. Therefore, z-anonymity acts as a performant proxy for k-anonymity in high-velocity environments, balancing privacy protection with the strict low-latency requirements of IoT architectures [10].

2.4 Data Preprocessing Techniques for Privacy

While formal privacy models like k- or z-anonymity provide the criteria for protection, preprocessing techniques are the mechanisms used to transform raw data to meet these criteria. In the context of IoT architectures, these transformations are crucial for striking a balance between privacy (indistinguishability), utility (analytical precision), and system performance (bandwidth and latency). This section formally defines the three primary preprocessing steps investigated in this thesis: Data Generalization, Temporal Aggregation, and Local Prefiltering.

2.4.1 Data Generalization

Data generalization is a non-perturbative technique that replaces specific, precise values with broader, less specific categories or intervals. The fundamental goal is to increase the probability that a specific tuple belongs to a larger equivalence class (or meets the z-threshold) by reducing the granularity of the data. Formally, generalization relies on a Value Generalization Hierarchy (VGH). A VGH is a tree structure where leaf nodes represent raw values (e.g., 1.234 kWh) and root nodes represent the most general state (e.g., "Any Energy Value"). In between are intermediate levels of granularity [19].

There are two primary approaches to defining these hierarchies:

1. **Static Generalization (Fixed-Intervals)**: The domain of an attribute is divided into pre-defined intervals (e.g., 0-5 kWh, 5-10 kWh) or precision levels (e.g., rounding to the nearest integer). For categorical attributes, such as ZIP codes, this typically involves masking the least significant digits (e.g., replacing the last digit with *) [9]. This method is computationally lightweight (O(1)) and deterministic, making it highly suitable for the resource-constrained Edge layer of IoT devices [13].

2. **Dynamic Binning**: As noted in literature, static intervals typically rely on rigid hierarchies that may fail to capture the actual data distribution, resulting in unnecessary information loss. To address this, dynamic partitioning techniques such as Mondrian [20] were proposed to adapt the intervals based on the local density of the data. While dynamic binning often yields higher utility—measured, for instance, by the Normalized Certainty Penalty (NCP) defined in [21], it typically requires

sorting or multiple passes over the data, which introduces computational overhead that may not be viable for real-time edge processing.

For this thesis, generalization is modeled through precision reduction (rounding). This acts as a static VGH where moving up the hierarchy corresponds to reducing the number of decimal places, effectively grouping precise sensor readings into discrete "bins" to facilitate indistinguishability.

2.4.2 Temporal Aggregation

Temporal aggregation serves as a dimensionality reduction technique along the time axis. In high-frequency smart metering (e.g., reading every second), the timestamp itself acts as a quasi-identifier that can reveal minute-by-minute lifestyle patterns [6].

This technique replaces a sequence of fine-grained data points $d_1, d_2, \ldots, d_n$ occurring within a time window W with a single aggregate value V_{agg} (typically the mean, sum, or max). The windowing strategy determines how data is grouped:

1. **Tumbling Windows**: These are fixed, non-overlapping windows (e.g., 10:00–10:15, 10:15–10:30). In this approach, every data point contributes to exactly one aggregate value.

2. **Sliding Windows**: These are overlapping intervals (e.g., 10:00–10:15, 10:05–10:20), where a single data point contributes to multiple aggregate values.

In this research, tumbling windows are employed to calculate average energy consumption. This approach significantly reduces the data volume transmitted to the central entity while retaining the coarse-grained load profile necessary for forecasting applications [3].

2.4.3 Local Prefiltering (Suppression)

Suppression is the most drastic form of data protection, where specific attribute values are entirely removed from the dataset [9]. In a centralized architecture, suppression typically occurs after the data has reached the central server (global suppression). However, Local Prefiltering shifts this decision to the Gateway or Edge layer [11].

This technique operates on a rule-based logic to identify "outliers" locally. In the context of z-anonymity, an outlier is a value that appears with insufficient frequency to be safe [10]. By implementing a local threshold (z_{local}), the Gateway can assess the rarity of a value before transmission.

Rule: If Frequency(value)$<z_{\text{local}}$, the value is suppressed locally.

Effect: This prevents the transmission of data that is statistically likely to fail the global privacy check at the Central Entity anyway.

While prefiltering significantly reduces Communication Overhead (message count), it introduces a risk to Data Utility (Publication Ratio). If the local view of the data at the Gateway does not perfectly reflect the global distribution at the Central Entity, the Gateway might suppress a value that—had it been sent—would have found matches from other Gateways. Investigating this trade-off between bandwidth savings and false-positive suppression is a key objective of this study.

Bibliography

[1] M. Farooq, M. Waseem, S. Mazhar, A. Khairi, and T. Kamal, "A review on internet of things (iot)," *International Journal of Computer Applications*, vol. 113, pp. 1–7, 03 2015.

[2] M. Rasoulnia, E. Yaghoubi, E. Yaghoubi, A. Hussain, and I. Kamwa, "A comprehensive systematic and bibliometric review of technologies and measurement tools for power quality events detection, classification, and fault location in smart grids," *Renewable and Sustainable Energy Reviews*, vol. 226, p. 116302, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136403212500975X

[3] P. Siano, "Demand response and smart grids—a survey," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461–478, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032113007211

[4] G. Strbac, "Demand side management: Benefits and challenges," *Energy Policy*, vol. 36, no. 12, pp. 4419–4426, 2008, foresight Sustainable Energy Management and the Built Environment Project. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301421508004606

[5] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397–410, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544214011748

[6] M. A. Lisovich, D. K. Mulligan, and S. B. Wicker, "Inferring personal information from demand-response systems," *IEEE Security Privacy*, vol. 8, no. 1, pp. 11–20, 2010.

[7] "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance)," pp. 1–88, May 2016. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj

[8] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 2008, pp. 111–125.

[9] L. Sweeney, “k-anonymity: a model for protecting privacy,” *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, p. 557–570, Oct. 2002. [Online]. Available: https://doi.org/10.1142/S0218488502001648

[10] N. Jha, T. Favale, L. Vassio, M. Trevisan, and M. Mellia, “z-anonymity: Zero-delay anonymization for data streams,” 12 2020, pp. 3996–4005.

[11] C. Brunn, “dezent: Decentralized z-anonymity with privacy-preserving coordination,” 2025.

[12] U. P. Networks, “Smartmeter energy consumption data in london households,” 2011-2014, dataset from the Low Carbon London project, covering Nov 2011 – Feb 2014. [Online]. Available: https://data.london.gov.uk/dataset/smartmeter-energy-consumption-data-in-london-households-vqm0d/

[13] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[14] S. Yi, C. Li, and Q. Li, “A survey of fog computing: Concepts, applications and issues,” in *Proceedings of the 2015 Workshop on Mobile Big Data*, ser. Mobidata ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 37–42. [Online]. Available: https://doi.org/10.1145/2757384.2757397

[15] e. a. Zhou, X., “Introducing edge intelligence to smart meters via federated split learning,” *Nature Communications*, 2024, explicitly discusses the memory/computation constraints of current smart meter hardware.

[16] L. Sweeney, “Simple Demographics Often Identify People Uniquely,” 6 2018. [Online]. Available: https://kilthub.cmu.edu/articles/journal_contribution/Simple_Demographics_Often_Identify_People_Uniquely/6625769

[17] K. Guo and Q. Zhang, “Fast clustering-based anonymization approaches with time constraints for data streams,” *Knowledge-Based Systems*, vol. 46, pp. 95–108, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705113000877

[18] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, p. 211–407, Aug. 2014. [Online]. Available: https://doi.org/10.1561/0400000042

[19] B. C. M. Fung, K. Wang, and P. S. Yu, “Top-down specialization for information and privacy preservation,” in *Proceedings of the 21st International Conference on Data Engineering*, ser. ICDE ’05. USA: IEEE Computer Society, 2005, p. 205–216. [Online]. Available: https://doi.org/10.1109/ICDE.2005.143

[20] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, 2006, pp. 25–25.

[21] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization for privacy preservation with less information loss," *SIGKDD Explor.*, vol. 8, pp. 21–30, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:207162632