Analysis of the Impact of Data Granularity on Privacy for the Smart Grid

Analysis of the Impact of Data Granularity on Privacy for the Smart Grid

Valentin Tudor
Department of Computer Science and Engineering, Chalmers University of Technology
Göteborg, Sweden
tudor@chalmers.se

Magnus Almgren
Department of Computer Science and Engineering, Chalmers University of Technology
Göteborg, Sweden
magnus.almgren@chalmers.se

Marina Papatriantafilou
Department of Computer Science and Engineering, Chalmers University of Technology
Göteborg, Sweden
ptrianta@chalmers.se

ABSTRACT

The upgrade of the electricity network to the "smart grid" has been intensified in the last years. The new automated devices being deployed gather large quantities of data that offer promises of a more resilient grid but also raise privacy concerns among customers and energy distributors.

In this paper, we focus on the energy consumption traces that *smart meters* generate and especially on the risk of being able to identify individual customers given a large dataset of these traces. This is a question raised in the related literature and an important privacy research topic. We present an overview of the current research regarding privacy in the Advanced Metering Infrastructure. We make a formalization of the problem of de-anonymization by matching low-frequency and high-frequency smart metering datasets and we also build a threat model related to this problem. Finally, we investigate the characteristics of these datasets in order to make them more resilient to the de-anonymization process.

Our methodology can be used by electricity companies to better understand the properties of their smart metering datasets and the conditions under which such datasets can be released to third parties.

Keywords

Smart grid data privacy; Advanced Metering Infrastructure (AMI) data characteristics; Smart meter privacy; Smart metering data

1. INTRODUCTION

In any new domain where significantly more data starts being produced, the privacy of the customer who produces these data may be at risk. This is also the case in the new *smart grid* which is the name used for the modern electrical grid. One of the main differences between the traditional electrical grid and the new smart grid is the large number of computing and communication devices being installed in different parts of the grid and that are connected through an overlay communication network; their main purpose is to make the grid monitoring and operational processes more accurate and more efficient.

These computing and communication devices are deployed in all of the three main sections of the electrical network: the generation section, the transmission section and the distribution section. Specifically, in the distribution section, the traditional electro-mechanical meters that used to monitor the electrical energy consumed by the end customers are replaced by the new so-called *smart meters*. The smart meters, together with other devices that monitor, gather and send their data to the energy distributor's central location form the *Advanced Metering Infrastructure* (*AMI*). The AMI offers two-way communication between the central control system and the smart meters, resulting in better remote functionality of the smart meters, such as remote shut-off commands and control of demand-side electricity load and generation. Figure 1 presents an overview of the AMI, together with an exemplification of the different types of communication media (radio, wired, fiber-optics) and protocols used (Ethernet, Power Line Communication, ZigBee, GPRS) in suggested deployments.

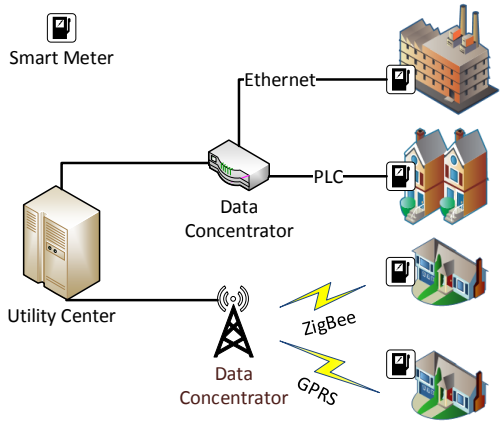

Figure 1: The Advanced Metering Infrastructure (AMI)

As a consequence of the upgrade to the smart grid, significantly more data is collected and analyzed, for example in the AMI where more parameters than just the the electrical energy consumed by customers are recorded, at a higher frequency than before. It is estimated that the size of the smart grid will be larger than the size of the Internet[1] and the quantity of data produced will be considerable. These data are expected to play a key role in the development of the smart grid and will improve the balance between energy production and energy consumption by making a significant contribution in improving electrical grid stability and energy efficiency.

However, there are concerns that these benefits may come at the cost of privacy: the large quantity of data produced and the granularity with which individual items are collected raise privacy concerns regarding the information that can be inferred about the lifestyle of the customers. In some countries, the debate regarding customer's privacy has even slowed down the deployment of smart meters [8].

Therefore, the main question is whether data can be collected in such a way as to keep an adequate privacy level and still be useful for billing and grid operational purposes. Using the terminology from [17], any solution should offer *anonymity* (the state of being not identifiable in a set of subjects) and also temporary *unlinkability* (the relation of two items based on the adversary's observations) of the customer with the quantity of electrical energy used in that specific unit of time. However, in the smart grid full unlinkability is almost impossible to be attained because the customer needs to be billed at some point for the resources used. The same goes for the *unobservability* (usage of a resource without someone to be able to observe that the resource is being used); the aggregated consumption is known at all times as the energy used for a group of customers is monitored at the substation level.

In this paper we first provide an *overview* of the current research regarding privacy in the AMI where we present some of the current privacy problems and privacy enhancing technologies proposed in the literature, motivating also the contributions presented subsequently in the paper. We construct a *formalization* of the de-anonynimization problem present in the AMI. The problem is caused by matching two types of datasets collected in the AMI, the low-frequency dataset (mainly used for billing of customers) and the high-frequency dataset (mainly used for grid operation). We build a theoretical model that describes this problem and also a threat model presenting a possible de-anonymization scenario performed by an adversary. We perform an *investigation* of the characteristics of these datasets in order to make them more resilient to the de-anonymization process. In our investigation, we concentrate on the data collected in the distribution network from the AMI where we focus mainly on the data granularity and timespan.

The rest of the paper is structured as follows: in Section 2 we present the general considerations of data privacy in the AMI as well as the different data types that can be collected. We give an overview of the current literature, present the main questions regarding the privacy concerns raised by AMI data and describe the characteristics of the two types of datasets mentioned earlier. Section 3 formally describes the de-anonymization problem, followed by the development of the theoretical framework and the threat model. Section 4 describes the investigation conducted and a discussion of the results obtained. This paper concludes with Section 5 which summarizes our results and their implications.

[1] `http://news.cnet.com/8301-11128_3-10241102-54.html`

2. DATA PRIVACY IN THE ADVANCED METERING INFRASTRUCTURE

As mentioned in the previous section, the main improvement introduced by the smart grid in the distribution section is the replacement of the traditional electromechanical meters with the new smart electrical meters which are the main producers of data from the AMI. Before the smart meters, energy consumption readings were usually made every month or even less frequently, usually by a human operator visiting each customer individually, so the quantity of data gathered was not even comparable with the one today.

2.1 Data from the Advanced Metering Infrastructure

Data from the Advanced Metering Infrastructure are primarily used for *billing purposes* and consist of the *index of energy consumption* in kWh. The modern smart meters offer the possibility to extract much more information about the well-being of the electrical distribution network. For billing of residential customers, only the quantity of the so-called *active energy* consumed is required. For high industrial consumers the quantity of *reactive energy* used may also be billed; grid operation may require information about instantaneous values of voltage, current, active/reactive power, power outage logs, errors in the metering equipment, and much more. Table 1 shows a short list of useful data types that can be gathered from the AMI.

Billing data	Operational Data
Active energy	Power (active, reactive, power factor)
Reactive energy	Voltage (value, phase angle)
	Current (value, phase angle)
	Power outage logs, Alarms

Table 1: Data from the AMI

Efthymiou et al. [7] use the term *high-frequency data* for data used for grid operational purposes and *low-frequency data* for data used for billing purposes. We will keep the same definitions throughout this paper. Low-frequency data need to be collected seldomly (every month or every few months) but the law dictates that such data need to be identifiable to a specific customer for correct billing and to prevent fraud, both from the customer side and from the utility provider side. High-frequency data, used for grid operations, need to be collected very often (every few minutes) in order to give an accurate overview of the electrical distribution network. Previous research [1, 16] shows that fine-grained data can infer information about the lifestyle of the inhabitants such as electrical device usage patterns and presence or absence from the premises. Although the utility of these data for

grid operation is evident, the privacy concerns that may be raised cannot be ignored. In an ideal case, these data should not be identifiable with a specific customer [7], but with a group of customers served by the same electrical transformer or distribution station.

2.2 Data usage in the Advanced Metering Infrastructure

As mentioned in the previous section, there are several different types of data that can be collected in large quantities from the Advanced Metering Infrastructure. The main consumer of these data is the Distributor System Operator (DSO), followed by other third parties, each of them having different purposes. Data can be used by the DSO for billing, processed for fraud detection, operational purposes (grid stability and security) or marketing. Third parties (researchers, other companies, malicious entities) may also be interested in these data for benign activities (research, marketing) or for malign ones (fraud, invasion of privacy or even attacks against the critical infrastructures).

The privacy preserving techniques (PPTs) are usually implemented at a large scale by the DSO, or a legal trusted third party and at a small scale by the customers. When thinking about a specific privacy preserving technique it is important to remember the complexity of the parties that may have access and use the data produced. For example, we should be able to answer the following (not exhaustive) list of questions:

- does the PPT offer privacy protection against DSOs?
- does it offer privacy protection against third parties?
- does it provide availability of billing data?

If interested in providing privacy for its customers, the DSO may prefer to employ a solution that offers privacy protection against third parties but which first provides availability of the billing data.

The customer may prefer a solution that offers privacy protection against both the DSO and other third parties, while availability of billing data might come in a later position in the customer's priority list. Thus, the DSO's and the customer's visions of privacy might be different and even conflicting. In an ideal case, the customer's data privacy should be protected against both the DSO and other third parties.

2.3 Overview of smart grid privacy mechanisms in the literature

As a concept, Warren and Brandeis [21] give in 1890 the definition of "privacy" as the "right to be let alone". More recently, Pfitzmann and Hansen [17] define the terminology to be used when talking about privacy by data minimization.[2] From a legal point of view, to the best of our knowledge, there is no specific European Directive which covers smart metering data privacy. Thus, only the general European Directive, EU Data Protection Directive 95/46/EC [6], would cover these types of data. However, the German Federal Office for Information Security developed the Protection Profile for the Gateway of a Smart Metering System; closely related to this, Stegelmann and Kesdogan [19] propose an architecture called GridPriv that includes a non-trusted k-anonymity service for pseudonymised meter data.

Siddiqui et al. [18] make an overview of some of the proposed solutions towards preserving privacy in the smart grid and divide these into the following categories: *anonymous credentials, third party escrow mechanisms, load signature moderation, smart energy gateway and privacy-preserving authentication.*

Anonymous credentials are based on blind signatures (similar to the ones used in the e-cash payment systems) and have the advantage to offer privacy protection against both DSO and third parties. The disadvantage of this solution is that it does not provide availability of billing data and it can only be used for pre-paid energy.

Third party escrow mechanisms [1, 7, 20] require the presence of a trusted third party entity whose role is to anonymize the data collected from the customers and then present it to the DSO or to aggregate the data and present it in an anonymized form. As mentioned in Section 2.1, Efthymiou and Kalogridis [7] present a solution based on separation of data into attributable low-frequency data, collected seldom and mainly used for billing, and anonymized high-frequency data, collected very often and used for grid operation. Each of these will be reported using a different pseudonym (one public and one private) and only the trusted third party is supposed to know the connection between the anonymous pseudonym and the public one. Their solution offers privacy protection against other third parties and also provides availability for billing data. The open question that remains is if the DSO can later recreate low-frequency data from the high-frequency and match it with the already available low-frequency data and so breaking the privacy. We will return to this question in Section 3.

Load signature moderation [10, 11] is a good privacy preserving method that can be used by customers. It requires the presence of an energy storage facility at the customer premises, such as an old battery from an electrical vehicle. The customer can then even out her external load signature by drawing energy from the battery in the high-load periods or by charging it during the low consumption periods or when energy is cheaper. This method offers protection both against DSOs and other third parties and also provides availability of billing data, because the Smart Meter will register only the energy used from the electricity network. However, the method has the disadvantage of requiring extra hardware.

The last two categories proposed by Siddiqui et al. [18] are *smart energy gateway* and *privacy-preserving authentication.* In the same way as load signature moderation, these also require the presence at the customer premises of a ded-

[2]The terminology includes: anonymity, unlinkability, linkability, undetectability, unobservability, pseudonymity, identifiability, identity, partial identity, digital identity and identity management.

icated system. In the first case the system is responsible to manage data released from the smart meter on some internal rules based on the data requester, while in the second case its role is to create trusted pseudo-identities that are used in requesting different energy amounts. In the first case privacy protection and availability of billing data can be enforced by setting up proper rules; the second one can only be used in a pre-paid energy scenario.

Hiding in the crowd is another method used to preserve privacy. Borges et al. [2] present a solution based on anonymity networks in which a customer uses two different identities to send his billing data and grid-operational data. While the billing data is directly attributable to him, the grid-operational data is forwarded to the DSO through an anonymity network, so the customer cannot be directly identified in a group of customers from the same network.

Data aggregation can also be used as a privacy-preserving solution. Before data is aggregated, one initial step in order to prevent unlawful disclosure of information is to perform mutual authentication [22, 23] between the entities involved in the process. Following this, privacy against the DSO and third parties can be obtained by using homomorphic cryptography [2, 12, 13, 23], or by adding random noise from a known distribution of zero mean [12, 13], but unfortunately aggregating methods do not provide availability of billing data and techniques based on homomorphic cryptography can be expensive on devices with reduced processing power and low resources such as the currently deployed smart meters.

Privacy enhancing techniques should also be resistant to attacks. Jawurek et al. [9] present the problem of breaking smart meter privacy by using de-pseudonymization. They propose a framework based on machine learning with support vector machines for the analysis of consumption traces and tracking consumption traces across different pseudonyms by using two linking procedures. Linking by Behaviour Anomaly (LA) tries to link a real ID to a consumption trace or two consumption traces together by correlating anomalies that happen in the same time, for example consumption spikes or blackouts. Linking by Behaviour Pattern (LB) tries to link different pseudonyms for one consumer and their method can be applied even if the consumption profiles do not overlap in time. In this paper we show that even simpler functions may also work quite well in identifying customers.

Buchmann et al. [3] show that identification of individual houses based on their energy-consumption records is possible even by using simple statistical tools such as means and standard deviations on a reduced number of data features. They show that 68% of the records coming from a set of 180 houses can be re-identified by using these simple methods.

So far we presented an overview of the current literature regarding privacy in the smart grid context. Next we will present the research papers that are close connected to our work.

Related work especially relevant to this paper:

Out of the presented papers above, the ones that are most closely related to ours are [3, 4, 7, 9]. Efthymiou and Kalogridis [7] set up the terminology on which we build our framework. Their solution is based on a trusted third party that takes care of the private IDs used in the process of high-frequency data anonymization and also of the connections between the high-frequency ID and the low-frequency one. Jawurek et al. [9] present a de-pseudonymization framework based on machine learning and are focusing mainly on anomalies in data consumption that happen in the same time. For their solution, fine-grained data is required, because such anomalies can be missed if aggregated daily or monthly values are used. Compared with their solution, we are focusing mainly on aggregated consumption where we try to identify uniqueness. Buchmann et al. [3] use simple statistical tools on a reduced number of consumption features and also on external information sources such as physical observation of people habits. Focusing on demand-response schemes, Cárdenas et al. [4] present the problem of appropriate sampling intervals in AMI as a trade-off between keeping a good level of customer privacy and gains in the demand-response scheme properties. They focus on the economics behind this problem as a parameter into the proper sampling scheme.

2.4 Advanced Metering Infrastructure data characteristics and problem formulation

Summarising regarding the AMI data characteristics on the two types of active energy consumption data reported by the smart meters in AMI, high-frequency (HF) data and low-frequency (LF) data, the question that arises is how these data should be reported and gathered in order to keep an adequate level of privacy against both the DSO and third parties? The level of privacy is measured as a reduced number of uniquely identifiable customers based on these two types of reported data.

There are a number of questions to which the research community tries to find the answers:

- Can customers be identified based on their energy consumption reported by the smart meters?
- How similar are customers with each other based on their energy consumption trace?

There are three characteristics of these data that were identified in the literature that determine the privacy level: number of *pseudonyms* for the same customer used in reporting/storing data, the *timespan* of data stored by the utility provider and the *granularity* of reported/stored data. The investigation presented in Section 4 focuses on the last two of these characteristics and on their role in making the datasets more resilient to the de-anonymization process.

Reporting high-frequency data under different pseudonyms and making sure that connections between pseudonyms are extremely hard to find and/or known by only a trusted third party [2, 7, 12] have been proposed earlier in the research literature. Using one pseudonym can be useful, if the connection between this pseudonym and the real customer ID is secret, but reporting or storing data from the same smart meter under different pseudonyms for shorter timespans can

be very efficient [2]. Although useful, generating multiple pseudonyms can be expensive for the smart meter device, because they need to create them through the use of a cryptographic algorithm, or they need to be provided when shipped from the factory.

The timespan of data stored is also very important, because longer periods of stored data for a smart meter (under the same pseudonym) can infer much more information about the energy consumption that took place. The question here is what the window for stored data that is useful for billing/grid operation is but which is also, at the same time, privacy preserving?

The last characteristic taken into consideration is the granularity of reported and stored data. Low-frequency data must be reported in fine-grained detail for accurate billing and to prevent fraud. Customers want, naturally, only to be billed for what they consumed, and the utility company wants to know exactly how much is consumed in order to level production and to better operate the grid. Unfortunately, loss occurs in the distribution grid due to transformers and old equipment, and are taken into consideration [5]. The question is whether the reported high-frequency data can be altered in a minor way such that the modification will not affect the grid operation, but making it hard to identify each customer uniquely by, for example, making the data from different customers more uniform? Figure 2 presents these three characteristics in relation to the adequate privacy level that is desired.

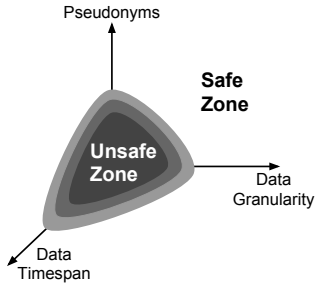

Figure 2: Characteristics of AMI data

3. METHODOLOGY

We will now formally describe the de-anonymization process by linking a *low-frequency dataset* with a *high-frequency dataset*. We will also present a threat scenario featuring an adversary which attempts to uniquely identify as many customers as possible and learn as much as possible, for example about their habits and living conditions, using the information from the high-frequency dataset.

Based on this scenario, in Section 4 we will conduct an investigation using a large real dataset on which we will study the influence of data timespan and granularity in the de-anonymization process. Our methodology can be used to better understand the limits of what is safe and what is not with regard to releasing datasets to third parties.

3.1 Formal framework

Assume that there exists a dataset, $\mathcal{C} = \{(identifier, timestamp, value)\}$, collected from the smart meters in an advanced metering infrastructure. This dataset contains identifiers ($identifier$) that can be used to identify individual customers, as well as high-frequency data ($value$) of the form described in Section 2.1, each marked with a specific $timestamp$. As mentioned, the high-frequency data can be used to infer habits of households.

There are two functions, $f_H(\cdot)$ and $f_L(\cdot)$ such that we can derive two new datasets by letting

$$\begin{cases} \mathcal{H} &= f_H(\mathcal{C}) \\ \mathcal{L} &= f_L(\mathcal{C}) \end{cases}$$

where $\mathcal{H}$ and $\mathcal{L}$ are related but have slightly different properties. In a scenario within the smart grid, $\mathcal{H}$ would be a dataset with, for example, the originally collected high-frequency data but where all customer identification would be replaced with untraceable labels (one simple way to obtain untraceable labels is to use a random label generator and check for possible collisions). This dataset could then be used for grid operation and optimization as it would not be possible to use it to identify individual customers. The set $\mathcal{L}$, on the other hand, would retain the original identifiers making it possible to identify customers but instead the data in this dataset would be aggregated (under the original identifier) so as to be less privacy invasive. This dataset could then be used for billing of monthly consumption, for example. The complete dataset, $\mathcal{C}$ is then discarded.[3]

We further assume that finding f_H^{-1} and f_L^{-1} is intractable, as information is deliberately discarded in each transform. Thus, if an adversary obtained either $\mathcal{H}$ or $\mathcal{L}$ it would be difficult to recreate $\mathcal{C}$ and each dataset in isolation would not be interesting. This is similar in vein to the indirect assumptions for the solution presented by Efthymiou and Kalogridis [7].

However, as $\mathcal{H}$ and $\mathcal{L}$ originate from the same dataset, we assume that there exists another function, $g(\cdot)$, such that $\mathcal{H}'\mathcal{L}' = g(\mathcal{H})$.[4] The data in $\mathcal{H}'\mathcal{L}'$ would retain the identifying labels from $\mathcal{H}$ and be aggregated in a similar fashion to $\mathcal{L}$. If we could then link any entries between these two datasets, $\mathcal{H}'\mathcal{L}' \sim \mathcal{L}$, we might partially be able to recreate $\mathcal{C}$ by relabeling the entries in $\mathcal{H}$. The problem, though, is that many of the aggregated values in $\mathcal{L}$ may not be unique but would be the same across a number of customers meaning that we cannot easily infer which labels should be linked as there will be a set of possible matches. Intuitively, we would expect customers with a very uncommon behavior to maybe be re-identified but that a majority of customers would belong to clusters that behave in a similar nature and thus not be uniquely identified.

Formally, the question we would like to answer is whether it is at all possible to link these two datasets, given a large realistic scenario. If so, we want to measure how well $\mathcal{C}$ can be

[3]The complete dataset, $\mathcal{C}$, might not ever exist if the transforms are run continuously in the smart meters.

[4]The existance of $g(\cdot)$ would depend on how $f_H(\cdot)$ and $f_L(\cdot)$ were constructed. Based on our survey of existing methods, we say that it is likely $g(\cdot)$ exists.

recreated and provide boundaries on what an adversary can achieve if she has access to both $\mathcal{H}$ and $\mathcal{L}$. Can we limit the information gained by the adversary by changing the properties of either of these datasets, for example by using more pseudonyms or storing less data as discussed in Section 2.4?

3.2 Adversarial strategy

A possible adversarial strategy algorithm, that is also implied by the previous literature, is presented in Algorithm 1, and it is used to derive the associated adversary model. The adversary gets hold of two sets of data, one containing high-frequency ($\mathcal{H}$) data and one containing low-frequency data ($\mathcal{L}$) with the properties described in Section 3.1. The individual smart meters that produced these data are labeled differently in these two datasets; to simplify the presentation, we assume each smart meter has only one identifier in each of the sets, being equivalent to using only one pseudonym. The algorithm can easily be extended with sets of more pseudonyms.

Algorithm 1: Adversarial strategy algorithm

```
Requirement: adversary has obtained H and L
Goal: recreate as much as possible of C
begin algorithm
    create H'L' = g(H);
    while IDlink = findLink(H'L', L) do
        recreate one entry in C;
        remove identified trace from H'L' and L
    end
end
begin function findLink
    /* Version 1:  find *unique* consumption traces
       in a time period in L                          */
    foreach timeperiod j in L do
        if any unique consumption traces exists then
            extract identifying ID from L;
            find corresponding entry in H'L';
            extract identifying ID' from H'L';
            return <ID, ID'>;
        end
    end
    /* no more links can be made                      */
    return false;
end
```

As stated above, we assume the adversary wants to be able to recreate the dataset $\mathcal{C}$, where she can label the high-frequency data with the identity of the individual customers from the low-frequency dataset. By analyzing the low-frequency datasets, she tries to find as many "unusual" customers as possible, i.e. customers that at some point in time have data values that differ from the norm so that she can create a link between $\mathcal{H}'\mathcal{L}'$ and $\mathcal{L}$. In the algorithm, this is performed in the function `findLink()`. This analysis can range in its sophistication. In the first version of the algorithm, we have chosen to implement a method that only looks for unique values in a time period to show that even a relatively simple and fast analysis can be surprisingly efficient. As we will show in Section 3.3, the simplicity of the function also allows us to model it as a game of *balls and bins* so that we can estimate the probabilities of the success of the scenario.

Note that our discussion so far has been of a general nature; the datasets can contain a diverse set of data as described in Section 2.1. However, in the following we are going to concentrate on consumption traces. These types of datasets are often used in the literature (please see Table 2 for an overview).

3.3 Probabilistic framework and analysis

In this section, we model the adversarial strategy algorithm in a probabilistic framework to be able to reason formally about the adversary's capabilities and possibilities of success in de-anonymizing customers. Given the properties of the `function findLink()` shown in Algorithm 1 it is possible to model the algorithm as a *game of balls and bins* [15]. In the following discussion, we assume the datasets contain energy consumption data e.g. kWh consumption indexes.

The energy index data from m_j smart meters (balls) in one time period, $j \in T$, can be sorted into a set of n different intervals (bins), where the width of the bins corresponds to a range of energy consumption units (multiples of kWh). We let the width of the bins be an integer, w, that can vary from 1 to W. The number of bins is then $n = \frac{M}{w}$, where $M = max(m_T)$ is the maximum index consumption value for all the time periods considered. At each round, the number of balls in all the bins is equal with the number of balls at the beginning of the round e.g. $\sum_{i=1}^{n} m_{w_i} = m_j$.

Any ball that falls alone in a bin is considered to be uniquely identified and it is removed. This is then repeated; each round of the game uses data from a time period where index data for the m_j smart meters exist. The game ends when either all the balls are removed or when there exists no more time periods with new data. The percent of eliminated balls at the end of the game is then equivalent to the percent of uniquely identifiable consumption indexes from that specific dataset.

In the analysis of the bins and balls game presented by Mitzenmacher and Upfal [15] the probability that a bin receives a number of r balls when m balls are thrown independently and uniformly at random into n bins is given as a Poisson distribution of mean $\frac{m}{n}$.

$$P[a\,bin\,has\,r\,balls] = \frac{e^{-\frac{m}{n}} \times (\frac{m}{n})^r}{r!} \tag{1}$$

Note that we assume a Poisson distribution of the balls into bins. This will be further discussed in Section 4.2. In our specific case with $r = 1$ (only one ball per bin so that we can identify the customer), the probability above becomes the following.

$$P[a\,bin\,has\,1\,ball] = e^{-\frac{m \times w}{M}} \times \frac{m \times w}{M} \tag{2}$$

For the Poisson case, the number of balls in each bin must be independent random variables. In our case, the number of balls in the last bin is known as we have m balls and we

know the number of balls in the first $n-1$ bins. Corollary 5.9 from [15] states the following.

COROLLARY 3.1. *[15] Any event that takes place with probability p in the Poisson case takes place with probability at most $pe\sqrt{m}$ in the exact case.*

Thus, any event that happens with a small probability in the Poisson case also happens with a small probability in the exact case, where balls are thrown into bins [15] and justifies the Poisson analysis for the bins and balls game. The *expected number of bins with only* 1 *ball* becomes the following.

$$E[bins\,with\,1\,ball] = e^{-\frac{m\times w}{M}} \times m \tag{3}$$

Let the number of balls available at the beginning of the game be m_0. If we consider two consecutive rounds in the game, the expected number of balls m_j at the beginning of round j can be computed as:

$$m_j = m_{j-1} - E_{j-1}[bins\,with\,1\,ball] \tag{4}$$

where if we substitute the expression for the expected value:

$$m_j = m_{j-1} \times (1 - \exp(-\frac{m_{j-1}\times w}{M})) \tag{5}$$

The game ends when either all balls have been removed from the game ($m_j = 0$) or all the time periods with available data, T, have been used ($j > T$).

The adversary would win the game when the percentage of extracted balls is above a specific threshold, λ, meaning that a large percentage of the smart meters have been identified uniquely ($\frac{\sum_{1\le j\le T} m_j}{m_0} > \lambda$). The utility company wins the game when the percentage of uniquely identified smart meters is low (m_T is very close to m_0). By investigating the parameters (m, λ, w), we can explore the limits of the capabilities of the adversary to make sure that she cannot identify a large set of customers.

4. EVALUATION STUDY

As mentioned in Section 2.4, the *granularity* and *data timespan* play an important role in the data de-anonymization process. We investigate these two characteristics by using a *simulation* based on the probabilistic framework presented in Section 3.3 and an *evaluation* based on a dataset, described in Section 4.1. We expect to identify the influence of the characteristics in the context of the adversarial strategy algorithm presented in Section 3.2. The last part of this section presents a discussion of the results obtained.

4.1 Description of the dataset

To see how well the adversarial strategy algorithm works in a real setting we use a dataset consisting of smart meter readings from a large number of consumers in a medium-sized city. The original data have hourly smart meter index readings for a period of seven non-contiguous months. The data originates from a range of smart meters serving very small consumers (summer cottages) to large consumers (industrial customers). The data have gone through a two-step anonymization process, once by the utility provider and once by us, to make sure it is not possible to physically identify any customers in the set. Each record has the $<ID_{anon}$, *timestamp*, *value*$>$ format. The *timestamp* and the index *value* remain in clear. The data from each smart meter in the set can be identified by a unique numerical identifier (ID_{anon}), that remains the same over the seven months. This is equivalent of having a single pseudonym for each of the smart meters for the whole time period.

Dataset	Number of meters	Number of readings
Kalogridis et al. [7]	N/A	N/A
Jawurek et al. [9]	53	281,112
Buchmann et al. [3]	180	60,480
Daisuke and Cárdenas [14]	108	*1,890,000
Tudor et al. (this paper)	19,334	99,355,998

Table 2: Datasets from AMI
The * value was estimated based on values in [14]

As the data comes from a real AMI, where problems with missing values sometimes exist, we also sanitized the data by creating a smaller set where we removed a number of collection artifacts. Mainly, we removed any smart meters that had gaps in the hourly reporting (values lost), double conflicting records for the same timestamp or decreasing index values for increasing timestamp values.

After the sanitization process, the dataset contained 19,334 unique smart meters with 99,355,998 hourly energy consumption readings. This set is considered to be the high-frequency dataset ($\mathcal{H}$). From this dataset, we then created the low-frequency dataset ($\mathcal{L}$), which is similar for each customer to the energy consumption values printed in the electrical bill. This resulted in 4,156,810 daily values and 135,338 monthly values. As can be seen from Table 2, our dataset is significantly larger than the ones previously used in literature.

4.2 The Poisson distribution assumption

In the game of bins and balls presented in Section 3.3 we assume that the balls are thrown independently and uniformly at random so that they can be modeled as a Poisson distribution.

The balls signify specific smart meters. These smart meters belong to the same households with the same number of people with habits that will probably not change on a monthly basis. Regardless of what bin a ball lands in for a round, the theoretical model assumes that it is equally likely that the ball falls into any of the bins in the next round. However, in the real case it is likely that the energy consumption pattern would be somewhat similar across months, so that it is more likely that the ball falls into a bin close to the bin from the last month. We say that the balls in the real case are somewhat *sticky* as they tend to favor, across months, bins that are closely located.

This also implies that if two balls fell into the same bin one month, it is likely that they will do so also the following month in the real case. For that reason, we expect that the

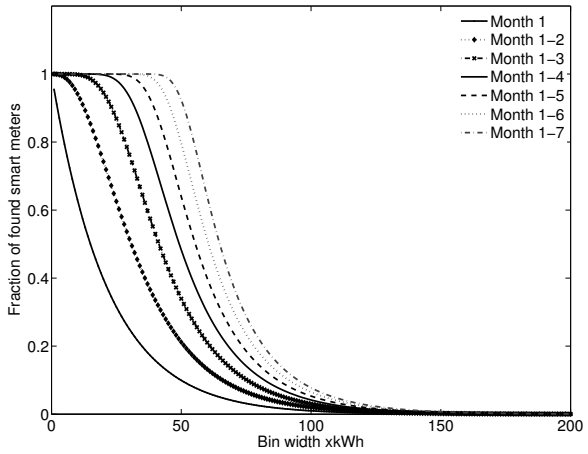

Figure 3: Fraction of unique smart meters - seven months of data - estimation case

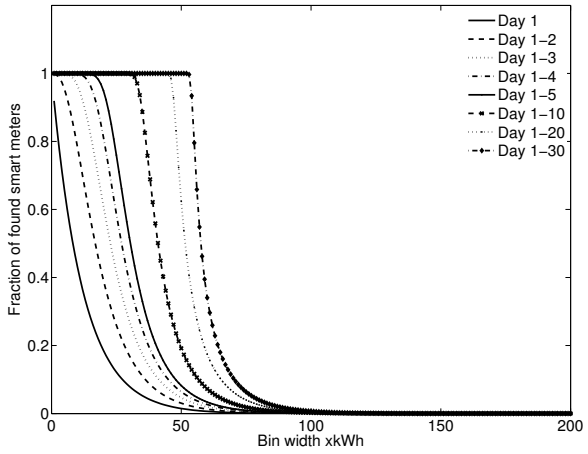

Figure 4: Fraction of unique smart meters - 30 days of data - estimation case

identification process, using `function findLink()` in Algorithm 1, will be more difficult in the real case compared to the probabilistic model.

Furthermore, the customers may not be divided uniformly at random across the consumption bins. A typical household in Europe hosts about 2.3 people[5]. Our dataset reflects a large number of such domestic customers, agglomerated in the low consumption zone, while the probabilistic model assumes a random spread in the bins. This also strengthens our expectations regarding the difficulty to identify a majority of the customers in the real case.

Even with these assumptions, the formal framework let us reason about the characteristics of the AMI datasets and their influence on the de-anonymization process.

[5]`http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_lvph01&lang=en`

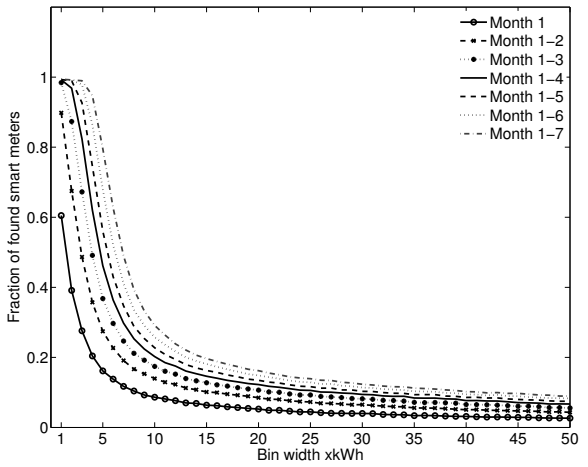

Figure 5: Fraction of unique smart meters - seven months of data - dataset case

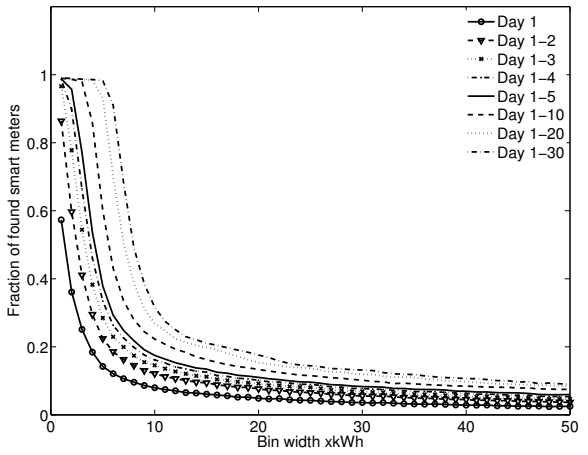

Figure 6: Fraction of unique smart meters - 30 days of data - dataset case

4.3 Results from the probabilistic framework

The *estimation* run of the adversarial strategy algorithm is based on the formulas presented in Section 3.3 with the parameters adapted to match the dataset of real consumption traces. The initial number of smart meters is selected to be 19,334 in order to exactly match the number present in the dataset, and the number of rounds in the estimation is the same as the number of time periods in the dataset – seven for the monthly case and 30 for the daily case. The granularity is varied from 1 to 200 kWh. The value for M is selected to be the same as in the dataset and is the highest index value for the time periods considered. The expected number of identified smart meters (balls) at each round is computed by using Relation (3), while Relation (4) is used to compute the number of remaining smart meters (balls) at the beginning of each round.

Figure 3 presents the fraction of uniquely identified smart

meters in the monthly case (seven time periods) obtained in the estimation, while varying the granularity with which energy consumption index is reported and the available time periods. Similarly, Figure 4 presents the estimation results for the daily case, using 30 time periods (only periods 1-5, 10, 20 and 30 are presented in the figure).

4.4 Results of the adversarial strategy algorithm

An *evaluation* of the effectiveness of the adversarial strategy algorithm is performed, for two characteristics presented in Section 2.4, data granularity and timespan. The starting number of smart meters is 19,334 and the number of monthly values is 135,338. There are seven time periods, equivalent to the seven months of data and the maximum index value is computed from the dataset.

Figure 5 presents the fraction of the uniquely identified smart meters by using the dataset presented in Section 4.1 for the monthly case, while varying the granularity of the reported energy consumption index from 1 to 200 kWh. The figure presents only the results from 1 to 50 kWh, after this point the values continue on what seems to be a linear trend.

The simulation is repeated by using daily datasets, the equivalent of one month of recordings (30 days - 593,830 values), where the first month of data is used. Similarly to the monthly case, the granularity is varied between 1 and 200 kWh. Figure 6 shows the results based on the dataset for a granularity between 1 and 50 kWh.

4.5 Dicussion of results

Table 3 presents the expected number of uniquely identified smart meters at each round of the simulation and the evaluation of the adversarial strategy algorithm in the monthly case for a granularity of 1 kWh. The difference between the simulation and the evaluation is that in the case of the simulation more smart meters are identified in the first round compared with the case of the evaluation based on the dataset. This makes the identification in the next rounds easier because a smaller number of smart meters needs to be divided into bins so the probability of having more than one smart meter in a bin decreases. This result can be explained through the assumptions that are discussed in Section 4.2. For the 1 kWh monthly case, the estimation ends after three rounds, when all the smart meters are identified. In the evaluation case, the algorithm runs for all the seven rounds, and 125 smart meters remain unidentified at the end, but the percent of identified smart meters is still above 99%.

Table 4 holds the results for the 10 kWh monthly case and we can observe that in the simulation case the algorithm took one more round compared to the 1 kWh case, but the percent of identified smart meters at the end is still 100%. The evaluation results for the 10 kWh monthly case show that the number of unidentified smart meters after all the rounds is 13,706 and the percent of identified smart meters is 29.1%. This is a much better result than for the 1 kWh monthly case.

Figure 5 shows that varying the granularity under which

	Newly found smart meters		**Total found smart meters %**	
Time period	**Simulation**	**Evaluation**	**Simulation**	**Evaluation**
m_1	18461	11698	95.4%	60.5%
m_2	871	5655	99.9%	89.7%
m_3	2	1669	100 %	98.3%
m_4	0	155	100 %	99.1%
m_5	0	11	100 %	99.2%
m_6	0	11	100 %	99.3%
m_7	0	10	100 %	99.3%
Total	19334	19209	100 %	99.3%

Table 3: Expected number of identified smart meters for a reporting granularity of 1 kWh

data are reported can drastically reduce the fraction of identified smart meters. For example, reporting the index without the last digit, at a 10 kWh scale, can reduce in this case the percent of identified smart meters at under 10% for one period and under 30% for all periods. This result justifies a reporting scheme in which electrical energy consumed is rounded to the next 10 kWh value, before it is reported and billed, instead of being reported with 1 kWh accuracy. This will provide a good and cheap anonymity solution for the other 70% of the customers, in the case that everyone opts for such a reporting scheme. The same result can be observed in Figure 6 where for the daily reporting with 10 kWh granularity, the percent of identified customers is brought down to almost 10% for one period and to almost 40% for all periods. This simple reporting solution offers a better degree of privacy, but it may still not be feasible in regions where the law requires that energy reporting should be done with kWh accuracy.

We can see that high-frequency datasets contain so much information so that the re-identification process is possible even with simple means. In our analysis we have assumed that the adversary would have access to the complete high-frequency dataset and the complete low-frequency dataset.

The evaluation results show that reporting energy consumption indexes with kWh accuracy makes the datasets prone to re-identification, because a large percent of the customers can be identified uniquely, solely based on their energy consumption. The results closely tie together the granularity and the timespan of the data and show their common effect in the re-identification process. They show that reducing the granularity used for reporting consumption data can be a very simple and beneficial solution that increases the privacy level of the datasets. Results from Tables 3 and 4 strengthen this assumption and show a significant decrease of the percent of uniquely identified smart meters from 99.3% to 29.1%, for a decrease of granularity from 1 kWh to 10 kWh. Also, Figures 5 and 6 show that further reduction of the granularity may significantly reduce the percent of identified customers, making the datasets more resilient to the de-anonymization process.

As a general consideration, data timespan and granularity should be taken into consideration before releasing any AMI

	Newly found smart meters		Total found smart meters %	
Time period	**Simulation**	**Evaluation**	**Simulation**	**Evaluation**
m_1	12182	1670	63.0%	8.6%
m_2	6029	1027	94.1%	13.9%
m_3	1093	671	99.8%	17.4%
m_4	30	543	100 %	20.2%
m_5	0	487	100 %	22.7%
m_6	0	579	100 %	25.7%
m_7	0	651	100 %	29.1%
Total	19334	5628	100 %	29.1%

Table 4: Expected number of identified smart meters for a reporting granularity of 10 kWh

consumption data to third parties, as these two characteristics greatly influence the anonymity of the datasets.

5. CONCLUSION

It is almost unquestionable that the smart grid will produce more and more data regarding the electrical energy consumed and the well-being of the electrical grid. Harnessing and processing these large quantities of data will make the electrical grid more resilient to faults, provide a better balance between the production and the consumption, but as we saw, these datasets also raise privacy concerns. In this paper we presented an overview of research regarding smart grid data privacy. We constructed a formalization of the problem of de-anonymizing AMI data by matching two different types of smart metering datasets. We take into account two main properties of smart metering data: the granularity of the data reported and its timespan. We argue that these two, together with the number of pseudonyms used in the reporting process play a significant role in a three-way balance towards obtaining better customer anonymity. We consider a class of adversarial strategies that can be formulated as combinatorial and probabilistic problems and used it to evaluate characteristics of these datasets (granularity and timespan) in an investigation process towards better resilience against the de-anonymization process; our results show that this process should be taken into consideration before releasing AMI datasets. Future research directions include extending the theoretical framework and the adversarial strategy model and also to be able to limit the theoretical maximum number of customers that can be identified. Related research issues refer to billing models; it is interesting to investigate the possibilities and limitations in managing trade-offs between customer incentives for improving their usage of electricity and privacy issues regarding the data in the billing system, as these two imply different needs in the granularity of the data.

6. ACKNOWLEDGEMENTS

This work has been partially supported by the European Comission Seventh Framework Programme (FP7/2007-2013) through the SysSec Project, under grant agreement 257007, through the FP7-SEC-285477-CRISALIS project and through the collaboration framework of Chalmers Energy Area of Advance project "SN7: Algorithms for Adaptiveness and Robustness in Electricity Networks".

7. REFERENCES

[1] J.-M. Bohli, C. Sorge, and O. Ugus. A privacy model for smart metering. In *Communications Workshops (ICC), 2010 IEEE International Conference on*, pages 1 –5, may 2010.

[2] F. Borges, L. Martucci, and M. Mühlhäuser. Analysis of privacy-enhancing protocols based on anonymity networks. 2012.

[3] E. Buchmann, K. Böhm, T. Burghardt, and S. Kessler. Re-identification of smart meter data. *Personal and Ubiquitous Computing*, 17(4):653–662, 2013.

[4] A. A. Cárdenas, S. Amin, and G. Schwartz. Privacy-aware sampling for residential demand response programs. 2012.

[5] J. Dickert, M. Hable, and P. Schegner. Energy loss estimation in distribution networks for planning purposes. In *PowerTech, 2009 IEEE Bucharest*, pages 1–6, 2009.

[6] E. Directive. 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the EC*, 23:6, 1995.

[7] C. Efthymiou and G. Kalogridis. Smart grid privacy via anonymization of smart metering data. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 238 –243, oct. 2010.

[8] R. Hoenkamp, G. B. Huitema, and A. J. de Moor-van Vugt. The neglected consumer: the case of the smart meter rollout in the netherlands. *Renewable Energy Law and Policy Review*, 2011(4):269–282, 2011.

[9] M. Jawurek, M. Johns, and K. Rieck. Smart metering de-pseudonymization. In *Proceedings of the 27th Annual Computer Security Applications Conference*, pages 227–236. ACM, 2011.

[10] G. Kalogridis, R. Cepeda, S. Denic, T. Lewis, and C. Efthymiou. Elecprivacy: Evaluating the privacy protection of electricity management algorithms. *Smart Grid, IEEE Transactions on*, 2(4):750 –758, dec. 2011.

[11] G. Kalogridis, C. Efthymiou, S. Denic, T. Lewis, and R. Cepeda. Privacy for smart meters: Towards undetectable appliance load signatures. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 232 –237, oct. 2010.

[12] K. Kursawe, G. Danezis, and M. Kohlweiss. Privacy-friendly aggregation for the smart-grid. In *Privacy Enhancing Technologies*, pages 175–191. Springer, 2011.

[13] F. Mármol, C. Sorge, O. Ugus, and G. Pérez. Do not snoop my habits: preserving privacy in the smart grid. *Communications Magazine, IEEE*, 50(5):166 –172, May 2012.

[14] D. Mashima and A. A. Cárdenas. Evaluating electricity theft detectors in smart grid networks. In *Research in Attacks, Intrusions, and Defenses*, pages 210–229. Springer, 2012.

[15] M. Mitzenmacher and E. Upfal. *Probability and*

computing: Randomized algorithms and probabilistic analysis. Cambridge University Press, 2005.

[16] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin. Private memoirs of a smart meter. In *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building*, pages 61–66. ACM, 2010.

[17] A. Pfitzmann and M. Hansen. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. *URL: http://dud. inf. tu-dresden. de/literatur/Anon_Terminology_v0*, 34, 2010.

[18] F. Siddiqui, S. Zeadally, C. Alcaraz, and S. Galvao. Smart grid privacy: Issues and solutions. In *Computer Communications and Networks (ICCCN), 2012 21st International Conference on*, pages 1 –5, 30 2012-aug. 2 2012.

[19] M. Stegelmann and D. Kesdogan. Gridpriv: A smart metering architecture offering k-anonymity. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, pages 419 –426, june 2012.

[20] S. Wang, L. Cui, J. Que, D.-H. Choi, X. Jiang, S. Cheng, and L. Xie. A randomized response model for privacy preserving smart metering. *Smart Grid, IEEE Transactions on*, 3(3):1317 –1324, sept. 2012.

[21] S. D. Warren and L. D. Brandeis. The right to privacy. *Harvard law review*, 4(5):193–220, 1890.

[22] Y. Yan, Y. Qian, and H. Sharif. A secure and reliable in-network collaborative communication scheme for advanced metering infrastructure in smart grid. In *Wireless Communications and Networking Conference (WCNC), 2011 IEEE*, pages 909 –914, march 2011.

[23] Y. Yan, Y. Qian, and H. Sharif. A secure data aggregation and dispatch scheme for home area networks in smart grid. In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pages 1 –6, dec. 2011.