Data privacy issues in the AMI

ICT Support for Adaptiveness and (Cyber)security in the Smart Grid
DAT300/DIT668

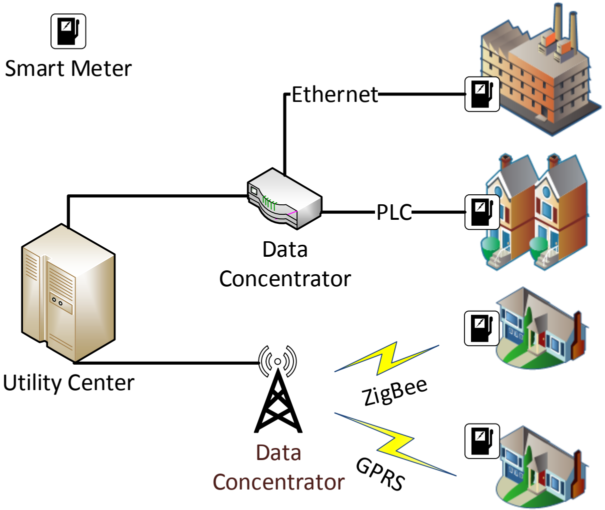
Valentin Tudor
Computer Science and Engineering Department
Chalmers University of Technology
Gothenburg, Sweden

22nd of September 2016

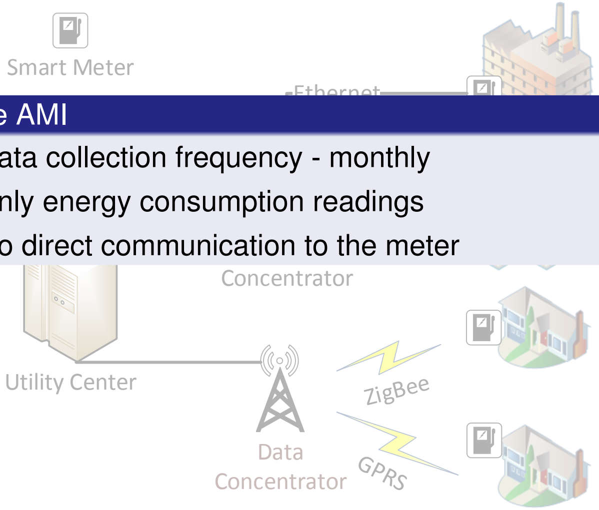
Outline

- The Advanced Metering Infrastructure (AMI)
- AMI data - utility and privacy
- Privacy issues - de-anonymization and de-pseudonymization
- Differential-privacy and AMI data
- AMI data application - energy load forecast using DP-aggregated AMI data
- Conclusion

The Advanced Metering Infrastructure (AMI)



Collecting energy related data



Before AMI

- data collection frequency - monthly
- only energy consumption readings
- no direct communication to the meter

Collecting data from AMI

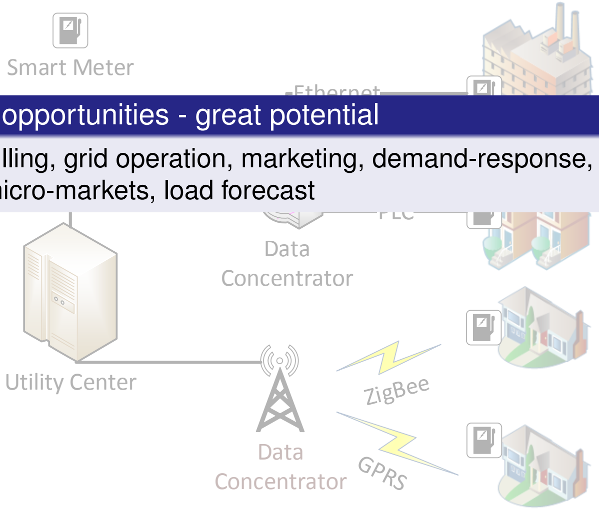
Before AMI

- data collection frequency - monthly
- only energy consumption readings
- no direct communication to the meter

With AMI

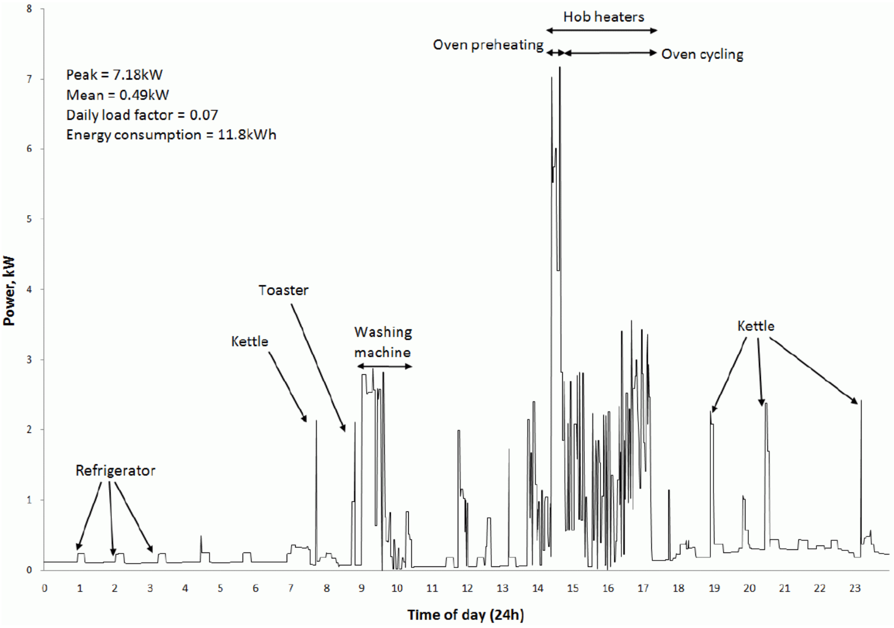
- data collection frequency - minutes or less
- energy consumption, power quality info., alarms
- two-way communication with the meter

Utilizing large data from AMI



Many opportunities - great potential

- billing, grid operation, marketing, demand-response, micro-markets, load forecast


Hob heaters
Oven preheating
Oven cycling
Peak = 7.18kW
Mean = 0.49kW
Daily load factor = 0.07
Energy consumption = 11.8kWh
Toaster
Kettle
Washing machine
Kettle
Refrigerator
Power, kW
Time of day (24h)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Utilizing large data from AMI

Many opportunities - great potential

- billing, grid operation, marketing, demand-response, micro-markets, load forecast

Privacy challenges

- fine-grained AMI data - sensitive information[a]
- privacy invasive techniques against anonymization [b,c,d]

---

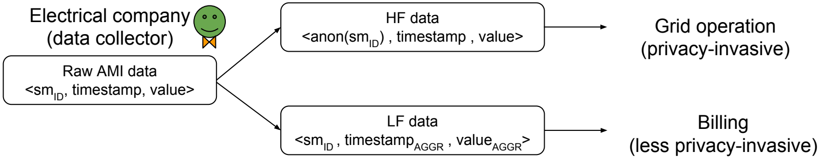
[a] Mármol, F.G., Sorge, C., Ugus, O., and Pérez, G. M. 2012. Do not snoop my habits: preserving privacy in the smart grid. In IEEE Communications Magazine, 50(5), pp.166-172.

[b] Jawurek, M., Johns, M. and Rieck, K., 2011, December. Smart metering de-pseudonymization. In Proceedings of the 27th Annual Computer Security Applications Conference (pp. 227-236). ACM.

[c] Tudor, V., Almgren, M. and Papatriantafilou, M., 2013. Analysis of the impact of data granularity on privacy for the smart grid. In Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society.

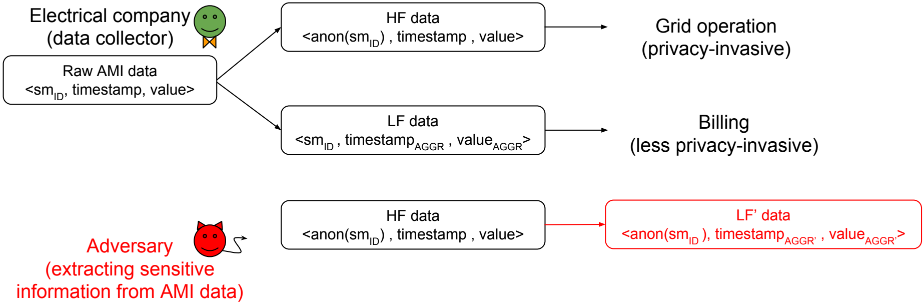
[d] Tudor, V., Almgren, M., and Papatriantafilou, M., 2015. A study on data de-pseudonymization in the smart grid. In Proceedings of the Eighth European Workshop on System Security

AMI data anonymization - remove PII[1] and aggregation

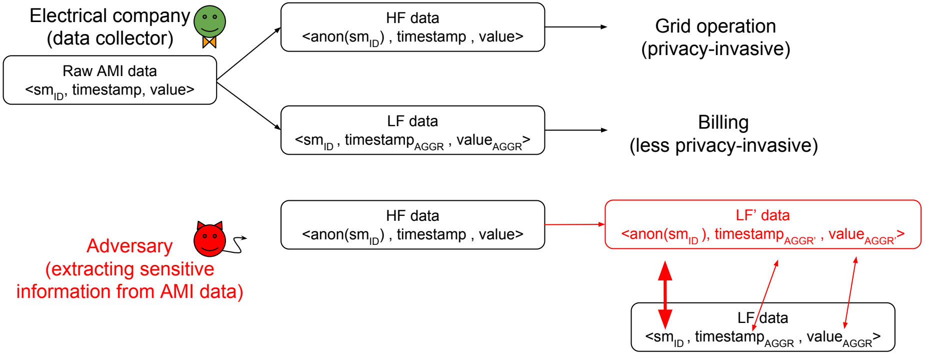


[1] PII - Personal Identifiable Information

De-anonymization



De-anonymization



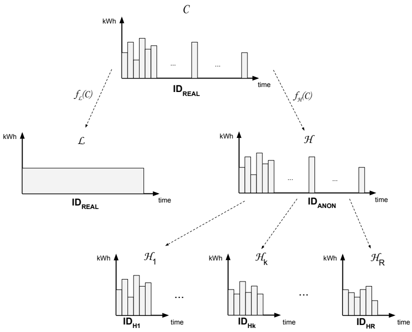
The adversary de-anonymizes AMI datasets using information extracted from AMI data itself.

A bit of formalism

$\mathcal{C} = \{(identifier, timestamp, value)\}$ - original dataset

There are two functions such that two new datasets can be derived

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



De-anonymization

$\mathcal{C} = \{(identifier, timestamp, value)\}$ - original dataset

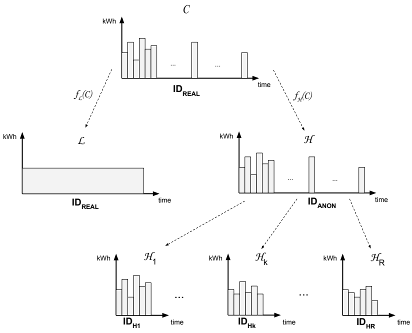
There are two functions such that two new datasets can be derived

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

An adversary is interested in matching the identities in $\mathcal{H}$ with $\mathcal{L}$.

There is a function $g(\cdot)$, such that $\mathcal{L}' = g(\mathcal{H})$.

Link entries $\mathcal{L}' \sim \mathcal{L}$.



De-pseudonymization

$\mathcal{C} = \{(identifier, timestamp, value)\}$ - original dataset

There are two functions such that two new datasets can be derived

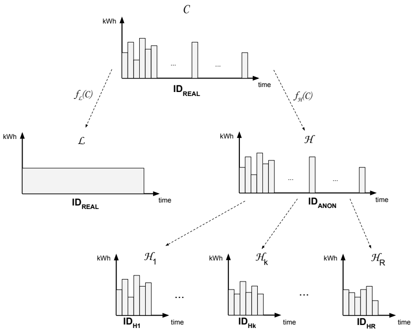
$$\begin{cases} \mathcal{H} & = & f_H(\mathcal{C}) \\ \mathcal{L} & = & f_L(\mathcal{C}) \end{cases}$$

An adversary is interested in matching the identities in $\mathcal{H}$ with $\mathcal{L}$.

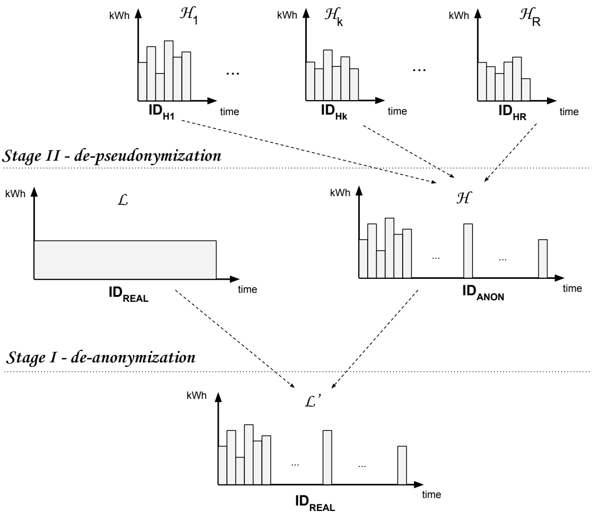
There is a function $g(\cdot)$, such that $\mathcal{L}' = g(\mathcal{H})$.

Link entries $\mathcal{L}' \sim \mathcal{L}$.

There is a function $r(\cdot)$, such that $\mathcal{H}_{k,k+1} = r(\mathcal{H}_k, \mathcal{H}_{k+1})$



Complete adversarial picture



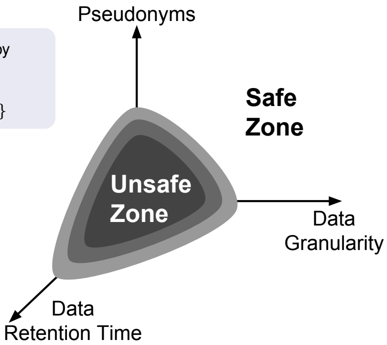
Road so far

- The Advanced Metering Infrastructure (AMI)
- AMI data - utility and privacy
- **Privacy issues - de-anonymization and de-pseudonymization**
- Differential-privacy and AMI data
- AMI data application - energy load forecast using DP-aggregated AMI data
- Conclusion

AMI data characteristics

Reporting sensitive data by using different identities

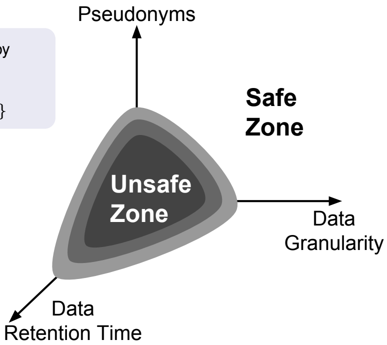
$$A \rightarrow \{A', A'', A'''\}$$



AMI data characteristics

Reporting sensitive data by using different identities

$A \rightarrow \{A', A'', A'''\}$



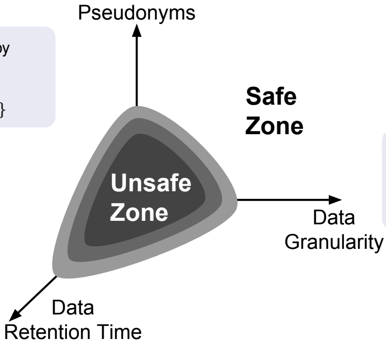
The time window for keeping a client's data

…days, *months*, *years*, …

AMI data characteristics

Reporting sensitive data by using different identities

$A \rightarrow \{A', A'', A'''\}$



The level of precision on which consumed energy is reported

Wh, kWh, $10^n \times kWh$, …

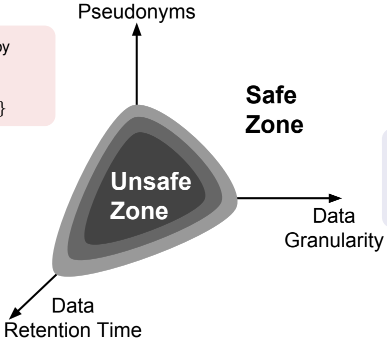
The time window for keeping a client's data

…$days$, $months$, $years$, …

AMI data characteristics

Reporting sensitive data by using different identities

$$A \rightarrow \{A', A'', A'''\}$$



The level of precision on which consumed energy is reported

Wh, kWh, $10^n \times kWh$, $\ldots$

The time window for keeping a client's data

$\ldots days$, $months$, $years$, $\ldots$

How do the properties of the **HF** datasets influence the effectiveness of utilizing pseudonyms?

De-pseudonymization problem

Linking together data produced by the same source (household), but stored under different pseudonyms.

We assume a scenario in which an adversary:

1. Gets hold of two HF datasets.
2. For each household computes a number of features based on the data in the two datasets.
3. Uses the features to link together the identities used in the two HF datasets.
4. For each correctly linked identity the adversary obtains an extended HF dataset.

De-pseudonymization problem

Linking together data produced by the same source (household), but stored under different pseudonyms.

We assume a scenario in which an adversary:

1. Gets hold of two HF datasets.
2. For each household computes a number of features based on the data in the two datasets.
3. Uses the features to link together the identities used in the two HF datasets.
4. For each correctly linked identity the adversary obtains an extended HF dataset.

What is the **effect** of the dataset **size** and collection **season** on the **de-pseudomynization ratio**?

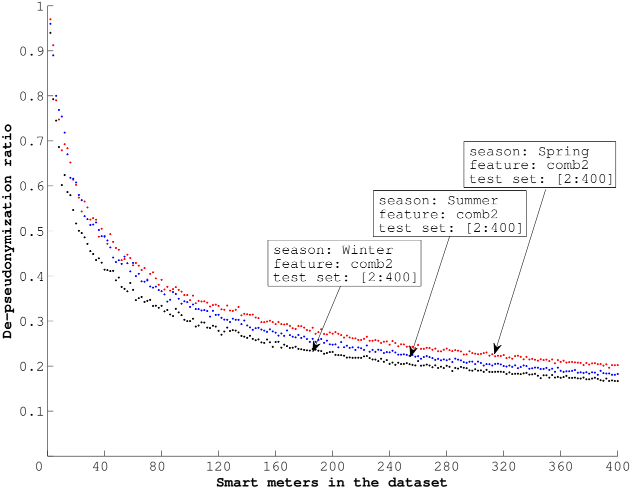
Choosing features

Statistical features that can be computed efficiently on HF data:

#	Feature name	Abbrev.	Description
1	Standard deviation	`std`	variation of energy consumption
2	Mode	`mode`	most common consumption value
3	Mean consumption	`meanc`	average energy consumption
4	Max consumption	`maxc`	maximum energy consumption
5	Coefficient of variation	`cv`	ratio of standard deviation to mean

Each household data record becomes a point in a multi-dimensional space $\rightarrow$ a distance metric can be used to find similar households.

Seasonal results - combined features `std, mode, meanc`



Random guessing

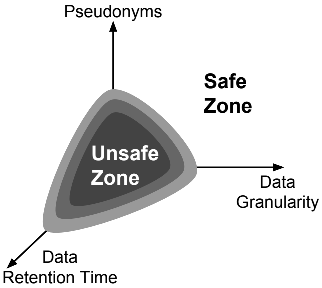
Assume that we have two HF datasets using different pseudonyms for the N households stored in each.

Probability to match:

- 1 pseudonym: $P[1 \text{ out of } N \text{ matched}] = \frac{1}{N}$
- 2 pseudonyms: $P[2 \text{ out of } N \text{ matched}] = \frac{1}{N} \times \frac{1}{N-1}$
- ...
- k pseudonyms:
  $P[k \text{ out of } N \text{ matched}] = \frac{1}{N} \times \frac{1}{N-1} \times \ldots \times \frac{1}{N-k+1}$
  $P[k \text{ out of } N \text{ matched}] = \prod_{l=1}^{k} \frac{1}{N-l+1}$
- N pseudonyms: $P[N \text{ out of } N \text{ matched}] = \frac{1}{N!}$

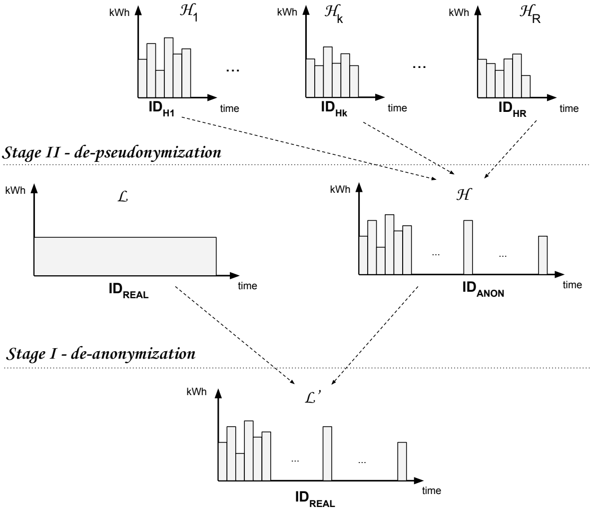
Random guessing might work for datasets with a small number of households, but it becomes harder as the size of datasets increases.

Summary - de-pseudonymization



- The **number of households** in the dataset and **collection season** influences the efficiency of the de-pseudonymization.
- The number of re-identified households is not proportional with the size of the dataset.
- The **characteristics** of the Advanced Metering Infrastructure **dataset** should be taken into consideration when evaluating or developing Privacy Enhancing Technologies for this domain.

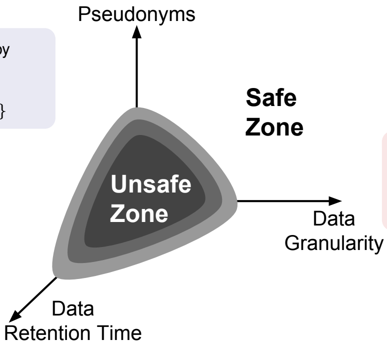
Complete adversarial picture



AMI data characteristics

Reporting sensitive data by using different identities

$$A \rightarrow \{A', A'', A'''\}$$



The level of precision on which consumed energy is reported

$$Wh, kWh, 10^n \times kWh, \ldots$$

The time window for keeping a client's data

$$\ldots days, months, years, \ldots$$

Adversarial strategy

1. The adversary gets hold on two datasets, one LF and one HF.
2. For each round (time period of data):
   - the adversary identifies unique smart meters based on their energy consumption index values
   - the values for the identified smart meters are removed from the future rounds
3. The adversary repeats this for each round until she has identified all smart meters or she has used all time periods of data.

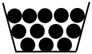
Her purpose is to identify uniquely a large number of customers.

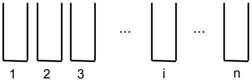
$$\text{HF} \rightarrow \text{LF'} \rightarrow \text{LF}$$

Probabilistic model - Game of balls and bins

Begin

The bins are created based on their width (**data granularity**).

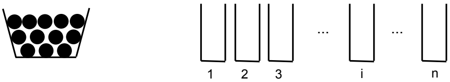




Probabilistic model - Game of balls and bins

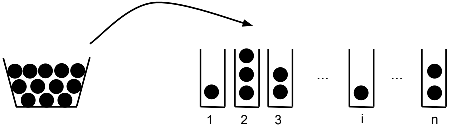
Begin

The bins are created based on their width (**data granularity**).



Round 1

The balls (energy consumption index values) from this specific period are distributed in the bins.



Probabilistic model - Game of balls and bins

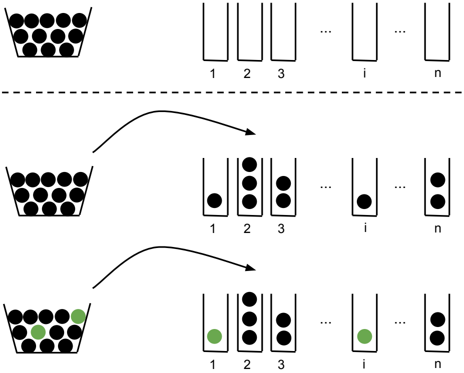
Begin

The bins are created based on their width (**data granularity**).

Round 1

The balls (energy consumption index values) from this specific period are distributed in the bins.

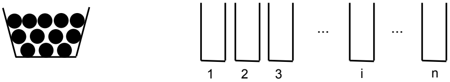
The balls that fall alone in their bin signify smart meters that are uniquely identified.



Probabilistic model - Game of balls and bins

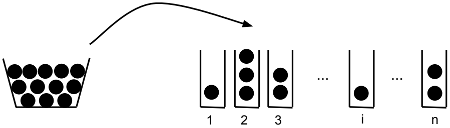
Begin

The bins are created based on their width (**data granularity**).

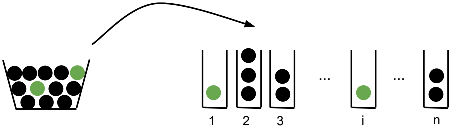


Round 1

The balls (energy consumption index values) from this specific period are distributed in the bins.

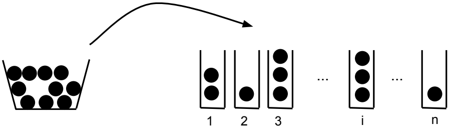


The balls that fall alone in their bin signify smart meters that are uniquely identified.



Round 2

The values for the smart meters identified last round are removed from this round data. The identification is repeated.



Probabilistic framework

Assume a Poisson distribution of balls into bins[2]

- the expected number of consumption indexes identified uniquely at round j:
  $E_j[\textit{bins with } 1 \textit{ ball}] = e^{-\frac{m_j \times w}{M}} \times m_j$

For two consecutive rounds $_{j-1,j}$

- at each round we remove the identified consumption indexes: $m_j = m_{j-1} - E_{j-1}[\textit{bins with } 1 \textit{ ball}]$
- the number of consumption indexes at current round depends only on the number of consumption indexes at previous round and the number of bins considered: $m_j = m_{j-1} \times (1 - \exp(-\frac{m_{j-1} \times w}{M}))$

The game ends when

- all balls have been removed from the game: $m_j = 0$
- all the time periods with available data have been used: $j > T$

[2] Adapted from: M. Mitzenmacher and E. Upfal - *Probability and computing: Randomized algorithms and probabilistic analysis*, Cambridge University Press, 2005

Experiments

Adversarial strategy modeled as a game of balls and bins

- We use the sizes of the different LF datasets as input for the **game of balls and bins**

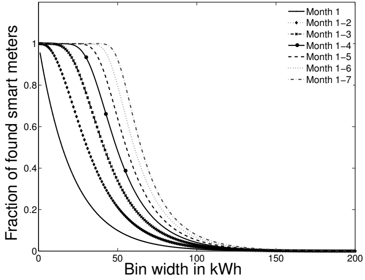
Actual execution of adversarial strategy

- We run the **adversarial strategy algorithm** on the different LF datasets

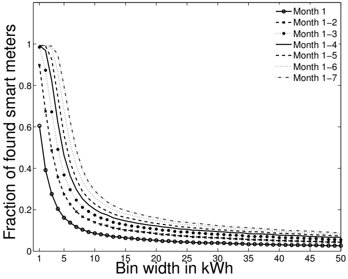
What is the effect of data **granularity** and data **timespan** on the ability of the adversary to identify a large number of customers?

Results - 7 months period

Bins and balls model



Real world data (RW data)



What is the effect of data **granularity** and data **timespan** on the ability of the adversary to identify a large number of customers?

Data granularity and data timespan

Granularity 1 kWh

	Newly found smart meters		Total found smart meters %	
Time period	**Model**	**RW data**	**Model**	**RW data**
m_1	18,461	11,698	95.4%	60.5%
m_2	871	5,655	99.9%	89.7%
m_3	2	1,669	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	19,334	19,209	100 %	**99.3%**

Granularity 10 kWh

	Newly found smart meters		Total found smart meters %	
Time period	**Model**	**RW data**	**Model**	**RW data**
m_1	12,182	1,670	63.0%	8.6%
m_2	6,029	1,027	94.1%	13.9%
m_3	1,093	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	19,334	5,628	100 %	**29.1%**

- A change in the granularity of data reported monthly can significantly reduce the number of identified smart meters
- If laws and regulations allow $\rightarrow$ customers can opt for this type of reporting to gain extra privacy

Summary - de-anonymization

- Data granularity and data timespan have an important influence in AMI data privacy
- These two characteristics should be taken in consideration when releasing datasets to 3^{rd} parties
- Even with the simple model a large number of smart meters can be identified uniquely based on their energy consumption

Road so far

- The Advanced Metering Infrastructure (AMI)
- AMI data - utility and privacy
- Privacy issues - de-anonymization and de-pseudonymization
- **Differential-privacy and AMI data**
- AMI data application - energy load forecast using DP-aggregated AMI data
- Conclusion

Differential privacy[3] (DP)

General Definition

A randomized function M gives ϵ-differential privacy for all data sets D and D' differing in at most 1 element, and all $S \subseteq \text{Range}(M)$, if

$$Pr[M(D) \in S] \leq exp(\epsilon) \times Pr[M(D') \in S]$$

Noise addition

For $f : D \rightarrow R^d$, the mechanism M, which adds independently generated noise following the Laplace distribution $\mathcal{L}(\Delta f/\epsilon)$ to each of the d output terms, enjoys ϵ-differential privacy.

Mechanism's Sensitivity

For $f : D \rightarrow R^d$, the L_1 sensitivity of f is $\Delta f = max_{D,D'} \parallel f(D) - f(D') \parallel_1$ for all D, D' differing in at most 1 element.

[3] Dwork, C., Naor, M., Pitassi, T. and Rothblum, G.N., 2010. Differential privacy under continual observation. In Proceedings of the forty-second ACM symposium on Theory of computing (pp. 715-724). ACM.

Differential privacy - maximizing the utility

For $f : D \rightarrow R^d$, the L_1 sensitivity of f is $\Delta f = max_{D,D'} \parallel f(D) - f(D') \parallel_1$ for all D, D' differing in at most 1 element.

- for binary data ($\{0, 1\}$) the sensitivity is at most 1
- for real data ($\mathcal{R}$) the sensitivity is $\infty \rightarrow$ infinite noise, no utility

Differential privacy - maximizing the utility

For $f : D \rightarrow R^d$, the L_1 sensitivity of f is $\Delta f = max_{D,D'} \parallel f(D) - f(D') \parallel_1$ for all D, D' differing in at most 1 element.

- for binary data ($\{0, 1\}$) the sensitivity is at most 1
- for real data ($\mathcal{R}$) the sensitivity is $\infty \rightarrow$ infinite noise, no utility

Solution?

Differential privacy - maximizing the utility

For $f : D \rightarrow R^d$, the L_1 sensitivity of f is $\Delta f = max_{D,D'} \parallel f(D) - f(D') \parallel_1$ for all D, D' differing in at most 1 element.

- for binary data ($\{0, 1\}$) the sensitivity is at most 1
- for real data ($\mathcal{R}$) the sensitivity is $\infty \rightarrow$ infinite noise, no utility

Solution?

Limit the noise by bounding [a]
the sensitivity to a value B.
$\mathcal{L}(\Delta f / \epsilon) \rightarrow \mathcal{L}(B/\epsilon)$

[a] Gulisano, V., Tudor, V., Almgren, M. and Papatriantafilou, M., 2016, May. BES: Differentially Private and Distributed Event Aggregation in Advanced Metering Infrastructures. In Proceedings of the Second ACM International Workshop on Cyber-Physical System Security (pp. 59-69). ACM.

Simple DP aggregation - employing Δf

Simple Aggregation

$$S = \sum_{i=1}^{n} y_t^i, \text{ where } y_t^i \in [0, E].$$

Noise addition

$$\mathcal{L}(k\Delta f/\epsilon), \text{ where } k = \lceil WS/WA \rceil.$$

Measuring the error

$$\left| \frac{S-(S+\mathcal{L}(k\Delta f/\epsilon))}{S} \right| = \left| 0 - \underbrace{\frac{\mathcal{L}(k\Delta f/\epsilon)}{S}}_{Err_{noise}} \right|$$

Simple DP aggregation - employing Δf

Simple Aggregation

$$S = \sum_{i=1}^{n} y_t^i, \text{ where } y_t^i \in [0, E].$$

Noise addition

$$\mathcal{L}(k\Delta f/\epsilon), \text{ where } k = \lceil WS/WA \rceil.$$

Measuring the error

$$\left| \frac{S-(S+\mathcal{L}(k\Delta f/\epsilon))}{S} \right| = \left| 0 - \underbrace{\frac{\mathcal{L}(k\Delta f/\epsilon)}{S}}_{Err_{noise}} \right|$$

if $\Delta f \to \infty$ (very large consumption values) then the noise introduced by $\mathcal{L}(k\Delta f/\epsilon)$ will be very large

Bounded DP Aggregation - employing B

Bounded Aggregation

$$S_B = \sum_{i=1}^{n} \min(y_t^i, B), \text{ where } B \in [0, E].$$

Noise addition

$$\mathcal{L}(kB/\epsilon), \text{ where } k = \lceil WS/WA \rceil.$$

Measuring the error

$$\left|\frac{S-(S_B+\mathcal{L}(kB/\epsilon))}{S}\right| = \left|\underbrace{\frac{S - S_B}{S}}_{Err_{approx}} - \underbrace{\frac{\mathcal{L}(kB/\epsilon)}{S}}_{Err_{noise}}\right|$$

Bounded DP Aggregation - employing B

Bounded Aggregation

$$S_B = \sum_{i=1}^{n} \min(y_t^i, B), \text{ where } B \in [0, E].$$

Noise addition

$$\mathcal{L}(kB/\epsilon), \text{ where } k = \lceil WS/WA \rceil.$$

Measuring the error

$$\left| \frac{S - (S_B + \mathcal{L}(kB/\epsilon))}{S} \right| = \left| \underbrace{\frac{S - S_B}{S}}_{Err_{approx}} - \underbrace{\frac{\mathcal{L}(kB/\epsilon)}{S}}_{Err_{noise}} \right|$$

Limit the noise by bounding the sensitivity to B.
How to choose B?

Bes - choosing a bound *B*

Using open data repositories

Compute B on data from an already public dataset or which can easily be made public.

Bes - choosing a bound *B*

Using open data repositories

Compute B on data from an already public dataset or which can easily be made public.

Use a differentially private mechanism to compute B

Choose the bound B among a set of candidate values $\mathscr{B} = \{B_1, \ldots, B_o\}$ with a given mechanism M run over a dataset (D_{explore}) containing the events used to quantify the utility of each individual bound in $\mathscr{B}$.

Bes - choosing a bound B

Using open data repositories

Compute B on data from an already public dataset or which can easily be made public.

Use a differentially private mechanism to compute B

Choose the bound B among a set of candidate values $\mathscr{B} = \{B_1, \ldots, B_o\}$ with a given mechanism M run over a dataset (D_{explore}) containing the events used to quantify the utility of each individual bound in $\mathscr{B}$.

Most Common B (MCB)

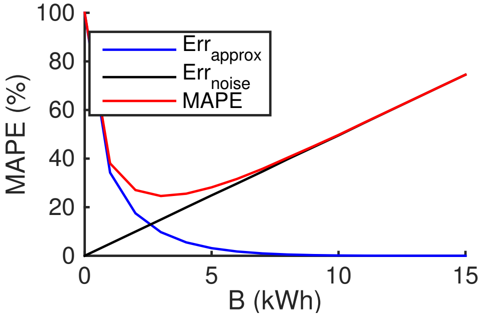
MCB aims at **finding the bound** $B \in \mathscr{B}$ resulting in the **minimum error for the majority of the SMs.**

High Enough B (HEB)

HEB looks for the bound $B \in \mathscr{B}$ for which **at least** p **of the** n smart meters observe an **error lower** than the one observed for **any higher bound** B_j.

Error composition - DP bounded aggregation

$$\left|\frac{S-(S_B+\mathcal{L}(kB/\epsilon))}{S}\right| = \left|\underbrace{\frac{S-S_B}{S}}_{Err_{approx}} - \underbrace{\frac{\mathcal{L}(kB/\epsilon)}{S}}_{Err_{noise}}\right|$$

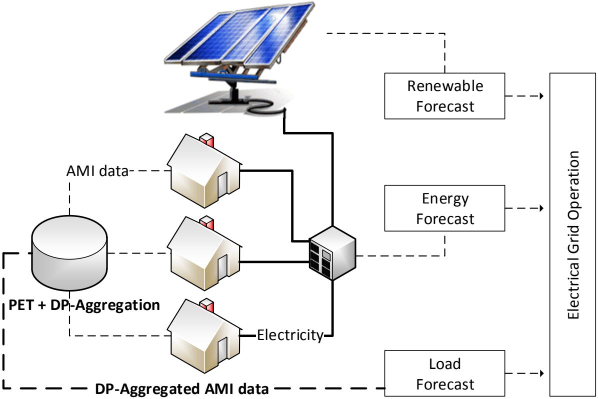


for $B = \{0, 1, ..., 15\}$ kWh.

Road so far

- The Advanced Metering Infrastructure (AMI)
- AMI data - utility and privacy
- Privacy issues - de-anonymization and de-pseudonymization
- Differential-privacy and AMI data
- **AMI data application - energy load forecast using DP-aggregated AMI data**
- Conclusion

AMI load forecast scenario

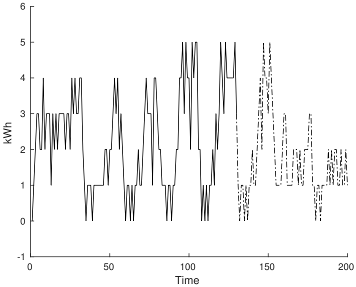


Using data for load forecast while protecting customers' privacy.

Load forecast

Load forecast

Predicting energy load based with the help of a model built on historical records.



Short term load forecast

Aims to predict consumption for short time frames, typically one hour to one week.

Load forecast - Model parameters

$$\hat{y} = \beta Y + \gamma W + \delta D + \alpha A$$

Y - energy load data,
W - weather-related data,
D - calendar-related data,
A - anthropological data.

Variable	**Characteristic**	**Forecast**	**Privacy**
Y	Granularity	+	-
Y	Data collection periodicity (sampling)	-	+
Y	Dataset size (aggregated # customers)	+	+
Y	Training window size (duration)	+	-
Y	Test window size (duration)	+	-
Y	Predicted horizon (duration)	-	-
W	Temperature	+	neutral
D	Day of week	+	neutral*
A	Anthropological data	++	- -

Load forecast - linear models

Persistence model (Seasonal Naïve) - PM

$$\hat{y}_t = y_{t-24}$$

Linear regression model 1 - LR1

$$\hat{y}_t = \beta_1 y_{t-24} + \beta_2 y_{t-48} + \beta_3 y_{t-72}$$

Linear regression model 2 - LR2

$$\hat{y}_t = \beta_1 y_{t-24} + \beta_2 y_{t-48} + \beta_3 y_{t-72} + \beta_4 \hat{T}_t + \beta_5 D(t)$$

LR1 and LR2 - Adapted from Y. Iwafune et al., "Short-term forecasting of residential building load for distributed energy management," in *Energy Conference (ENERGYCON), 2014 IEEE International*, May 2014, pp. 1197–1204.

Load forecast - data aggregation

Simple aggregation

$$y_t = \sum_{i=1}^{N} y_t^i$$

DP-aggregation

$$y_{t_{DP}} = \sum_{i=1}^{N} y_t^i + \mathcal{L}(\mathrm{B}/\epsilon)$$

DP-aggregation deters de-anonymization attacks[a].

[a] Tudor, V., Almgren, M., and Papatriantafilou, M., 2015. A study on data de-pseudonymization in the smart grid. In Proceedings of the Eighth European Workshop on System Security (p. 2).

Evaluation

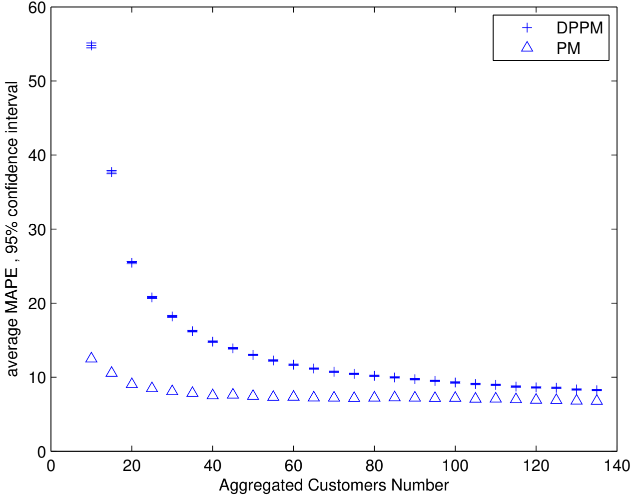
Evaluate the accuracy of three forecast models

- using simply aggregated AMI data
- using DP-aggregated AMI data
- on a variable number of customers

#	Name	Train Window	Forecast Horizon	Forecast Window	Number of Customers
1	PM	-	24h	1440h	10—135
2	DPPM	-	24h	1440h	10—135
3	LR1	1440h	24h	1440h	10—135
4	DPLR1	1440h	24h	1440h	10—135
5	LR2	1440h	24h	1440h	10—135
6	DPLR2	1440h	24h	1440h	10—135

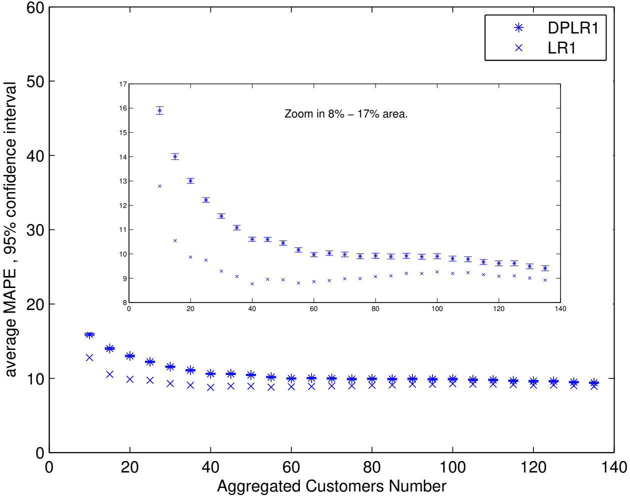
Average Mean Absolute Percentage Error (MAPE) for Persistent Method (PM)

24h forecast horizon, 100 tests/day and 60 predicted days



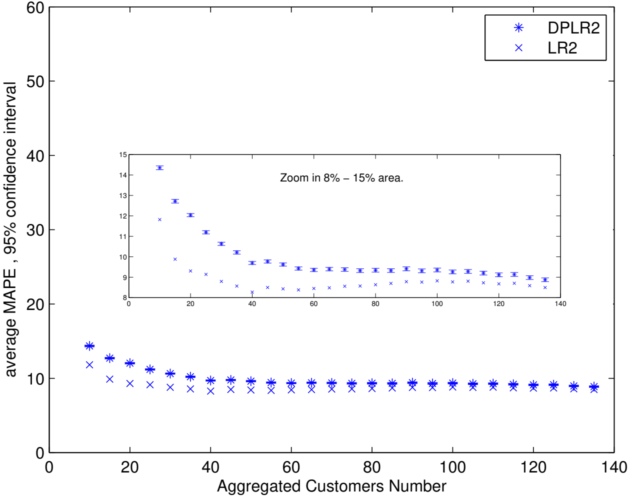
Average MAPE for Linear Regression model 1 (LR1)

24h forecast horizon, 100 tests/day and 60 predicted days



Average MAPE for Linear Regression model 2 (LR2)

24h forecast horizon, 100 tests/day and 60 predicted days



Which forecast method to use?

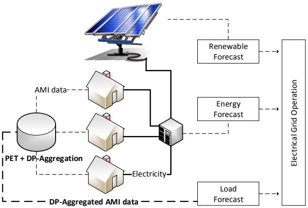
Number of Customers	PM	DPPM	LR1	DPLR1	LR2	DPLR2
10	12.50	54.80	12.79	15.90	**11.82**	**14.35**
35	7.84	16.22	9.07	11.08	8.57	10.21
50	**7.43**	13.00	8.94	10.45	8.43	**9.62**
100	7.17	9.30	9.26	9.91	8.82	9.35
135	**6.78**	**8.25**	8.92	9.41	8.49	8.87

Average MAPE (60 predicted days, 100 tests/day for DP methods)

PM: $\hat{y}_t = y_{t-24}$

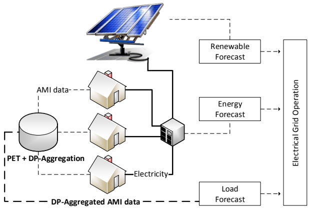
LR1: $\hat{y}_t = \beta_1 y_{t-24} + \beta_2 y_{t-48} + \beta_3 y_{t-72}$

LR2: $\hat{y}_t = \beta_1 y_{t-24} + \beta_2 y_{t-48} + \beta_3 y_{t-72} + \beta_4 \hat{T}_t + \beta_5 D(t)$



Summary - DP data application

- Differential privacy can successfully be employed in AMI data applications such as energy consumption forecasting.
- Short term forecast methods employing small-scale AMI data perform well and they can receive a boost in accuracy by further integrating **privacy neutral information.**
- Compared with their classical counterparts, the noise added by the forecast methods utilizing DP-aggregated data will introduce a **small prediction error.** This error increases with a decrease in the customers' group size.



That's all folks!

- The Advanced Metering Infrastructure (AMI)
- AMI data - utility and privacy
- Privacy issues - de-anonymization and de-pseudonymization
- Differential-privacy and AMI data
- AMI data application - energy load forecast using DP-aggregated AMI data

Bibliography

- Analysis of the impact of data granularity on privacy for the smart grid - V Tudor, M Almgren, M Papatriantafilou - Proceedings of the 12th ACM WPES 2013
- A study on data de-pseudonymization in the smart grid - V Tudor, M Almgren, M Papatriantafilou - Proceedings of Eurosec 2015
- BES: Differentially Private and Distributed Event Aggregation in Advanced Metering Infrastructures - V Gulisano, V Tudor, M Almgren, M Papatriantafilou - Proceedings of the 2nd ACM TCPS, 2016