The privacy implication of open-sourcing electricity consumption data

Antonin Voyez[1, 3], Tristan Allard[1], Gildas Avoine[1, 2], Pierre Cauchois[3], Élisa Fromont[1, 4, 5], Matthieu Simonin[4]

[1]Univ Rennes, CNRS, IRISA, France [2]INSA Rennes, CNRS, IRISA, France [3]ENEDIS, France [4]Inria, IRISA, France [5]IUF (Institut Universitaire de France), France

Outline

1. Enedis open data
2. Risks publishing single series.
3. Risks publishing aggregates.

Smart metering

- Enedis: The leading French electricity distribution company.
- Linky: The smart meter.
  - 35 million meters deployed.
  - Performing 1 measurement every day.
  - Performing 1 measurement every 30 minutes (for 3.5 million meters).
- Enedis has the legal obligation to collect and publish the measurements.[1]
  - The publications are anonymized using aggregates.

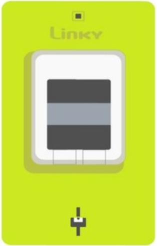


Figure: Linky smart meter

[1] Article 23, loi pour une république numérique, 2016; D111-59 à D111-66 du code de l'énergie

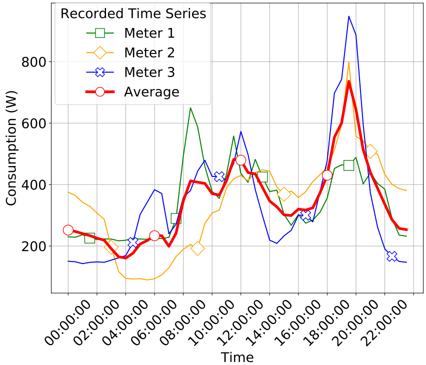
Data usage example

- Enedis open data:
  - https://data.enedis.fr/
- Observatory:
  - https://observatoire.enedis.fr/
  - Figure: Energy consumption in Rennes
- Inferring datacenters energy consumption:
  - >15 GWh / year for research on the city hosting the Jean Zay supercomputer.[2]

[2] https://data.enedis.fr/pages/consommation-electrique-par-secteur-activite/, NAF 72 per IRIS

Series sensitivity

- Figure: Example of electricity consumption series (Meter 1 to 3) and the mean-aggregate (Average).
- Devices used (heating devices).
- Socio-economical metrics.[3]
  - House occupancy.[4]
  - Number of residents.



[4] Beckel & al. "Automatic socio-economic classification of households using electricity consumption data". In: *International conference on Future energy systems*. 2013.

[4] Dong Chen et al. "Non-Intrusive Occupancy Monitoring using Smart Meters". In: *Workshop On Embedded Systems For Energy-Efficient Buildings*. 2013.

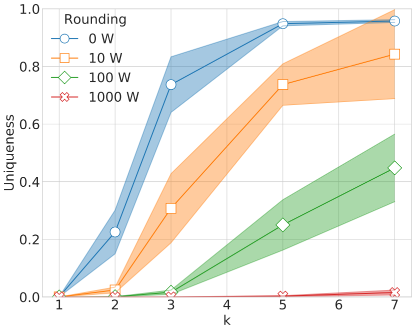
Uniqueness study

- Unique series are potentially identifiable.[5]
- Uniqueness: Proportion of unique series.
- A series is unique if it is the only one sharing a set of values.
  - For a timestamp and k.
  - k: The number of consecutive timestamps used to compute uniqueness.

[5] Yves-Alexandre De Montjoye et al. "Unique in the crowd: The privacy bounds of human mobility". In: *Scientific reports* (2013).

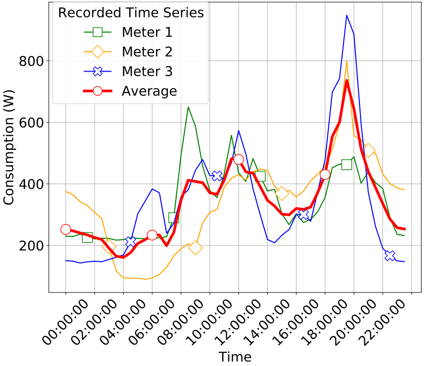
Enedis data uniqueness

- 2.5M residential half-hourly series over 1 year (Sept. 2020 to Sept. 2021).
- High uniqueness considering only a few timestamps: $> 70\%$ for $k = 3$ (1h30), $> 90\%$ for $k = 5$.
- Reducing the precision of the measurement (by rounding them) is not enough to protect the series.
  - Approx. 12k series are unique for $k = 7$ and a rounding to 1 kW.
  - Rounding to 1 kW renders the data useless: 80 % measurements below 1kW.



Aggregates sensitivity

- Figure: Example of electricity consumption series (Meter 1 to 3) and the mean-aggregate (Average).
- Aggregate: Sum or average per timestamp.
  - Legal threshold: at least 5000 series.
  - Published with an attribute (i.e., descriptive information).
- The aggregate masks individual information.
- Inferring if a series participates in the aggregate.
  - Labeling the series with the additional attributes published with the aggregate.



Membership Inference Attack (MIA)

Objective:

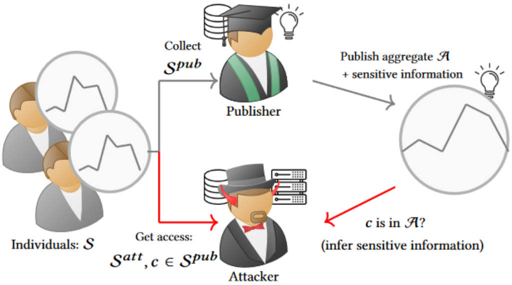
Inferring if a series is present in an aggregate.

Publication:

- An aggregate (sum or average, $\mathcal{A}$).
- Additional information labeling the aggregate (contract, devices used).

Attacker knowledge:

- The targeted series (c).
- A set of series similar to the ones in the aggregates: public data, supplier data, data leaks ($\mathcal{S}^{att}$).



MIA algorithm

Algorithm:[6]

- Build a set of fake aggregate (with and without the target).
- Train a classifier to detect the target in the aggregates.
- Test the classifier on test aggregates and evaluate using accuracy.
- Use the attack model against the real aggregate.

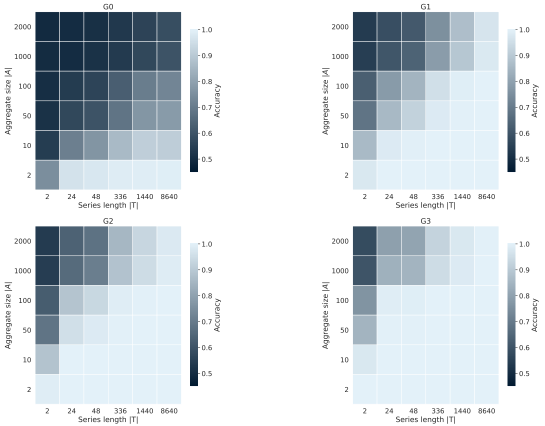
Experiments:

- Find the threshold (aggregate size) such as aggregates are no longer vulnerable.
- Vulnerable: accuracy > 0.6.
- Identify vulnerable series.

[6] Apostolos Pyrgelis & al. "Knock Knock, Who's There? Membership Inference on Aggregate Location Data". In: *NDSS*. 2018.

MIA Results

- CER-ISSDA: public dataset containing approx. 4500 Irish electricity consumption time series.
- The same period is in training and testing.
- Accuracy diminishes with the aggregate size $|A|$.
- More extended series ($|T|$) increase the accuracy.
- Atypical series (group G) leads to higher accuracy.



Take away

- Enedis: collect and publish electricity consumption time series.
  - Electricity consumption data impact the energy transition.
- Individual series are sensitive personal data.
  - Are easily identifiable
  - Are published using threshold aggregates.
- Aggregates vulnerability
  - Threat: inferring if a series participates in an aggregate.
  - Our work: find safe threshold and vulnerable series.