YoctoDB: A Data Stream Management System

Angela Gong, Max Hirschhorn, Kalpana Suraesh
Department of Computing and Mathematical Sciences, California Institute of Technology

Motivation

- Current implementations of data-stream management systems (DSMS) mainly focus on supporting low-latency, high-volume data processing
- Not many practical implementations of differentially-private streaming algorithms

Research Goals

Goals

- Implement DSMS with that supports various privacy-preserving operators
- Use push-based model for computing results of query via asynchronous message-passing

Features

- Typical relational operators found in RDBMS
- Stream-to-stream operations including selection, projection, and aggregation
- Windowing operators based on number of rows and time
- Privacy-preserving SUM, COUNT, and AVERAGE

Privacy Implementation

Uses an **ε-Differentially Private Mechanism** – when run on two neighboring streams, behaves about the same on both sets. The presence or absence of a single data point should not affect the final output.

- **ε** – the privacy parameter. A smaller ε increases the amount of noise added.

Future Work

Some improvements and additional features:

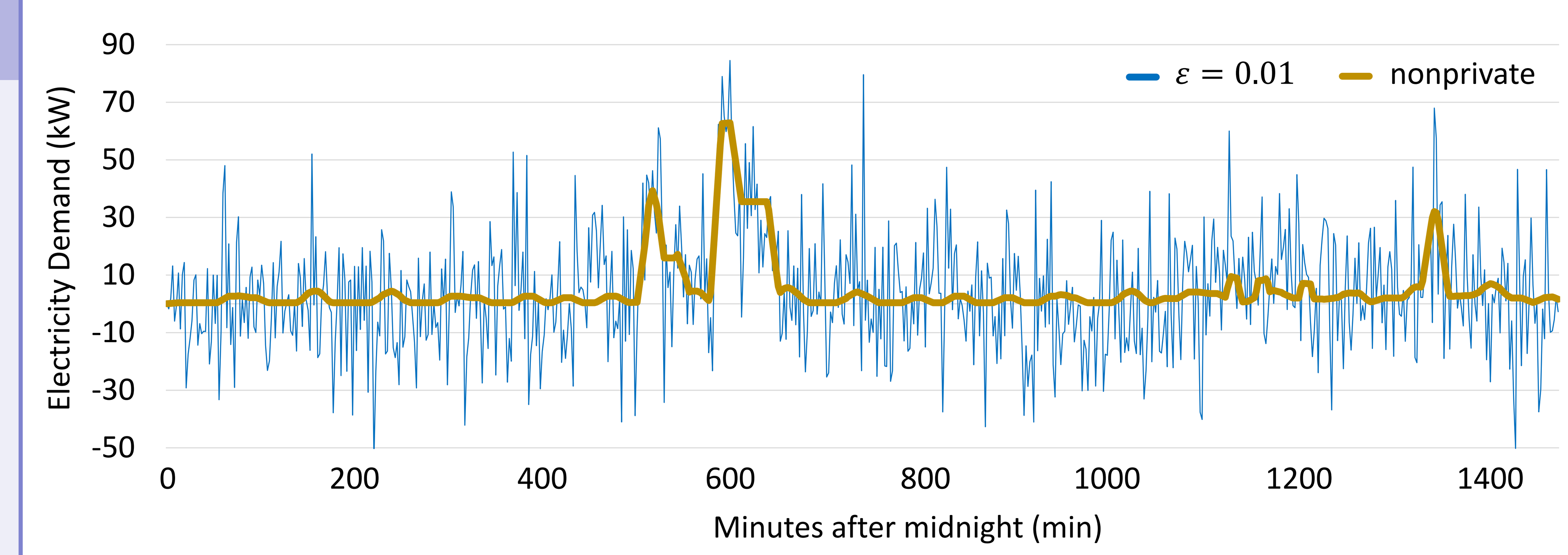
- **Parser and CQL –** Support for CQL (continuous query language) for ease of querying and interacting with the system.
- **Additional Private Operators –** Support more complex operators, such as variance.
- **Load-shedding** – Reduce the latency by randomly dropping tuples from the stream.

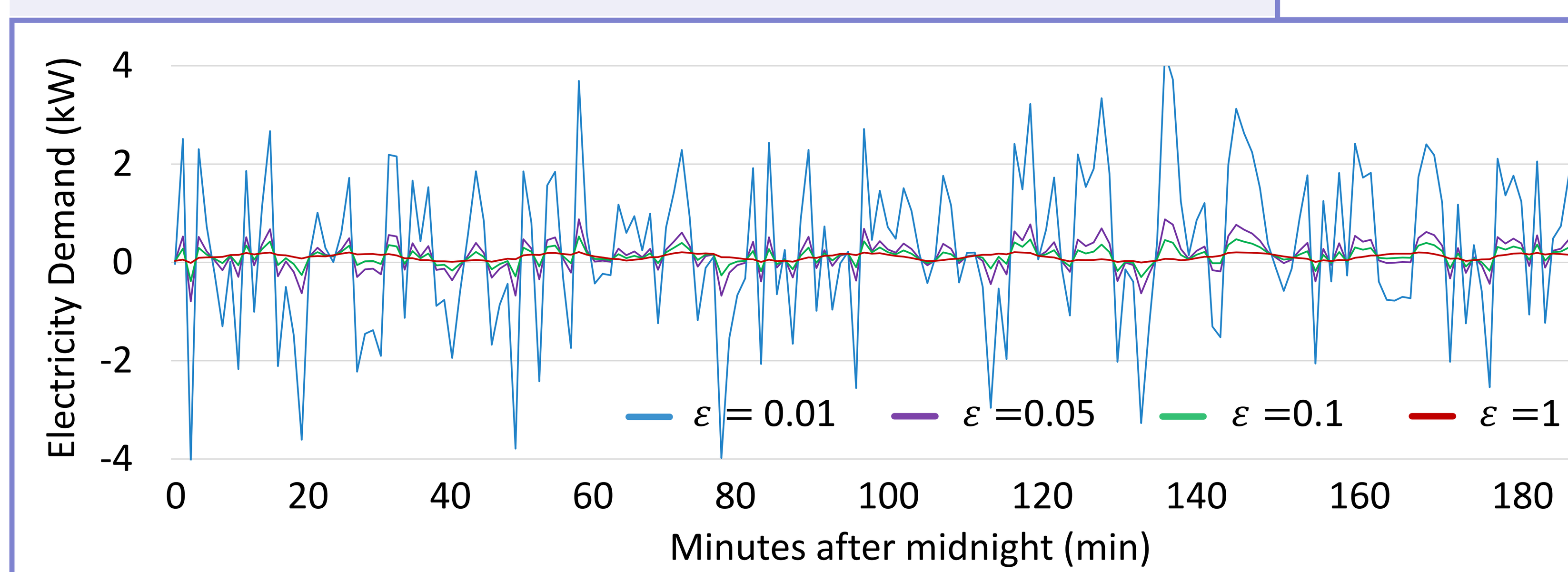
Case Study | Household and Neighborhood Electricity Demand

Using model from Richardson, et al, *Domestic electricity use: A high-resolution energy demand model.*

1. Can we hide household electricity demand?

- With smart metering, an adversary could easily tell when someone is home
- Want to make electricity demand ε-differentially private to protect this information



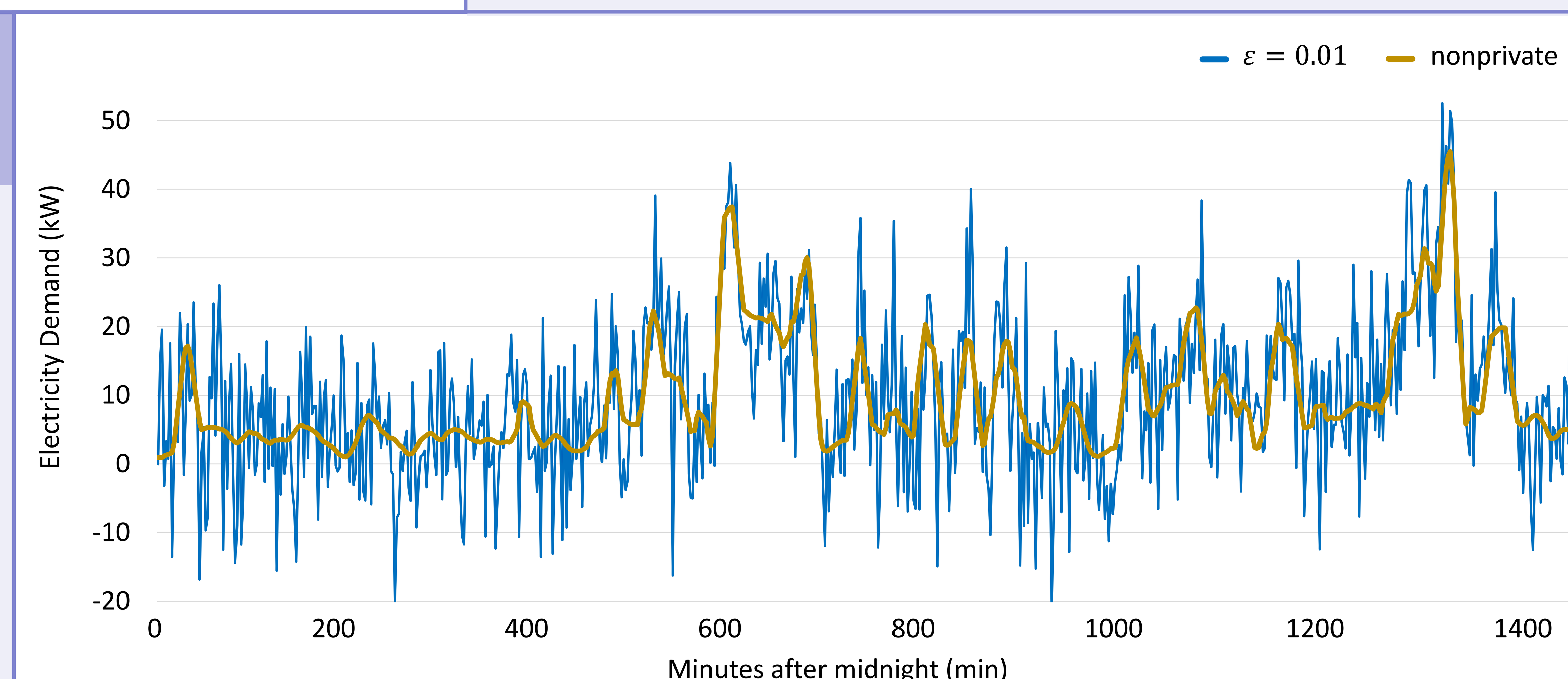


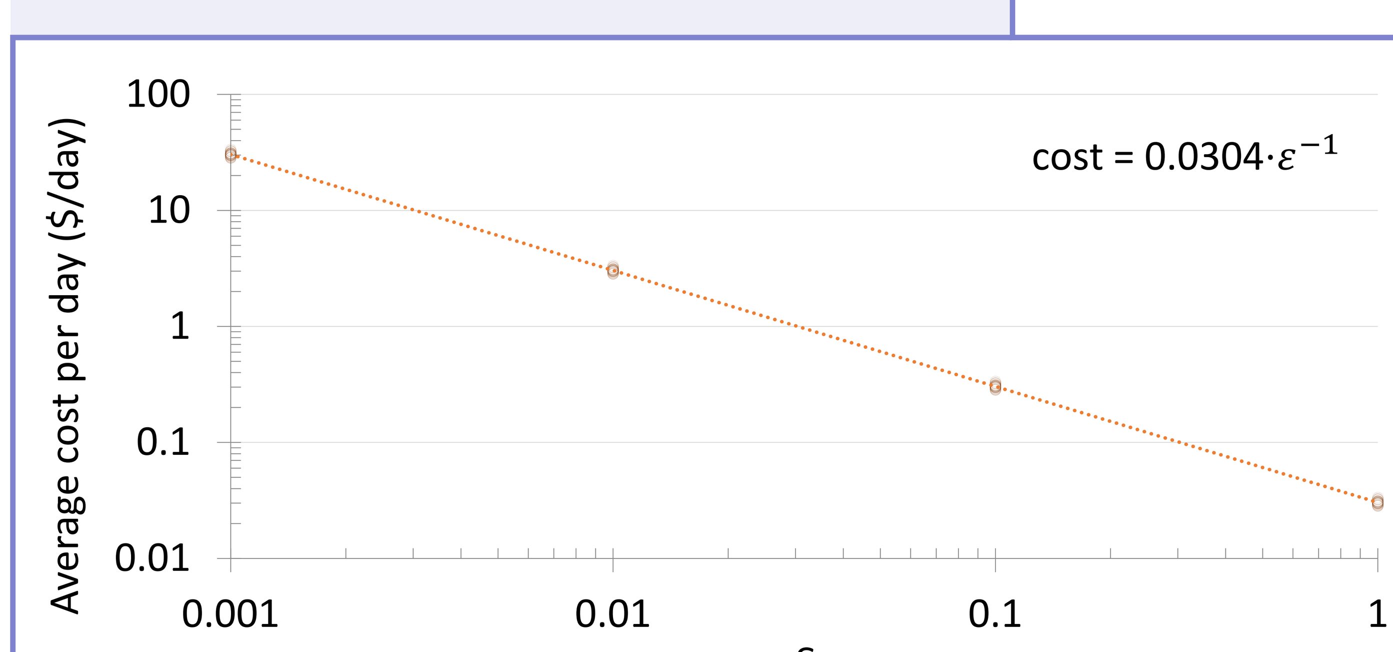
2. What is a good value for the privacy parameter, ε?

- Error amounts drawn from Laplace distribution
- Need to find the "sweet spot" of noise to add: too much and data is meaningless, too little and activity is still obvious
- Our chosen $\varepsilon = 0.01$

3. How accurate is the privacy-preserved neighborhood electricity demand?

- Simulated 30 households based on 5 distinct usage and occupancy patterns
- In expectation, the sum of error amounts is zero





4. What is the cost of privacy?

- Compute the difference integral between the actual and privacy-preserved electricity demand
- Use a battery to increase electricity demand, and use a generator to decrease electricity demand
- Assuming generator runs for 81.25 kWh/gal with cost of diesel at \$4.025/gal, and cost of electricity is 15.34¢/kWh per day (in California).