Model-Driven Data Acquisition in Sensor Networks

Amol Deshpande	U.C. Berkeley
Carlos Guestrin	Intel Research Berkeley
Samuel R. Madden	M.I.T.
Joseph M. Hellerstein	U.C. Berkeley
Wei Hong	Intel Research Berkeley

International Conference on Very Large Databases, 2004.

Sensor Network as a Database

- Use sensor network to answer queries about the environment
- Sensor readings so not exhaustively represent the physical reality
  - Non-uniform sensor distribution
  - Faulty sensors
  - Communication link failures
- Approximate queries are inefficient
  - Collecting all data may be an overkill

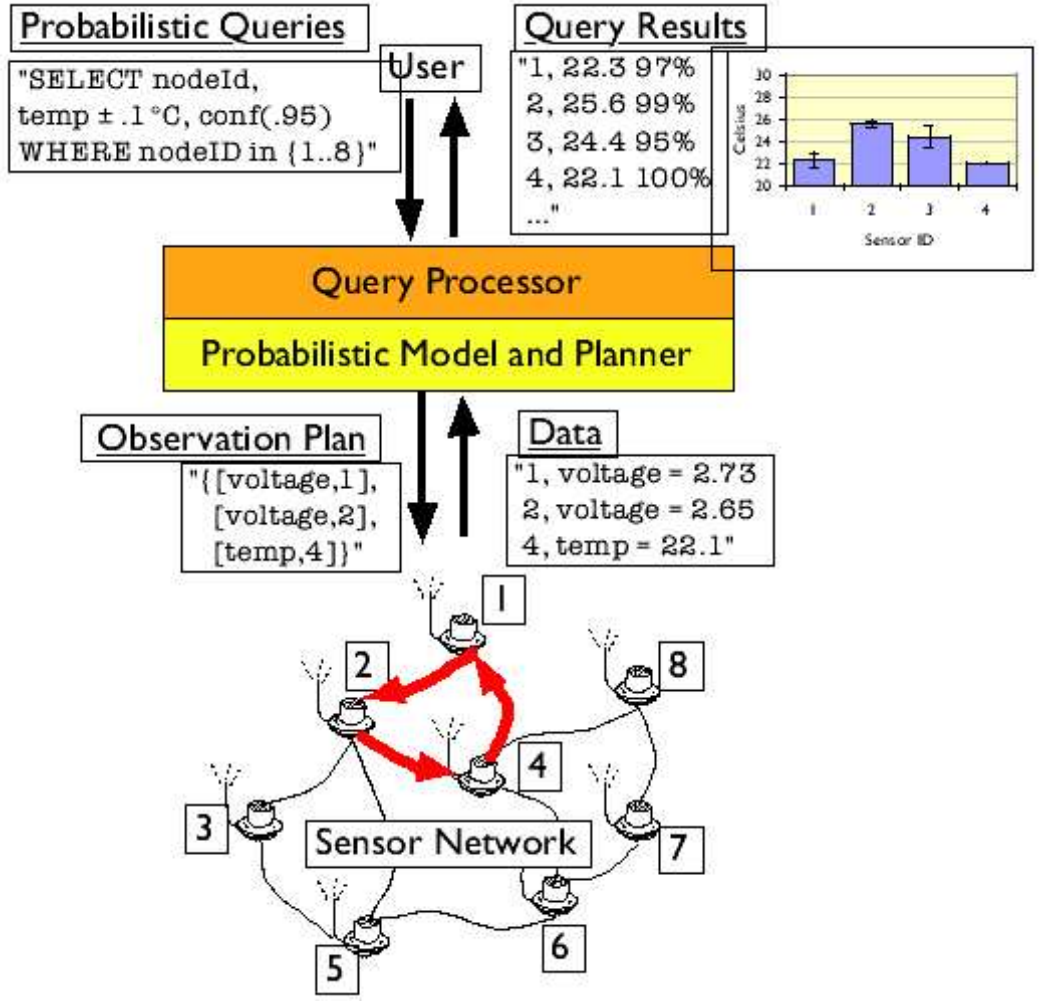
Contribution

- Use a *statistical model* of the environment
- Statistical model is a joint pdf

$$p(x_1, x_2, \ldots, x_n)$$

- Variables
  - Observable: temperature measurement of sensor #5
  - Hidden: sensor #12 is faulty
- Advantage
  - Captures correlations among attributes
  - Can predict readings
  - Save on sensing and communication

The BBQ System



Statistical Model

- Choice of model affects the planning phase
- BBQ: multivariate Gaussian model

$$p_{\mu,\Sigma}(x) \sim \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

- New observations

$$p(x \mid \mathcal{O} = o) = \frac{p(x, \mathcal{O} = o)}{p(\mathcal{O} = o)}$$

- Initial model obtained from training data

Answering Queries

- Result

$$\overline{y} = \mathrm{E}[y] = \int_x y(x)p(x)dx$$

- Confidence interval

$$\Pr[|y - \overline{y}| \leq \epsilon] = \int_{x:\ |y-\overline{y}|\leq\epsilon} p(x)dx \geq 1 - \delta$$

RANGE	$y(x) =$ indicator $(a_i \leq x_i \leq b_i)$
VALUE	$y(x) = x_i$
AGGREGATE	$y(x) = \frac{1}{\lvert\mathcal{I}\rvert} \sum_{i\in\mathcal{I}} x_i$

Answering Queries

- Result

$$\overline{y}(o) = \mathrm{E}[y|o] = \int_x y(x)p(x|o)dx$$

- Confidence interval

$$\Pr[|y - \overline{y}(o)| \leq \epsilon \mid o] = \int_{x:\ |y-\overline{y}(o)|\leq\epsilon} p(x|o)dx \geq 1 - \delta$$

Dynamic Models

- Exploit temporal correlations
- *Transition model* describes how system evolves over time

$$p(x^{t+1} \mid x^t, \ldots, x^0)$$

- Markovian assumption

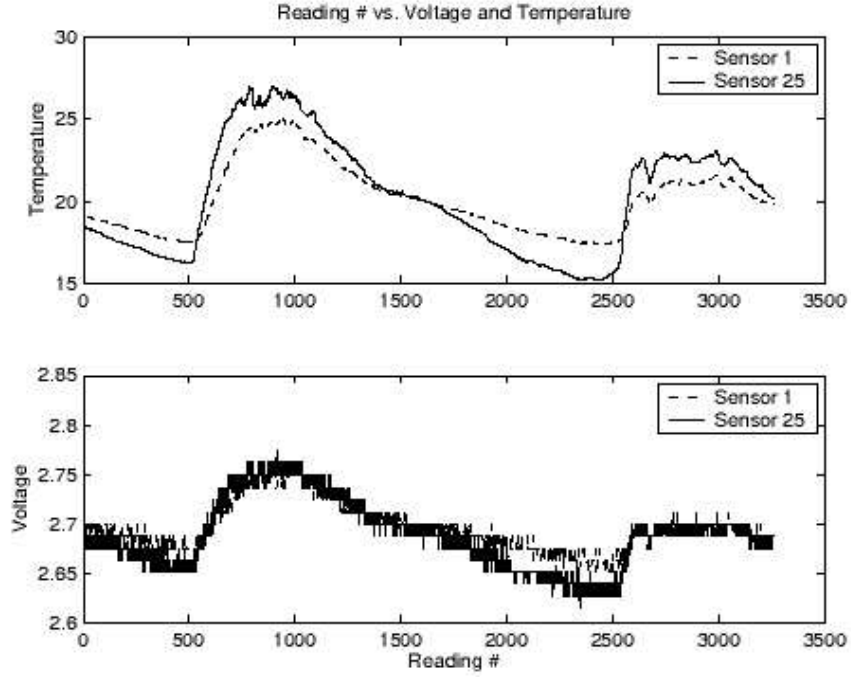
$$p(x^{t+1} \mid x^t, \ldots, x^0) = p(x^{t+1} \mid x^t)$$

STATIC	$p(x^{t+1} \mid o^t)$
DYNAMIC	$\int_{x^t} p(x^{t+1} \mid x^t) p(x^t \mid o^t) dx^t$

- Time-varying transition model

Observation Plan

- Goal
  - Set of attributes to observe
  - Path for data collection
- REQUESTED ATTRIBUTES $\neq$ OBSERVED ATTRIBUTES
  - System exploits correlation and cost differential



Observation Plan

- Benefit
  - Range

$$R(o) = \max\{\overline{y}(o), 1 - \overline{y}(o)\}$$

  - Value, aggregates...

$$R(o) = \Pr\{|y - \overline{y}(o)| \leq \epsilon \mid o\}$$

- Average benefit

$$R(\mathcal{O}) = \int_o R(o)p(o)do$$

Observation Plan

- Cost, $C(\mathcal{O}) = C_a(\mathcal{O}) + C_t(\mathcal{O})$
- Acquisition (sensing) cost

$$C_a(\mathcal{O}) = \sum_{i \in \mathcal{O}} C_a(i)$$

- Communication (transmission) cost
  - Known topology, unreliable transmission

$$C_t(\mathcal{O}) = \min_{p \in \mathcal{P}} \sum_{e \in p} \frac{1}{p_e}$$

  - TSP, use k-OPT heuristic

Observation Plan

- Attribute selection

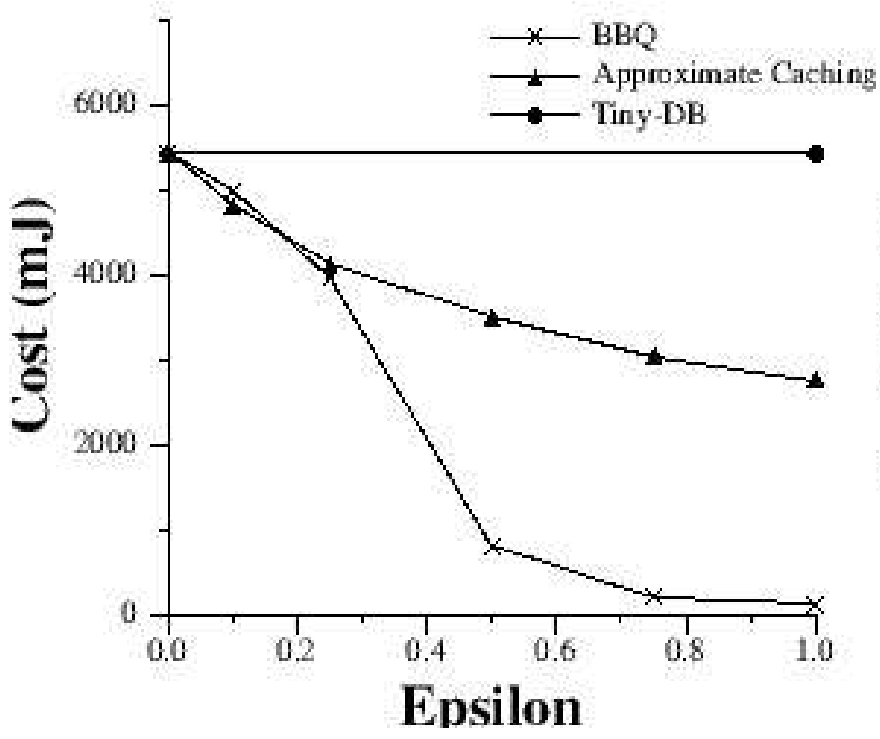
$$\min_{\mathcal{O}} C(\mathcal{O}) \qquad \text{subject to} \qquad R(\mathcal{O}) \geq 1 - \delta$$

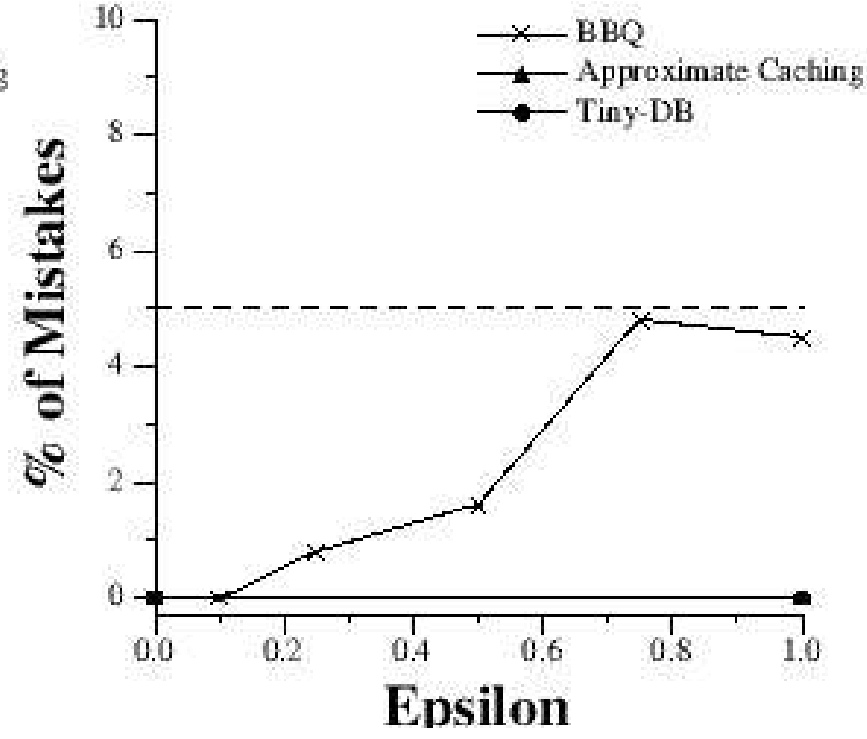
- Exhaustive search
- Greedy
  - Compute $\mathcal{G} = \{i \mid i \notin \mathcal{O} \ \wedge \ R(\mathcal{O} \cup i) \geq 1 - \delta\}$
  - If $\mathcal{G} \neq \varnothing$ add $i \in \mathcal{G}$ with smallest cost
  - Otherwise, add $i \notin \mathcal{O}$ with the largest benefit/cost ratio

Experiments

- Real-world data sets: training + test
- Query types
  - ALL VALUES: report all sensor values with accuracy ϵ and confidence $1 - \delta$.
  - RANGE: Report all sensors whose measurements are in the given range, with confidence $1 - \delta$
- Comparison systems
  - TINYDB – uses aggregation tree
  - APPROXIMATE CACHING – motes report significant changes in readings

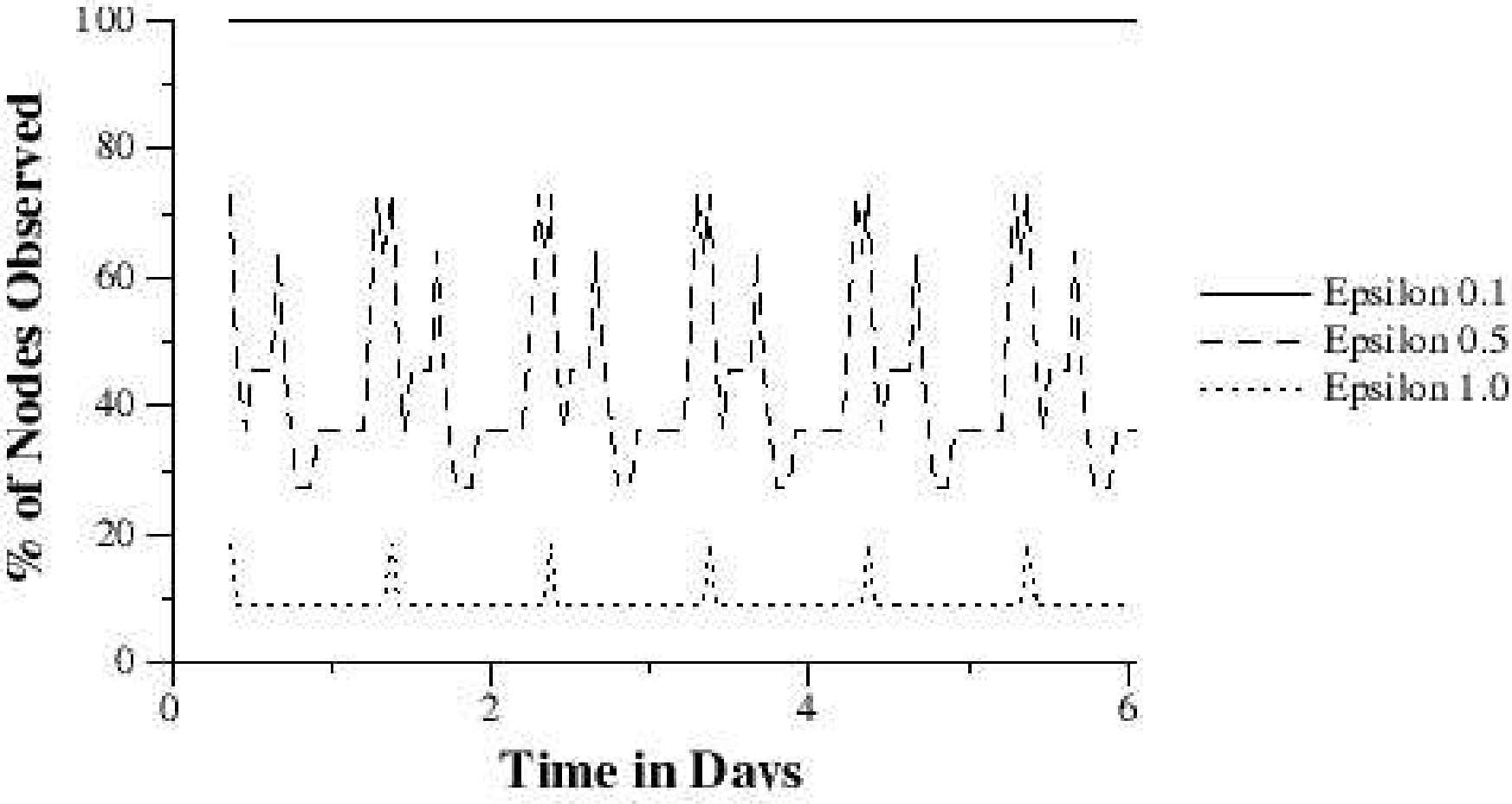
Experiments





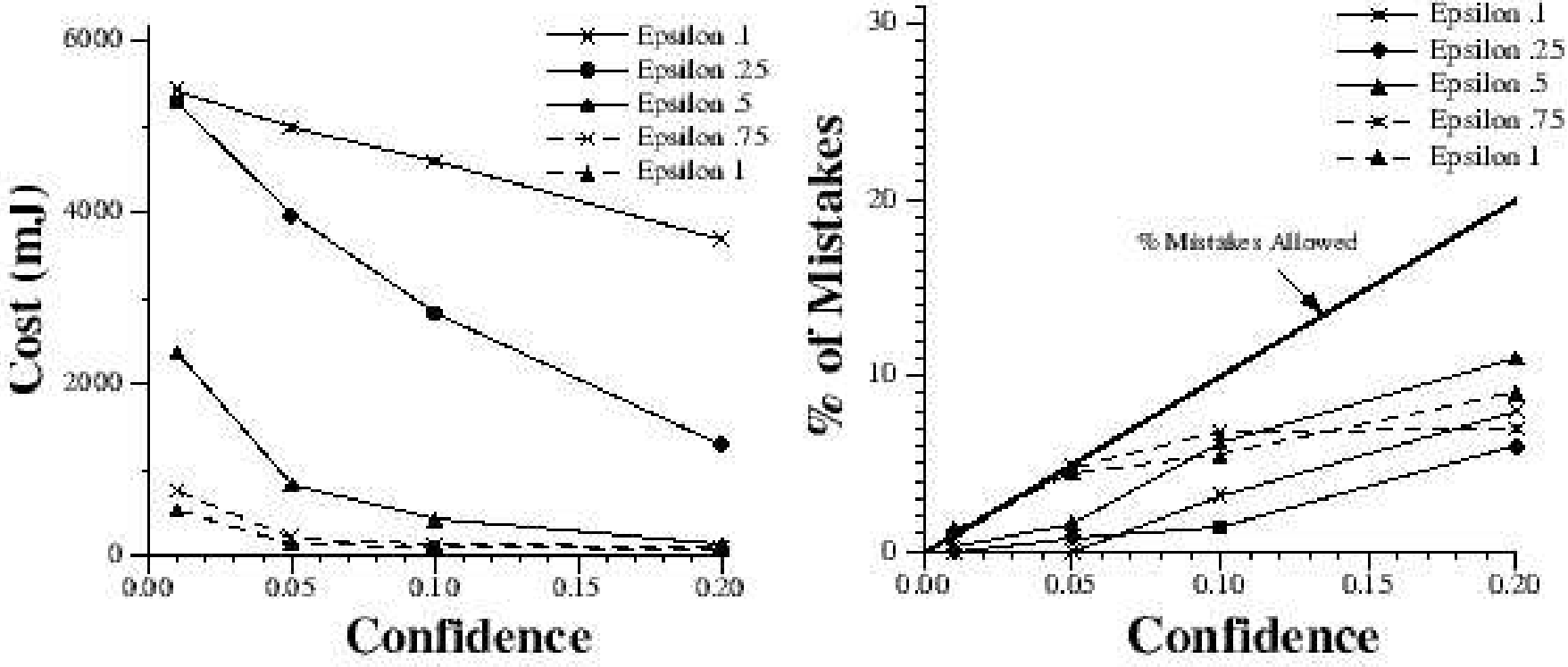
ALL VALUES query, 95% confidence, variable ϵ (in °C)

Experiments



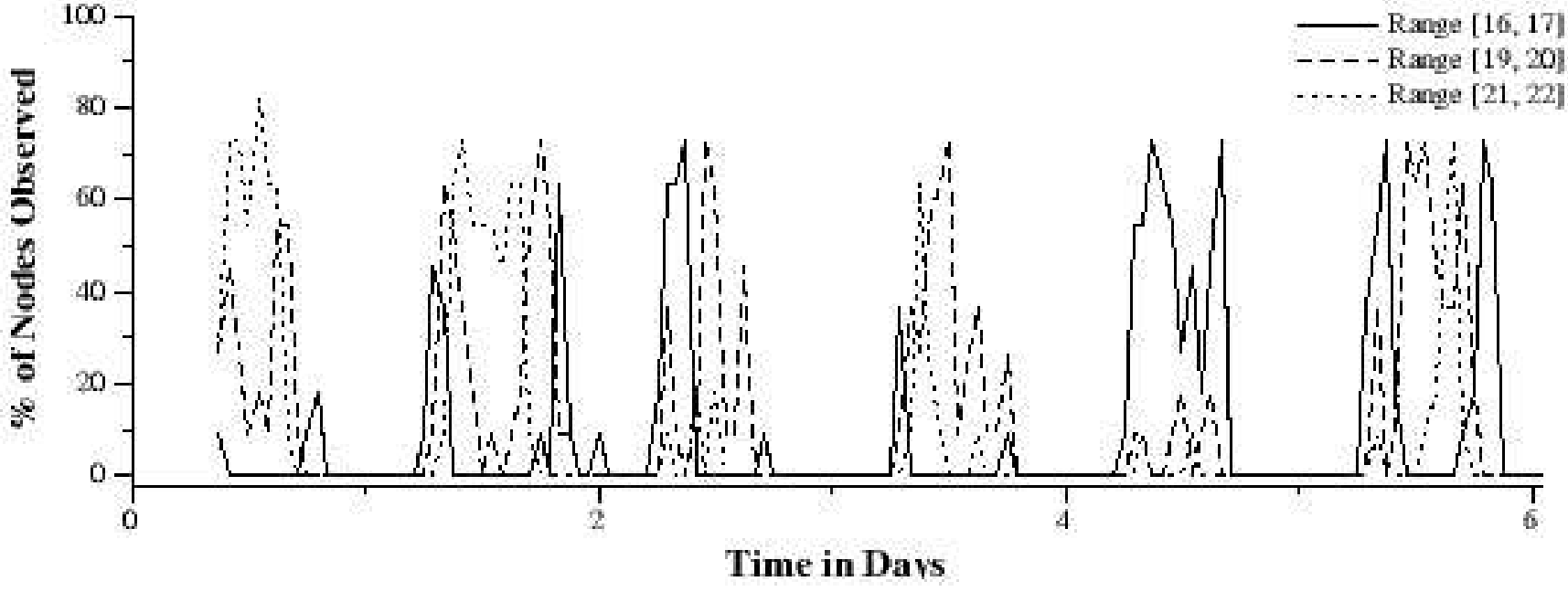
ALL VALUES query, 95% confidence, variable ϵ (in °C)

Experiments



ALL VALUES query, variable ϵ (in °C) and δ

Experiments



RANGE query, 95% confidence, varying range (in °C)

Extensions

- Conditional plans
- More complex models
- Outlier detection
- Continuous queries
- Support for dynamic networks