# MIST: Distributed Indexing and Querying in Sensor Networks

Ambuj K. Singh

(Arnab Bhattacharya, Anand Meka)
Department of Computer Science,
University of California, Santa Barbara

# Semantic Modeling: Motivation

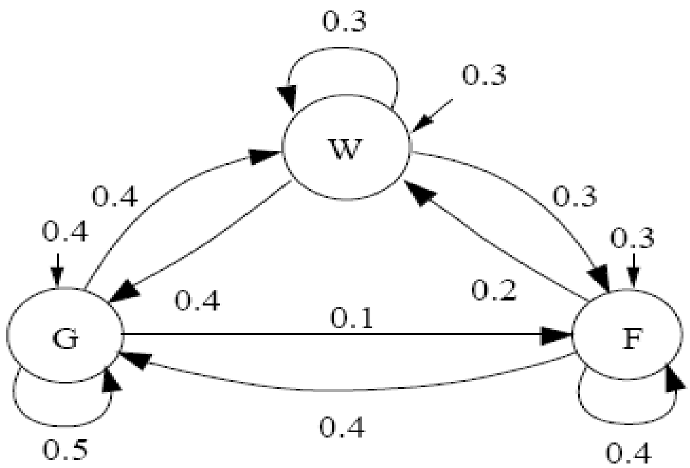
- Content-Summarization: Extracting high level semantic events from low-level sensor readings
  - What is the current traffic pattern?
  - Which room has the highest-likelihood of occupancy?
  - Is a storm approaching?
- Vocal Recognition of Acorn Woodpeckers by acoustic sensors
  - Which is this individual?
  - Where is this bird?
  - Are any birds alarmed?

Individual Recognition via models trained on sound

# Motivation

- Track movement to study zebra behavior & social patterns
- Zebra's movement characterized into (G) Grazing, (W) Walking, and (F) Fast Moving
- Temporal behavior is a sequence of states
  - Ex: GWWWWFGG
- Training of Markov Chains over such sequences
  - Start state vector (π)
  - Transition vector (Γ)
  - P(GWW) = π(G) Γ(G,W) Γ(W,W)

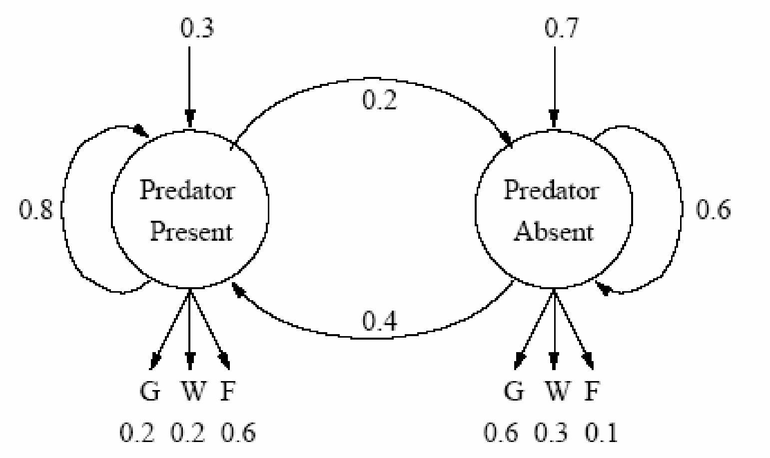
Zebranet Project




Zebra Mobility: Markov model

# Semantic Modeling

- Hidden Markov Models (HMMs)
  - Additional Emission Vector (ξ)
  - Probability of state path: Akin to MC
  - Probability of observing sequence from a particular state path:
    - $P(GWW \mid S_1S_2S_1) = \pi(S_1)\ \xi(S_1,G)\ \Gamma(S_1,S_2)\ \xi(S_2,W)\ \Gamma(S_2,S_1)\ \xi(S_1,W)$
  - Probability of observing the sequence is the summation of sequence probability over all possible state-paths
    - Viterbi Algorithm: $O(n^2k)$ computation for a n-state HMM, and k-length sequence




Zebra Mobility: HMM model

# Goals

- Identify interesting behaviors in the network
  - Ex: Which zebranet sensors observed FFFFF sequence with a likelihood of 0.85? (denoting a possible predator attack)
- Sequence Queries
  - Range Query: Return sensors which observed a particular pattern with a likelihood of at least $\Phi$?
  - Top-1 Query: Which is the sensor that is most likely to exhibit a given behavior?
- Model Query
  - 1-NN Query: Which sensor *model* is the most similar to the given *model*?
    - Where is the woodpecker?

# Centralized Solution

- Each sensor trains a model on the observation sequence
  - Transmits the model to the Base Station (BS)
- Queries are answered at BS
- Each update of model is transmitted to the BS

# A Better Solution: Slack-based Centralized Scheme

- Slack parameter maintained at local models
- Updates are not transmitted if the change is within the slack
- If query cannot be answered by the BS using the cached models, it is transmitted to each node

# Talk Outline

- Motivation & Preliminaries
- MIST: An In-network Model based Index Structure
- Query Algorithms
- Experiments
- Conclusions

# MIST Index Structure

- Overlay a tree on the network
- Each sensor trains a MC/HMM on the observed sequences
- Bottom-up aggregation of index structure
- Types of Index Models
  - Average Model
  - Min-Max Model
- Index models capture correlation between constituent models

# Modeling Correlations

- Models at nodes $S_1$ and $S_2$ are (1-ε)-correlated if for all corresponding parameters $\sigma_1$ of $S_1$, and $\sigma_2$ of $S_2$ :

$$1-\varepsilon = \min_i \{ \frac{\min(\sigma_1^i, \sigma_2^i)}{\max(\sigma_1^i, \sigma_2^i)} \}$$

- $\varepsilon \to 0$ : high similarity

Example: 0.75 correlation

$S_1$

| | |
|---|---|
| 0.4 | 0.6 |
| 0.7 | 0.3 |

($\Gamma_1$)

| | |
|---|---|
| 0.5 | 0.5 |

($\pi_1$)

$S_2$

| | |
|---|---|
| 0.3 | 0.7 |
| 0.6 | 0.4 |

($\Gamma_2$)

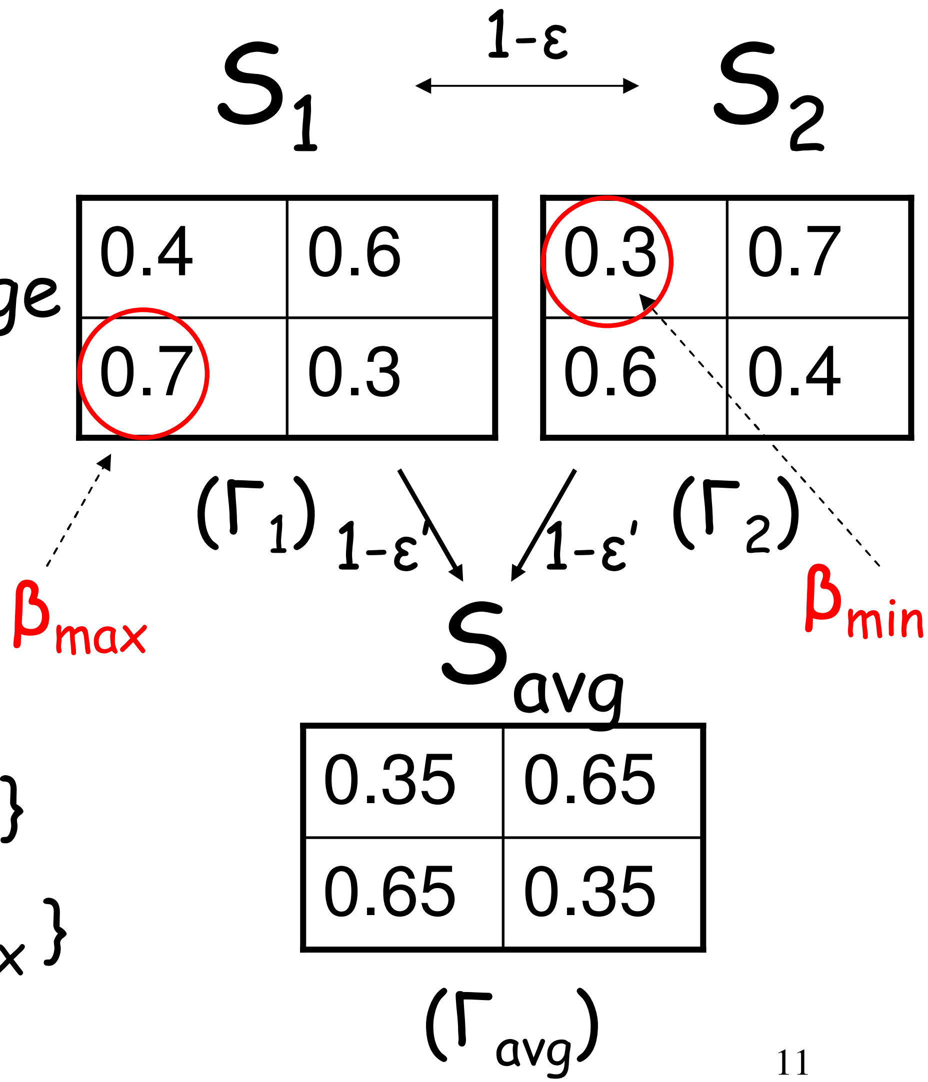
| | |
|---|---|
| 0.6 | 0.4 |

($\pi_2$)

$$0.75 = \min \{ \frac{0.3}{0.4}, \frac{0.6}{0.7}, \frac{0.4}{0.5}, \frac{0.5}{0.6} \}$$

# MIST: Average Index Models

- Index maintains
  - Model whose parameters are average of the constituent models
  - $\beta_{max}$, $\beta_{min}$, and $\varepsilon'$
- $\varepsilon' = \varepsilon/(2-\varepsilon)$
- $\sigma^i \geq \max\{\sigma^i_{avg}(1-\varepsilon'),\ \beta_{min}\}$
- $\sigma^i \leq \min\{\sigma^i_{avg}/(1-\varepsilon'),\ \beta_{max}\}$



# MIST: Min-Max Index Models

- Index maintains
  - Min Model: parameters are minimum of the constituent models
  - Max Model is similarly defined

$S_1$

| | |
|---|---|
| 0.4 | 0.6 |
| 0.7 | 0.3 |

$(\Gamma_1)$

$S_2$

| | |
|---|---|
| 0.3 | 0.7 |
| 0.6 | 0.4 |

$(\Gamma_2)$

$S_{min}$

| | |
|---|---|
| 0.3 | 0.6 |
| 0.6 | 0.3 |

$(\Gamma_{min})$

$S_{max}$

| | |
|---|---|
| 0.4 | 0.7 |
| 0.7 | 0.4 |

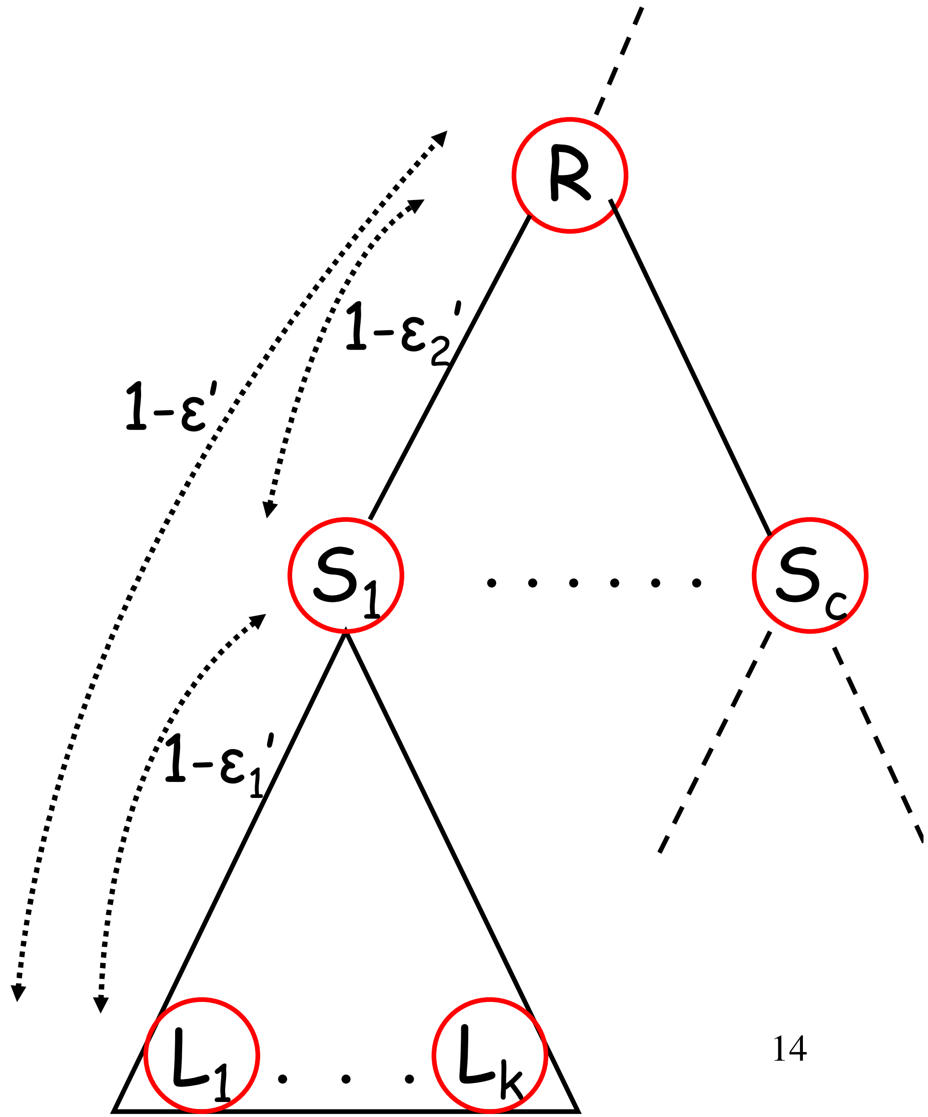
$(\Gamma_{max})$

# Index Models for HMMs

- Correspondence of states required to define index models
- Domain knowledge to infer correspondence between states

# MIST: Hierarchical Index Structure

- Average Model:
  - $(1-\varepsilon') = (1-\varepsilon_1')(1-\varepsilon_2')$
  - $\beta_{max}$, $\beta_{min}$ aggregated at node R
- Min-Max model aggregation similar to an R-Tree



# Dynamic Maintenance

- After every period ‘d’, each sensor trains a new model on the recent observation sequence
- Update protocol: Child does not update its parent if the new model is (1-δ)–correlated with the model maintained at the parent
- Correlation maintained at the parent
  - $(1 - \varepsilon_{slack}) = (1- \delta)\,(1- \varepsilon_{no\text{-}slack})$
  - Optimal slack analysis in paper

# Talk Outline

- Motivation & Preliminaries
- MIST: An In-network Model based Index Structure
- Query Algorithms
- Experiments
- Conclusions

# Range Querying

- Return all nodes that have observed a particular sequence of symbols q: $q^1 \ldots q^i \ldots q^k$ with a probability > $\Phi$
- Assume $P(q) = \sigma^1 \ldots \sigma^i \ldots \sigma^k$
- Pruning by Average model $S_{avg}$:
  - lb: $\prod^k [\max\{\sigma^i_{avg}(1 - \varepsilon_s), \beta_{min}(1-\delta)\}]$
  - ub: $\prod^k [\min\{\sigma^i_{avg}/(1 - \varepsilon_s), \beta_{max}/(1-\delta)\}]$
- Pruning by Min-Max models $S_{min}$ & $S_{max}$
  - lb: $P(q|S_{min})(1-\delta)^k$
  - ub: $P(q|S_{max})/(1-\delta)^k$

# Top-1 Query

- Return the sensor that has the highest probability of observing the query sequence
- Pruning employed by index models
  - compute lower and upper bounds on the query probability for each child-subtree
  - Prune the child-subtree if the upper bound from this subtree is smaller than the lower bound from another child-subtree

# Model Querying

- Return the sensor model most similar to the query model.
- Similarity measure: $L_2$ distance between corresponding model parameters
- Average Models
  - Create an M-Tree index
  - Index node: routing object $S_{avg}$ and covering radius
- Min-Max Models
  - Build an R-Tree based index
  - Index node: MBR in the vector space

# Semantic Modeling & Querying

- Motivation & Preliminaries
- MIST: An In-network Model based Index Structure
- Query Algorithms
- Experiments
- Conclusions

# Experimental Results

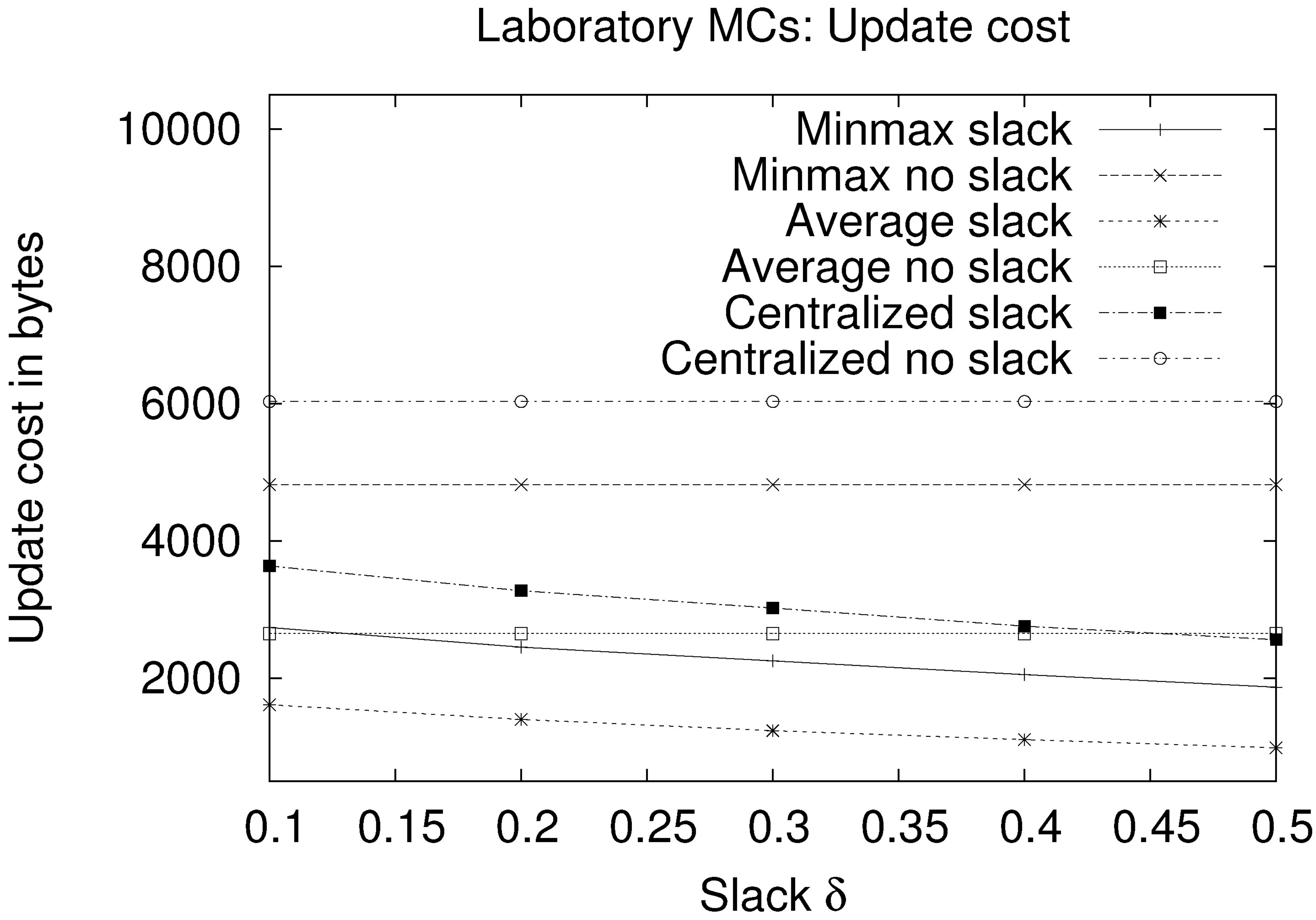
- Datasets
  - Lab-data:
    - 4 rooms, 4 sensors in each room.
    - Temperature readings every 30s for 10 days.
    - Symbols: C (cold), P (Pleasant), H (Hot)
    - Example semantic queries: CCHHCC , PPPPP

  - Synthetic data
    - Network size varied between 16-512
    - State size varied between 3-11
    - Correlation parameter ε varied between 0.001 -0.5

# Compared Techniques

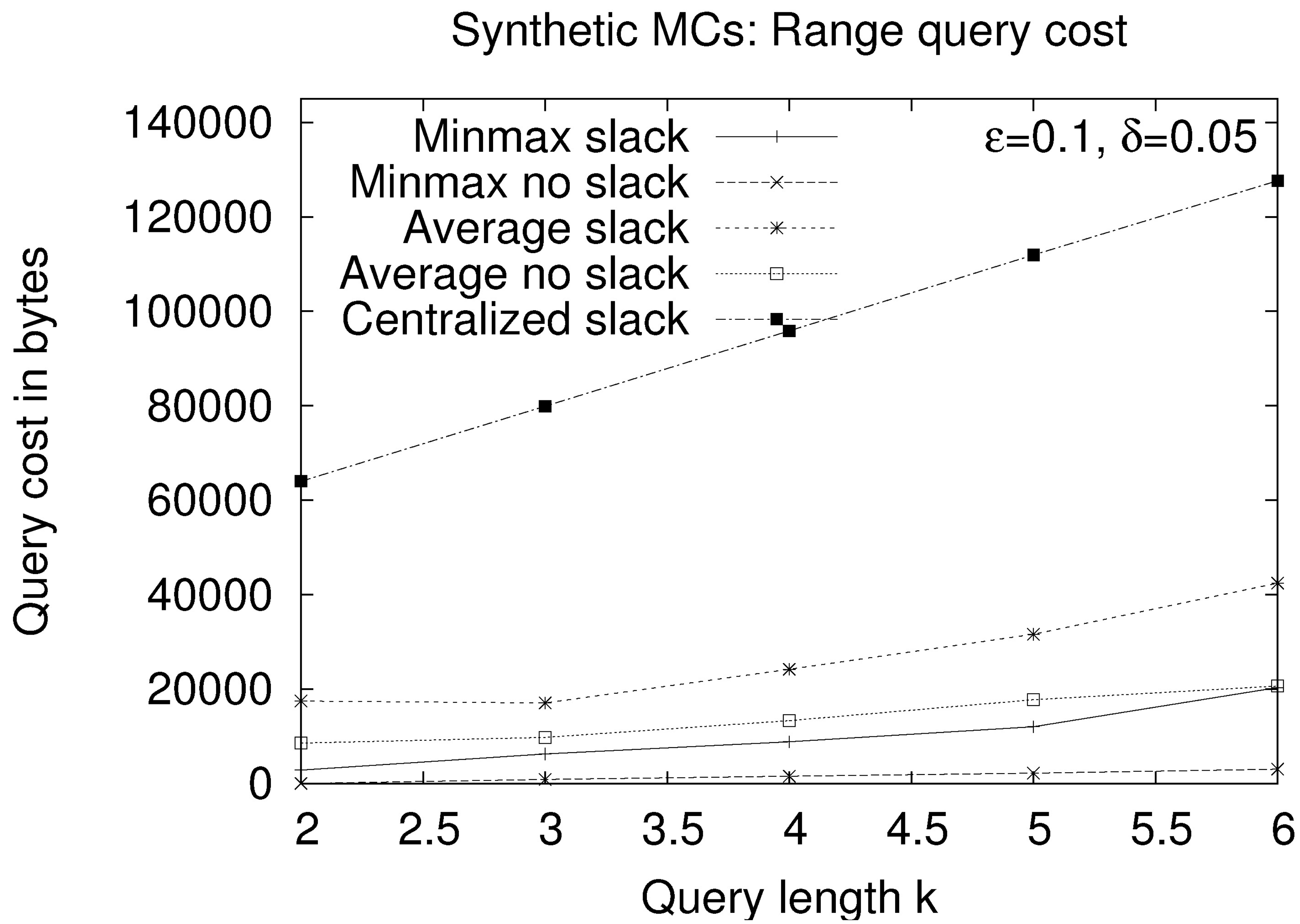
- Centralized scheme with no slack
  - A node transmits each parameter update to the base station
  - Zero querying cost
- Centralized scheme with slack
  - Slack maintained at base station
  - Updates transmitted if they exceed the slack
  - If query cannot be answered by the BS using cached models, it is transmitted to the nodes.
- MIST schemes
  - Without slack
  - With slack at every level of the tree

# Update cost



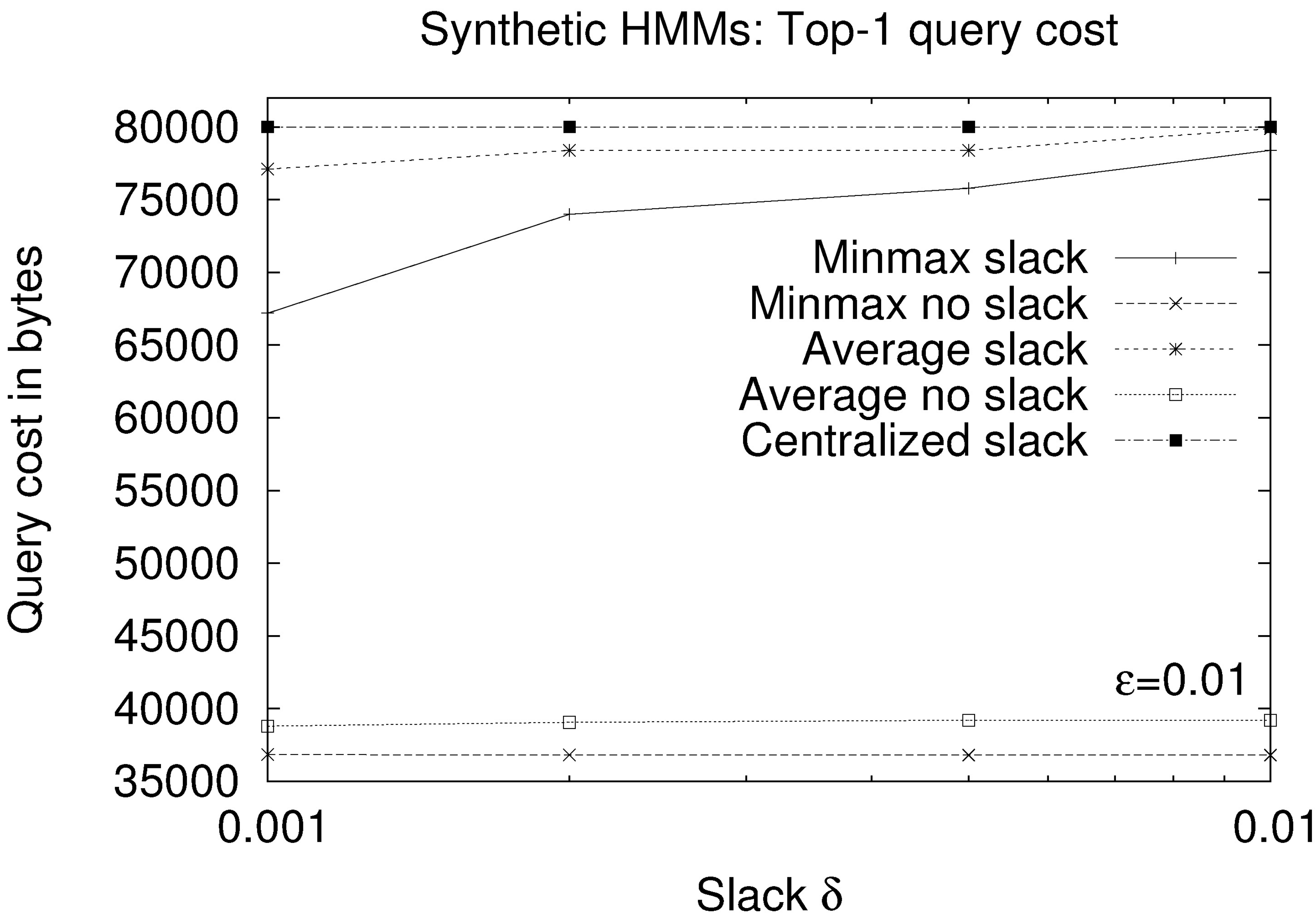
Laboratory MCs: Update cost
Minmax slack
Minmax no slack
Average slack
Average no slack
Centralized slack
Centralized no slack
10000
8000
6000
4000
2000
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Update cost in bytes
Slack δ

# Range query



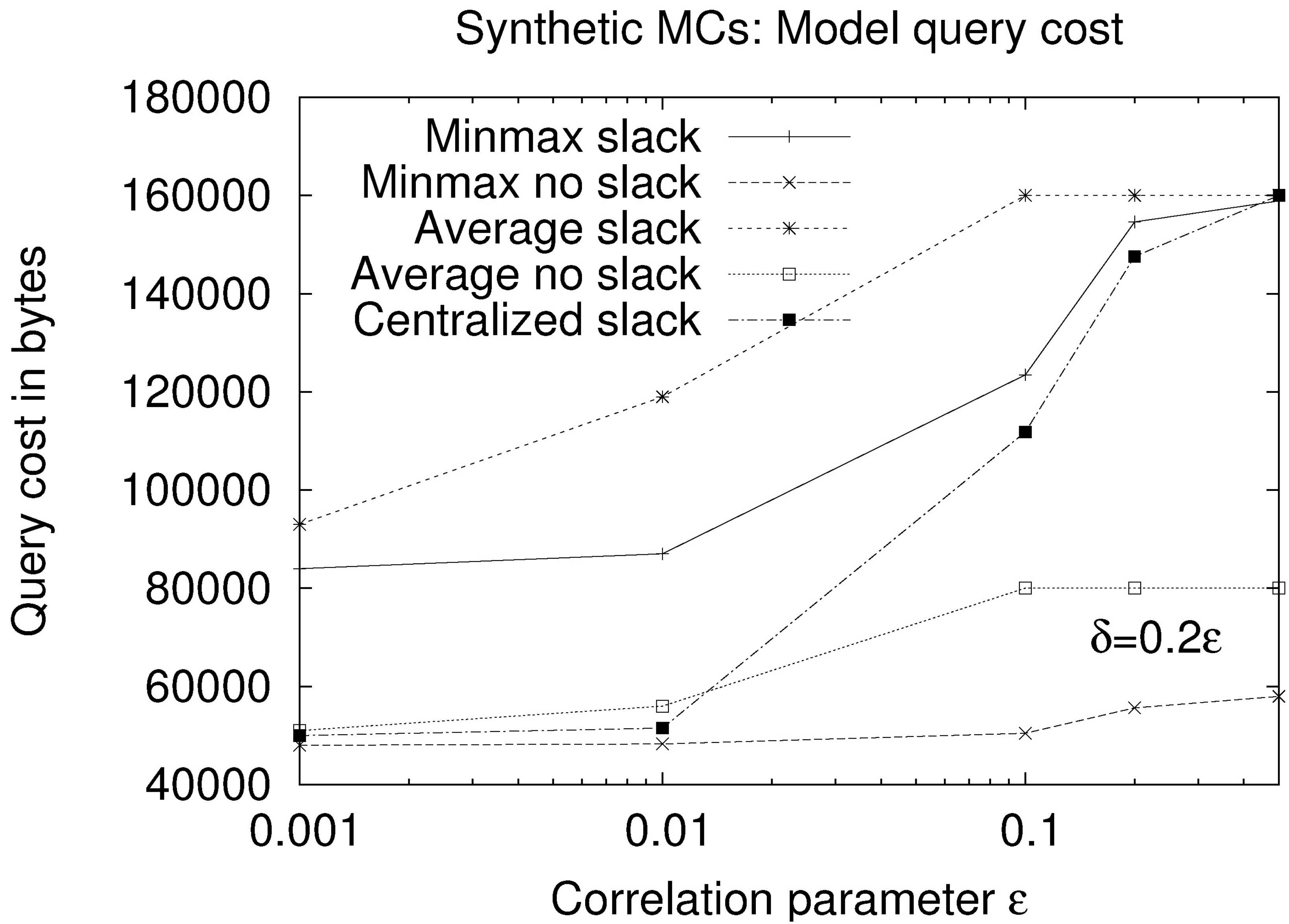
Synthetic MCs: Range query cost
Query cost in bytes
Query length k
ε=0.1, δ=0.05
Minmax slack
Minmax no slack
Average slack
Average no slack
Centralized slack
140000
120000
100000
80000
60000
40000
20000
0
2
2.5
3
3.5
4
4.5
5
5.5
6

# Top-1 query

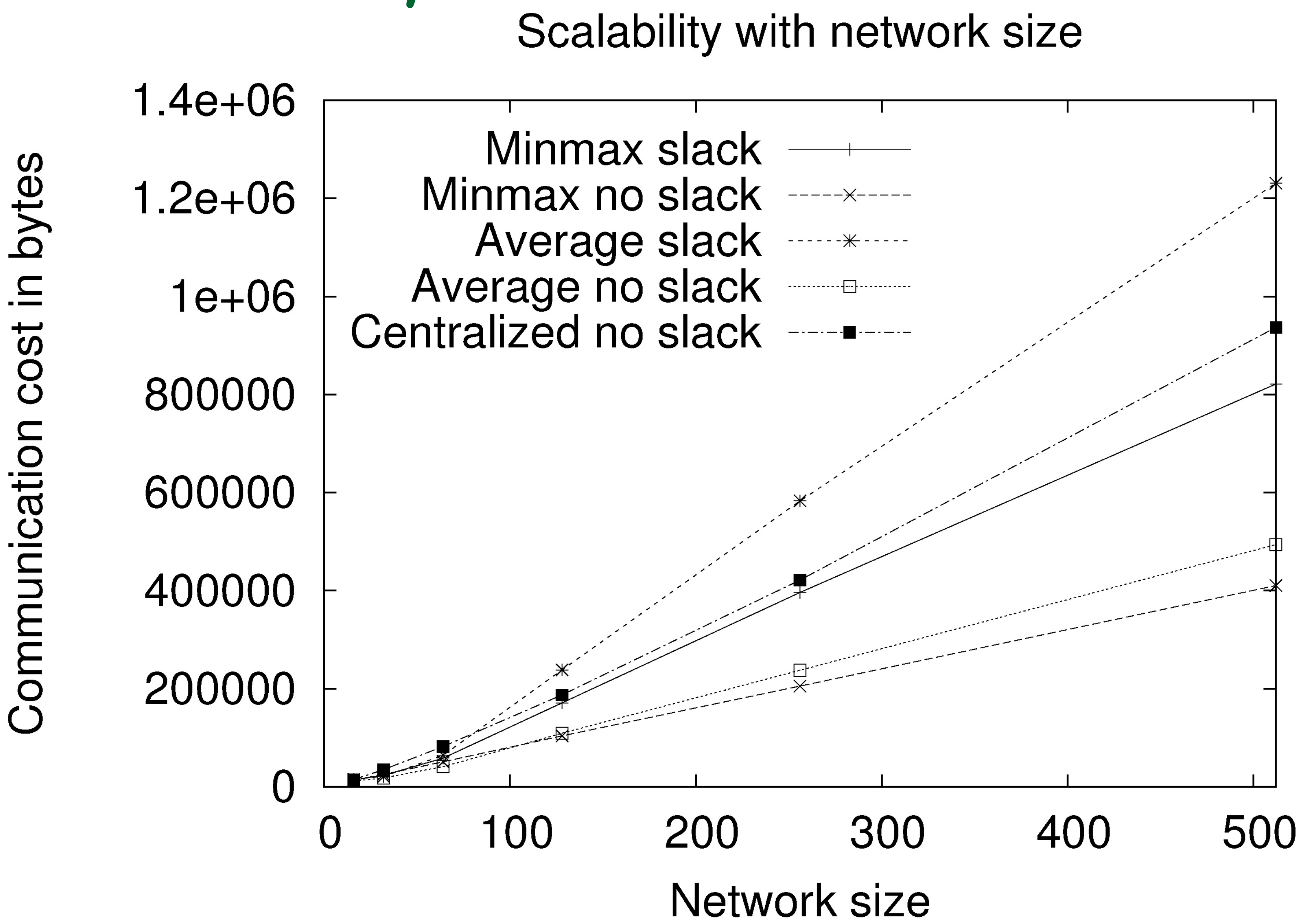


Synthetic HMMs: Top-1 query cost
Query cost in bytes
80000
75000
70000
65000
60000
55000
50000
45000
40000
35000
0.001
0.01
Slack δ
Minmax slack
Minmax no slack
Average slack
Average no slack
Centralized slack
ε=0.01

# Model query


Synthetic MCs: Model query cost
Minmax slack
Minmax no slack
Average slack
Average no slack
Centralized slack
Query cost in bytes
180000
160000
140000
120000
100000
80000
60000
40000
0.001
0.01
0.1
δ=0.2ε
Correlation parameter ε

# Scalability with network size



# Semantic Modeling & Querying

- Motivation & Preliminaries
- MIST: An In-network Model based Index Structure
- Query Algorithms
- Experiments
- Conclusions

# Semantic Modeling & Querying

- Content summarization using MCs and HMMs
- Semantic queries
  - Sequence-based: Range and Top-1 queries
  - Model-queries
- MIST: In-network index structure
  - Average model and $\varepsilon$, $\beta_{max}$, $\beta_{min}$
  - Min-Max model
- Efficient pruning of queries via MIST
- MIST shows superior scalability than centralized schemes in update, query and total communication costs

# Future work

- Other models
- Other ways of summarizing parameters
  - $\beta_{max}$, $\beta_{min}$
- Other query algorithms
- State correspondence problem
- Application domains
- Learning the model

Questions?