A General Framework for Adaptive and Distributed Data Collection in Wireless Sensor Networks

Hongbo Jiang and Shudong Jin

EECS Department
Case Western Reserve University

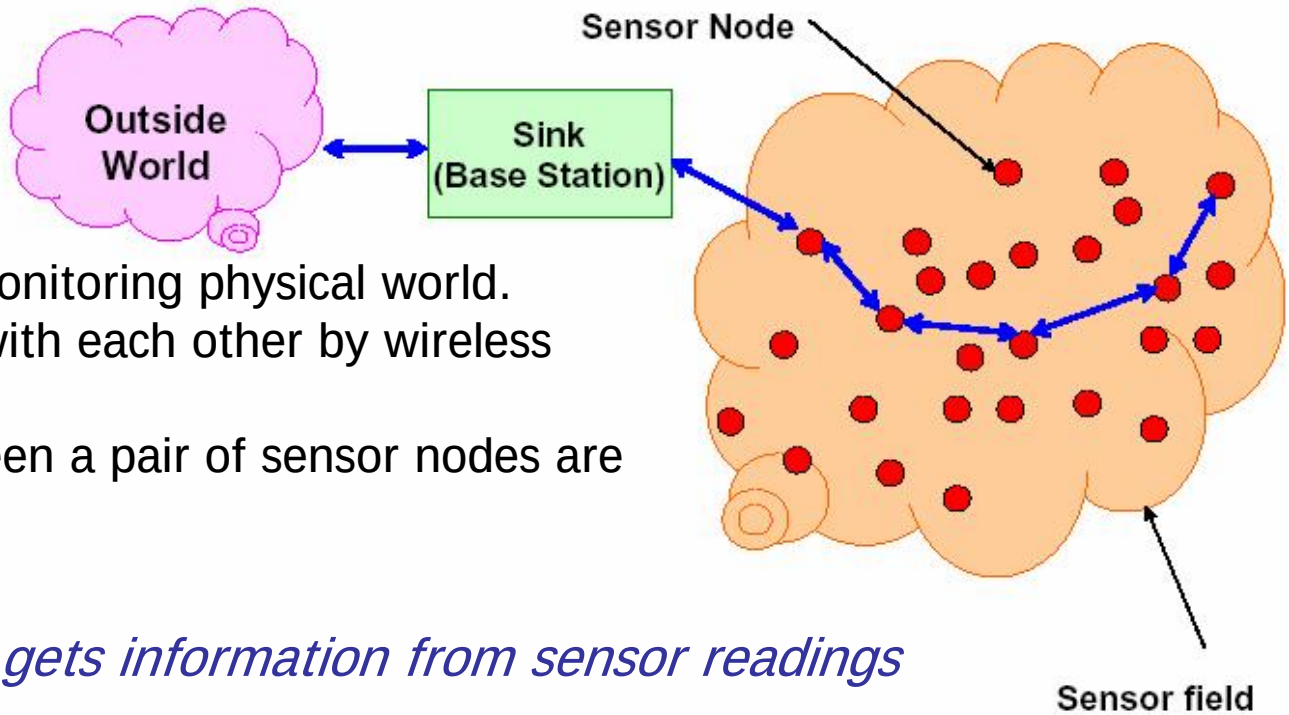
Northeast Ohio Networking Workshop (NEONet 2008)

Outline

- **Introduction**
- Related Work
- Framework
- Results
- Conclusion

Problem Statement

- Two kinds of nodes:
  - Sensor nodes, which are inexpensive, low-power and low computational capacity devices and communicating by means of wireless
  - Base station, which is interface to outside networks and receiving data from sensor nodes



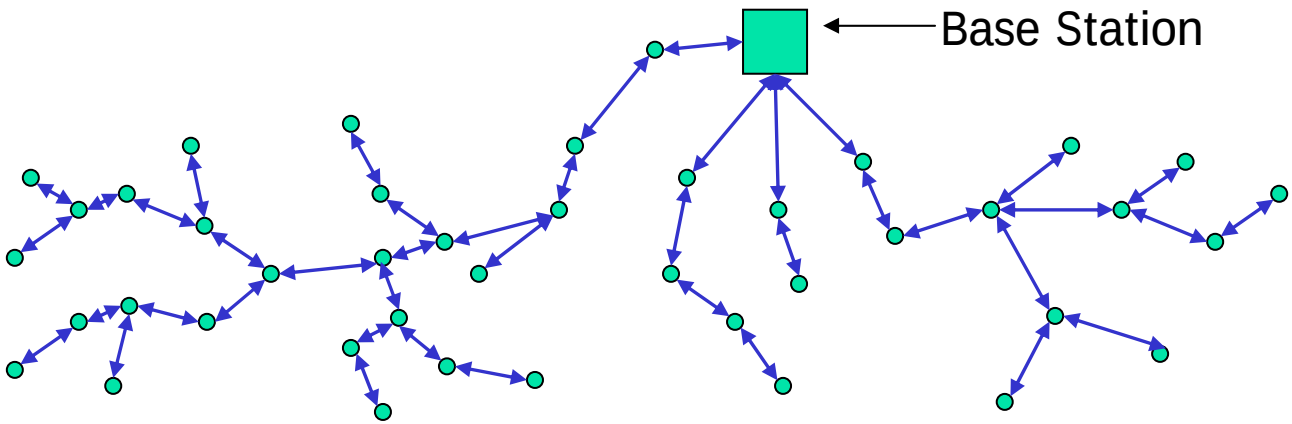
1. Continuously monitoring physical world.
2. Communicate with each other by wireless connection
3. The links between a pair of sensor nodes are lossy

- Task:

  Base station gets information from sensor readings

Problem Statement

- Task:
  - Base station gets information from sensor readings
- Cons:
  - The communication cost is very large if all nodes send packets back to base station one by one
    - At each time, number of packets: O(NlogN)
    - All the time, from 1,2, 3,…,T: O(T*NlogN)

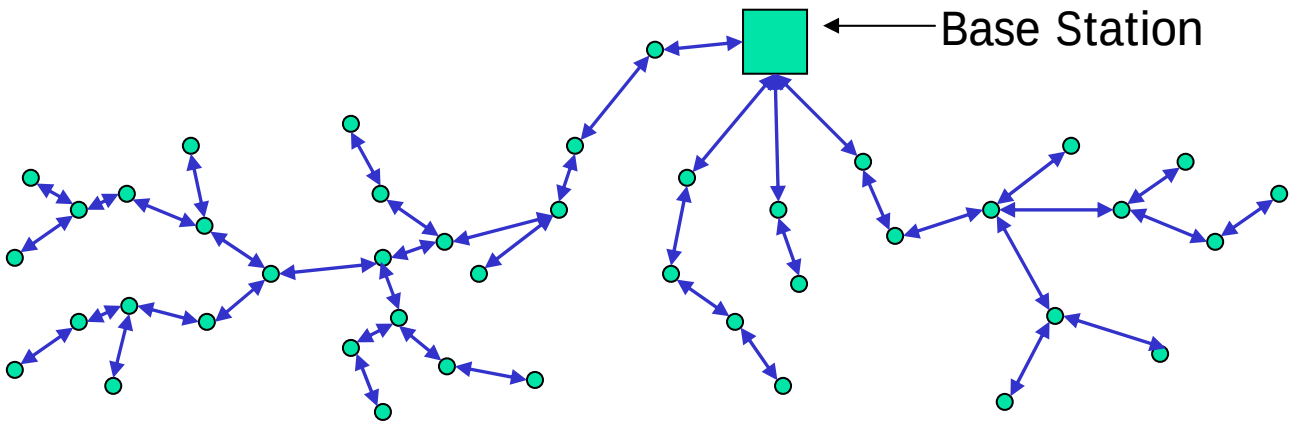


Outline

- Introduction
- **Related Work**
- Framework
- Results
- Conclusion

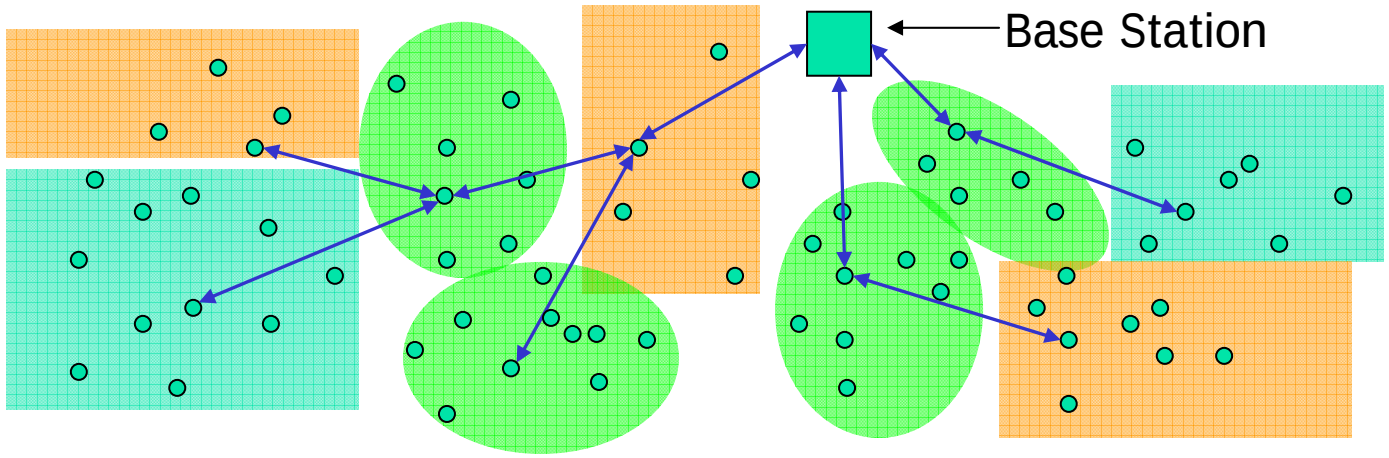
Related Work

- Motivation: Aggregation – I
  - *TAG, Q-Digest (OSDI'02, SenSys'04)*
  - Aggregation at intermediate nodes
    - Number of packets: O(N)
  - Prob. continuously reporting. O(T*N) < O(T*NlogN)



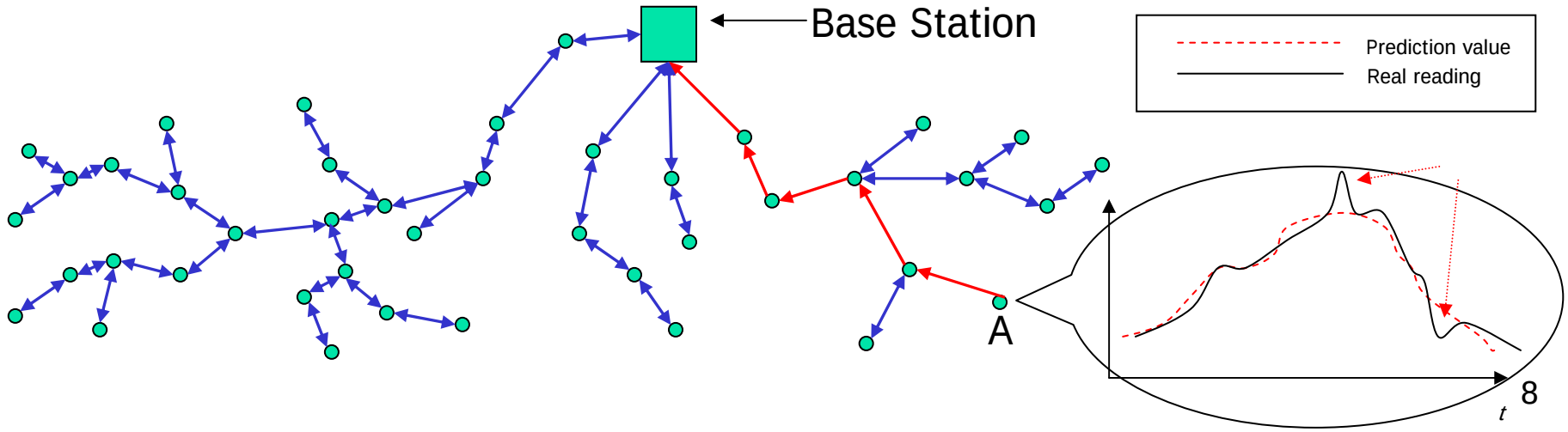
Related Work

- Motivation: - Aggregation – II Clustering
  - *HEED(INFOCOM'04)*
  - Aggregation at cluster head
    - Number of packets: O(N) + O(M*log(N))
    - M: number of cluster heads (M<<N)
  - Prob. O(T*N) + O(T*M*log(N))
    - Continuously reporting.



Related Work

- Motivation – Cont.:
  - *(CIDR'07)*
  - Prediction
    - Number of packets: O(T*E(Inaccuracy of the prediction model)*NlogN) < O(T*NlogN)
    - Each node transmits packet back to the station
      - If and only if the predicted value is far away from the real reading.
  - Prob. Path length is considerable long.

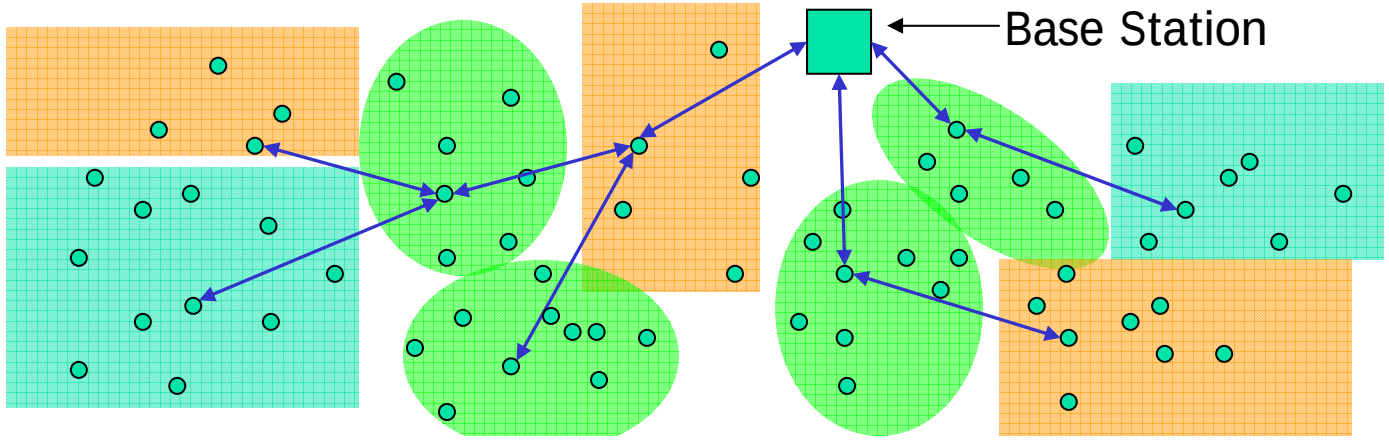


Outline

- Introduction
- Related Work
- **Framework**
- Results
- Conclusion

Framework

- High-level Framework
  - A sensor network is partitioned into many clusters.
  - Cluster head obtains the data by predicting or receiving. It estimates the data distribution within cluster.
  - All cluster heads form a routing tree. The sink obtain the overall distribution via inter-cluster data aggregation over the routing tree.
  - Total # of packets: O(T*E(Inaccuracy of the prediction model)*N*1) + O(T*M*log(N))
    - O(T*NlogN) - Naive
    - O(T*N) – Aggregation
    - O(T*N) + O(T*M*log(N)) – Aggregation II
    - O(T*E(Inaccuracy of the prediction model)*NlogN) - Prediction

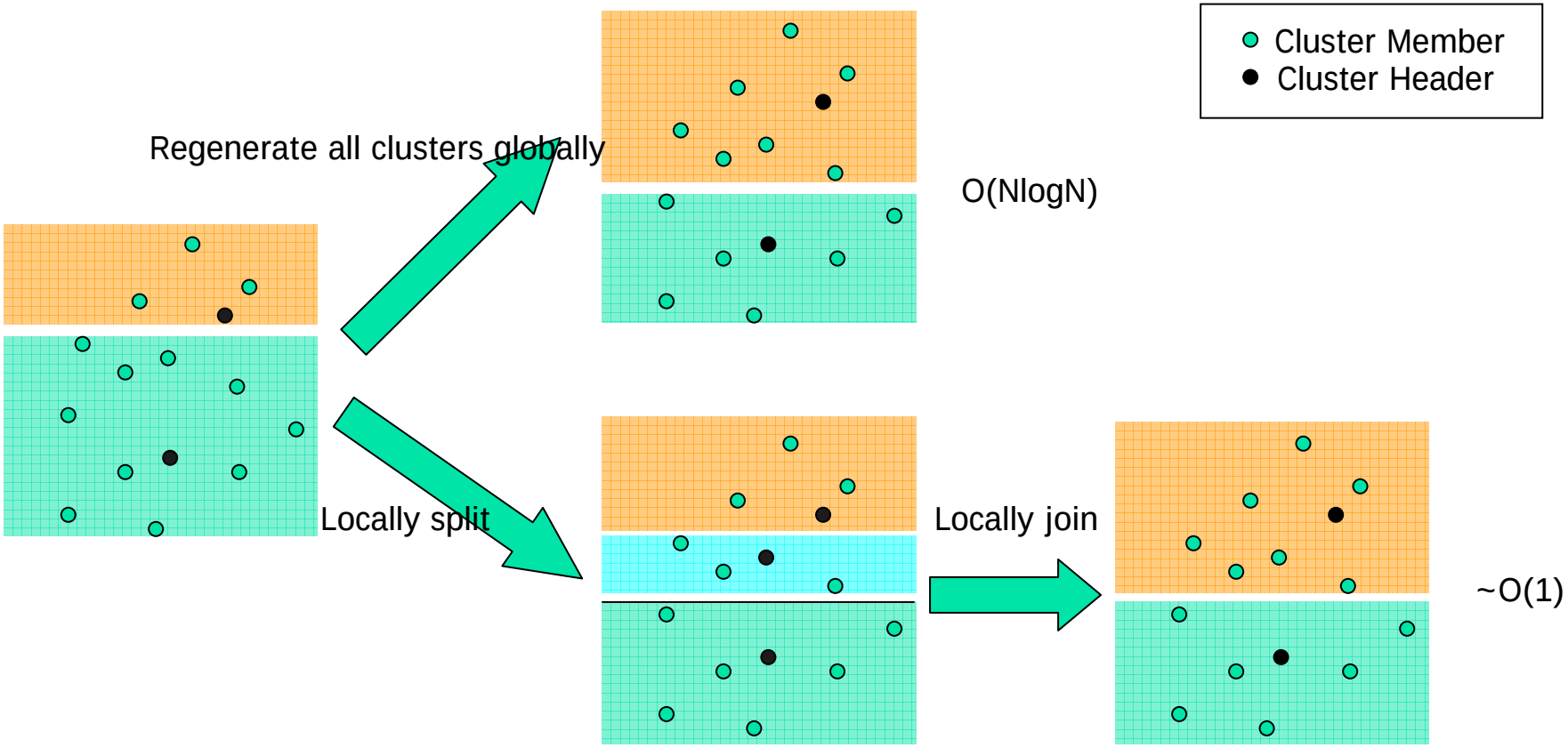


Clustering

- Clustering
  - Clustering Algorithm
    - Initial global clustering algorithm - EDBT'06
    - Each node has a prediction model.
    - Measure the distance of nodes in a neighborhood
      - The distance of parameters for two nodes' prediction model.
    - Associate two neighbor nodes whose distance is less than a threshold in to the same cluster.
  - Cluster Head Selection
    - The node close to sink
  - Update of Clustering
    - Our contribution – Local operations

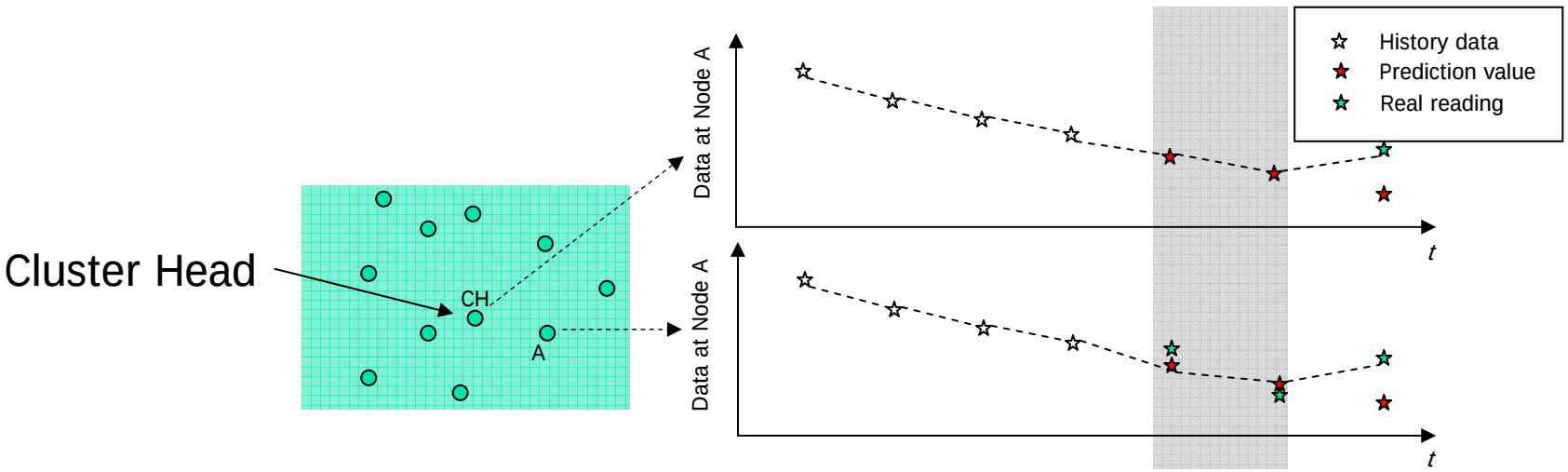
Clustering - Adaptive Update of Clustering

- Main ideas (to maintain locality patterns)
  - To avoid frequently regenerating clusters, locally determine whether to join or split clusters.



Intra-Clustering Prediction

- Combining prediction-based approach and statistical information extraction
  - Prediction model can provide error-bounded continuous data.
    - Node A does not need to continuously report readings.
  - Data distribution (within cluster) to capture statistical information.



Prediction-based Information Processing

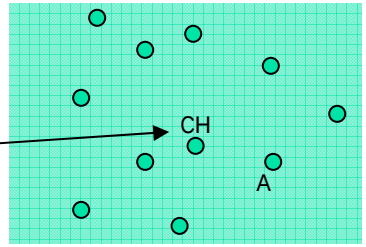
- Prob.
  - Since the node is close to cluster head (within 1 or 2 hops), the prediction computing cost is comparable to the communication cost.
  - An extreme case:
    - The sensor data is totally unpredictable. That is, the node A always has to predict, as well as transmit the data.
    - Disabling prediction will be better.
  - Periodically and adaptively enable/disable prediction
    - Error bound threshold, correlation, the ratio between communication cost and prediction cost.

Corollary 1. *If the error tolerant ϵ and correlation coefficients $\rho_x(j)$ satisfy*

$$\frac{\epsilon}{\sqrt{1-\sum_{j=1}^{p}\phi_j\rho_x(j)}} > m\sigma_x\Phi^{-1}\left(\frac{k+1}{2k}\right), \qquad (2)$$

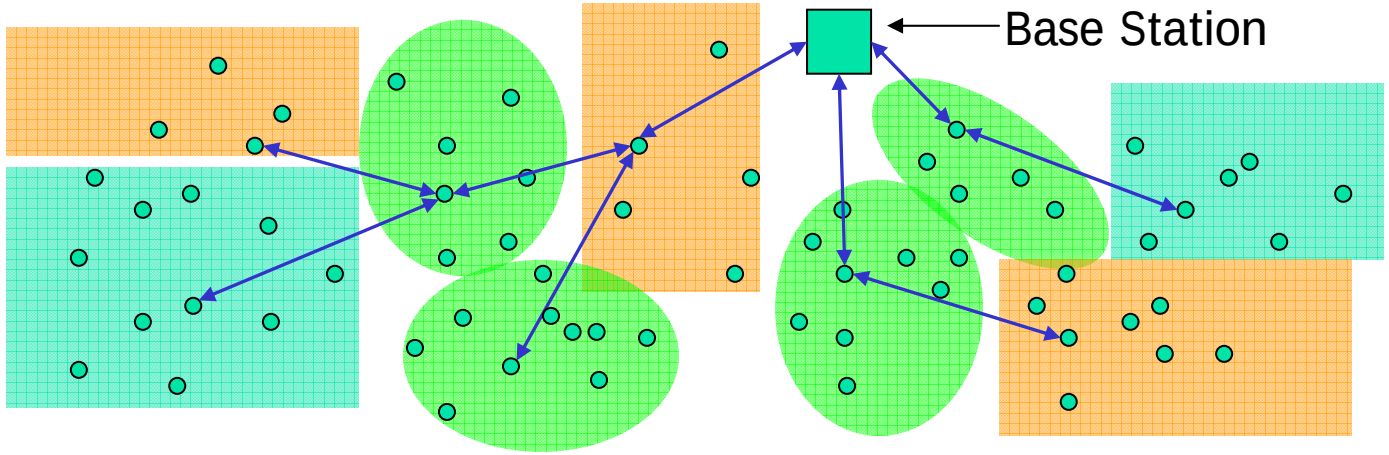
the scheme with local prediction is more energy-efficient.

Cluster Head



Data Aggregation

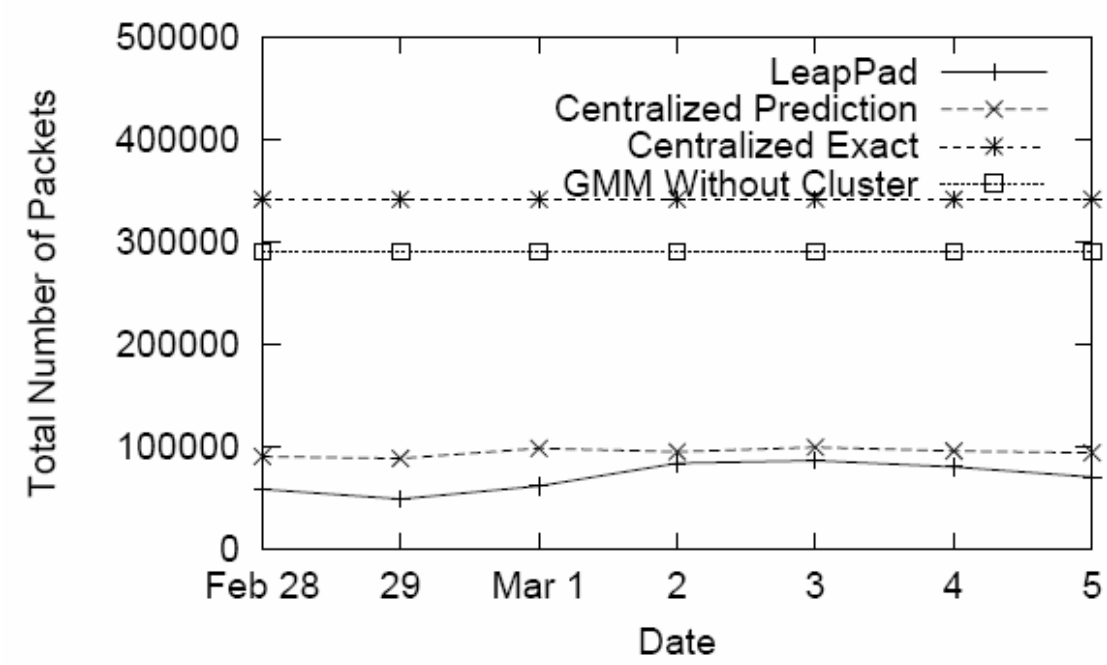
- Inter-Cluster Data Aggregation
  - Cluster head forwards the data (data distribution) over the routing path.
    - CH receives packets from its children.
    - CH aggregates them (and its own data)
      - Combine all input data distributions.
      - Use EM algorithm to estimation parameters
    - CH sends packets to its parents



- Experiment Setup
  - Data-sets:
    - Intel Berkeley Lab Data: 54 nodes, time epoch 30 sec, 1 week.
    - Synthetic data: we also generated synthetic networks and synthetic data.
      - The size of synthetic networks ranges from 100 nodes to 1,000 nodes.
      - Random placement of nodes with a uniform probability distribution.
- Alternative Techniques (we label ours to be “Leap”)
  - Centralized exact:
    - All sensor values are always reported to the sink.
  - Centralized prediction:
    - This technique caches the last a few reported data values at the sink and sensors
    - The sensor does not report a data value if it is within an error bound.
  - Data aggregation without clustering

Experimental Results

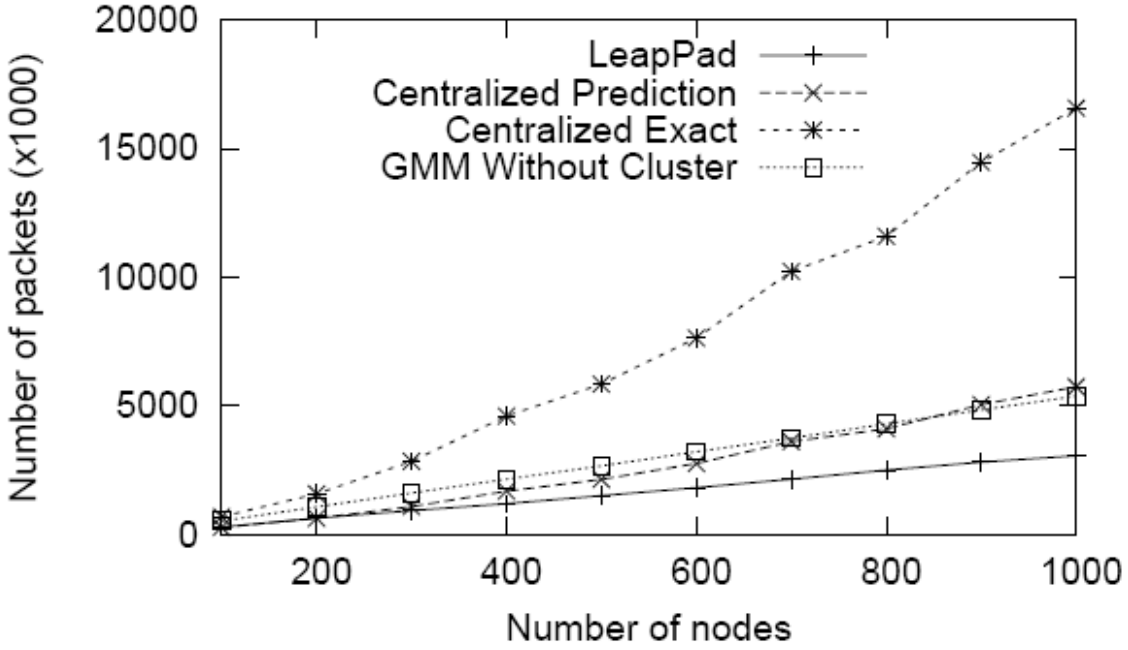
- Overall Communication Cost
  - Intel Berkeley Lab data. (Feb 28 – Mar 5)
  - The total # of transmitted packets.



Total number of packets required for a variety of algorithms

Experimental Results

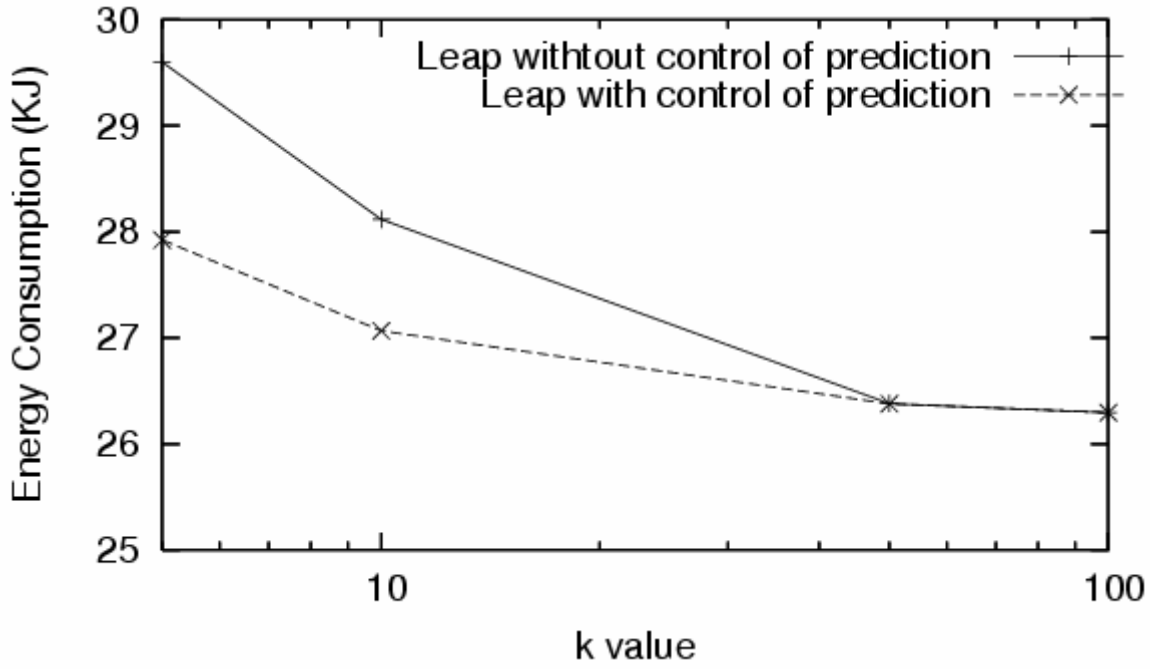
- Overall Communication Cost – Cont.
  - Synthetic data.
  - The total # of transmitted packets.



Total number of packets required for a variety of algorithms

Experimental Results

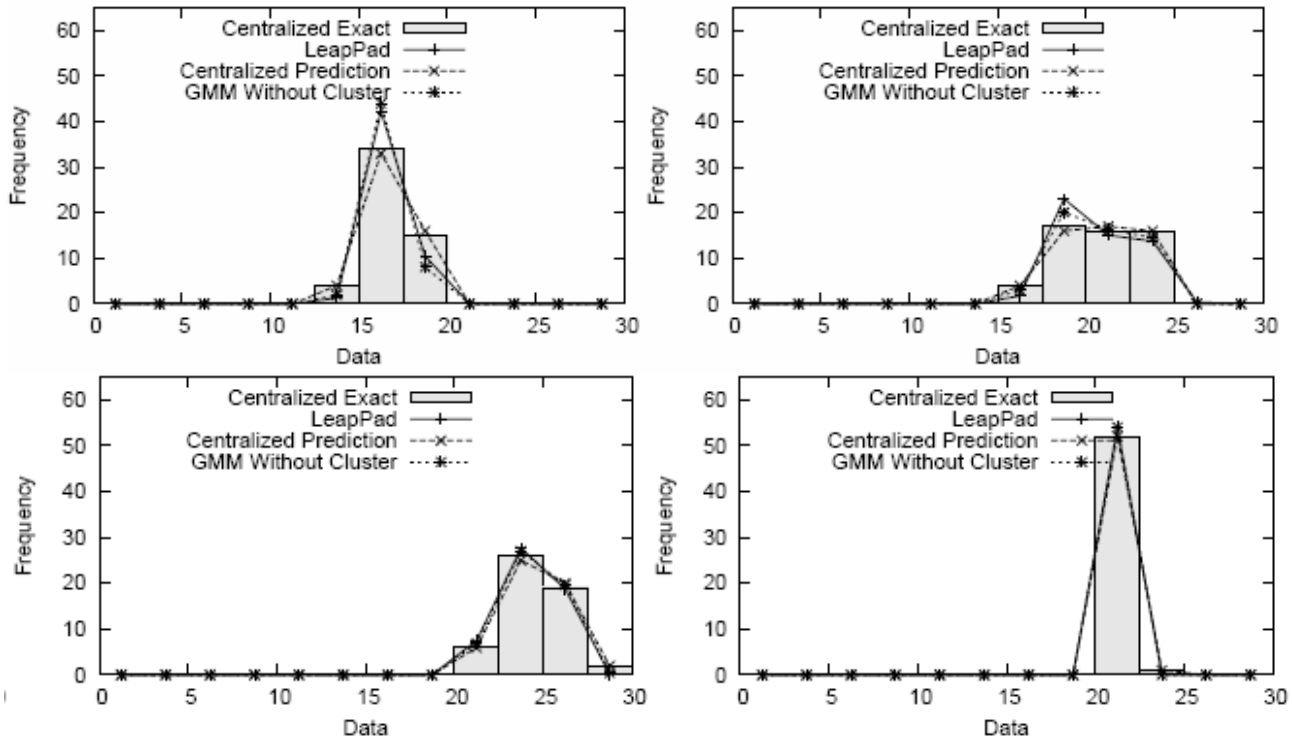
- Benefits of Energy Awareness.
  - Intel Berkeley Lab data. (Feb 28)
  - Overall energy consumption of cluster members.



Energy consumption with / without energy-awareness

Experimental Results

- Aggregation Quality.
  - Intel Berkeley Lab data. (Feb 28)



Comparison of aggregation quality using data histogram. (a) 4am (b) 8am (c) 12pm (d) 4pm

Conclusion

- A framework for continuous data collection in wireless sensor network.
  - It is cluster-based and provide dynamic update.
  - Within cluster, adaptive energy-aware scheme is used to control prediction scheme.
  - Inter-cluster data aggregation is used for data collection.
- Simulation results shows that it outperforms competing algorithms.
- Future directions
  - More efficient algorithms to reduce the computational overhead of our prediction and aggregation techniques.
  - Further improve the framework in order to collaborate with other realistic applications such as object tracking.

Thanks

A General Framework for Adaptive and Distributed Data Collection in Wireless Sensor Networks

Hongbo Jiang and Shudong Jin

EECS Department
Case Western Reserve University