Exploring Data Aggregation

- Aggregate data as it flows through the network
- Data must meet certain criteria
- Not all data can be aggregated
- Lossy & Lossless aggregation

Overview

- What is aggregation?
- Data collection Taxonomy
- Some aggregation algorithms for collecting data from all nodes
  - Escan
  - TAG
  - Cascading timeouts
  - Isolines

Data Aggregation

Data Collection Taxonomy

- When is data generated?
  - Sporadic, Periodic
- Where is data generated?
  - all-producing, selected-producing, single-producing
- Where is data collected?
  - full-site, selected-site, single-site
- How is data packaged? Lossy or Lossless?
  - Concatenation, fusion, addition, reduction
All-producing aggregation schemes

- Characteristics:
  - All nodes produce periodically
  - Data can be aggregated via addition
  - Tree construction is done by simple broadcast. (Not the focus)
- Types
  - No-aggregation
  - Periodic aggregation
  - Per-hop aggregation

Aggregation Topology

eScan

An eScan

- Node constructs eScan with its information. Only reports on significant change
- Gateways express INTERESTS
- INTERESTS construct trees.
- eScans flow through trees.
- Nodes receiving scans can aggregate them if they are near each other and of similar values.
**TAG**

- In-network processing of aggregates
  - Common data analysis operation
    - Aka *gather* operation or *reduction* in || programming
  - Communication reducing
    - Operator dependent benefit
  - Across nodes during same epoch

- Exploit semantics improve efficiency!

---

**Query Propagation**

```
SELECT COUNT(*) ...
```

**Aggregation timing Schemes**

- Simple
- Per-Hop
- Slotted
Cascading Timers

Simulation Parameters
- NS2
- 802.11-like MAC
- 100 nodes
- No movement
- Transmission range = 100m
- Area 500x500
- 3 scenarios, corner sink, center sink, random sink. Random sink is presented here.

Data Freshness

Weighted Freshness
Collection under Packet loss
- Can't spend time recovering
- Loosing aggregated packages is a risk
- FEC
- Double-Send
- Max-Send
- Adaptive-Send

Packets sent per round

Readings collected per round

Packets per reading
“Real” Freshness

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>1s Period</th>
<th>10s period</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.56s</td>
<td>5.00s</td>
</tr>
<tr>
<td>Tag</td>
<td>0.41s</td>
<td>4.23s</td>
</tr>
<tr>
<td>Cascading Timeouts</td>
<td>0.28s</td>
<td>0.28s</td>
</tr>
<tr>
<td>Periodic</td>
<td>2.04s</td>
<td>19.13s</td>
</tr>
<tr>
<td>Per-Hop</td>
<td>1.54s</td>
<td>15.35s</td>
</tr>
</tbody>
</table>

Average delay per packet from being sent to being “considered” at sink.

Aggregating Data by Grouping

- When to aggregate is not enough for collecting all the data on the network
- We need to exploit the correlation between the data being collected.
- Can we group nodes? How do we determine the grouping? How do we represent them?

Isoclusters: Grouping by value

- Group nodes into isoclusters, where all the members have a sensed variable in the same range (i.e. contour maps, isotherms)
- When collecting data we can focus on isolines (isopleths)

Iso-lines
Isolines

- Isolines are lines which pass through our network and have the same value. Nodes detect them by comparing the value they are sensing with their neighbors.
- When nodes detect a nearby isoline they send a report to the data sink.
- Only nodes detecting lines report.

Isoclustering Example

- NS-2, 400 x 400 meter
- 16 x 16 sensor nodes in grid pattern
- 40 meter communication range
- We map reality, no aggregation and isoclusters.
- Isoclusters sends 1/3 of the readings no aggregation sends.
- We map with GMT mapping tools
Isoclustering Example (cont)

Continuous Monitoring
- Monitor an area for events (mapping a continuous variable, like monitoring temperature)
- We want to exploit both spatial-correlation and temporal-correlation
- Report only when data has changed significantly
- Be careful about dead nodes and packet drops

Front monitoring

Front monitoring @ T=7

None
None Optimized
Polygon
Isocluster
Other aggregation

- Aggregate results for a specific output rate (see Directed Diffusion)
- Keep a global estimate from neighbors broadcasts (see An Energy-Accuracy Trade-off)
- Calculate the best aggregation tree (see On Network Correlated Data Gathering)