Distributed Clustering for Robust Aggregation in Large Networks

Ittay Eyal, Idit Keidar, Raphi Rom

Technion, Israel
Aggregation in Sensor Networks – Applications

- Temperature sensors thrown in the woods
- Seismic sensors
- Grid computing load
Aggregation in Sensor Networks – Applications

- Large networks, light nodes, low bandwidth
- Target is a function of all sensed data
- Multidimensional information

Average temperature, max location, majority…
What has been done?
Tree Aggregation and Gossip

Hierarchical solution
Fast - $O(\text{height of tree})$

Hierarchical solution

- Fast - $O(\text{height of tree})$
- Limited to static topology
- No failure robustness


Gossip:
Each node maintains a synopsis

Gossip:
Each node maintains a synopsis
Occasionally, each node contacts a neighbor and they improve their synopses

Gossip:
Each node maintains a synopsis
Occasionally, each node contacts a neighbor and they improve their synopses

- Indifferent to topology changes
- Crash robust
- Proved convergence
- No data error robustness

A closer look at the problem
The Implications of Irregular Data

A single erroneous sample can radically offset the data.

The average (47°) doesn’t tell the whole story.
Sources of Irregular Data

Sensor Malfunction
- Short circuit in a seismic sensor

Software bugs:
- In grid computing, a machine reports negative CPU usage

Sensing Error
- An animal sitting on a temperature sensor

Interesting Info:
- DDoS: Irregular load on some machines in a grid
- intrusion: A truck driving by a seismic detector
- Fire outbreak: Extremely high temperature in a certain area of the woods
Distribution Estimation

Data distribution estimation solutions

○ One dimensional data only [1,2]
○ No data error robustness. [1,2]
Or
○ High complexity [3,4]

Definition: **Outliers**

Samples deviating from the distribution of the bulk of the data

**Previous solutions:**
- 😊 Fast aggregation in a dynamic network
- 😞 No data error robustness

**Our solutions:**
- 😊 Fast aggregation in a dynamic network
- 😊 Data error robustness by outlier detection
Outlier Detection Challenge

A double bind:

Regular data distribution \( \sim 26^\circ \) \quad \leftrightarrow \quad \text{Outliers} \quad \{98^\circ, 120^\circ\}

No one in the system has enough information
Aggregate Clusters

- Each cluster has its own **mean** and **mass**
- A bounded number \((k)\) of clusters is maintained

\[
\begin{align*}
&\text{Original samples} \\
\text{Cluster a and b} \\
\text{Cluster a, b and c} \\
\text{Cluster all}
\end{align*}
\]
But what does the mean mean?

The variance must be taken into account
Distribution is described as $k$ clusters

Each cluster is described by:
- Mass
- Mean
- Covariance matrix
Gossip Aggregation of Gaussian Clusters

(a)

Merge

(b)
Our solution:

- Aggregate a mixture of **Gaussian clusters**
- Merge when necessary

Recognize outliers
Simulation Results:

1. Data error robustness
2. Crash robustness
3. Elaborate multidimensional data
It works where it matters

Not Interesting

Easy
It works where it matters

With outlier detection

No outlier detection
Simulation Results:

1. Data error robustness
2. Crash robustness
3. Elaborate multidimensional data
Protocol is Crash Robust

- Simulation round: each node performs one gossip step
- After each round, 5% crash probability
- No message loss or corruption
Protocol is Crash Robust

No outlier detection, 5% crash probability

No outlier detection, no crashes

Outlier detection
Simulation Results:

1. Data error robustness
2. Crash robustness
3. Elaborate multidimensional data
Describe Elaborate Data

[Graph showing the relationship between Temperature, Distance, and Fire/No Fire]
Robust Aggregation requires **outlier detection**

We present outlier detection by Gaussian clustering:
Summary – Our Protocol

Outlier Detection (where it’s important)

Crash Robustness

Elaborate Data
Future Directions

- Prove convergence properties
- Consider other clustering schemes
- Analyze elaborate data estimation
Thank you

Ittay Eyal, Idit Keidar, Raphael Rom. *Distributed Clustering for Robust Aggregation in Large Networks*, Technion, 2009