Distributed Clustering for Robust Aggregation in Large Networks

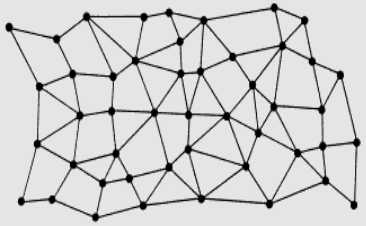
Ittay Eyal, Idit Keidar, Raphi Rom

Technion, Israel

Aggregation in Sensor Networks – Applications

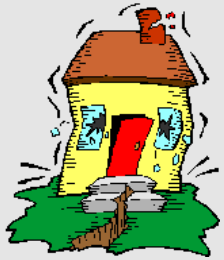
Temperature sensors thrown in the woods

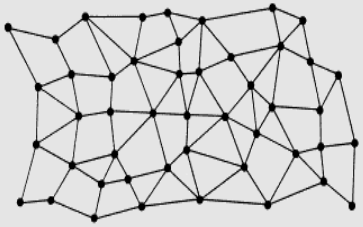
Seismic sensors



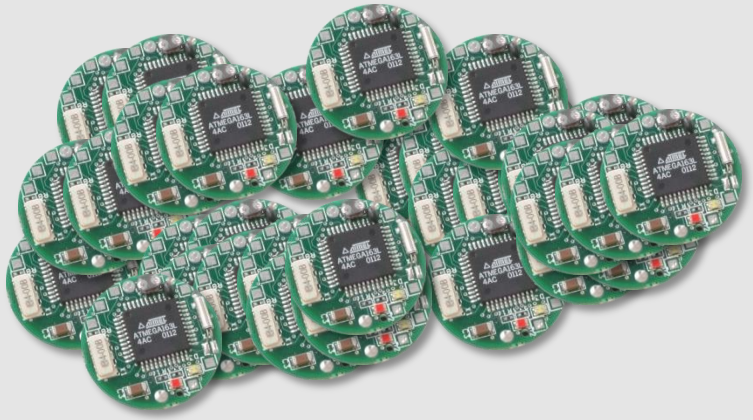
Grid computing load

Aggregation in Sensor Networks – Applications





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

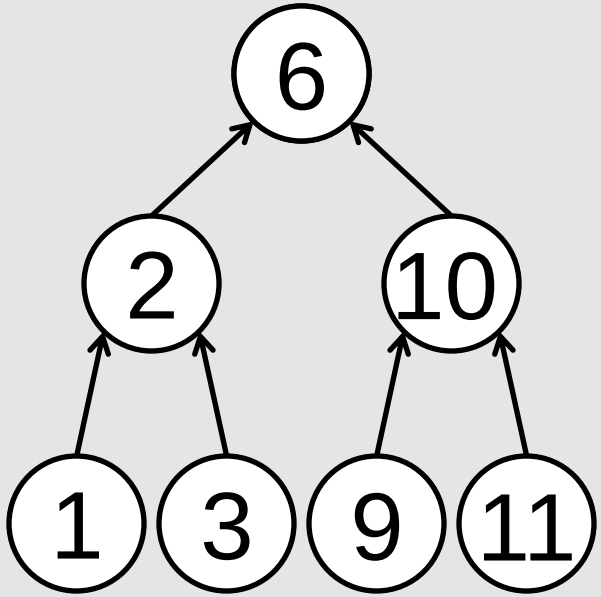


Average temperature, max location, majority…

What has been done?

Hierarchical solution

Fast - O(height of tree)



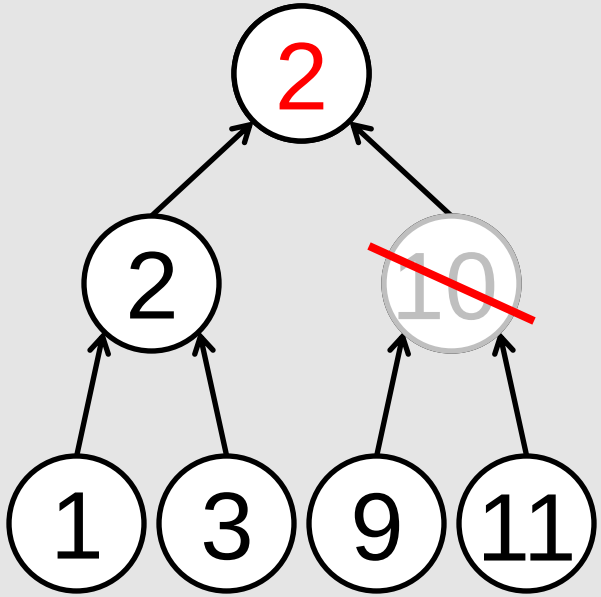
- D. Kempe, A. Dobra, and J. Gehrke. *Gossip-based computation of aggregate information*. In FOCS, 2003.
- S. Nath, P. B. Gibbons, S. Seshan, and Z. R. Anderson. *Synopsis diffusion for robust aggregation in sensor networks*. In SenSys, 2004.

Tree Aggregation and Gossip

Hierarchical solution

Fast - O(height of tree)

☹ Limited to static topology
☹ No failure robustness

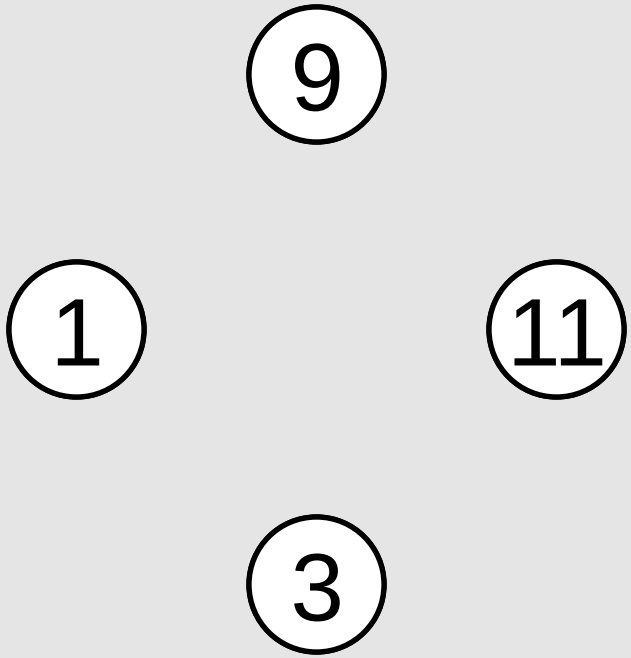


- D. Kempe, A. Dobra, and J. Gehrke. *Gossip-based computation of aggregate information*. In FOCS, 2003.
- S. Nath, P. B. Gibbons, S. Seshan, and Z. R. Anderson. *Synopsis diffusion for robust aggregation in sensor networks*. In SenSys, 2004.

Tree Aggregation and Gossip

Gossip:

Each node maintains a synopsis



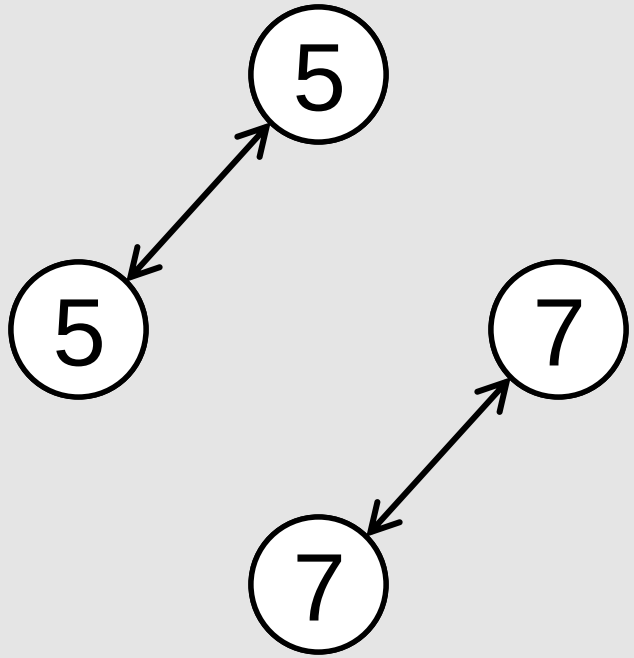
- D. Kempe, A. Dobra, and J. Gehrke. *Gossip-based computation of aggregate information*. In FOCS, 2003.
- S. Nath, P. B. Gibbons, S. Seshan, and Z. R. Anderson. *Synopsis diffusion for robust aggregation in sensor networks*. In SenSys, 2004.

Tree Aggregation and Gossip

Gossip:

Each node maintains a synopsis

Occasionally, each node contacts a neighbor and they improve their synopses



- D. Kempe, A. Dobra, and J. Gehrke. *Gossip-based computation of aggregate information*. In FOCS, 2003.
- S. Nath, P. B. Gibbons, S. Seshan, and Z. R. Anderson. *Synopsis diffusion for robust aggregation in sensor networks*. In SenSys, 2004.

Tree Aggregation and Gossip

Gossip:

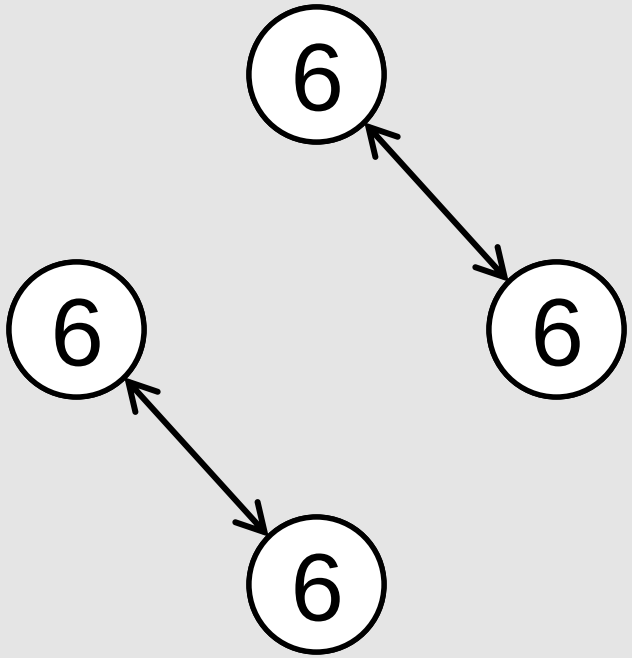
Each node maintains a synopsis

Occasionally, each node contacts a neighbor and they improve their synopses

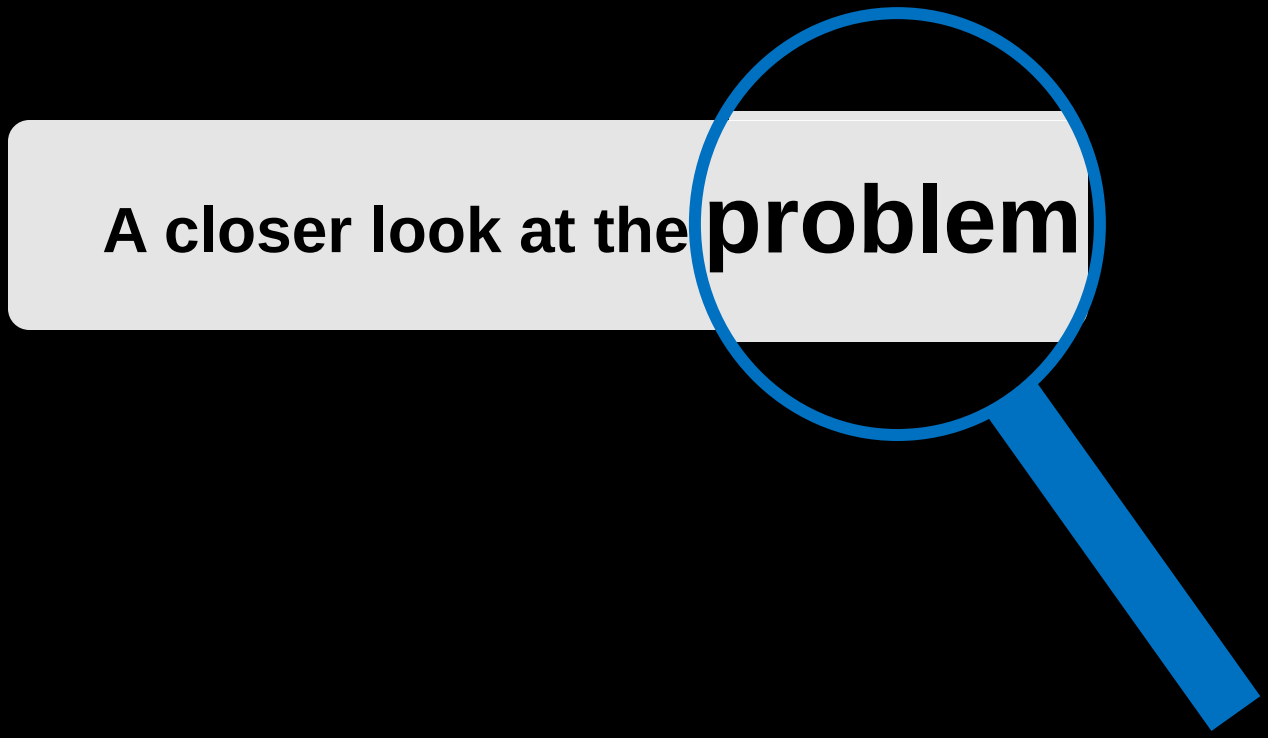
☺ Indifferent to topology changes
☺ Crash robust

Proved convergence

☹ No data error robustness

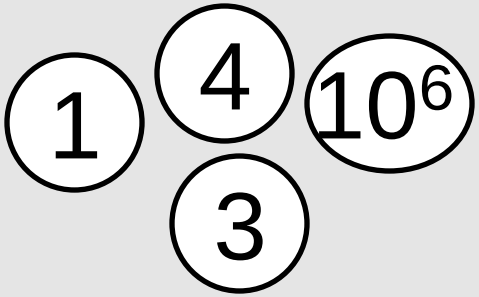


- D. Kempe, A. Dobra, and J. Gehrke. *Gossip-based computation of aggregate information*. In FOCS, 2003.
- S. Nath, P. B. Gibbons, S. Seshan, and Z. R. Anderson. *Synopsis diffusion for robust aggregation in sensor networks*. In SenSys, 2004.


A closer look at the problem

The Implications of Irregular Data

A single erroneous sample can radically offset the data



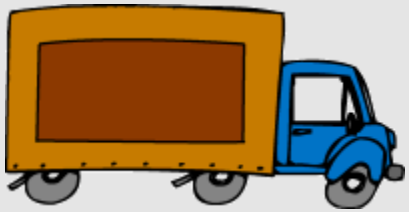
The average (47º) doesn't tell the whole story

27º 25º 26º 25º 28º 98º 120º 27º

Sources of Irregular Data

Sensor Malfunction

Short circuit in a seismic sensor



Interesting Info:
intrusion: A truck driving by a seismic detector

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

Interesting Info:
DDoS: Irregular load on some machines in a grid

Sensing Error

An animal sitting on a temperature sensor

Interesting Info:
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]

1. M. Haridasan and R. van Renesse. *Gossip-based distribution estimation in peer-to-peer networks*. In InternationalWorkshop on Peer-to-Peer Systems (IPTPS 08), February 2008.
2. J. Sacha, J. Napper, C. Stratan, and G. Pierre. *Reliable distribution estimation in decentralised environments*. Submitted for Publication, 2009.
3. W. Kowalczyk and N. A. Vlassis. Newscast em. In Neural Information Processing Systems, 2004.
4. N. A. Vlassis, Y. Sfakianakis, and W. Kowalczyk. *Gossip-based greedy gaussian mixture learning*. In Panhellenic Conference on Informatics, 2005.

Previous solutions:

- ☺ Fast aggregation in a dynamic network
- ☹ No data error robustness

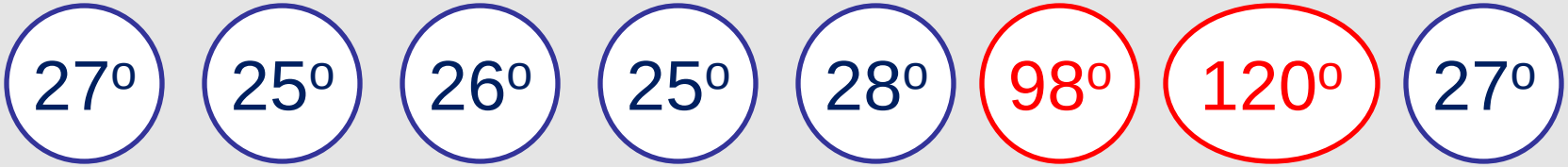
Our solutions:

- ☺ Fast aggregation in a dynamic network
- ☺ Data error robustness by outlier detection

Definition: Outliers

Samples deviating from the distribution of the bulk of the data

Outlier Detection Challenge



Outlier Detection Challenge

27º 25º 26º 25º 28º 98º 120º 27º

A double bind:

Regular data distribution

~26º

⇒

⇐

Outliers

{98º, 120º}

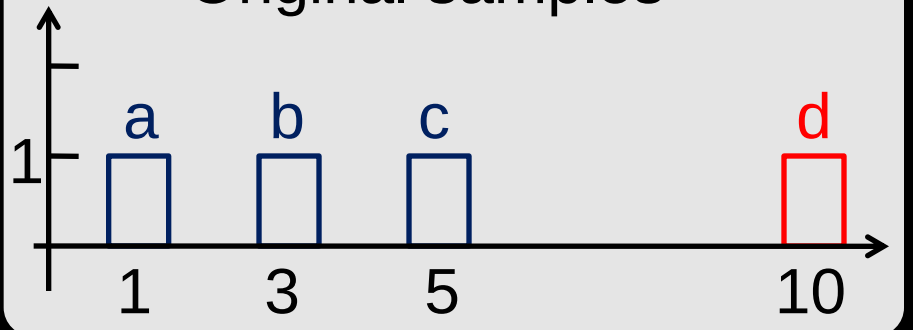
No one in the system has enough information

Aggregate Clusters

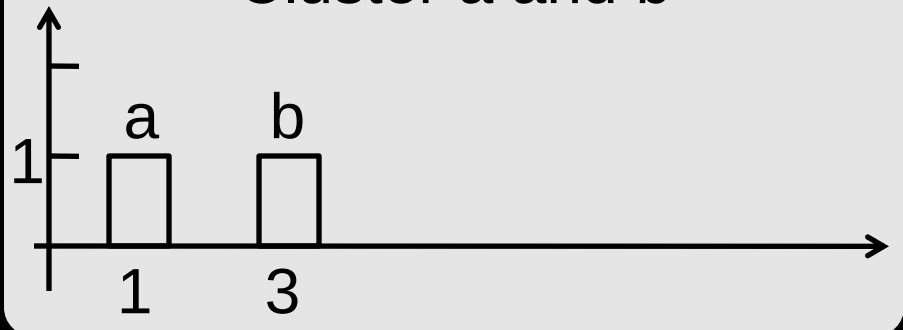
- Each cluster has its own **mean** and **mass**
- A bounded number (k) of clusters is maintained

Here $k = 2$

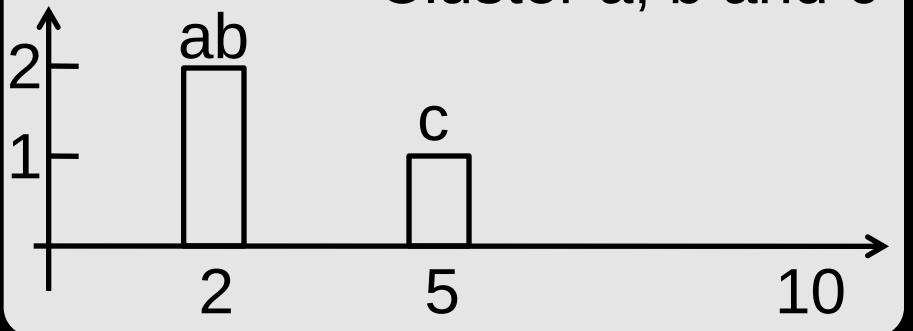
Original samples



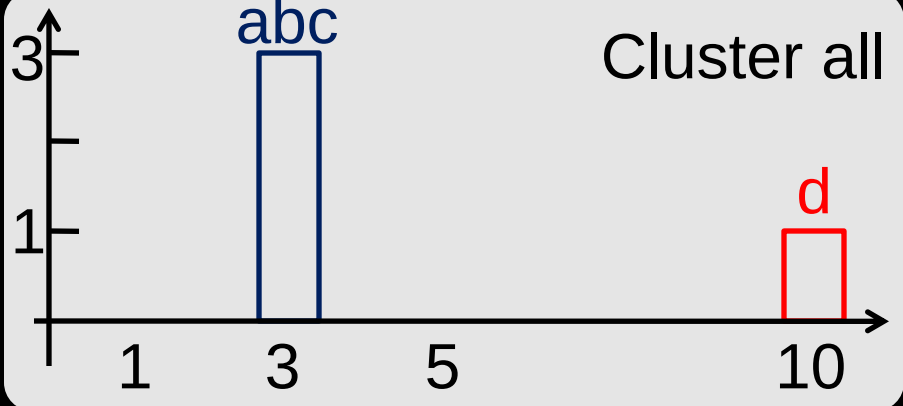
Cluster a and b



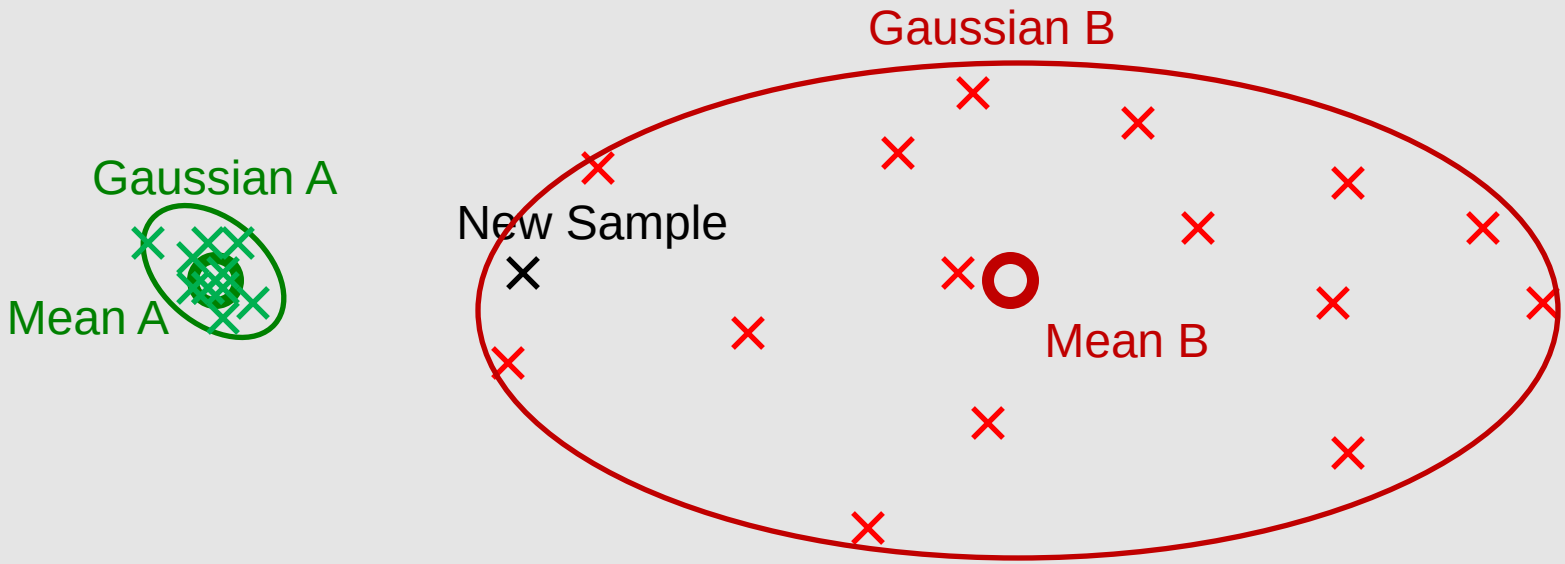
Cluster a, b and c



Cluster all



But what does the mean mean?



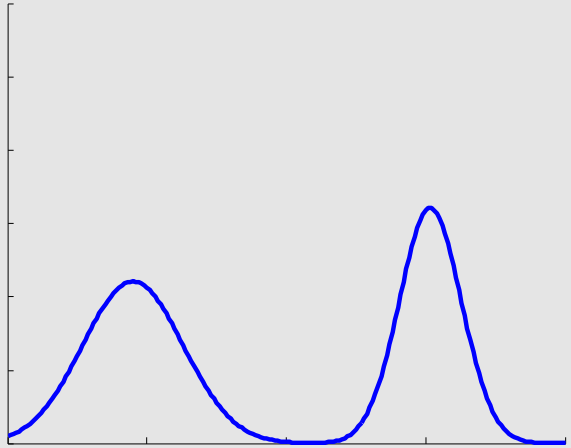
The variance must be taken into account

Gossip Aggregation of Gaussian Clusters

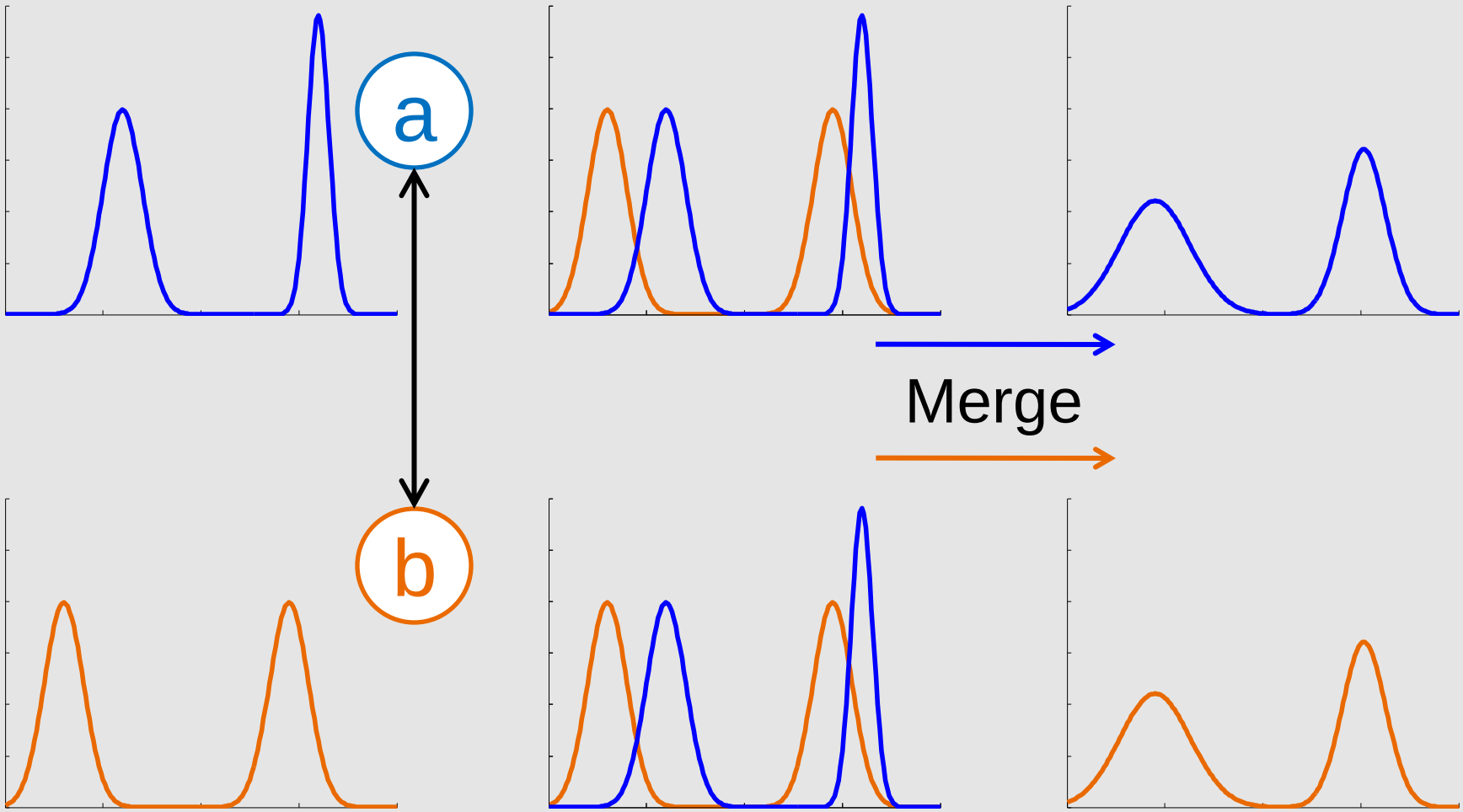
Distribution is described as k clusters

Each cluster is described by:

- Mass
- Mean
- Covariance matrix

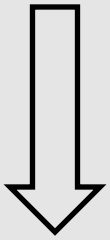


Gossip Aggregation of Gaussian Clusters



Our solution:

- Aggregate a mixture of Gaussian clusters
- Merge when necessary

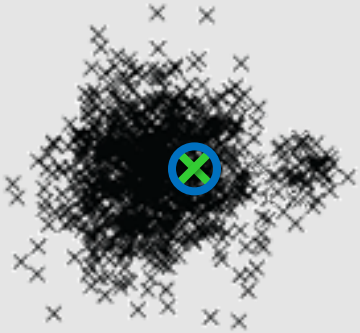


Recognize outliers

Simulation Results:

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

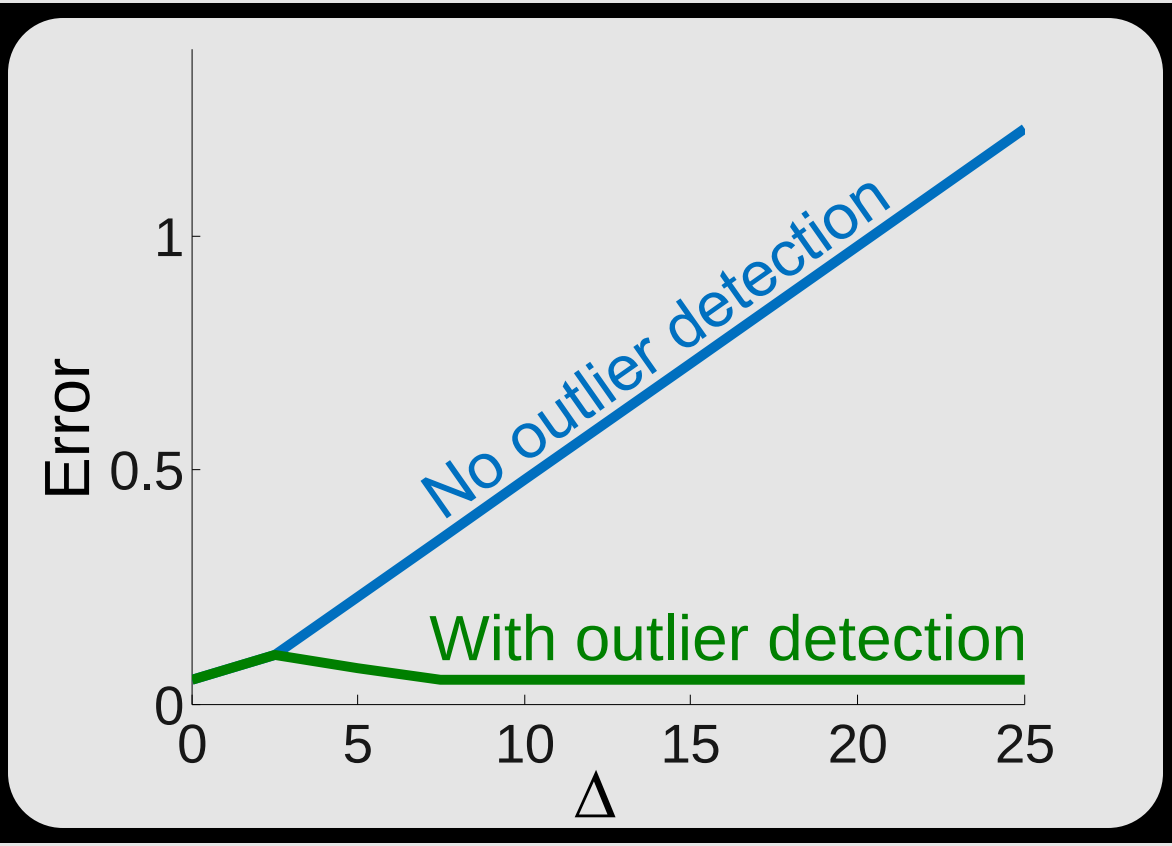
It works where it matters

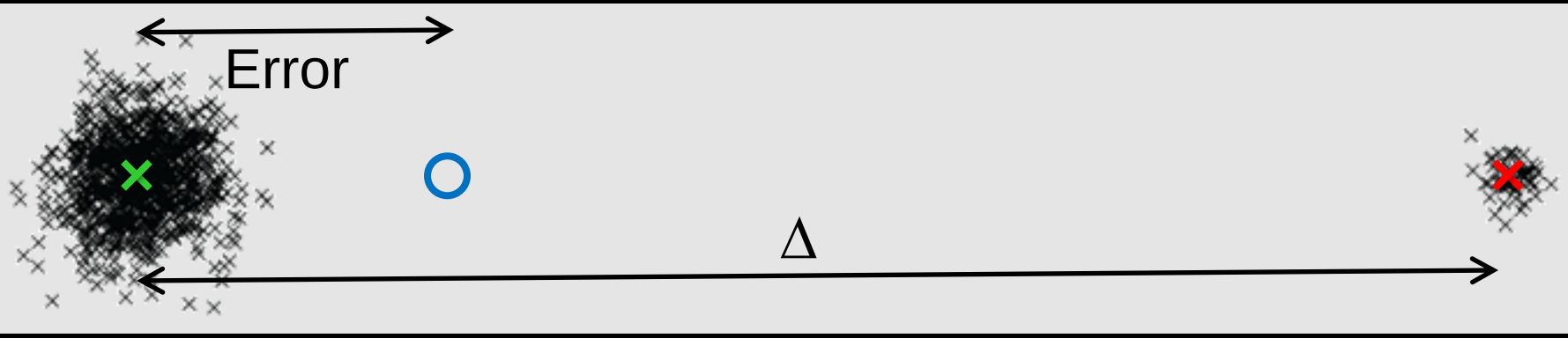


Not Interesting

Easy

It works where it matters





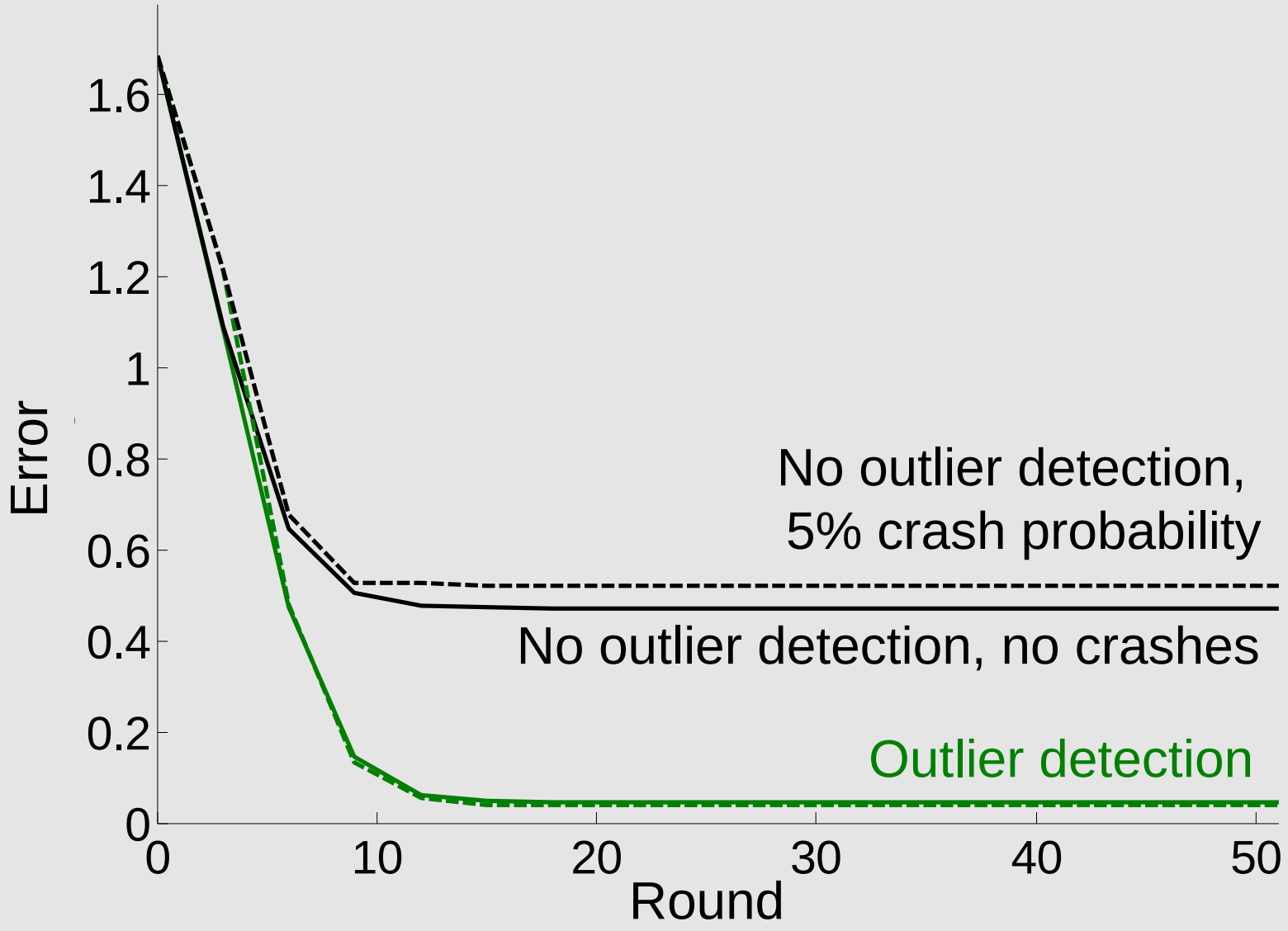
Simulation Results:

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

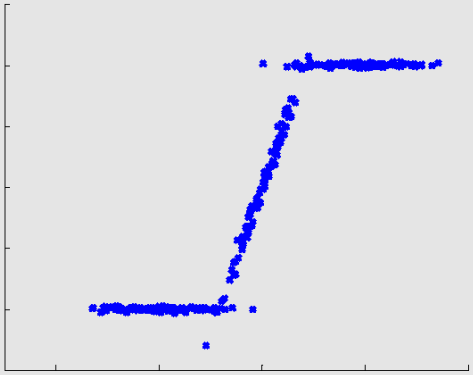
Protocol is Crash Robust

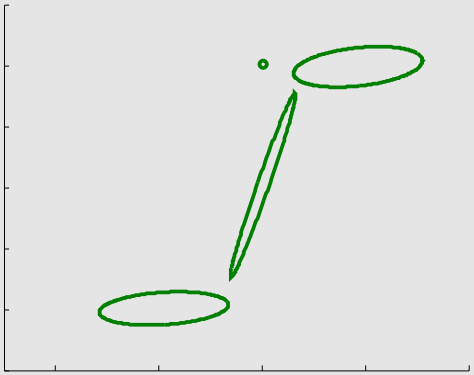


Simulation Results:

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

Describe Elaborate Data



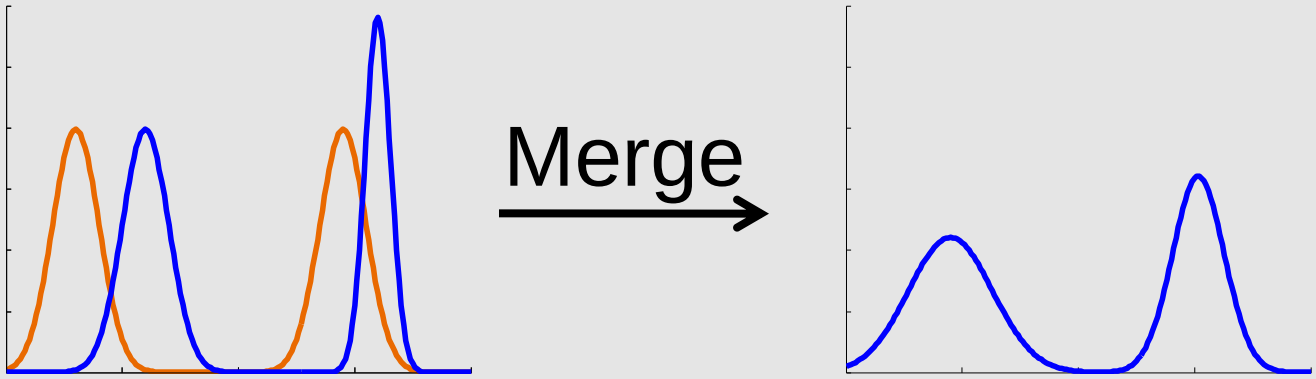


Summary

Robust Aggregation requires outlier detection

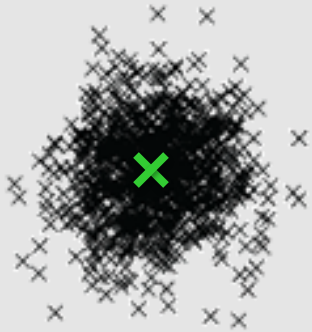
27° 27° 27° 27° 98° 120° 27°

We present outlier detection by Gaussian clustering:

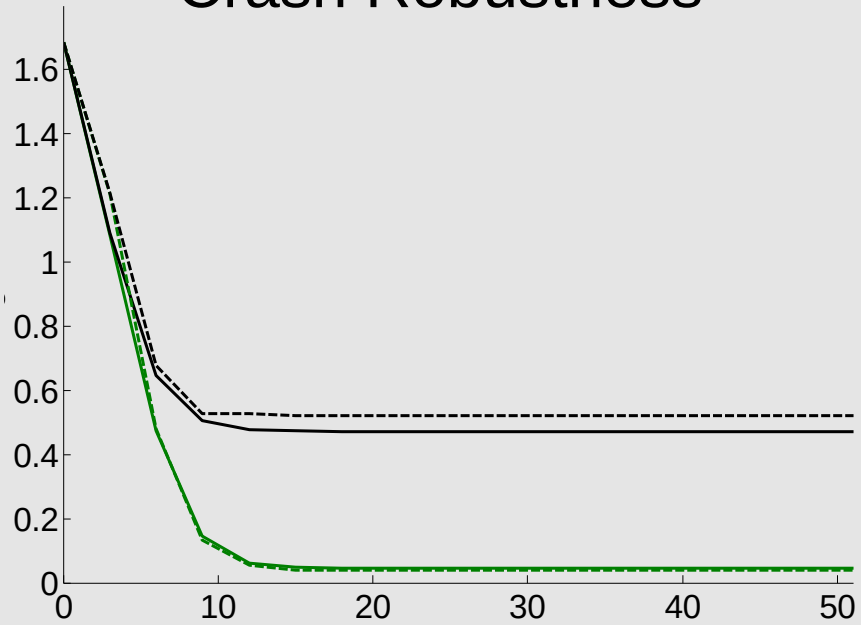


Summary – Our Protocol

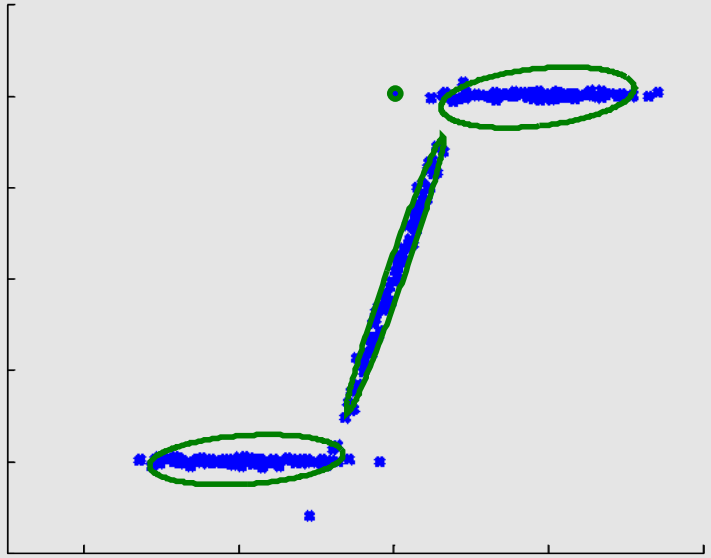
Outlier Detection (where it's important)



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