

Robust Distributed Data Aggregation

MAP-I – Theme Proposal for the 2007/2008 Edition

Carlos Baquero and Paulo Sérgio Almeida – CCTC-DI – U. Minho
cbm@di.uminho.pt, psa@di.uminho.pt

1 Context

Data aggregation is a core technique in the design of efficient sensor networks and scalable systems in general. An introduction to the subject area can be found on Robbert Van Renesse paper on the *The Importance of Aggregation*

... We define aggregation as the ability to summarize information. In the area of sensor networks it is also referred to as data fusion. It is the basis for scalability for many, if not all, large networking services. ...

Many basic distributed paradigms and consistency mechanisms are based on aggregation. For example, synchronization based on voting requires votes to be counted.

One application of aggregation is decentralized network size estimation algorithms. Systems, such as Skype, estimate the number of active users by centralized algorithms, enacting a high demand on central resources and creating single points of failure. Aggregation introduces scalable decentralized algorithms that allow counting (or summing in the general case) over a distributed network of nodes.

A particular application is found on sensor networks, where data collection is often only practicable if aggregation is performed. Aggregation-based techniques can be used to detect how many sensors are deployed in a given area or what is the average temperature in the covered area.

The current state of the art in this area depicts a rich variety of algorithms, showing different accuracies, time and communication tradeoffs. In particular, averaging algorithms are adequate for high precision estimation of aggregates but are highly sensitive to message loss.

2 Objectives

The research will focus on the construction of robust aggregation algorithms that can operate in environments that are prone to message loss, and possibly high levels of churn (variations in the number of nodes). A particular attention will be given to the introduction of fault tolerance in averaging based algorithms, and their deployment in mesh network topologies.

3 References

- [1] R. Van Renesse. The Importance of Aggregation. *Future Directions in Distributed Computing*. Springer LNCS, volume 2584, pages 87-92. 2003.
- [2] D. Kempe, A. Dobra and J. Gehrke. Gossip-based Computation of Aggregate Information. *44th Annual IEEE Symposium on Foundations of Computer Science*. 2003.
- [3] M. Jelasity, A. Montresor and O. Babaoglu. Gossip-based Aggregation in Large Dynamic Networks. *ACM Transactions on Computer Systems*. 23(3):219-252. 2005.
- [4] J. Chen, G. Pandurangan and D. Xu. Robust Computation of Aggregates in Wireless Sensor Networks: Distributed Randomized Algorithms and Analysis. *IEEE Transactions on Parallel and Distributed Systems*. 17(9):987-1000. 2006.
- [5] M. Durand and P. Flajolet. Loglog Counting for Large Cardinalities. *ESA*. Springer LNCS, volume 2832, pages 605-617. 2003
- [6] M. Bawa, H. Garcia-Molina, A. Gionis and R. Motwani. Approximate Aggregation Techniques for Sensor Databases. *ICDE, IEEE*, pages 449-460. 2004.
- [7] K. Horowitz and D. Malkhi. Estimating Network Size from Local Information. *Information Processing Letters*. 88(5):237-243. 2003.
- [8] P. S. Almeida, C. Baquero, N. Preguiça, D. Hutchison. Scalable Bloom Filters. *Information Processing Letters*. 101(6):255-261. 2007.