Gossip-based aggregate computation: computing faster with non address-oblivious schemes

ROBERTO DI PIETRO
Università di Roma Tre, Roma – Italy;
UNESCO Chair in Data Privacy, Tarragona – Spain
and
PIETRO MICHIARDI
Institut Eurécom, Sophia-Antipolis – France.

In this paper, we sketch a novel gossip-based scheme that allows all the nodes in an n-node overlay network to compute a common aggregate (MAX) of their values using $O(n \log \log n)$ messages within $O(\log n)$ rounds of communication. The proposed algorithm can be intuitively extended to compute other aggregates such as MIN, SUM, AVERAGE, and RANK. In contrast, the best known gossip-based algorithms for computing these aggregates require either $O(n \log n)$ messages and $O(\log n)$ rounds or $O(n \log \log n)$ messages and $O(\log n \log \log n)$ rounds. Preliminary simulations confirm our analytical findings.
Our result is achieved relaxing the hypothesis that nodes are address-oblivious, raising the interesting question whether this paradigm (address-aware) is more expressive than the address-oblivious one.

Categories and Subject Descriptors: H.1.1 [**Systems and Information Theory**]: Subject—*Information theory*; F.1.2 [**Modes of Computation**]: Subject—*Parallelism and concurrency*

General Terms: Algorithms, Design, Theory.

1. INTRODUCTION

Many large-scale distributed applications require aggregate statistics (e.g., MIN, MAX, SUM, AVERAGE) to be computed over data stored at individual nodes. Depending on the application, the aggregate computation procedure must satisfy some of the following requirements: scale to a large number of nodes; be robust in the presence of link and node failures; incur low communication overhead.
Gossip-based schemes are one solution to the above-mentioned scalability and reliability problems. Researchers have proposed decentralized gossip-based schemes for computing various aggregates in overlay networks [Boyd et al. 2006; Kempe et al. 2003]. In gossip-based protocols, each node exchanges information with a randomly chosen communication partner in each round. The main contribution of this paper is to show that we can compute MAX in just $O(\log n)$ rounds and $O(n \log \log n)$ messages, relaxing the hypothesis that node are address-oblivious. Note that it is possible to extend the same algorithm to compute other aggregates.
Our result is interesting when compared to [Karp et al. 2000]. In that work the authors proved that a single message cannot be spread in a network using less than $O(n \log \log n)$ message exchanges for a class of algorithms referred to as *address-oblivious* algorithms. Our algorithm is not address-oblivious; we are leaving for future work to prove whether the proposed algorithm is optimal, or it is just a (good) heuristic.

2. NON ADDRESS-OBLIVIOUS GOSSIP-BASED AGGREGATION

In this section we sketch our algorithm, focusing on the computation of the MAX value. The algorithm can be intuitively extended to compute other aggregate values. The main characteristic of our approach is that, relaxing the address-oblivious hypothesis, our algorithm can reduce both the required rounds and messages to have all the nodes to share the value MAX, with high probability.

We assume a *bootstrap phase* wherein, based on the address of the nodes: the network is divided into approximately equally sized *clusters* (C_j) of cardinality $\Theta(\log^2 n)$; within each cluster, some nodes are selected to be cluster-head for the cluster: the *cluster head sets* (CHS_j) will have approximately the same cardinality ($\Theta(\log n \log\log n)$). In the extended version of this paper we show that: selecting clusters and cluster heads can be practically achieved leveraging the node address [Pietro et al. 2004] and hashing; the requirements on the cardinality of both clusters and cluster heads are meet (with high probability) via the Chernoff-Bounds [Alon et al. 1992].

Our proposed algorithm is divided into three phases. In this section we discuss the general idea and we will then focus on a specific method to compute the MAX.

—Phase 1: *Intra-cluster communication*. In this first phase, nodes communicate their values to their respective cluster heads. Formally, $\forall n_i \in C_j, \notin CHS_j$ a node n_i sends $val(n_i)$ to a randomly chosen $n_k \in CHS_j$. This phase is carried out simultaneously for all clusters in which the network is divided into. At the end of this phase, the cluster head set of every cluster holds the aggregate information for the corresponding cluster.

—Phase 2: *Inter-cluster communication*. Only nodes belonging to cluster head sets are involved in this phase. During this phase a node leverage the gossip-based aggregation algorithm presented in [Kempe et al. 2003], with the constraint that the address space used to place independent and uniformly random calls is restricted to the nodes belonging to some cluster head set. In other words, all nodes $n_i \in CHS_j$, $\forall i, j$ select independently and uniformly at random a remote node $n_k \in CHS_l$, $\forall k, l$ and send a message to it. At the end of this phase, nodes that belong to the cluster head sets hold the final aggregate value that accounts for all values initially stored at all nodes in the network. Note that at the end of this phase, all cluster heads share the same value MAX.

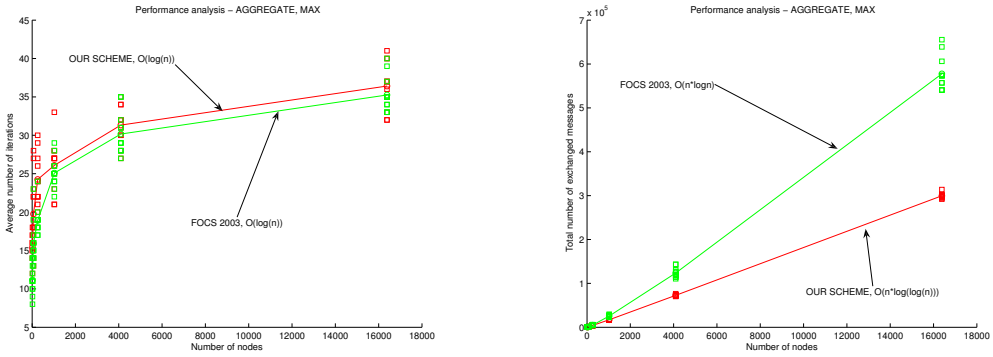
—Phase 3: *Intra-cluster communication*. This last phase is used to diffuse the final aggregate value from the cluster head set to all members of the same cluster. In this phase, $\forall n_i \in CHS_j$ send the final aggregate value to a randomly chosen node $n_j \in C_j$.

Our approach has two key features: (i) it *shifts the cost* of message complexity of address-oblivious schemes to memory: nodes have to store a (reasonable) amount of information on cluster heads. Note that it is also possible to trade-off the memory cost with computations: nodes could compute the needed information on clusters and cluster heads on the fly. Indeed, clusters and cluster heads are generated according to the address (ID) of the nodes; (ii) although we introduce a *structure* (clusters) to the original unstructured gossip-based approach, this has little impact on the resilience of our scheme to node and link failures, as preliminary analysis and simulation results show.

While a rigorous analysis of the proposed protocol will be provided in the extended version of this paper, in the following we show a simpler version (Simple MAX) of our algorithm that provides an intuitive characterization of the round and message complexity of our approach.

Simple MAX: Specific to the problem of computing the MAX, it is possible to modify the three phases described above as follows. In every phase, the gossip based aggregation algorithm proposed in [Kempe et al. 2003] is executed among cluster members, then among members of cluster head sets, and finally among cluster members again. Besides the simplification introduced by this variant to analyze the complexity of our approach, the other advantage of SIMPLE-MAX is that it is very simple to implement and the reduced memory footprint of the corresponding byte-code can easily fit on resource constrained devices such as sensors node of a wireless sensor network.

It has been proven in [Kempe et al. 2003] that the round and message complexity of the gossip algorithm is $O(\log M)$ and $O(M \log M)$ respectively, where M is the size of the group in which the algorithm is executed. We have that the cluster size resulting from our bootstrap phase is $O(\log^2 n)$, where n is the total number of nodes in the network. Hence, we have $O(\frac{n}{\log^2 n})$ clusters in the network. For every cluster, we have a cluster head set of size $O(\log n \log\log n)$. Hence, it is possible to show that the complexity of the first phase of our approach is $O(\log\log n)$



(a) Round complexity (b) Message complexity

Fig. 1. Message and round complexity of gossip-based aggregation: MAX.

rounds and $O(n \log \log n)$ messages; the complexity of the second phase is $O(\log n)$ rounds and $O(\log n \log \log n)$ messages; finally, the complexity of the last phase is the same as for phase 1. Thus, the complexity of the whole process is $O(\log n)$ rounds and $O(n \log \log n)$ messages.

3. NUMERICAL EVALUATION AND DISCUSSION

In this section we present a numerical evaluation of the Simple-MAX variant of our algorithm and we study the round and message complexities for experiments based on the *peak scenario* (that is just one node has value 1, and all the other nodes have a zero). The following figures are obtained averaging the results over 50 algorithm runs when the link failure probability is $\delta = 0.3$. Our results compare both round and message complexity, as a function of the total number of nodes in the network, for both the original approach presented in [Kempe et al. 2003] and our approach.

Figure 1(a) indicates that our scheme is equivalent to the one presented in [Kempe et al. 2003] in terms of the number of rounds for the algorithm to converge to the MAX value, whereas Figure 1(b) shows that our approach outperforms previous approaches in terms of message complexity.

4. CONCLUSIONS

In this paper we have relaxed the assumption that nodes in an overlay network are address-oblivious. Leveraging nodes' address, we have introduced a probabilistic protocol to compute and to disseminate the MAX that requires, just $O(n \log n)$ messages and $O(\log n)$ rounds. Preliminary simulation results confirm our theoretical findings.

Note that our result achieves the minimum theoretical overhead predicted for the address-oblivious case. However, being our proposal not address-oblivious, an interesting research issue is to prove whether our result is optimal, or it is just a (good) heuristic. In this latter case, it would be interesting to find out a lower bound for the non address-oblivious case.

REFERENCES

Alon, N., Spencer, J. H., and Erdős, P. 1992. *The Probabilistic Method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley and Sons.

Boyd, S. P., Ghosh, A., Prabhakar, B., and Shah, D. 2006. Randomized gossip algorithms. *IEEE Transactions on Information Theory 52*, 2508–2530.

Karp, R. M., Schindelhauer, C., Shenker, S., and Vöcking, B. 2000. Randomized rumor spreading. In *Proc. of IEEE FOCS*.

Kempe, D., Dobra, A., and Gehrke, J. 2003. Gossip-based computation of aggregate information. In *Proc. of IEEE FOCS*.

Pietro, R. D., Mancini, L. V., and Mei, A. 2004. Efficient and resilient key discovery based on pseudo-random key pre-deployment. In *IPDPS*.