Indexing data-oriented overlay networks using belief propagation

Danny Bickson[1], Danny Dolev[1] and Yair Weiss[1]
Karl Aberer[2], Manfred Hauswirth[2]

[1] The Hebrew University of Jerusalem, Israel
[2] École Polytechnique Fédérale de Lausanne, Switzerland *

Abstract. In this paper we discuss the problem of data-oriented partitioning in large-scale overlay networks as required by peer-to-peer databases or peer-to-peer information retrieval. The goal is to partition a large set of nodes into k partitions with the additional requirement of meeting certain load-balancing constraints without global knowledge of the network's parameters, i.e., the desired number of partitions and the partition distribution function are not known in advance and change dynamically as the network evolves. This key problem in large-scale decentralized systems has so far received only very limited attention. The novel contributions described in the following are the definition of a distributed algorithm for local estimation of the partitioning distribution function, which does not preclude the network's topology and a distributed method for performing the actual partitioning. As additional advantages, the algorithms do not require global knowledge and work highly parallel which provides a low latency of the partitioning process. Both algorithms are based on the max-product belief propagation algorithm and give exact results on trees, and quite accurate approximations on graphs containing cycles. We show the accuracy of the proposed algorithms in terms of number of nodes per partition and the good load balancing of partitions in the network by simulation. Our algorithms are scalable and the accuracy of the partitioning improves with growing network sizes. Having shown the efficiency of our proposed algorithms, we discuss a natural application for our algorithm in the data-oriented P2P system P-Grid (http://www.p-grid.org/). Using P-Grid's underlying tree abstraction, we can apply our algorithm to achieve optimal partitioning results in short times relative to the tree diameter.

* The work presented in this paper was supported (in part) by the European project Evergrow No 001935.