Technical Perspective: Efficient Signal Reconstruction for a Broad Range of Applications

Zachary G. Ives
University of Pennsylvania
zives@cis.upenn.edu

When problems are scaled to "big data," researchers must often come up with new solutions, leveraging ideas from multiple research areas — as we frequently witness in today's big data techniques and tools for machine learning, bioinformatics, and data visualization. Beyond these heavily studied topics, there exist other classes of general problems that need to be rethought at scale. One such problem is that of *large-scale signal reconstruction* [4]: taking a set of observations of relatively low dimensionality, and using them to reconstruct a high-dimensional, unknown *signal*. This class of problems arises when we can only observe a subset of a complex environment that we are seeking to model — for instance, placing a few sensors and using their readings to reconstruct an environment's temperature, or monitoring multiple points in a network and using the readings to estimate end-to-end network traffic, or using 2D slices to reconstruct a 3D image.

This *signal reconstruction problem* (SRP) is typically approached as an optimization task, in which we search for the high-dimensional signal that minimizes a *loss function* comparing it to the known properties of the signal. Prior solutions to the SRP make use of linear algebra techniques [4] or expectation maximization [2] to find a solution. However, at scale, the dimensionality of the signal is high enough to render such optimization techniques too costly. In "Efficient Signal Reconstruction for a Broad Range of Applications," Asudeh et al. show that algorithmic insights about SRP, combined with database techniques such as similarity joins and sketches, can be used to scalably solve the signal reconstruction problem. The paper creatively integrates query processing, approximation, and linear algebra techniques.

The authors start by noting that SRP is a special case of quadratic programming, which they exploit by solving the Lagrangian dual formulation of the original problem. Building upon this, they make a connection to query processing: the key part of the algorithm computes the product of a (typically very sparse) matrix A with its transpose, AA^T. In turn, that computation derives most of its value from a small number of elements from A.

The authors creatively leverage this observation to handle huge matrices, by implementing matrix multiplication via a set-intersection primitive. They build upon set-similarity joins and apply threshold-based techniques [3] to bound the values of the matrix product, thus developing a fast approximation algorithm. Finally, they show how to use min-hash sketches [1] to approximate the sets, allowing further trade-offs of accuracy vs performance (and space). Experimental analysis shows these techniques scale well enough to to predict end-to-end routes in a large P2P network, which is several orders of magnitude larger than prior solutions could handle.

This paper is notable because it scalably addresses an under-served problem with practical impact, and does so in a clean, insightful, and systematic way. It makes several key contributions. First, it shows how insights into the linear algebra computation can be used for greater efficiency (the connection to quadratic programming, which allows it to be solved via the Lagrangian dual). Subsequently, it makes insightful connections to techniques from query processing and sketches, to develop approximation algorithms. Finally, the paper conducts an experimental study demonstrating high performance at scale. The paper illustrates the potential benefits of connecting important optimization problems with database approximate query processing techniques.

1. REFERENCES

[1] Andrei Z Broder. On the resemblance and containment of documents. In *Compression and Complexity of SEQUENCES 1997*, pages 21–29. IEEE, 1997.

[2] Jin Cao, Drew Davis, Scott Vander Wiel, and Bin Yu. Time-varying network tomography: router link data. *Journal of the American statistical association*, 95(452):1063–1075, 2000.

[3] Surajit Chaudhuri, Venkatesh Ganti, and Raghav Kaushik. A primitive operator for similarity joins in data cleaning. In *ICDE*, pages 5–5. IEEE, 2006.

[4] Curtis R Vogel. *Computational methods for inverse problems*, volume 23. SIAM, 2002.