Reminiscences on Influential Papers

Kenneth A. Ross, editor

I continue to invite unsolicited contributions. See http://www.acm.org/sigmod/record/author.html for submission guidelines.

Minos Garofalakis, Bell Labs, Lucent Technologies, minos@research.bell-labs.com.

[Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. Proceedings of the 28th Annual ACM Symposium on the Theory of Computing, Philadelphia, Pennsylvania, May 1996, pp. 20–29. (Extended version in the STOC'96 Special Issue of the Journal of Computer and System Sciences, 58(1), February 1999.)]

Long before data-streaming was in vogue, Alon, Matias, and Szegedy (AMS) had the foresight to attack the difficult problem of estimating the frequency moments of a large data set in one pass with limited memory. The end result was this very elegant and truly seminal STOC'96 paper, that has had and will continue to have enormous impact on the theory and practice of data-stream management.

I first came across the AMS paper when I started getting interested in the data-streaming area, in the spring of 2001. Reading this paper was a real eye-opener for me. It was just amazing to see how simple randomization ideas and basic probabilistic tools (like the Chebyshev inequality and the Chernoff bound) can come together to provide elegant, space-efficient randomized approximation algorithms for estimation problems that, at first glance, would seem impossible to solve. The second-moment method described in the AMS paper is essentially the father of all "sketch-based" techniques for data-stream management. Even though the idea of randomized linear projections (a.k.a., sketches) was known for some time in the domain of functional analysis (dating back to the famous Johnson-Lindenstrauss Lemma), Alon, Matias, and Szegedy were the first to exploit sketches for small-space data-stream computation, through the use of limited-independence random variates that can be constructed in small space and time. Of course, in addition to small-space sketching, the AMS paper also makes a number of other fundamental contributions in data streaming, including practical approximation algorithms for other frequency moments (e.g., the number of distinct values in a stream), as well as several inapproximability results (i.e., lower bounds) based on beautiful communication-complexity arguments.

Another amazing fact is that this is not an intimidating theory paper — the exposition is simply excellent with all key ideas thoroughly analyzed and explained, even for the non-specialist. In retrospect, the elegance of the AMS results is one of the key things that got me excited about the data-streaming area; further, the AMS sketching techniques have been and continue to be the foundation of most of my work on approximate SQL query processing over streams. And, of course, I never forget to express my appreciation for this beautiful piece of work to Yossi Matias whenever I meet him at SIGMOD/VLDB gatherings. The AMS paper is absolute "must" reading not only for researchers working on data streaming, but also for anyone wanting to appreciate the power and elegance of randomized algorithms.

Jeffrey Naughton, University of Wisconsin, Madison, naughton@cs.wisc.edu.

[Wen-Chi Hou, Gultekin Ozsoyoglu, and Baldeo K. Taneja. Statistical Estimators for Relational Algebra Expressions. PODS 1988, pp. 276–287.]

This wonderful paper foreshadows and establishes the framework for a great deal of work related to sampling from database systems in order to provide approximate answers to queries. I was a young assistant professor at the time of its publication, working on the evaluation of Datalog queries, and this was the paper that convinced me to look at an entirely different class of problems. (Some cynics would argue that moving someone away from working on Datalog is by itself reason enough to consider a paper valuable and influential, but that is a topic for a different note altogether!)

The key insight in this paper is so well established that now it may seem obvious, but that is just because we have had 15 years to think about it. Prior to this paper, there was a deeply rooted assumption that when faced with running a relational query, either you wait for it to run to completion, or you give up. In this paper, the authors give a third alternative - you spend as much time as you have available, and then give an approximate answer to the query. This approximate answer could be a random subset of the complete answer, or (in the case of an aggregate query) a statistical approximation of the true result. With this insight, the authors proceeded to consider how best to use system resources to compute the approximation, and how to prove statistical properties about the quality of the resulting approximation. This framework is so rich and runs so deep that there is still work being done today that traces its lineage (whether explicitly or not) to this original paper.
