Big Data as a powerful source of randomness

Periklis Papakonstantinou
Rutgers Business School

9/23/2015 at 11:00 am
Core A (Room 301)

Abstract

True randomness is indispensable for every natural science where numerical simulations are performed, and is also of paramount important for secure communication. Obtaining true random bits is a difficult process, one that must reply on uncertain events. We take a non-standard view of Big Data, viewing them as big sources of low quality randomness.

Now, the problem is to extract true uniform random bits out of big data. For this we develop the first streaming extractor for sources of linear min-entropy, circumventing a number of provable limitations that include a general impossibility result by Bar-Yosef, Reingold, Shatiel, and Trevisan. New techniques and tools are developed towards proving that this extractor works. The extractor itself is perhaps the conceptually simplest generic extractor.

We implement our extractor as a streaming algorithm and we run a very extensive empirical validation study. Before our work randomness extraction was a vastly theoretical area of study. No known extractor could work in practice on statistical sources of moderate size, in the sense that a few 1,000s of years were needed for such extraction, where on sources of size about 20GBs a conservatively estimated running time exceeds 100,000 years. In this sense, our extractor is the first practical such construction.

This is joint work with Guang Yang and David Woodruff.