**Diamond Sampling for Approximate Maximum All-pairs Dot-product (MAD) Search (*)**

Tammy Kolda, Sandia National Laboratories

Given two sets of vectors, $A = \{ \vec{a_1}, \ldots, \vec{a_m} \}$ and $B = \{ \vec{b_1}, \ldots, \vec{b_n} \}$, our problem is to find the top-$t$ dot products, i.e., the largest $|\vec{a_i} \cdot \vec{b_j}|$ among all possible pairs. This is a fundamental mathematical problem that appears in numerous data applications involving similarity search, link prediction, and collaborative filtering. We propose a sampling-based approach that avoids direct computation of all $mn$ dot products. We select diamonds (i.e., four-cycles) from the weighted tripartite representation of $A$ and $B$. The probability of selecting a diamond corresponding to pair $(i, j)$ is proportional to $(\vec{a_i} \cdot \vec{b_j})^2$, amplifying the focus on the largest-magnitude entries. Experimental results indicate that diamond sampling is orders of magnitude faster than direct computation and requires far fewer samples than any competing approach. We also apply diamond sampling to the special case of maximum inner product search, and get significantly better results than the state-of-the-art hashing methods. We show results on Amazon purchasing data, music song pairing, and movie recommendations.

This is joint work with Grey Ballard, Ali Pinar, and C. Seshadhri.

(*) Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.