DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
UNIVERSITY OF SOUTHERN DENMARK, ODENSE

COMPUTER SCIENCE COLLOQUIUM

Making Large-scale Machine Learning work in Theory and in Copenhagen's Coolest Start-up

Søren Dahlgaard
PhD in computer science and co-founder of SupWiz

Friday, 25 May, 2018 at 14:15

Auditorium U164

Abstract:

This talk consists of two parts with the majority focusing on my work on hashing and Machine Learning during my PhD studies followed by a few minutes about SupWiz, a Copenhagen based startup created by world leading researchers focused on using machine learning to improve customer service and support.

Hashing is an important computational primitive used in many popular algorithms and data structures for tackling problems in many areas of computer science when working with large data volumes. One such area is machine learning, where e.g. Minwise Hashing by Broder et. al and Feature hashing by Weinberger et al. are prominent examples used in many machine learning algorithms.

Analyzing such algorithms is often done under the assumption that truly random unit cost hash functions are available, but using a weak hash function in practice may lead to bias and inconsistency. Furthermore, hashing is often employed as an "inner-loop" operation and evaluation time is thus of utmost importance.

We provide efficient families of hash functions with strong theoretical guarantees for many influential algorithms including Minwise hashing, HyperLogLog counters, One Permutation Hashing, and more. Furthermore, our techniques lead to the most efficient known hashing schemes for the power of two choices, approximately minwise independence, constant moment bounds, and more. We also present a new similarity sketch with properties similar to the seminal MinHash sketch, but with much faster running time. This problem has previously been considered from a practical perspective, but the proposed solutions fail to give strong concentration bounds. We complement our theoretical result with experiments demonstrating that our new results systematically outperform simpler hashing schemes for similarity estimation and feature hashing on both synthetic and real-world data.

The talk is based on joint work with Mathias Bæk Tejs Knudsen, Eva Rotenberg, and Mikkel Thorup appearing at SWAT'14, FOCS'15, SODA'16, FOCS'17, and NIPS'17