RUTGERS
School of Arts and Sciences

Department of Statistics and Biostatistics
Rutgers, The State University of New Jersey
110 Frelinghuysen Road
Piscataway, NJ 08854

Web: www.stat.rutgers.edu
E-mail: office@stat.rutgers.edu
Phone: (848) 445-2690
Fax: (732) 445-3428

RUTGERS UNIVERSITY

Department of Statistics, Department of Computer Science

www.stat.rutgers.edu

Workshop on Big Data:
Theoretical Foundation of Big Data

Speaker: **Dr. Mikkel Thorup**
University of Copenhagen

Title: **The Power of Tabulation Hashing**

Time: **1:05pm – 2:00pm, *Thursday*, October 16, 2014**

Place: ***CoRE Auditorium***

Abstract

Randomized algorithms are often enjoyed for their simplicity, but the hash functions used to yield the desired theoretical guarantees are often neither simple nor practical. Here we show that the simplest possible tabulation hashing provides unexpectedly strong guarantees. The scheme itself dates back to game playing programs of Zobrist from 1970. Keys are viewed as consisting of c characters. We initialize c tables T1,…,Tc mapping characters to random hash codes. A key x=(x1,…,xc) is hashed to T1[x1] xor … xor Tc[xc]. While this scheme is not even 4-independent, we show that it provides many of the guarantees that are normally obtained via higher independence, e.g., we get reliable performance in hash tables based on linear probing and cuckoo hashing, and in min-wise hashing for estimating set intersection. We will also discuss a twist to simple tabulation that leads to reliable statistics with Chernoff-type concentration. Finally, we show how simple tabulation applied recursively gives high indepedence.

Bio:

Mikkel Thorup (born 1965) has a D.Phil. from Oxford University from 1993. From 1993 to 1998 he was at the University of Copenhagen. From 1998 to 2013 he was at AT&T Labs-Research, and since 2013 he has been back as Professor at the University of Copenhagen.

Mkkel Thorup is a Fellow of the ACM and of AT&T, Member of the Royal Danish Academy of Sciences and Letters, and co-winner of the 2011 MAA Robbins Award. His main work is in algorithms and data structures and he is the editor of this area for the Journal of the ACM. One of his best-known results is a linear-time algorithm for the single-source shortest paths problem in undirected graphs.