UCF DEPARTMENT OF COMPUTER SCIENCE

Fall 2016 Seminar Series

A BRIEF INTRODUCTION TO SUCCINCT DATA STRUCTURES AND TEXT INDEXING

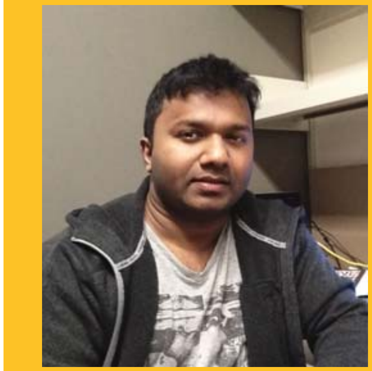
TUESDAY, OCTOBER 4, 2016

1:00 PM – HEC 356

Pattern matching is a fundamental problem in Computer Science with applications in web-data, texts and biological sequences. In the data structural sense, the text T (of n characters) is pre-processed and an index is built to answer pattern matching queries for a pattern P. Both text and pattern come from alphabet set ∑ of size σ. In the basic pattern matching query, all \occ occurrences of P in \T, identified by their location in \T, are reported. Suffix trees are the most powerful and ubiquitous data structures for such purpose. They find myriad applications in sequence analysis for many different applications. In the era of budget, one of the negative aspects of suffix tree was seen to be its space utilization -- about 50 times the text for DNA sequences. In the theoretical sense, although considered linear in terms of words, the suffix trees take Θ(n log n) space in terms of bit. However, the optimal is n log σ bits, leading to a complexity gap. The advent of succinct data structures and compressed text indexing, where the goal is to have data structure in the space equal to the information theoretical minimum, presented us with new indexes like Compressed Suffix Array (CSA) and FM-Index, eventually leading to a wonderful data structure called fully-functional compressed suffix tree (i.e., suffix trees encoded in optimal bits). In practice, these achieved a remarkable breakthroughs saving orders of magnitude of space. By presenting the crux of these elegant theoretical results (with sufficient technical details), I will introduce the exciting and combinatorically rich field of Succinct/Compressed Data Structures.

DR. SHARMA V. THANKACHAN

Georgia Institute of Technology



Sharma Thankachan will be joining the Computer Science Department at the University of Central Florida as an Assistant Professor in December 2016 and is currently a Research Scientist in the School of CSE, Georgia Institute of Technology, Atlanta. His research interests are in the area of algorithms and data structures, primarily for string searching/indexing/compression problems. Sharma holds a PhD in Computer Science from Louisiana State University.

Hosted by: Dr. Sumit Jha