Department of Computer Science

East Carolina University

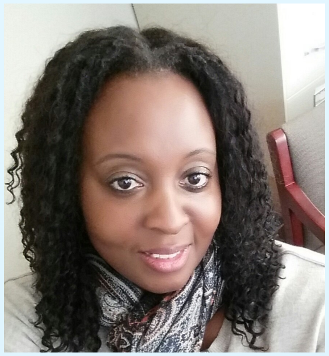
Seminar

Dealing with Semantic Heterogeneity –The Often-Overlooked Challenge in Big Data & Data Science

Abstract. Distributed computing has become fundamental to addressing challenges of processing Big Data. However, many of the existing techniques and technologies have focused on addressing the Big Volume and Big Velocity challenges. On the other hand, the challenge of Big Variety or heterogeneity has remained largely unaddressed. Dynamic data environments such as the Internet of Things (IoT) and Sensor Network are ever-evolving and links between data produced by different sources need to be done in dynamic and ad-hoc ways. Consequently, there does not exist an all-encompassing schema that can be used to model and organize data a priori. This has led to the increasing popularity of graph data models such as knowledge graphs for representing heterogeneous data. Knowledge graphs allow explicit modeling of types of entities and relationships between entities. In complex data environments such as the biomedical domain, data modeling not only captures relationships between entities but also between entity types as well as between relationships, e.g., capturing that a relationship is an inverse of another. The resulting graph models introduce additional processing challenges, e.g., semantic inferencing during data processing, when compared to the processing of traditional homogeneous graph models such as social networks or Web graphs.

Some efforts have been made to leverage distributed processing platforms such as Hadoop-based platforms like Apache Pig and Hive as well as other NoSQL platforms to enable scalable processing of knowledge graphs. However, significant scalability challenges remain which we posit is due to the semantics-oblivious nature of such platforms. We suggest that a strategy of enabling *semantics-awareness* in both the storage and the processing layers of distributed computing platforms based on inducing implicit *semantic relationships* from data and using that as a guide for data organization and storage and process execution. This will achieve better scalability by enabling more aggressive pruning of task-irrelevant data and computational search space, and minimizing the associated overhead. This talk shares our vision and preliminary work for a *Semantics-Aware Distributed Computing Platform.*

Biography. Dr. Anyanwu Ogan is an Associate Professor of Computer Science in the College of Engineering at North Carolina State University (NC State) – the largest university in the state of North Carolina. She is also a visiting Computer Science faculty member at the African University of Science and Technology (AUST) in Nigeria. At NC State, she teaches courses related to data management and directs a research lab on Semantic Computing and Big Data Analytics and her research has been supported by over $1.5M in research grants from the United States National Science Foundation, the United States Office of Naval Research and industry organizations like IBM and Bosh Global Services. She is frequent panelist on the National Science Foundation proposal peer review panels and a member of organizational and steering committees of several international conferences and workshops and editorial board of leading journals. She has several widely cited scientific research publications with over 1300 citations reported in Google scholar. Her Ph.D. and Masters students have gone on to work for top American computing organizations like Microsoft, Yahoo, IBM, Amazon.



Dr. Kemafor Ogan
Department of
Computer Science
NC State University

kogan@ncsu.edu
www.csc.ncsu.edu/people/kogan

Thursday April 13, 2017
Time: 1:00 – 1:50pm
Room: SCITEC 0144A

Contact: Dr. Kamran Sartipi
Dept. of Computer Science, ECU
www.cs.ecu.edu/sartipi/CSseminar/