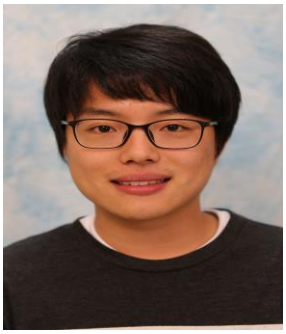
Carnegie Mellon University
Computer Science Department

THESIS ORAL



KIJUNG SHIN

Mining Large Dynamic Graphs and Tensors

Monday, February 4, 2019 – 11:00 a.m. – GHC 4405

Graphs are ubiquitous, representing a variety of information, ranging from who follows whom on online social networks to who reviews what on e-commerce sites. Many of these graphs are large and dynamic (i.e., changing over time). Moreover, they are with rich side information (e.g., e-commerce reviews with timestamps, ratings, and text) and thus naturally modeled as tensors (i.e., multi-dimensional arrays).

Given large dynamic graphs and tensors, how can we analyze their structure, detect interesting anomalies, and model behaviors of individuals in the data? To answer this question, which is fundamental to understand massive evolving data on user behavior, my thesis focuses on developing fast scalable algorithms for the following tasks:

- Structure Analysis: We build one-pass, sublinear-space algorithms for estimating the triangle count, which is an important connectivity measure, in large dynamic graphs. Especially, our distributed algorithm yields up to 39X more accurate estimates without speed reduction than a baseline. We also develop distributed and out-of-core algorithms for summarizing the structure of large graphs and tensors. They summarize 25-1000X larger data without quality loss than their best competitors.

- Anomaly Detection: We develop near-linear time approximation algorithms for detecting unusually dense subgraphs and subtensors, which signal notable anomalies such as 'edit wars' on Wikipedia and fake followers on Twitter. Especially, our tensor algorithm is up to 114X faster without accuracy loss than the previously best heuristic. We also extend it for distributed or dynamic data with the same approximation guarantee.

- Behavior Modeling: We design game-theoretic models for purchases of individuals in social networks and a fast algorithm for finding Nash equilibria of the models. In addition, we develop a stage model for the progression of individuals on social media and a distributed optimization algorithm that fits our model to behavior logs with trillions of records.

To achieve the highest performance and scalability, our algorithms employ mathematical techniques (e.g., approximation and sampling), use distributed computing frameworks (e.g., MapReduce), and/or exploit data patterns (e.g., power laws). We successfully apply them to massive datasets, including 20.6 billion connections on LinkedIn, 783 million hyperlinks between web pages, and 483 million edits on Wikipedia.

Thesis Committee:
Christos Faloutsos, Chair
Tom M. Mitchell
Leman Akoglu
Philip S. Yu, University of Illinois at Chicago