Spring 2020

CS 786-002: Special Topic - Real World Graph Analytics

David A. Bader

Course Syllabus

Jump to Today Edit

CS 786: Real-World Graph Analytics
Spring 2020
Instructor: Prof. David A. Bader

Course Description

Emerging real-world graph problems include detecting community structure in large social networks, graph representation learning in AI/data science, improving the resilience of the electric power grid, and detecting and preventing disease in human populations. Unlike traditional applications in computational science and engineering, solving these problems at scale often raises new challenges because of sparsity and the lack of locality in the data, the need for additional research on scalable algorithms and development of frameworks for solving these problems on high performance computers, and the need for improved models that also capture the noise and bias inherent in the torrential data streams. In this course, students will be exposed to the opportunities and challenges in massive data-intensive computing for applications in data science and AI, financial services, health informatics, and security.

This course will introduce students to designing high-performance and scalable algorithms for massive graph analysis. The course focuses on algorithm design, complexity analysis, experimentation, and optimization, for important “big data” graph problems. Students will develop knowledge and skills concerning:

- the design and analysis of massive-scale graph algorithms employed in real-world data-intensive applications, and
- performance optimization of applications using the best practices of algorithm engineering

Required Background

- Programming Skills
  - C/C++ or Java
- Prerequisite Courses
  - CS 610: Data Structures and Algorithms (or equivalent algorithms course)
  - Or permission of instructor

Coverage of Topics

An increasingly fast-paced, digital world has produced an ever-growing volume of petabyte-sized datasets. At the same time, terabytes of new, unstructured data arrive daily. As the desire to ask more detailed questions about these massive streams has grown, parallel software and hardware have only recently begun to enable complex analytics in this non-scientific space. In this course, we will discuss the open problems facing us with analyzing this "data deluge". Students will learn the design and implementation of algorithms and data structures capable of analyzing spatio-temporal data at massive scale on parallel systems. Students will understand the difficulties and bottlenecks in parallel graph algorithm design on current systems and will learn how multithreaded and hybrid systems can overcome these challenges. Students will gain hands-on experience mapping large-scale graph algorithms on a variety of parallel architectures using advanced programming models.

Network Analysis

1. Introduction to Graph Theory & Data Structures
2. Motivating Applications in Data
3. Opportunities & Challenges
4. Parallel, Multicore, & Multithreaded Architectural Support for Graph Processing
5. Mapping Graph Algorithms to Architectures
6. Open Discussion

Static Parallel Algorithms

1. Programming Models
2. Parallel Prex & List Ranking
3. Graph Search, Spanning Tree, Connected Components
4. Minimum Spanning Tree Matroid Algorithms
5. Social Networking Algorithms
6. Betweenness Centrality
7. Community Detection

Dynamic Parallel Algorithms

1. Streaming Data Analysis
2. Data Structures for Streaming Data
3. Tracking Clustering Coefficients
4. Tracking Connected Components
5. Anomaly Detection

Programming & Software

1. Programming Environments: OpenMP, MPI, MapReduce, CUDA/OpenCL
2. Graph Software: PBGL, MTGL, LEDA, SNAP, STINGER, HORNET, Pregel, Neo4j
3. Advanced Topics

Textbook
(https://www.amazon.com/Big-Data-Analytics-Enterprise-Integration-ebook/dp/B00EVSOWVA)

- **(https://www.amazon.com/Big-Data-Analytics-Enterprise-Integration-ebook/dp/B00EVSOWVA) Network science (http://networksciencebook.com/)** , A. Barabási, and M. Pósfai, Cambridge University Press, 2016.

- This course will also use current research reports

Resources

Additional reading materials including reference books and online resources will be assigned for some advanced topics as the course proceeds.

Evaluation

The Final Exam will be given remotely using the ProctorU platform.

Grading components:

Attendance	10%
Homework	10%
Project	20%
Midterm	30%
Final Exam	30%

Late Policy

Students are expected to complete work on schedule. Late work is not accepted unless prior arrangements are made with the instructor.

Academic Integrity and Student Conduct:

"Academic Integrity is the cornerstone of higher education and is central to the ideals of this course and the university. Cheating is strictly prohibited and devalues the degree that you are working on. As a member of the NJIT community, it is your responsibility to protect your educational investment by knowing and following the academic code of integrity policy that is found at: http://www.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf (http://www5.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf.)

Please note that it is my professional obligation and responsibility to report any academic misconduct to the Dean of Students Office. Any student found in violation of the code by cheating, plagiarizing or using any online software inappropriately will result in disciplinary action. This may include a failing grade of F, and/or suspension or dismissal from the university. If you have any questions about the code of Academic Integrity, please contact the Dean of Students Office at dos@njit.edu" (mailto:dos@njit.edu%E2%80%9D)

Course Summary:

Date	Details

Date	Details	
Mon Jan 27, 2020	Class 1: Motivation for Massive-Scale Graphs; Definitions, Data Structures, Graph Algorithms (https://njit.instructure.com/calendar?event_id=3940&include_contexts=course_10906)	2:30pm to 5:20pm
Mon Feb 3, 2020	Class 2: Graph Traveral and Breadth-First Search; Spanning Trees and Connected Components (https://njit.instructure.com/calendar?event_id=3941&include_contexts=course_10906)	2:30pm to 5:30pm
Mon Feb 10, 2020	Class 3: Guest Lecture: Dr. Zhihui Du (https://njit.instructure.com/calendar?event_id=3942&include_contexts=course_10906)	2:30pm to 5:20pm
	Homework 1 (https://njit.instructure.com/courses/10906/assignments/37474)	due by 11:59pm
Mon Feb 17, 2020	Class 4: Parallelism, Multicore, Multithreaded, GPU, Accelerators; Mapping Parallel Graph Algorithms to Architectures, Cost Models (https://njit.instructure.com/calendar?event_id=3943&include_contexts=course_10906)	2:30pm to 5:20pm
Mon Feb 24, 2020	Class 5: Parallel Breadth-First Search, Shortest Paths, Betweenness Centrality (https://njit.instructure.com/calendar?event_id=3944&include_contexts=course_10906)	2:30pm to 5:20pm
Mon Mar 2, 2020	Class 6: Map-Reduce Algorithms, Hadoop; Graph Twiddling and other analytics on Map-Reduce (https://njit.instructure.com/calendar?event_id=3945&include_contexts=course_10906)	2:30pm to 5:20pm
	Project proposal (https://njit.instructure.com/courses/10906/assignments/40316)	due by 11:59pm
Mon Mar 9, 2020	Class 7: Midterm (https://njit.instructure.com/calendar?event_id=3946&include_contexts=course_10906)	2:30pm to 5:20pm
	Homework: Modeling Epidemic Networks (https://njit.instructure.com/courses/10906/assignments/40539)	due by 11:59pm
Mon Mar 23, 2020	Class 8: Massive-Graphs in Computational Biology, Genome Assembly (https://njit.instructure.com/calendar?event_id=3948&include_contexts=course_10906)	2:30pm to 5:20pm

Date	Details	
Mon Mar 30, 2020	Class 9: Finding Community Stucture in Large Graphs; Modularity and Conductance (https://njit.instructure.com/calendar?event_id=3949&include_contexts=course_10906)	2:30pm to 5:20pm
	Homework: Graph properties (https://njit.instructure.com/courses/10906/assignments/43671)	due by 11:59pm
Mon Apr 6, 2020	Class 10: Modularity Algorithms, Normalizations, and Resolution Limit; Parallel Modularity Approaches (https://njit.instructure.com/calendar?event_id=3950&include_contexts=course_10906)	2:30pm to 5:20pm
	Triangle Counting (https://njit.instructure.com/courses/10906/assignments/45259)	due by 11:59pm
Mon Apr 13, 2020	Class 11: Clustering in Weighted Networks; Graph Databases and Complex Queries (https://njit.instructure.com/calendar?event_id=3951&include_contexts=course_10906)	2:30pm to 5:20pm
	Network Science: Chapters 1 & 2 (https://njit.instructure.com/courses/10906/assignments/46879)	due by 11:59pm
Mon Apr 20, 2020	Class 12: Project Presentations (https://njit.instructure.com/calendar?event_id=3952&include_contexts=course_10906)	2:30pm to 5:20pm
	Network Science: Chapters 3 & 4 (https://njit.instructure.com/courses/10906/assignments/48255)	due by 11:59pm
Mon Apr 27, 2020	Class 14: GPU Graph Algorithms (https://njit.instructure.com/calendar?event_id=3953&include_contexts=course_10906)	2:30pm to 5:20pm
Mon May 4, 2020	Class 15: Web Knowledge Graph (https://njit.instructure.com/calendar?event_id=3954&include_contexts=course_10906)	2:30pm to 5:20pm
Mon May 11, 2020	Class 16: Final Exam (https://njit.instructure.com/calendar?event_id=3955&include_contexts=course_10906)	2:30pm to 5:20pm
	Final Exam (https://njit.instructure.com/courses/10906/assignments/48981)	
	Midterm Exam (https://njit.instructure.com/courses/10906/assignments/44724)	
	Roll Call Attendance (https://njit.instructure.com/courses/10906/assignments/36343)	