CAREER: Towards Portable and Self-Optimizing Graph Mining on Heterogeneous Multi-GPU Systems

Overview

Graph mining underpins scientific discovery, infrastructure analysis, fraud detection, and biomedical research. Over the past decade, GPU acceleration has dramatically improved graph mining performance, yet state-of-the-art systems remain *difficult to use, hardware-locked, and unable to scale*. Performance depends on extensive expert tuning; implementations are tightly coupled to a specific GPU architecture; and most frameworks are confined to a single device. As GPU architectures diversify (NVIDIA Hopper/Blackwell, AMD MI300, Intel Ponte Vecchio) and multi-GPU nodes become the norm, this usability and portability gap is widening rather than closing.

This CAREER project develops a **portable, hardware-agnostic, and self-optimizing runtime system for multi-GPU graph mining** that bridges high-level mining programs and heterogeneous GPU hardware. The PI will pursue three integrated research thrusts: (T1) a *hardware-agnostic abstraction layer* that decouples mining algorithms from device-specific optimizations; (T2) an *adaptive execution engine* that selects execution strategies online based on graph structure and pruning behavior, supported by a lightweight learning-based auto-tuner; and (T3) a *multi-GPU orchestration layer* that delivers scalable partitioning, communication, and load balancing across heterogeneous devices. Together, these thrusts establish the missing systems layer that makes GPU graph mining usable, reproducible, and portable across platforms.

The research is tightly integrated with an education plan that introduces GPU and graph systems concepts to undergraduates, prepares graduate students for cross-disciplinary HPC research, and broadens participation in computing through structured outreach.

Intellectual Merit

The proposed work advances the foundations of high-performance graph computing in three ways. **First**, it introduces a new abstraction model that separates the *logic* of graph mining (pruning, matching, enumeration) from the *mechanics* of GPU execution (memory layout, warp scheduling, kernel selection), enabling a single program to run efficiently across vendors. **Second**, it develops adaptive runtime techniques that exploit the structural heterogeneity of real-world graphs—skewed degree distributions, irregular pruning trees, and dynamic working-set sizes—to outperform fixed-strategy systems without manual tuning. **Third**, it formulates multi-GPU graph mining as a joint partitioning–scheduling problem and proposes communication-aware algorithms that scale to heterogeneous multi-GPU nodes. Expected outcomes include open-source software, reproducible benchmarks, and new algorithmic and systems insights published in top venues (SIGMOD, VLDB, SC, PPoPP, OSDI).

Broader Impacts

The project will *democratize* high-performance graph analytics by removing the expert-tuning barrier that currently restricts GPU graph mining to a small community of systems specialists. Released as open-source software with reproducible artifacts, the runtime will benefit researchers and practitioners in cybersecurity, computational biology, social network analysis, and scientific computing. The integrated **education and outreach plan** includes: (i) a new upper-level/graduate

course on *GPU Systems for Data-Intensive Computing* at Rowan University, with hands-on multi-GPU labs released publicly; (ii) curriculum modules introducing parallel graph algorithms into existing undergraduate database and algorithms courses; (iii) mentoring of graduate and undergraduate researchers, with explicit recruitment of women and students from groups historically underrepresented in computing through Rowan's CS Diversity initiatives and the CRA-WP/REU programs; and (iv) a yearly summer workshop introducing high-school students from the Philadelphia public school system to data science and GPU computing. All teaching materials, benchmarks, and software will be openly released to maximize community uptake.

Keywords: graph mining; GPU computing; heterogeneous systems; runtime systems; auto-tuning; multi-GPU; reproducibility.