KU-The Future

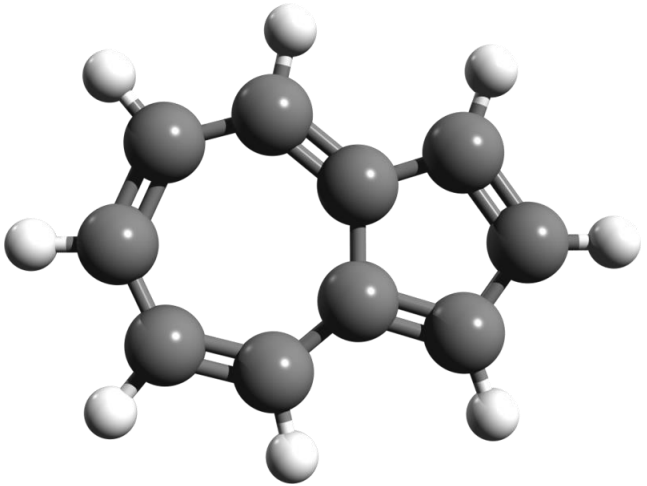
그래프 특성 벡터의 희소성에 따른 GCN 추론 커널의 성능 분석

Analyzing the Performance of GCN Inferences with respect to Sparsity of Graph Features

김인제º, 구건재
고려대학교 컴퓨터학과

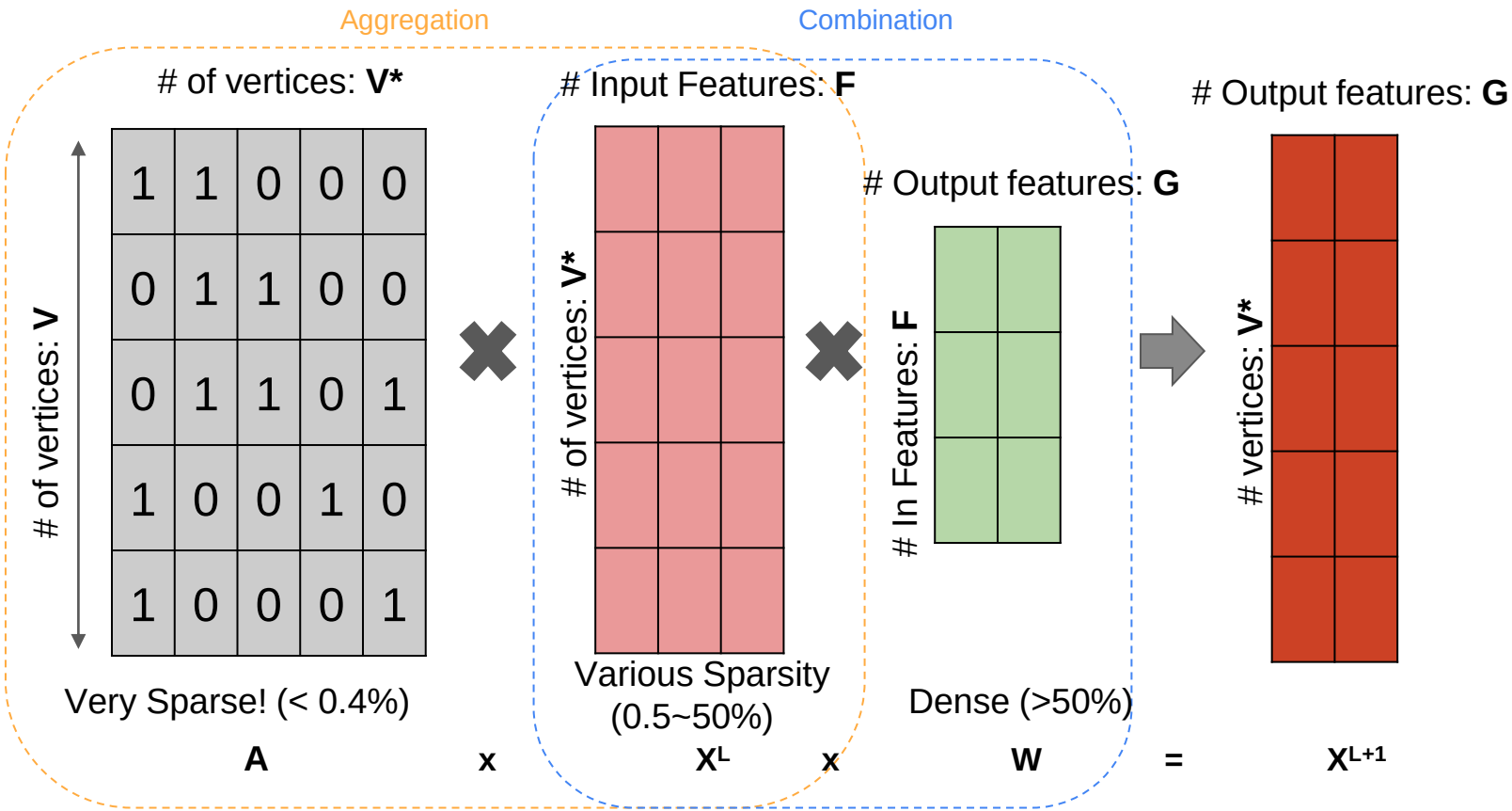
Graphs are everywhere

- Graph data structures represent ***non-euclidean*** data
- Widely used in recommendation systems, bio-chemistry analysis , etc ...



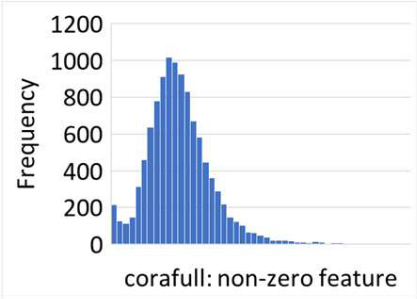
https://appsource.microsoft.com/es-es/product/power-bi-visuals/WA104381236?tab=Overview
https://bcachemistry.wordpress.com/2014/04/11/graph-theory-and-hydrocarbons/

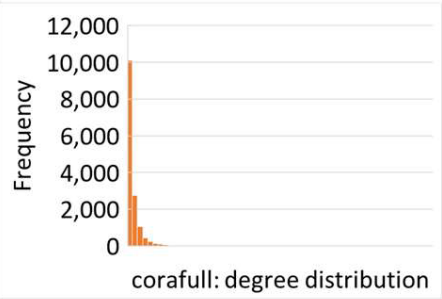
GCN Propagation Rule

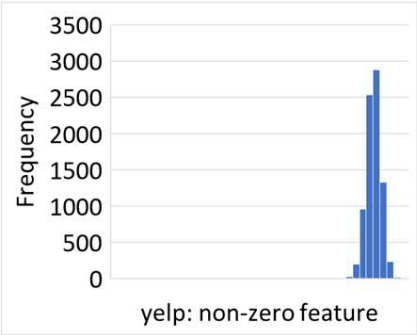


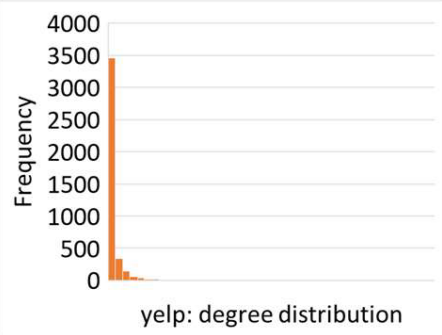
Methodology and Data Characteristics

- Graph adjacency matrices have long-tail distribution
- Graph feature matrices have asymmetric Gaussian-like distribution
- Kernels are constructed based on Fast-GCN









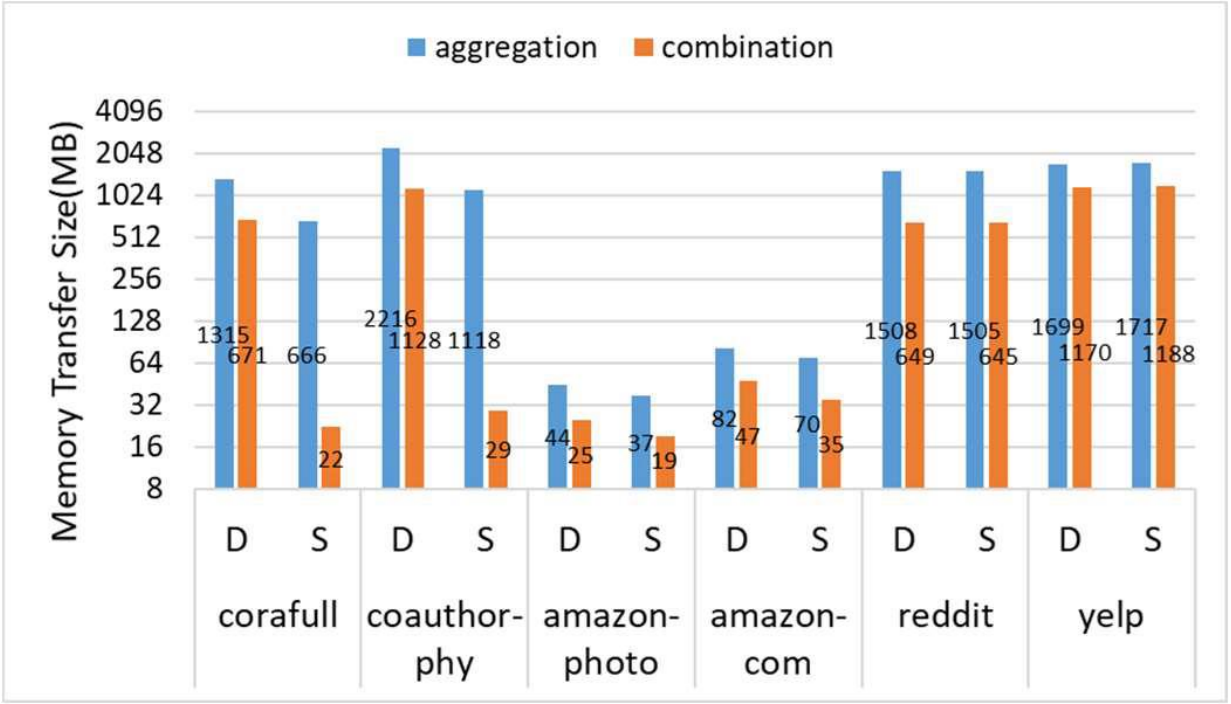
- **System**
  CUDA C/C++ using Cuda toolkit 10.2
  on Intel i7-9700 + NVIDIA rtx 2080
- **Profiler**
  NVIDIA Nsight Compute(2019.5) + NVProf

	Node	Edge	Feature Density(%)	Graph Density(%)
Corafull	19793	8710	0.65%	0.0374%
Coauthor-phy	34493	8415	0.39%	0.0446%
Amazon-photo	7650	745	34.72%	0.4250%
Amazon-com	13752	767	34.85%	0.2728%
Reddit	232965	602	49.55%	0.2100%
Yelp	716848	300	50.90%	0.0028%

Considering Storage Format

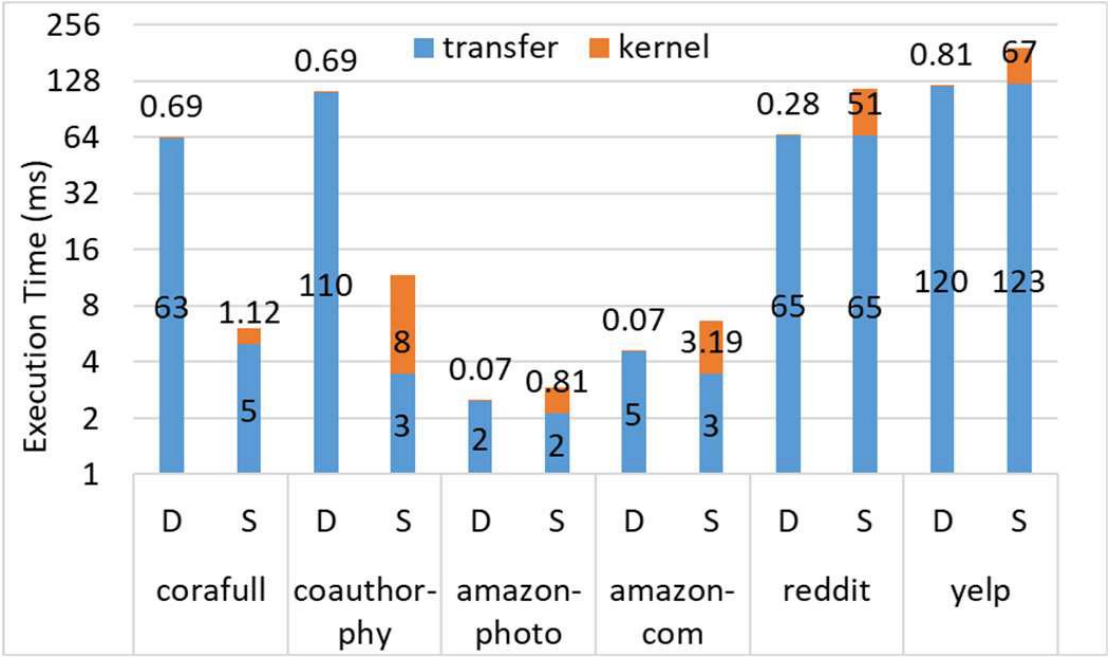
- Memory transfer size: **Aggreagtion > Combination**
- Feature matrix can affect on both of kernels

D: Dense format
S: Sparse format



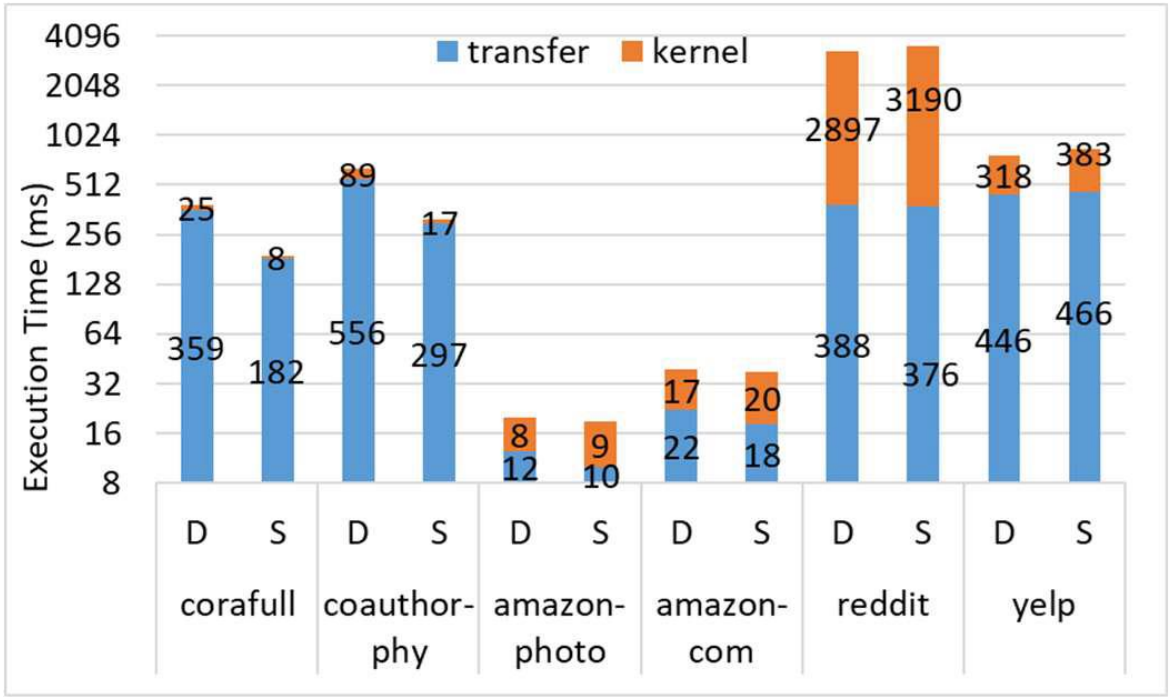
Execution Time: Combination

- The **memory transfer time** occupies a much larger part of the overall execution.
- Compressed format has **good trade-off gain**.



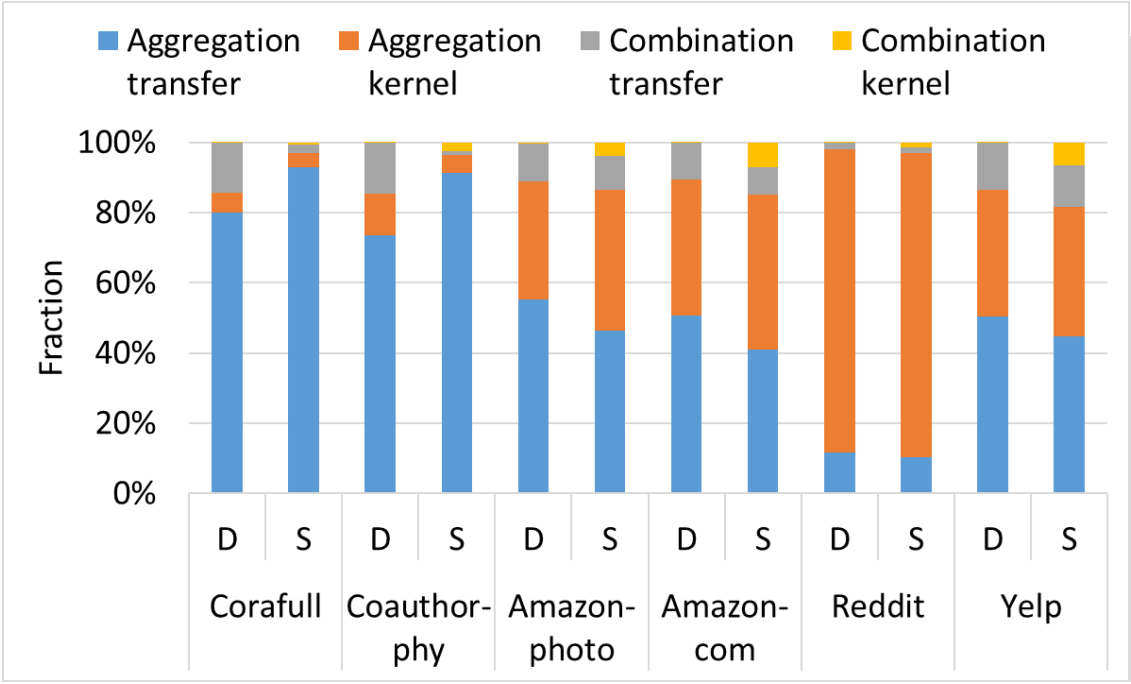
Execution Time: Aggregation

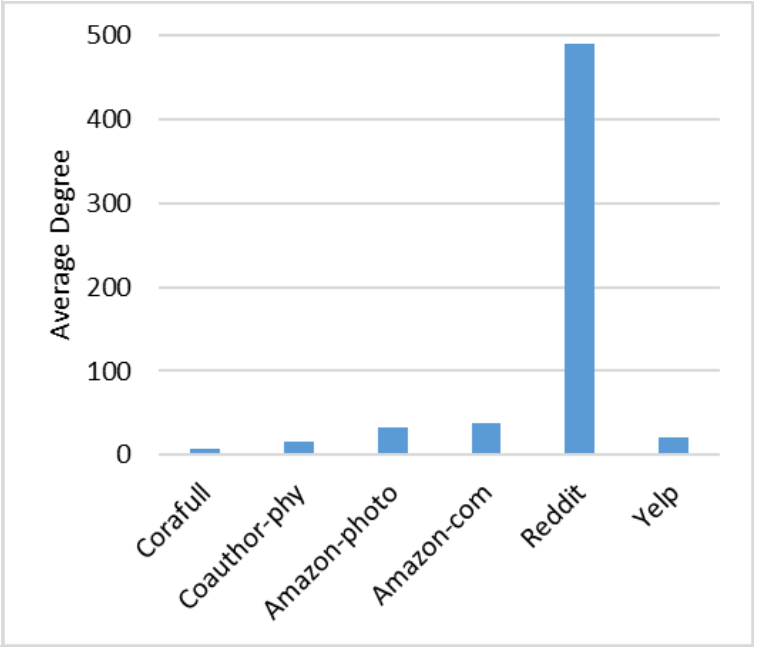
- Fractions of kernel execution time across datasets are varied
  - Corafull~Amazon-com: **Data transfer time** determines the total run-time
  - Reddit~Yelp: **Kernel execution time** determines the total run-time



Relationship between Kernel Execution and Data Properties

- Aggregation phase occupies most of the total execution time.
- The aggregation kernel execution time is proportional to the **average degree**.





Summary & Design Considerations

Summary

- The greater the **degree** (number of nodes x graph density),
  the higher the rate of **execution of the aggregation**.
- Design the kernel considering feature matrix improve the performance differently.

Design Considerations

- Data property-aware kernel execution
  - Selecting an appropriate kernel according to data characteristics.
- Improve the programming model of the kernel
  - Process multiple neighboring nodes at once.
  - Using shared memory to improve data locality.

Thank you