Towards Efficient Large-Scale Graph Neural Network Computing

Lingxiao Ma, Zhi Yang, Youshan Miao, Jilong Xue, Ming Wu, Lidong Zhou, Yafei Dai
Peking University, Microsoft Research Beijing
2018

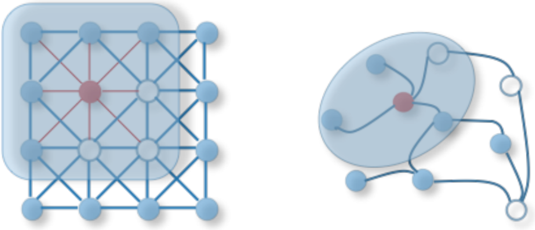
Presenter : Derrick Blakely

https://qdata.github.io/deep2Read

Outline

Outline

GNNs and GPUs

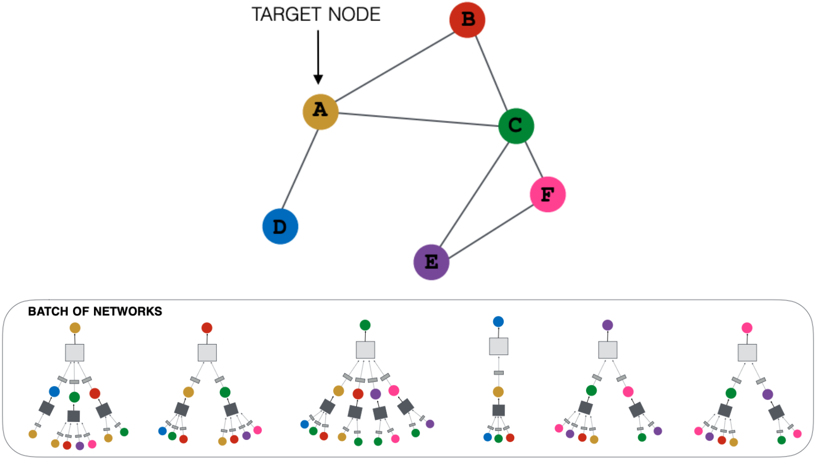


- Graphs too large to fit in GPU memory
- Irregular inputs make SIMD difficult
- Sparse matrices

GNNs and DL Frameworks

- No simple programming interface for training GNNs
- Lots of time spent allocating computation graphs
- Bad job mapping computation graphs to GPU
- Simple data-parallelism isn't the best way to parallelize GNNs

GNNs and DL Frameworks



Older Graph Libraries

- Pregel, PowerGraph, GraphLab, GraphX, etc
- Define vertex program, use Gather-Apply-Scatter (GAS)
- Don't allow users to define the "dataflow"
- Don't use efficient tensor libraries
- Don't support NN architectures
- Use scalar vertex features, not feature vectors

Outline

Example: Gated GCN

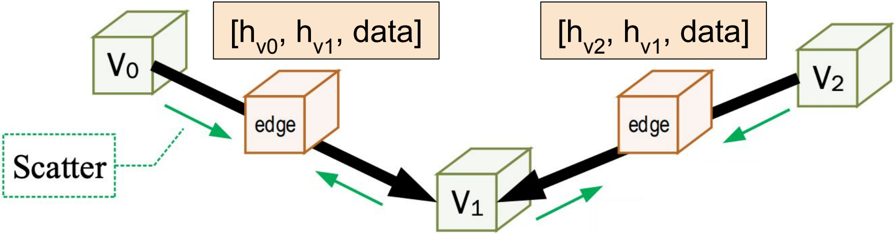
$$\mathbf{h_u^{l+1}} = ReLU(W^l(\sum_{v \to u} \sigma(W_H^l \mathbf{h_u^l} + W_C^l \mathbf{h_v^l}) \odot \mathbf{h_v^l}))$$

- Main observation: GNN layers can be split into edge functions and vertex functions
- Edge function: inside the summation
- Vertex function: outside the summation

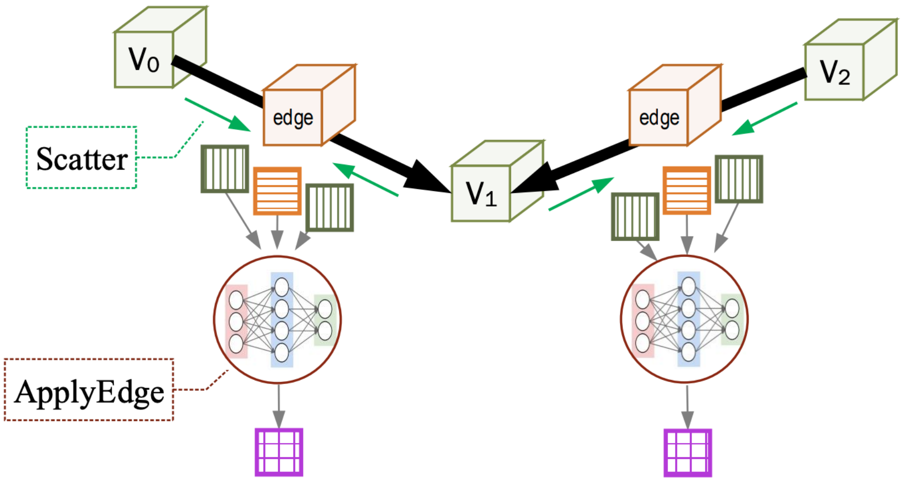
SAGA - four stages of computation

1. Scatter
2. ApplyEdge
3. Gather
4. ApplyVertex

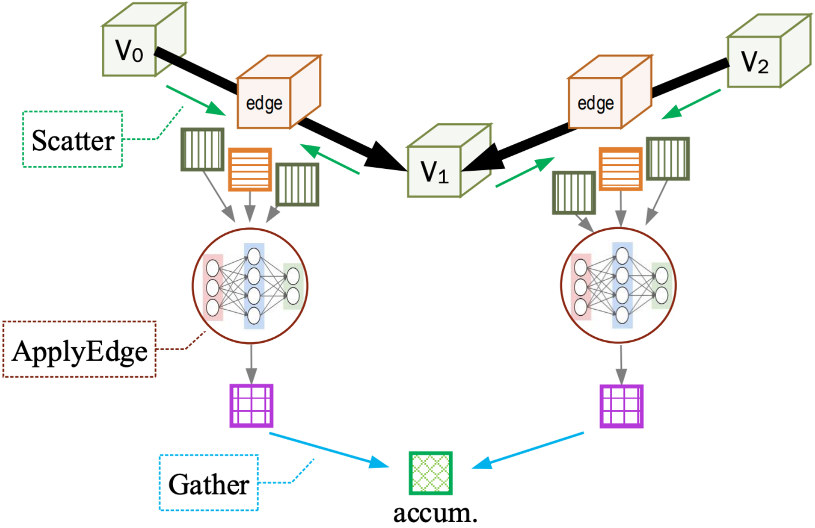
Scatter



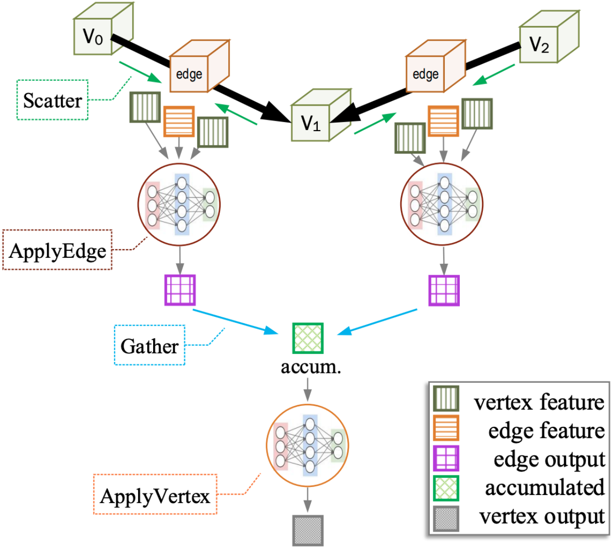
ApplyEdge



Gather

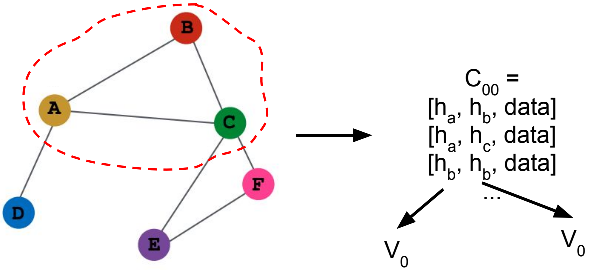


ApplyVertex



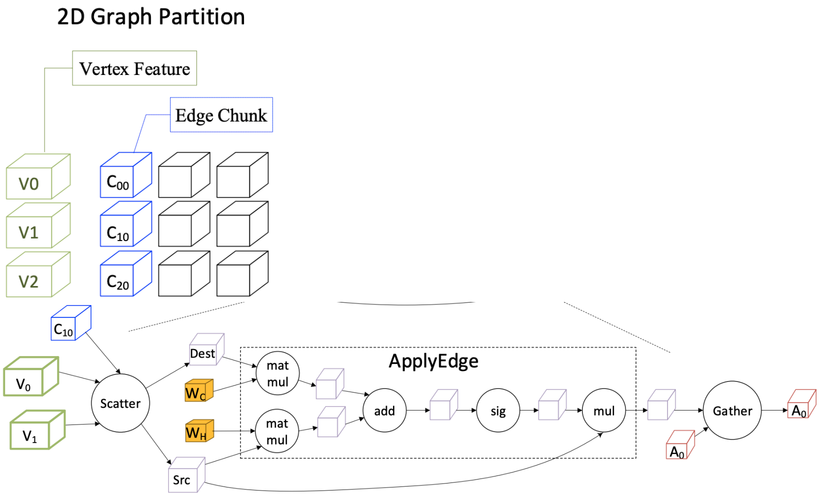
Outline

Chunk-Based Streaming Dataflow

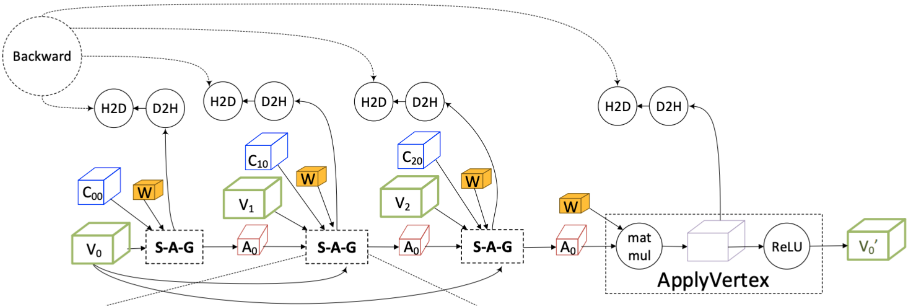


- Partition graph into chunks that fit in GPU memory
- C_{ij}: edge chunk connecting vertex chunks V_i and V_j

Chunk-Based Streaming Dataflow

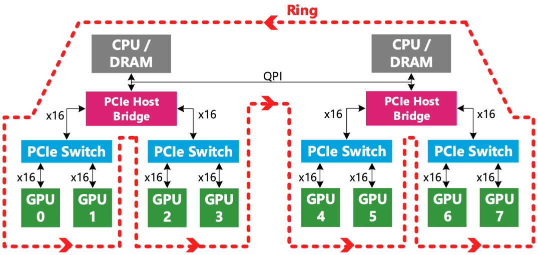


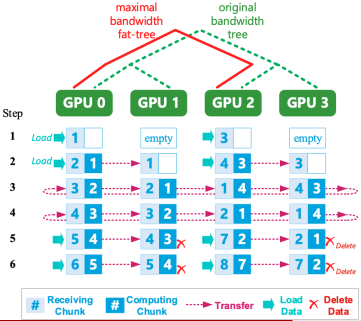
Dataflow Scheduling



Outline

Ring-Based Parallel Streaming





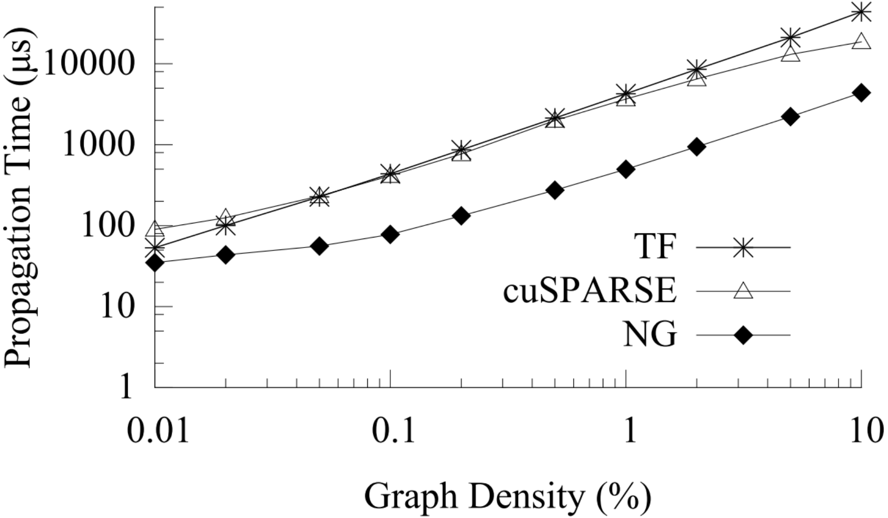
Outline

Datasets

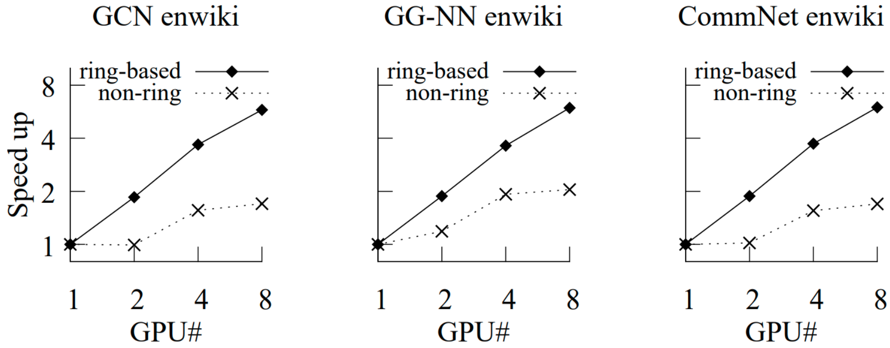
Dataset	vertex#	edge#	feature	label
pubmed	19.7K	108.4K	500	3
protein	43.5K	205.6K	29	3
BlogCatalog	10.3K	668.0K	128	39
reddit_small	46.6K	1.4M	602	41
reddit_middle	233.0K	23.2M	602	41
reddit_full	2.2M	571.0M	300	50
enwiki	3.2M	222.1M	300	12

- Pubmed citation network
- Protein-protein interaction graphs
- BlogCatalog Social network
- Reddit online discussion forum
- Wikidata

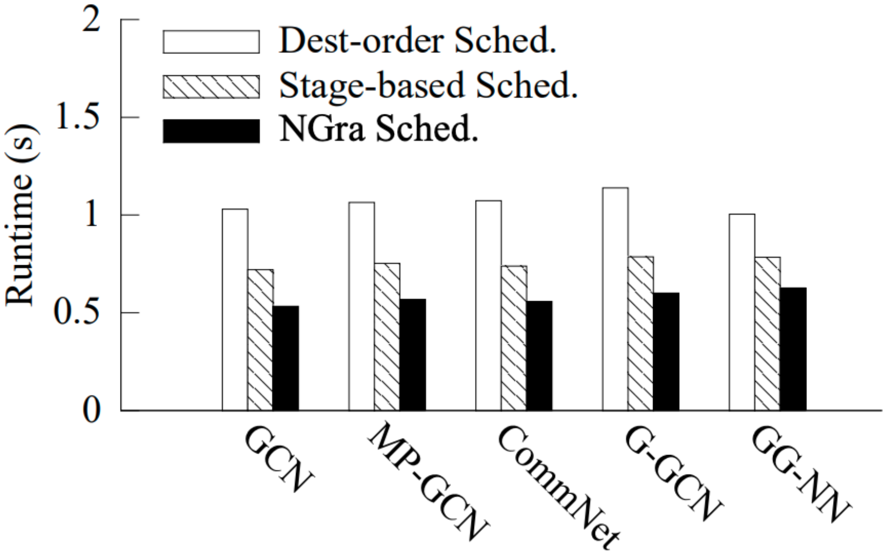
NGram and TF diverge with increasing sparsity



Ring-based streaming is effective



NGram scheduling strategy is effective



Outline

Weaknesses

- No available source code
- No reason to use their tool when we can use PyTorch Geometric
- Ring-based streaming is just a simple heuristic
- Doesn't scale to multi-host setting
- Evaluations aren't very exciting - we already knew TF wasn't great at this

Lessons Learned

- More exotic NN architectures can be handled by TF/PyTorch, provided we efficiently create and schedule their computation graphs
- Reducing sparsity important for efficient GPU use
- Vertex-centric programming model is becoming more and more prominent