Scaling K-Path Centrality using Optimized Distributed Data Structure

TEXAS A&M UNIVERSITY

Lawrence Livermore National Laboratory

Lance Fletcher (TAMU & LLNL), Trevor Steil (LLNL), Roger Pearce (LLNL & TAMU)

Introduction

- K-Path Centrality (KPC) measures information flow along simple paths in a graph [1].
- Estimation involves sampling numerous random paths of length at most K.
- KPC estimation effectively identifies high-betweenness vertices in many graphs [1].
- Used in various graph problems due to its computational efficiency [2],[3],[4].
- Requires a large number of path samples for large graphs.
- Introducing the RaNT-Graph: a distributed data structure for efficient path sampling.
- RaNT-Graph comprises three key components:
  1. Vertex delegation partitioning [5]
  2. Rejection sampling
  3. Asynchronous communication [6].

Methods

- RaNT-Graph uses vertex delegation partitioning to balance compute, memory, and communication [5].
- Vertex delegation partitioning distributes high degree vertices (*hubs*) amongst all processors.
- Rejection sampling is employed during path sampling to randomly select an unvisited neighbor.
- Stepping to a delegated vertex sends the path to the processor which owns a random edge of the current vertex.
- Stepping to an undelegated vertex v sends the path to the processor which owns v.

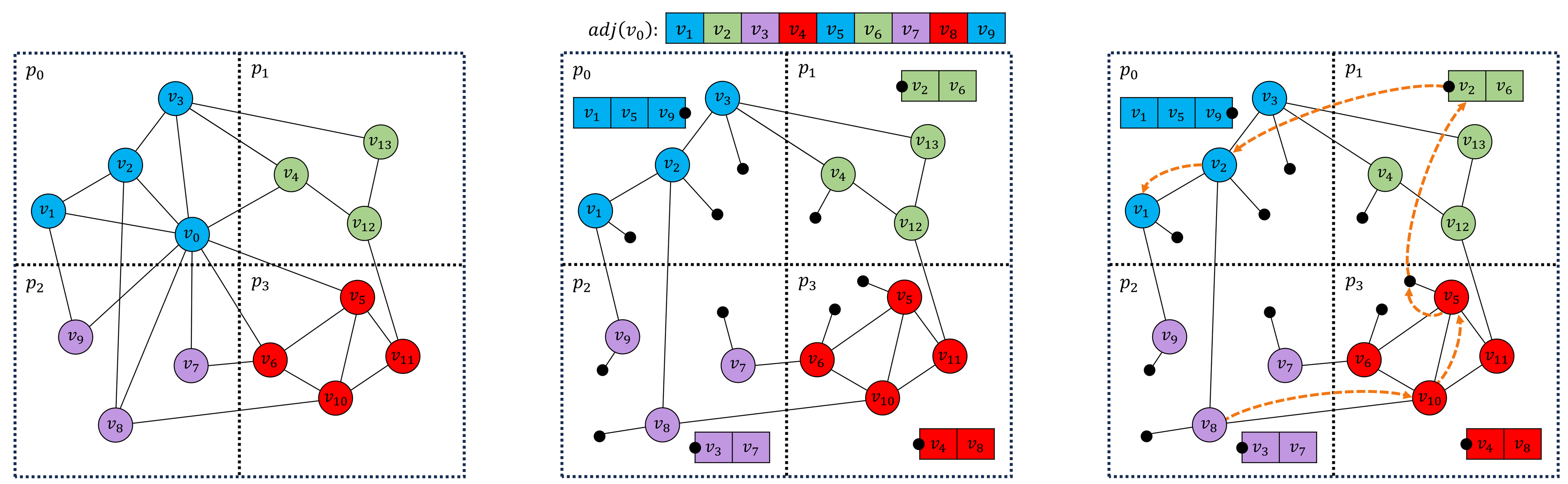

Fig. 1: (Left) A 1D partitioning of a graph with a hub vertex v_0. Colors indicate the processor which owns the vertex. (Middle) A vertex delegation partitioning of the same graph, where the distributed adjacency list of v_0 is shown at the top. (Right) An example of a simple path. The path of vertices is $v_8 \rightarrow v_{10} \rightarrow v_5 \rightarrow v_0 \rightarrow v_2 \rightarrow v_1$. The path of processors is $p_2 \rightarrow p_3 \rightarrow p_3 \rightarrow p_1 \rightarrow p_0 \rightarrow p_0$.

Discussion

- RaNT-Graph can sample large amounts of simple paths, but it can also be used to sample walks.
- This and extending RaNT-Graph to weighted graphs allows it to be used for
  - Personalized PageRank
  - DeepWalk
  - node2vec.
- Future optimizations include making use of co-located edges and other sampling techniques.

References

[1] T. Alahakoon, R. Tripathi, N. Kourtellis, R. Simha, and A. Iamnitchi, "K-path centrality: A new centrality measure in social networks," in Proceedings of the 4th Workshop on Social Network Systems, ser. SNS '11. New York, NY, USA: Association for Computing Machinery, 2011.

[2] A. Biswas and B. Biswas, "Community-based link prediction," Multimedia Tools and Applications, vol. 76, no. 18, pp. 18 619–18 639, Sep. 2017.

[3] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Enhancing community detection using a network weighting strategy," Information Sciences, vol. 222, pp. 648–668, Feb. 2013.

[4] —, "Mixing local and global information for community detection in large networks," Journal of Computer and System Sciences, vol. 80, no. 1, pp. 72–87, Feb. 2014.

[5] R. Pearce, M. Gokhale, and N. M. Amato, "Faster Parallel Traversal of Scale Free Graphs at Extreme Scale with Vertex Delegates," in SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, Nov. 2014, pp. 549–559, iSSN: 2167-4337.

[6] B. Priest, T. Steil, G. Sanders, and R. Pearce, "You've got mail (ygm): Building missing asynchronous communication primitives," in 2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2019, pp. 221–230.

Results

- Experiments written with YGM [5] to support MPI based communication.
- Utilized LLNL's Catalyst cluster with compute nodes having dual Intel Xeon E5-2695v2 processors (24 cores, 128GB DRAM).
- Table 1 presents graph details: vertices (n), edges (m), largest degree vertex (d_{max}), sampled paths (T), and max path length (K).
- Compared RaNT-Graph to 1D partitioning with rejection sampling (1D-Rej) and without (1D-No-Rej).
- Figure 2 illustrates weak scaling on R-MAT graphs. 1D-Rej, 1D-No-Rej, and RaNT-Graph-50K sample 50K paths per compute node. RaNT-Graph samples 1M paths per compute node.
- Figure 3 shows strong scaling using graphs from Table 1.

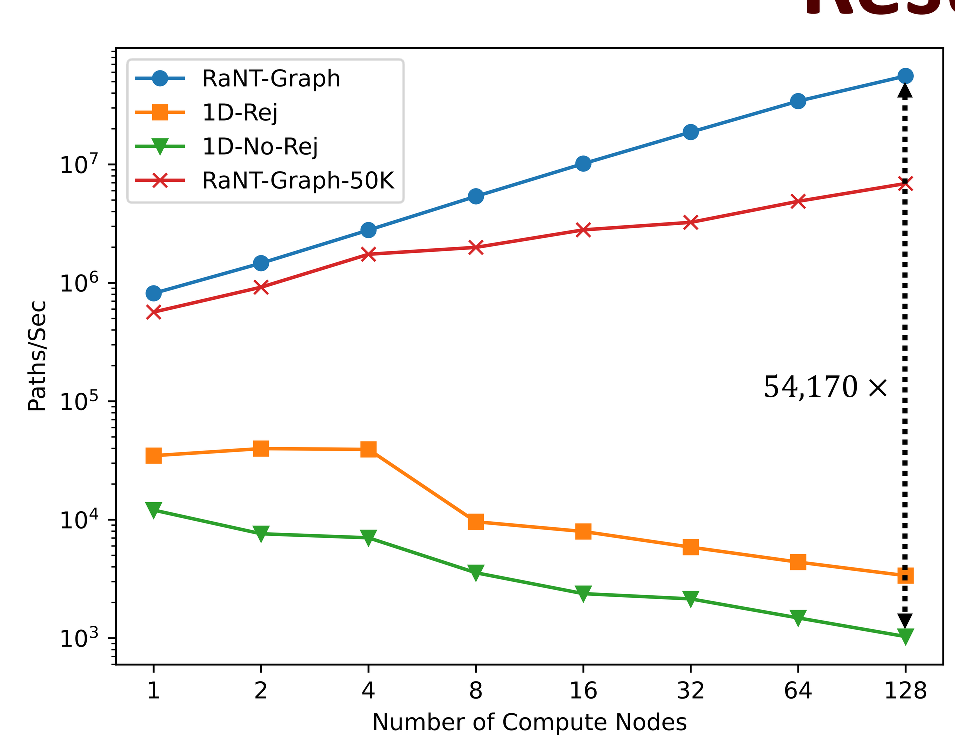

Fig. 2: Weak scaling experiment using R-MAT graphs starting at scale 26 up to scale 33.

Table 1

Graph	n	m	d_{max}	T	K
Orkut	3M	117M	33K	74M	18
LiveJournal	4.85M	43M	20K	102M	18
Twitter	42M	1.2B	3M	580M	21
Friendster	66M	1.8B	5.2K	857M	22
web-cc12-hostgraph	89M	1.9B	3M	1B	22
Uk-2007-05	106M	3.3B	975K	1.2B	22

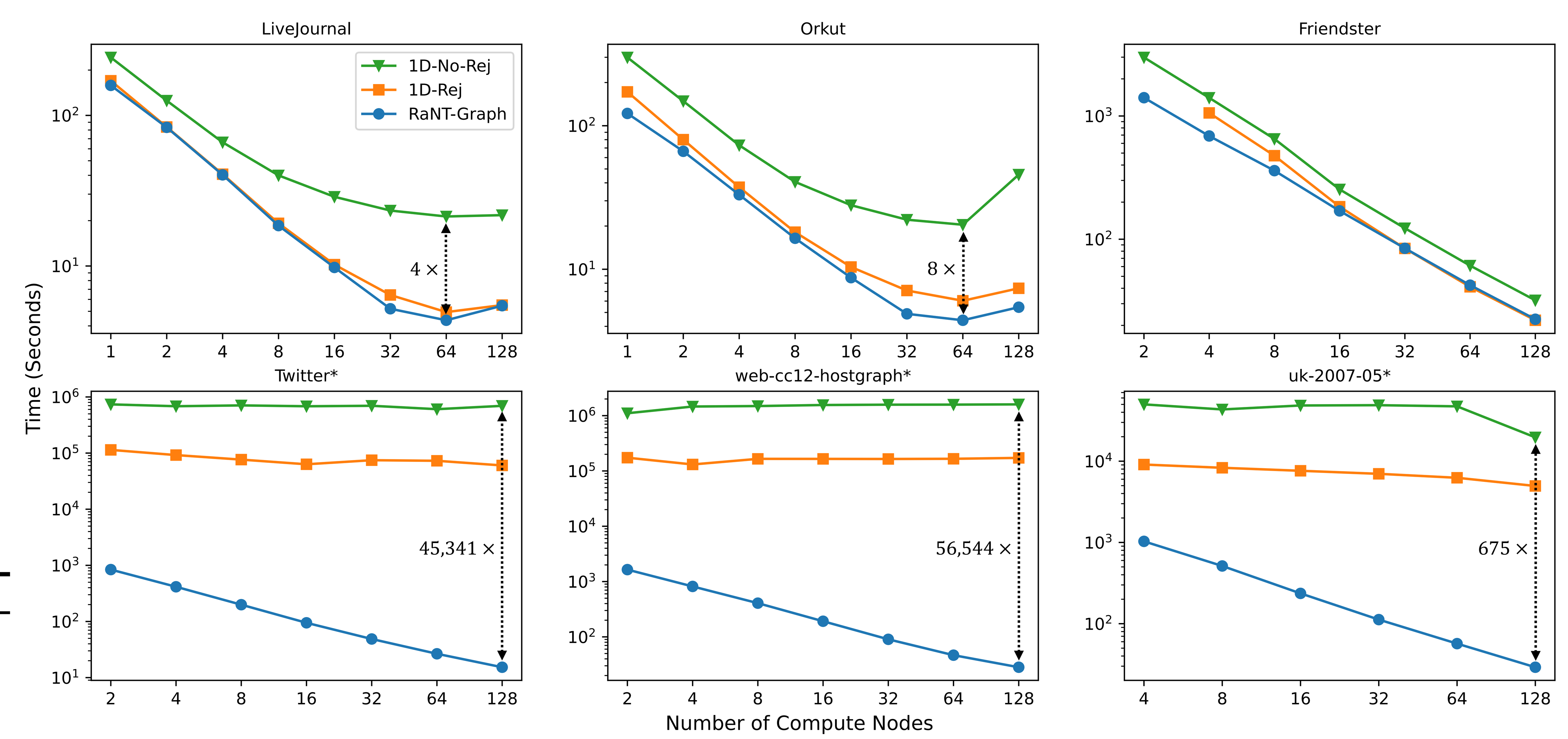

Fig. 3: Strong scaling experiment. *On these graphs 1D-Rej and 1D-No-Rej values estimated by sampling 1M paths and extrapolating based on desired number of paths T. RaNT-Graph provides a substantial speedup when the graph contains a large d_{max}.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52 07NA27344 (LLNL-POST-855041). Funding from LLNL LDRD project 21-ERD-020 was used in this work.