[Draft] Locality Optimization

for traversal-based queries on graph databases

A thesis submitted to the

Universität Konstanz

Department of Computer and Information Science

1st Reviewer:	Prof. Dr. Michael Grossniklaus
2nd Reviewer:	Dr. Michael Rupp

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer and Information Science

by

Fabian Klopfer

Konstanz, 2020

Abstract:

Some New Abstract Text

Contents

1 Introduction

Essay-like Intro.

Organisation

Contributions

2 Preliminaries

2.1 Graph Databases

2.1.1 Notations and Definitions

2.1.2 Architecture

2.1.3 Data Structures

2.1.4 Queries and Algorithms

2.1.5 Locality

2.2 Problem Definition

2.3 Related Work

2.3.1 Graphs and Locality in Relational Databases

2.3.2 Triplet Stores

2.3.3 Interaction Graphs

2.3.4 Neo4J

2.3.5 G-Store

3 Methods

3.1 G-Store: Multilevel Partitioning

3.1.1 Coarsening

3.1.2 Turnaround

3.1.3 Uncoarsening

3.1.3.1 Projection

3.1.3.2 Reordering

3.1.3.3 Refinement

3.2 ICBL: Diffusion Set Clustering

3.2.1 Diffusion Sets

3.2.2 Coarse Clustering

3.2.3 Hierarchical Subgraph Clustering

3.2.4 Block Layout Generation

3.3 Our Approach

3.3.1 Louvain method/Leiden

3.3.2 Kernighan-Lin Algorithm/N-Body Simulation/X?

3.3.3 Adjacency List Rearrangement

4 Experimental Evaluation

4.1 Experimental Setup

4.1.1 Implementation

4.1.2 Environment

4.1.3 Data Sets and Queries

4.2 Results

4.2.1 Modularity

4.2.2 Conductance and Cohesiveness

4.2.3 Tension

4.2.4 Block-based IOs

5 Conclustion

5.1 Discussion

Remention contributions

5.2 Future Work

5.3 Summary

Bibliography

[1] Renzo Angles and Claudio Gutierrez. “Survey of graph database models”. In: *ACM Computing Surveys (CSUR)* 40.1 (2008), pp. 1–39.

[2] L. Belady. “A Study of Replacement Algorithms for Virtual-Storage Computer”. In: *IBM Syst. J.* 5 (1966), pp. 78–101.

[3] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[4] Marek Ciglan and Kjetil Nørvåg. “SGDB–Simple graph database optimized for activation spreading computation”. In: *International Conference on Database Systems for Advanced Applications*. Springer. 2010, pp. 45–56.

[5] Allan M Collins and Elizabeth F Loftus. “A spreading-activation theory of semantic processing.” In: *Psychological review* 82.6 (1975), p. 407.

[6] Patricia Conde-Céspedes, Jean-François Marcotorchino, and Emmanuel Viennet. “Comparison of linear modularization criteria using the relational formalism, an approach to easily identify resolution limit”. In: *Advances in Knowledge Discovery and Management*. Springer, 2017, pp. 101–120.

[7] Peter J Denning. “The locality principle”. In: *Communication Networks And Computer Systems: A Tribute to Professor Erol Gelenbe*. World Scientific, 2006, pp. 43–67.

[8] Edsger W Dijkstra et al. “A note on two problems in connexion with graphs”. In: *Numerische mathematik* 1.1 (1959), pp. 269–271.

[9] Bura Gedik and Rajesh Bordawekar. “Disk-based management of interaction graphs”. In: *IEEE Transactions on Knowledge and Data Engineering* 26.11 (2014), pp. 2689–2702.

[10] Andrew V Goldberg and Chris Harrelson. “Computing the shortest path: A search meets graph theory.” In: In SODA (Vol. 5, pp. 156-165).

[11] Peter E Hart, Nils J Nilsson, and Bertram Raphael. “A formal basis for the heuristic determination of minimum cost paths”. In: *IEEE transactions on Systems Science and Cybernetics* 4.2 (1968), pp. 100–107.

[12] Jürgen Hölsch and Michael Grossniklaus. “An algebra and equivalences to transform graph patterns in neo4j”. In: *EDBT/ICDT 2016 Workshops: EDBT Workshop on Querying Graph Structured Data (GraphQ)*. 2016.

[13] Karl Pearson. “The problem of the random walk”. In: *Nature* 72.1867 (1905), pp. 342–342.

[14] Raghu Ramakrishnan and Johannes Gehrke. *Database management systems*. McGraw-Hill, 2000.

[15] Ian Robinson, Jim Webber, and Emil Eifrem. *Graph databases: new opportunities for connected data*. "O’Reilly Media, Inc.", 2015.

[16] Marko A Rodriguez and Peter Neubauer. “The graph traversal pattern”. In: *Graph Data Management: Techniques and Applications*. IGI Global, 2012, pp. 29–46.

[17] Robert Soulé and Bura Gedik. “RailwayDB: adaptive storage of interaction graphs”. In: *The VLDB Journal* 25.2 (2016), pp. 151–169.

[18] Robin Steinhaus, Dan Olteanu, and Tim Furche. “G-Store: a storage manager for graph data”. PhD thesis. University of Oxford, 2010.

[19] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Scientific reports* 9.1 (2019), pp. 1–12.

[20] Abdurrahman Yaar. “Scalable layout of large graphs on disk”. PhD thesis. Bilkent University, 2015.

[21] Abdurrahman Yaar, Bura Gedik, and Hakan Ferhatosmanolu. “Distributed block formation and layout for disk-based management of large-scale graphs”. In: *Distributed and Parallel Databases* 35.1 (2017), pp. 23–53.

[22] Konrad Zuse. “Über den allgemeinen Plankalkül als Mittel zur Formulierung schematisch-kombinativer Aufgaben”. In: *Archiv der Mathematik* 1.6 (1948), pp. 441–449.