

© 2022 Samiran Kawtikwar

ENHANCEMENTS IN HIGH PERFORMANCE SUBGRAPH
ENUMERATION ON GRAPHICS PROCESSORS

BY

SAMIRAN KAWTIKWAR

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Industrial Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Advisers:

Professor Rakesh Nagi
Professor Wen-mei Hwu

ABSTRACT

Subgraph enumeration is an important problem in graph theory with wide range of applications. With rapidly increasing graph sizes due to advent of internet and smartphones, subgraph enumeration needs high performing implementations. Being NP-complete, this problem poses significant scalability challenges and needs efficient implementations for practical solutions. Fortunately, this problem is highly amenable to parallelization. There are already many solutions in the multi-core and distributed computing community. GPU (Graphics Processing Unit)-based solutions are recently gaining recognition as they offer massive parallelism without network delays.

Most GPU solutions use Breadth First Traversal to utilize underlying parallelism and impose expensive restrictions on hardware due to huge memory requirements. PARSEC (Parallel Subgraph Enumeration and Counter) [1] is the first GPU-based solution that uses Depth First Search and performs in-memory subgraph enumeration. In this thesis, PARSEC is improved by leveraging insights from traditional sequential solutions and advanced parallel programming techniques.

The performance of Subgraph Enumeration is limited by the computational cost of adjacency list intersection operations. To tackle this, a smart preprocessing technique was developed for detecting opportunities for intersection reuse, which, in turn, reduces the number of intersections by up to $3.87\times$. A two-phase pruning technique was developed which shrinks the search space to further reduce the number of intersections by up to $6.6\times$. An in-depth analysis of PARSEC was conducted to overcome its load imbalance and limited hardware utilization. A hybrid parallelization scheme was developed that improves the load balance by up to $14\times$. Altogether, these improvements provide a geometric mean time speedup up to $4.6\times$ across data graphs and up to $3.7\times$ across all queries with max speedups up to $14.6\times$.

To Aai, Baba and Tai for their unconditional love and support.

ACKNOWLEDGMENTS

I express my deepest gratitude to my parents for their endless love and support, for their hardships and sacrifices enabling me to pursue my dreams.

I would like to thank my advisor Professor Rakesh Nagi for great mentorship and support. His energy and modesty have been a great source of motivation and humility. I would like to extend my sincerest acknowledgments to Professor Wen-mei Hwu and Dr. Jinjun Xiong for their excellent insights and support. I am grateful to have found such great mentors to work with.

A special thanks to my colleagues at the C3SR group: Mohammad, Yen-Hsiang, Samhita, and Vibhor for their excellent support and ideas. I learned a lot from you all through this project. A huge appreciation to IBM C3SR for partially funding my appointment and to the ISE admin staff especially Ms. Lauren Redman for being so helpful and accommodating through the pandemic.

My gratitude is also extended to my previous mentors: Professor Anand Dev Gupta and Dr. Varun Ramamohan for inspiring me to pursue research.

Lastly, I would also thank my friends in Urbana Champaign for all the fun and homely environment.

This research is part of the Delta research computing project, which is supported by the National Science Foundation (award OCI 2005572), and the State of Illinois. Delta is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.

TABLE OF CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	vii
LIST OF ABBREVIATIONS	viii
CHAPTER 1 INTRODUCTION	1
CHAPTER 2 BACKGROUND	4
2.1 GPU Overview	4
2.2 Graph Storage	10
CHAPTER 3 LITERATURE REVIEW	12
3.1 Notation and Definitions	12
3.2 Related Work	13
CHAPTER 4 BASIC TECHNIQUES	17
4.1 Query Graph Preprocessing	17
4.2 Data Graph Preprocessing	20
4.3 Search Tree Traversal	23
CHAPTER 5 IMPROVEMENTS	28
5.1 Intersection Reuse	28
5.2 Hybrid Symmetry Breaking	32
5.3 Hybrid Parallelism for Load Balance	37
CHAPTER 6 RESULTS	41
6.1 Experimental Setup	41
6.2 Performance Criteria	42
6.3 Final Results and Analysis	43
CHAPTER 7 CONCLUSIONS AND FUTURE WORK	47
REFERENCES	49

LIST OF TABLES

5.1	Effect of reuse on number of intersections for com-youtube . .	33
5.2	Intersection speedup with degree based symmetry breaking . .	34
5.3	Comparison of hybrid strategy with pure, best and worst possible strategies using intersection speedup criteria for data graph cit-patents	36
5.4	Time speedups with hybrid Symmetry Breaking	36
6.1	Data graphs used for experiments	41
6.2	Speedups across all queries and data graphs	43

LIST OF FIGURES

1.1	Subgraph Enumeration Example	2
2.1	NVIDIA Revenue by End Market	5
2.2	CUDA Heterogeneous Programming Model	6
2.3	GPU Thread Hierarchy	7
2.4	GPU Memory Hierarchy	8
2.5	Graph Storage Formats for Data Graph in Fig. 1.1	11
4.1	Example	18
4.2	Query Sequencing Output of G_q using $VF3$	18
4.3	Automorphism Group of G_q	19
4.4	Peeling Example	21
4.5	Generating Priority Sorted Column Index Array	22
4.6	Induced Subgraph for d_{12}	22
4.7	Step 1	24
4.8	Step 2	25
4.9	Step 3	26
4.10	Step 4	26
4.11	Step 5	27
5.1	Stepwise Fraction Time Spent	29
5.2	Reuse Detection Example	30
5.3	Flowchart With and Without Reuse	31
5.4	Parallelization Schemes	37
5.5	Degree vs Run Time for com-youtube	38
5.6	Load Balance for com-youtube	39
5.7	Load Balance for Nodes with Degree ≤ 256	39
5.8	Run Time Fractions vs Cutoff for com-youtube	40
5.9	Run Times with Hybrid Parallelism for com-youtube	40
6.1	Query Graphs used for Experiments	42
6.2	Execution Times and Speedups Compared to PARSEC [1]	44
6.2	Execution Times and Speedups Compared to PARSEC [1]	45
6.2	Execution Times and Speedups Compared to PARSEC [1]	46
6.3	Geo Mean Speedups across Data Graphs	46

LIST OF ABBREVIATIONS

BFS	Breadth First Search
CPU	Central Processing Unit
DFS	Depth First Search
GPGPU	General Purpose Graphics Processing Unit
GPU	Graphics Processing Unit
CUDA	Compute Unified Device Architecture
SIMD	Single Instruction Multi Data
SM	Streaming Multiprocessor
HPC	High Performance Computing
FLOPS	Floating Point Operations Per Second

CHAPTER 1

INTRODUCTION

Graphs are widely used to represent complex information such as social networks, chemical compounds, World Wide Web, logistic networks, bioinformatics, etc. Graph algorithms have applications in diverse areas like molecular engineering, social sciences, operations research, computer science, and many more. Formally, a graph G is defined as a collection of vertices V and edges $E \subseteq V \times V$ represented as $G = (V, E)$, where vertices represent entities and an edge represents relation between a pair of entities that are its endpoints.

Subgraph Enumeration is a fundamental problem in graph theory that aims at finding all instances of a given query graph G_q in a larger data graph G_d . Formally, given a query graph G_q and data graph G_d , the objective is to find all instances of subgraphs of G_d that are *isomorphic* to G_q . Figure 1.1 illustrates an instance of subgraph enumeration.

Isomorphic graphs are of interest since they represent similarity between inherent topology of the graph. For graphs G_1 and G_2 graph isomorphism is a bijection between the vertex sets $\mathcal{V}(G_1)$ and $\mathcal{V}(G_2)$ $f : \mathcal{V}(G_1) \rightarrow \mathcal{V}(G_2)$ with the condition: *any two vertices u and v of G_1 are adjacent in G_1 if and only if $f(u)$ and $f(v)$ are adjacent in G_2* . In this work we focus on undirected query graphs and data graphs. Subgraph enumeration entails the subgraph isomorphism problem which is known to be NP-complete [2].

Subgraph Enumeration has wide range of applications. For example, it can be used for plagiarism detection [3], network motif counting [4], comparing similarity between large graphs [5], designing bioinformatics networks [6], chemical target structure synthesis [7], analyzing insights in recommendation networks [8], etc.

With the advent of *Internet of Things* (IoT) and *Information Technology* (IT), the sizes of graphs representing underlying data are substantially bigger and pose significant challenges for data scalability. Increasing interests in

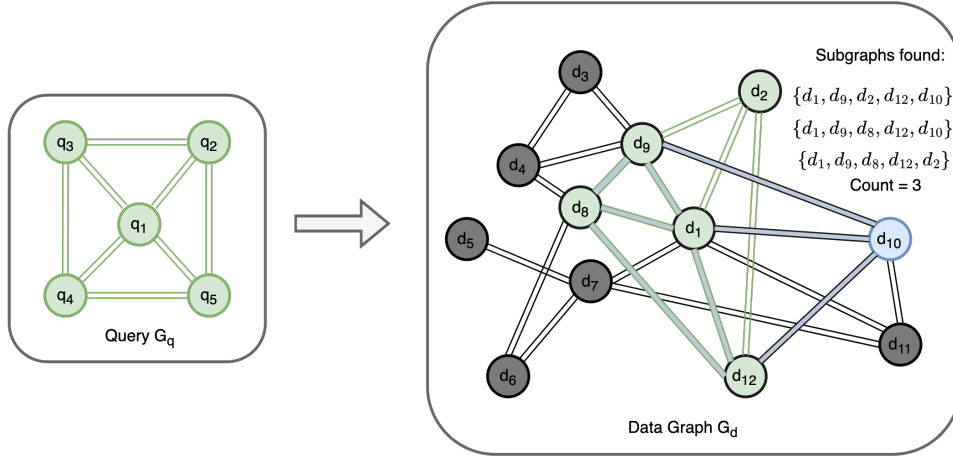


Figure 1.1: Subgraph Enumeration Example

graph analytics and improvements to computational hardware over the last two decades, particularly in CPUs and GPUs have helped develop practical solutions to this computationally challenging problem. Nevertheless, many existing solutions do not scale well with increasing template size due to issues like memory requirement, computational time, and load imbalance.

This work is focused on improving the existing state-of-the-art solution for subgraph enumeration: PARSEC, developed by Dodeja *et al.* [1]. A series of enhancements were developed to improve its computational performance which are discussed in later sections. These enhancements include:

1. Smart preprocessing techniques to detect intersection reuse and reduce computation.
2. Two-Phase strategy to improve symmetry breaking effectiveness.
3. Hybrid parallelism for better load balance.

The remainder of this thesis is organized as follows:

Chapter 2 provides an overview of Graphics Processing Units and graph storage formats.

Chapter 3 is devoted to a Literature Review that describes related work and defines mathematical notation used throughout this thesis.

Chapter 4 illustrates basic techniques used in the subgraph enumeration algorithm with the help of a running example.

Chapter 5 describes the aforementioned enhancements in detail and analyzes their individual contribution to performance.

Finally, Chapters 6 and 7 evaluate the combined improvement due to these individual enhancements and summarize the thesis with concluding remarks and scope for further work.

CHAPTER 2

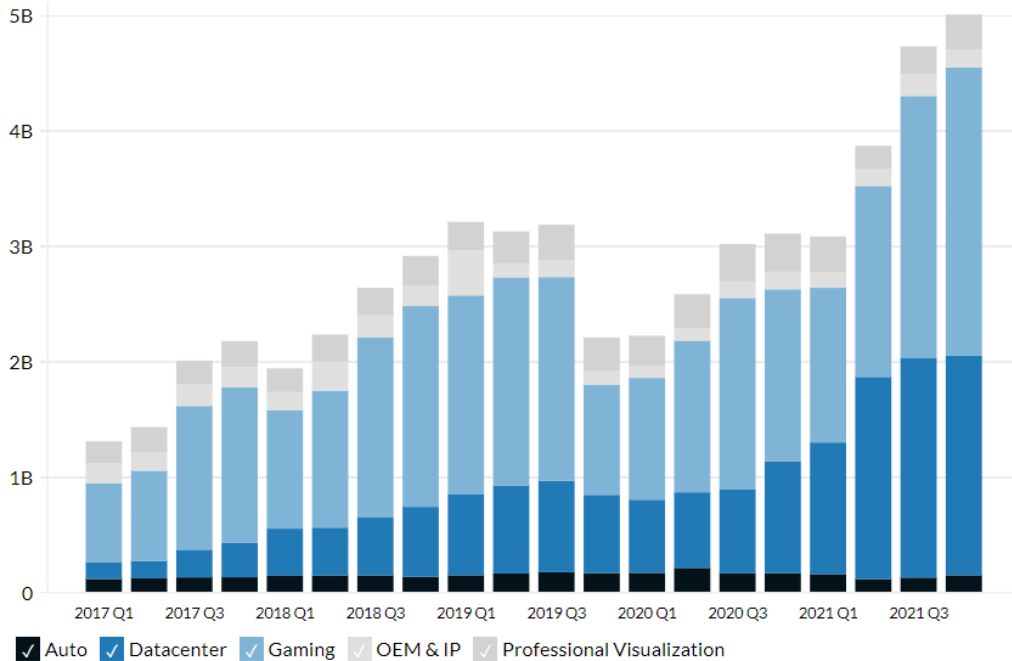
BACKGROUND

2.1 GPU Overview

2.1.1 History and Motivation

In the world of computational hardware, microprocessors based on single processing cores (traditional CPUs) have driven rapid performance improvements in floating-point operations per second (FLOPS) for nearly two decades (1985–2003). These microprocessors were capable of performing in the range of Giga-FLOPS (10^9) on desktops and Tera-FLOPS (10^{12}) on data centers [9]. This improvement drive flattened due to fundamental problems such as thermal stability, energy density, and quantum effects. Since then, all microprocessors have adapted architecture with multiple processing cores and thread virtualization to keep up with the improvements expected by the market. Naturally, it needed significant change in software to keep up with the reported performance and increasing computational requirements [10].

High-Definition graphics demand from the gaming industry was one of the biggest drivers for such high-performance hardware. The performance demanded here is throughput oriented, i.e., generating maximum frames per second while the computational requirement for each pixel is sufficiently low. Some chip manufacturers like NVIDIA took a different approach of single-instruction-multiple-thread (SIMT) hardware to cater to these requirements [11]. The hardware designed with this philosophy had threads capable of performing simple computational tasks with relatively lower clock speed to give an overall performance in Tera-FLOPS range. On top of that, the hardware gave more performance per watt which made it possible to add to a desktop machine [12]. Modern GPUs have also proven to be fast and cost-effective for deep learning and image processing tasks contributing to many



Source: Business Quant [13]

Figure 2.1: NVIDIA Revenue by End Market

advances in fields like computer vision and artificial intelligence. A modern data center GPU, for example (NVIDIA A100), now has more than 100,000 concurrent threads. These threads execute in many simple pipelines to give around 10 times faster raw performance than state-of-the-art CPUs [9].

Figure 2.1 shows that NVIDIA GPU data center market has improved significantly in recent years to become almost equivalent to the already huge gaming industry. This has fueled the hardware and software development of current server-grade GPUs which has enabled huge computational potential for deep learning. While HPC hardware improves at such rapid pace, software solutions for graph theory problems still need to move from the *multi-core* CPU paradigm to massively parallel GPU paradigm.

2.1.2 Heterogeneous Programming Environment

GPU is a throughput oriented computational hardware. A single thread of a GPU is significantly slower than a CPU. Hence, working with it for inherently sequential tasks can incur huge latencies. Since many basic programming tasks are inherently sequential and the whole operating system environment

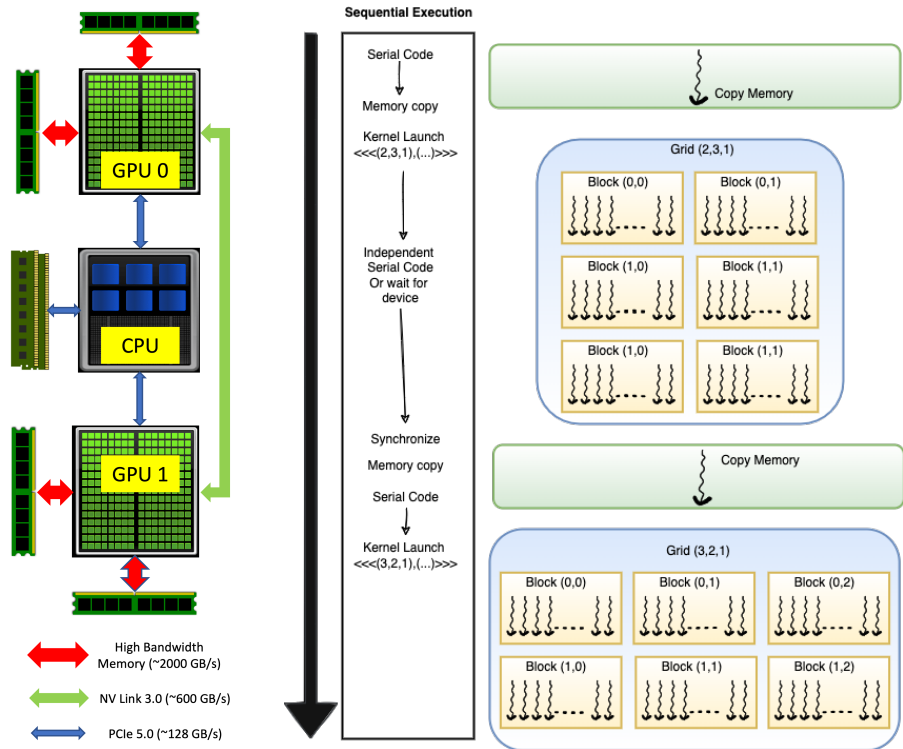
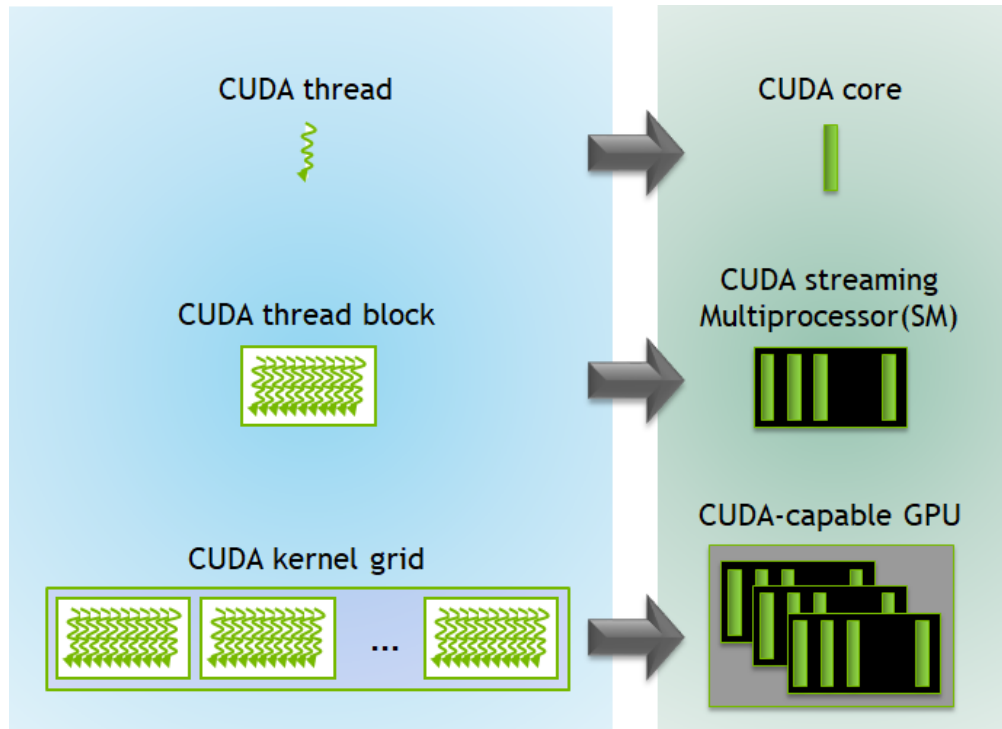


Figure 2.2: CUDA Heterogeneous Programming Model

is developed for latency optimized processors (CPUs), GPUs work better as hardware accelerator along with CPUs for computationally intensive tasks. Hence, GPUs are mostly used in a Heterogeneous programming environment with CPU (host) executing the governing code while offloading computationally intensive tasks as kernels to GPUs (device). Note that one host can be connected to multiple devices. Since the memory of the device and host are different, input and output data need to be copied before and after kernel execution. Figure 2.2 shows the high level hardware and software architecture. GPUs can communicate with CPU via PCIe (Peripheral Component Interconnect express) bus and with each other via NVLink™. The memory bandwidth of PCIe is the lowest and hence it is in best interests to avoid high data transfers between GPU and CPU. Native memory of the GPU or the VRAM, is High Bandwidth Memory (HBM) with the latest generation having bandwidth of around 2000 GB/sec. The NVIDIA proprietary GPU Interconnect with 12 links per GPU can provide up to 600 GB/sec. While the industry standard PCIe bus gen 5.0 with 32 links can provide up to 128 GB/sec of bandwidth.

2.1.3 Thread Hierarchy

As discussed in Section 2.1.1 GPUs have *SIMT* architecture with hundreds of thousands of concurrent threads. To have such massive hardware parallelism threads have to collaborate on usage of different resources for example, *cuda cores*. Since all threads do not share these resources, it is important to understand the thread Hierarchy in GPUs.



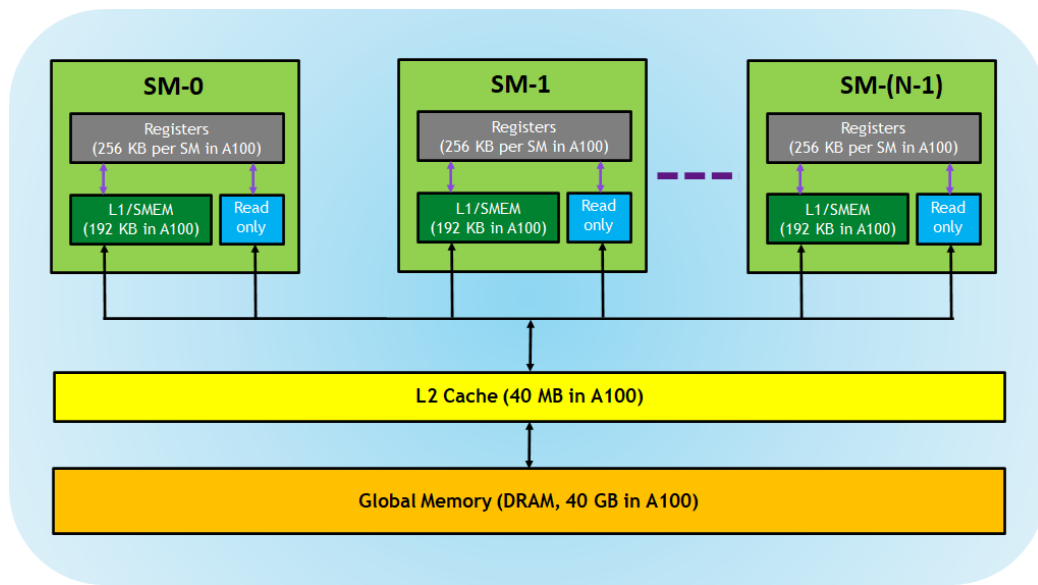
Source: NVIDIA Developer Blog [14]

Figure 2.3: GPU Thread Hierarchy

A thread is a single sequential flow of control within a program. A block is a chunk of threads while a grid is a chunk of blocks. Since threads are virtual they need to be optimally assigned hardware resources to perform tasks. Figure 2.3 shows how these virtual entities are mapped to the hardware. As discussed, a thread utilizes a *cuda core* for performing computation. A block of threads resides on a *Streaming Multiprocessor* (SM), which is a hardware block that consists of computational resources like *cuda cores*, *memory controllers*, *tensor cores*, etc. Thread block resides on an SM for the entirety of its execution. Hence, multiple threads in a block can utilize the same core. Multiple blocks can reside on an SM simultaneously within limits. GPU has

multiple SMs and hence many blocks can concurrently reside on the GPU during a context. If a grid requests more blocks than that can concurrently reside on the GPU, their execution is serialized. The latest server-grade GPU from NVIDIA (A100) has 108 SMs with a maximum of 32 blocks or 2048 threads residing per SM which means the maximum number of concurrent threads is 221,184. While on the hardware side, A100 has 6912 *cuda cores*, i.e., 64 cores per SM.

To add further, the thread instructions are processed in a Single Instruction Multi Data (SIMD) model. This SIMD chunk of threads is allocated resources at once; this chunk is called *warp*. If threads within a warp demand different operations, then all threads within that warp perform all operations with output of undesired threads being ignored. This causes a slowdown in execution and is referred as *warp divergence*. Hence, algorithms and implementations need to be designed in such a way that *warp divergence* is minimized. While the size of thread block and grid are configurable by the programmer, warp size is a fixed architectural parameter.



Source: NVIDIA Developer Blog [14]

Figure 2.4: GPU Memory Hierarchy

2.1.4 Memory Hierarchy

A GPU needs a high performance memory system as it can be easily stressed due to the underlying massive parallelism. If the available memory bandwidth is overwhelmed, threads need to wait for data to be able to perform computation even though some processing cores are free. This results in poor utilization and overall slower performance. Even though GPU memory is significantly faster than conventional RAM, it can still limit application performance. This problem is alleviated by building memory caches on the chip die itself, so accessing them is significantly faster (around 100 - 1000 times) but they have size restrictions. Modern compilers identify repeated memory accesses and can aggressively cache to reduce memory stress. The memory bandwidth problem is not as severe for CPUs as they have huge cache banks per thread core. But for GPUs, the cache per thread (L1 cache) is much smaller. For example, in the A100 GPU, up to 221,184 threads share an L2 cache of 40 MB, which is only 1.8 KB per thread.

GPU memory architects have attempted to solve this problem by adding hierarchical memory types. There are three memory types in a GPU system, namely *Global*, *Shared*, and *Register* memory. Figure 2.4 shows the memory hierarchy. As the name suggests global memory is accessible by all threads, it is equivalent to the Random Access Memory (RAM) on CPUs and is also sometimes referred as Video RAM (VRAM) or Device RAM (DRAM). Shared memory is accessible by all threads in a thread block. Register memory is private to each thread. The size restrictions of these memory banks are based on SM architecture and hardware. Programmers are also equipped with read-only constant memory (not shown in Figure 2.4) which is accessible by all threads. Unknown to the programmer, compilers can also utilize these memory banks by caching in register and shared memory banks if repeated accesses to global memory are detected. It is referred as L1 cache. Equivalent to CPU cache there is also an L2 cache for use by the compiler only and accessible to all threads.

As discussed, memory utilization plays a significant role in code performance; a program whose execution time is restricted by memory is called memory-bound. Memory-bound programs do not scale well with increasing number of devices as distributing computation to other workers provides no benefit and further increases latency due to communication. Some com-

mon techniques to reduce memory pressure are designing algorithms capable of using shared and register memory, consecutive threads accessing nearby memory, and reducing contention and/or latency for atomic operations on global memory [9].

2.1.5 Block Scheduling

As discussed in Section 2.1.3 multiple blocks can concurrently reside on a GPU. Generally, kernels are launched with grid sizes much higher to concurrently reside on a GPU so, the block execution needs to be scheduled. The block scheduler assigns hardware resources to maximum number of blocks that can occupy the GPU. The remaining blocks wait in a queue and are scheduled once a previously active block retires. This is done as block execution is non-preemptive. That is, once allocated resources, the block is never unallocated till completion. This scheduling provides scope for clever memory utilization as memory need not be allocated to the whole grid but only to the blocks that concurrently execute instructions. This concept of memory management is called persistent memory allocation and it is often used to reduce memory requirements [15].

It is important to design the parallelism in such a way that all blocks have similar amount of work. If not, blocks having longer work should be scheduled earlier to reduce makespan. To have better control over work allocation to a block, programmers often launch kernels with grid size equal to the number of blocks that can concurrently reside on the GPU. The blocks then pick from a common list of tasks in global memory; this approach is known as work stealing. Work stealing makes sure that work is allocated to a block as soon as it is free.

2.2 Graph Storage

Graphs can be stored in various formats. A suitable format is selected based on hardware and software constraints/requirements. *Adjacency Matrix* and *Adjacency list* representations are amongst the most elementary storage formats. As the name suggests, *adjacency list* format is stored as an array of lists where the i^{th} element of the array stores a list of neighbors of vertex

i. *Adjacency Matrix* format stores the relational information in a boolean matrix format. Let the underlying graph be $G = (V, E)$ and its adjacency matrix representation be A . Then $A = |V| \times |V|$ is defined as:

$$A_{i,j} = \begin{cases} 1 & (i,j) \in E, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

For most real world graphs, the underlying adjacency matrix is sparse. Hence, it is effective to be stored in sparse matrix formats. The computing community uses various sparse matrix formats such as *ELL-Pack* (ELL), *COOrdinate* (COO), *Compressed Sparse Column* (CSC), *Compressed Sparse Row* (CSR), *Blocked Compressed Sparse Row* (BSR), etc.

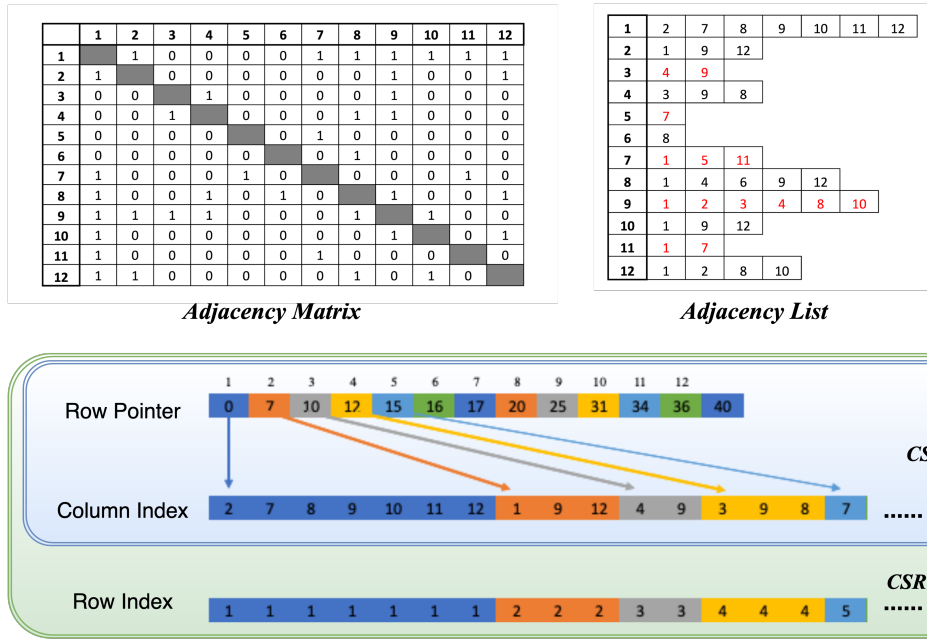


Figure 2.5: Graph Storage Formats for Data Graph in Fig. 1.1

We use a combination of CSR and COO formats in our implementation, which is referred as the CSRCOO format. This format uses three arrays for storing information of the underlying graph, namely, *Row Pointer* (length: $|V| + 1$), *Row Index* (length: $|E|$), and *Column Index*: (length: $|E|$). We maintain the *Row Index* and *Column Index* arrays sorted by their labels to enable binary search. Figure 2.5 shows comparison between different graph storage formats.

CHAPTER 3

LITERATURE REVIEW

This chapter describes relevant work in the area of subgraph enumeration in sequential (CPU) and parallel computing (multi-core/GPU) communities. First, some definitions and notation are established to develop formal understanding.

3.1 Notation and Definitions

A graph is a pair of sets $G = (V, E)$, where V is a set of elements called vertices and E is a set of vertex pairs whose elements are called edges. An edge is represented by the elements in the vertex pair (u, v) . u, v are called the endpoints of the edge. Edges of the graph can be directed or undirected. Undirected edges imply $\{(u, v) \in E \Leftrightarrow (v, u) \in E \quad \forall u, v \in V\}$. An edge $e = (x, y)$ is incident to a vertex v if $x = v$ or $y = v$. In this work, all graphs have undirected edges unless specifically mentioned. Graphs with directed Edge set are called directed graphs. $\mathcal{V}(G)$ and $\mathcal{E}(G)$ denote the vertex and edge set of graph G , respectively.

Neighbor set of a vertex v , denoted by $\mathcal{N}(v)$, is a set of all vertices u such that $\{u \in \mathcal{N}(v) \Leftrightarrow (u, v) \in E\}$. In case of a directed graph the neighbors are further categorized as *forward/outgoing* neighbors and *backward/incoming* neighbors based on the direction of the incident edge. Degree of a vertex v , denoted by $d(v)$ or *degree*(v) is the cardinality of its neighbor set, i.e., $d(v) = |\mathcal{N}(v)|$. Maximum degree of a graph is the maximum of all degrees from its vertex set. Oriented graph is a transformation of an undirected graph G to a directed graph \tilde{G} such that:

$$\{\forall (u, v) \ \& \ (v, u) \in \mathcal{E}(G), \text{ either } (u, v) \text{ or } (v, u) \in \mathcal{E}(\tilde{G})\}$$

A subgraph g of a graph G , denoted by $g \subseteq G$ is a graph $g = (V_g, E_g)$ such that $V_g \subseteq V$ and $E_g \subseteq E$ where, $\forall (u_g, v_g) \in E_g \Rightarrow u_g, v_g \in V_g$. Induced subgraph by a vertex v' in graph G , denoted by $G(v')$ is a subgraph $g = (V', E')$ of G such that $V' = \mathcal{N}(v') \cup \{v'\}$ and $\{(x, y) \in E' \Leftrightarrow x, y \in V'\}$.

For the rest of this thesis, the data graph is denoted by G_d and the query graph is denoted by G_q . $V_d = \mathcal{V}(G_d), V_q = \mathcal{V}(G_q)$. As introduced previously, the subgraph enumeration problem is to find all instances of G_q that are isomorphic to G_d . Isomorphism is a bijection f between vertex set of two graphs, here G_q and $g \subseteq G_d$, such that $(u, v) \in \mathcal{E}(G_q) \Leftrightarrow (f(u), f(v)) \in \mathcal{E}(g)$.

As we will see in Section 3.2, the traversal based algorithms for subgraph matching consist of routines for converting a given query graph to a directed graph and matching each vertex in the query graph to suitable vertices in the data graph. Whenever there is more than one match, each has to be considered independently. This creates branches and constructs a tree which is called the *search tree*. The order in which the data graph vertices are matched with the query graph is called the *match order*. Formally, given G_q , the match order π is the permutation of vertices V_q which reflects the order in which they will be matched.

3.2 Related Work

The underlying routine of subgraph enumeration is graph isomorphism. The first algorithm for detecting graph isomorphism was given by Ullman [16]. In this algorithm, an isomorphic subgraph was found using a brute force search tree traversal with elimination of successor nodes. The traversal used was DFS with each instance being represented as a binary table of $|V_q| \times |V_d|$ entries. For finding all instances, a DFS is recursively performed on all possible candidates with some of them pruned in advance. These search tree based algorithms are sometimes also referred to as *state space search* algorithms.

Some other promising techniques which perform traversal to find query instances are exploration [17], [18], [19], [20] and constraint programming [21], [22] based. Approaches which do not perform traversal are based on converting query instances to a multi-way join framework where the edges represent relations in G_q and evaluate the multi-way join to find all results

[23], [24]. However, these techniques perform better for labeled subgraph enumeration [25] and hence will be ignored in this review.

The algorithm of Ullman [16] formed the basis for various subsequent traversal based algorithms. Improvements came from different pruning heuristics and better match ordering. These algorithms are explained in Section 3.2.1. Most sequential algorithms perform state space search process at each root node. With advancements in multi-core processors and cloud computing servers, multiple scalable algorithms were developed, which are mentioned in Section 3.2.2. Most recently, the focus has shifted to GPUs because of increasing memory capacity and their advent in compute servers; these contributions are mentioned in Section 3.2.3.

3.2.1 Sequential Algorithms

Vento and Fogia (VF) [26] reduced the memory requirements of [16] by using a State Space Representation (SSR) and developed a deterministic matching method for simultaneously verifying isomorphism. VF2 [27] introduced five feasibility rules for early pruning of the search tree to shrink the search space. VF3 [28] developed multiple additional heuristics for early pruning and further reduced the search space. These heuristics were developed with the help of extensive computational experimentation. Grochow *et al.* [29] assigned partial ordering to vertices matched with symmetrical query vertex to reduce redundant work. Turbo_{iso} [30] improved match ordering by dynamically computing a match order for each candidate that made the search tree leaner and provided performance improvement. Boost_{iso} [31] in a novel contribution utilized vertex relationships to efficiently arrange candidate vertices to reuse adjacency list intersections.

These sequential algorithms offer great insights into the search tree exploration phase. Since the problem has so much independent work, a sequential processor tends to be slower due to repetition of tasks. Hence, sequential implementations are only practical for data graphs with up to 100,000 vertices. Many techniques used by parallel algorithms for shrinking search space arise from sequential algorithms.

3.2.2 Multi-core Algorithms

Parallel Subgraph Listing [32] was the first truly parallel search space traversal based algorithm on multi core hardware. It introduced a subgraph listing framework that statically divided work across multiple threads. Various load balancing strategies were developed to improve its performance. It was also proved that the problem of partial subgraph distribution for workload balance is strongly NP-Hard. CECI [33] proposed a novel framework for Subgraph Listing based on Compact Embedding Cluster Index, which divides the data graph into multiple embedding clusters for parallel processing. Despite orders of speedups and scalability, the technique uses BFS and needs huge amount of memory for storing intermediate results. LIGHT [34] eliminated the memory requirements of previous BFS parallel algorithms by developing a multi-threaded DFS framework. Dryadic [35] is the current state-of-the-art CPU subgraph enumerator that uses a robust computation tree structure which optimizes queries by mapping them to different backend hardware to perform custom compiled graph pattern matching.

To summarize, multi-core algorithms easily outperform sequential algorithms but are limited by the parallelism offered by CPUs. Memory bandwidth across different nodes makes it further difficult to scale. BFS-based strategies are prevalent in the parallel computing community, which impose huge hardware requirements due to significant memory usage.

3.2.3 GPU Algorithms

Due to the advancements in graph algorithms on GPUs like max-flow, shortest paths, and spanning trees, many state space traversal based subgraph enumeration algorithms were developed. To the best of our knowledge, GPSM [36] is the first subgraph enumeration implementation on GPU. It used advances in BFS on GPUs to implement the basic search tree traversal procedure. GPSM outperformed the most advanced multi-core algorithm of its time by an order of magnitude in speedup. RPS [37] focused on reducing the number of set intersection operations, since this is the most expensive computational step, by deploying a reuse framework in BFS traversal. PBE [38] developed a scalable framework by using graph partitioning strategies to enable multiple GPUs to work on each partition.

Since all BFS-based approaches explode in memory requirements, PARSEC [1] developed the first DFS-based subgraph enumeration on GPUs. It reduced the search space by limiting compatibility to queries with at least one fully connected node (*central node*).

PARSEC [1] is selected as a baseline for our computational experiments. It clearly showed that DFS-based techniques are superior to the traditional BFS approaches on GPUs. However, it performs redundant work due to the unavailability of intermediate candidates and multiple processing units working on the same subtree. The parallelism schemes used by PARSEC also have severe load imbalance that hampers its performance. We utilize the DFS framework developed by PARSEC along with different techniques from the traditional sequential and modern parallel algorithms to further improve its performance.

CHAPTER 4

BASIC TECHNIQUES

The Subgraph Enumeration algorithm can be decomposed into three major steps: (1) Query Graph Preprocessing, (2) Data Graph Preprocessing, and (3) Search Tree Traversal. This chapter covers each step in detail with a running example.

4.1 Query Graph Preprocessing

This section explains all processing done on the query graph to enable it for search tree traversal. *Query sequencing* and *Symmetry Detection* are the main tasks performed on the input query. Since the query graphs are very small and all preprocessing tasks are polynomial time complexity, this preprocessing workload is handled by the CPU. Query preprocessing is important to shrink the search tree exploration, the amount of work done at each node, and redundancy elimination.

4.1.1 Query Sequencing

The query graph input to the application is undirected. The first task is to convert this undirected graph to a directed acyclic graph (DAG). The ordering of the DAG governs the order in which these vertices are matched to the data graph. The query nodes are ordered based on their likelihood of matching with the data graph. Naturally, query nodes with lesser likelihood are prioritized over others.

The query sequencing algorithm used by VF3 [28] is utilized for this task. This algorithm uses multiple criteria to estimate the likelihood of matching and sequences the nodes in decreasing order of that likelihood. Interested readers are encouraged to read about these criteria in [28]. The pseudocode is

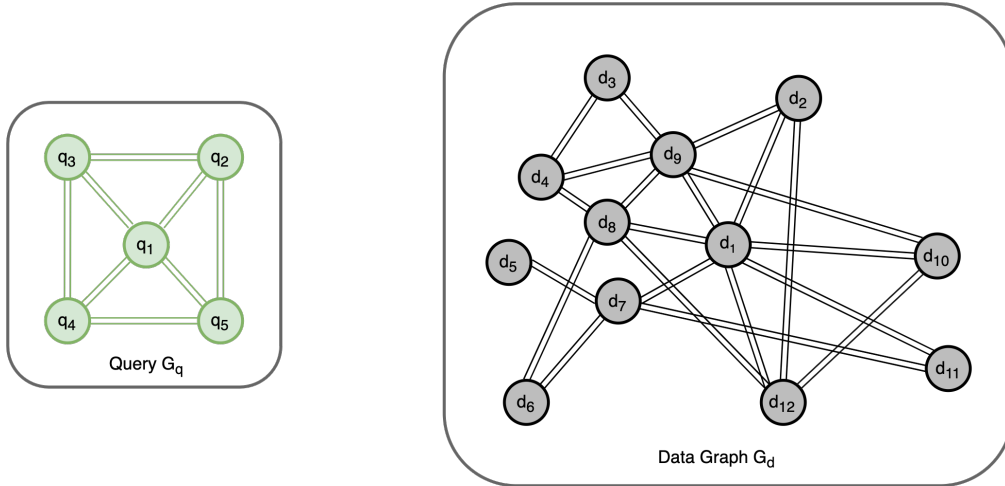


Figure 4.1: Example

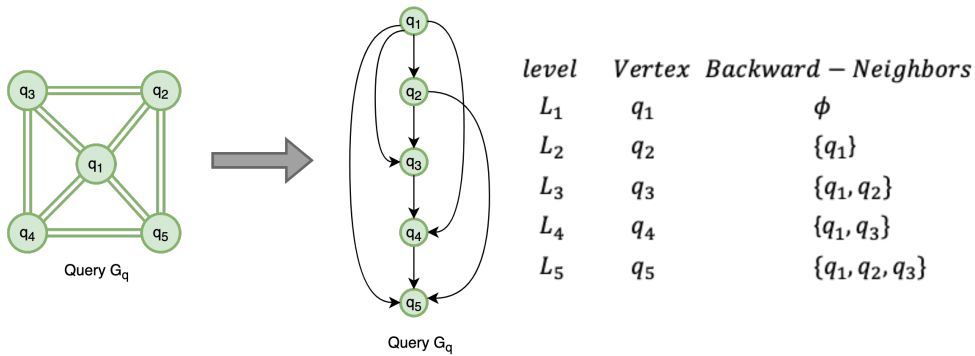


Figure 4.2: Query Sequencing Output of G_q using $VF3$

listed in Algorithm 1. The order generated after performing query sequencing on the do graph is shown in Figure 4.2.

4.1.2 Symmetry Detection and Breaking

Symmetry is the existence of automorphism in a graph. Automorphism, as the name suggests, is a graph that is isomorphic to itself. An automorphism of a graph is a permutation of vertices that maintains edges and non-edges. The set of automorphic permutations of the vertex set is called the *automorphism group*. Figure 4.3 shows the automorphism group of the query graph G_q . *Orbit* is an equivalence class of vertices in the automorphism group.

The presence of symmetry essentially causes the same subgraph of the data graph to be counted multiple times which results in extra work. The easiest

Algorithm 1: Query Sequencing

Input: Template Graph, G_q
Output: Node Query Sequence, S_q

```
1  $d_M[N_q] \leftarrow 0$ ;  
2  $S_q \leftarrow \phi$ ;  
3 for  $i \leftarrow 0$  to  $N_q$  do  
    // calculate likelihood for remaining nodes  
4   for  $j \leftarrow 0$  to  $N_q$  do  
5     if  $j$  not in  $S_q$  then  
6       likelihood[ $j$ ]  $\leftarrow d_M[j] \times N_q + \text{Degree}[j]$ ;  
7     end  
8   end  
   // Find node with maximum likelihood and it to  $S_q$   
9    $\text{idx} \leftarrow \text{argmax}(\text{likelihood})$ ;  
10  append  $\text{idx}$  to  $S_q$ ;  
   // Update  $d_M$   
11  foreach neighbor  $j$  of node  $\text{idx}$  do  
12     $d_M[j] \leftarrow d_M[j] + 1$ ;  
13  end  
14 end
```

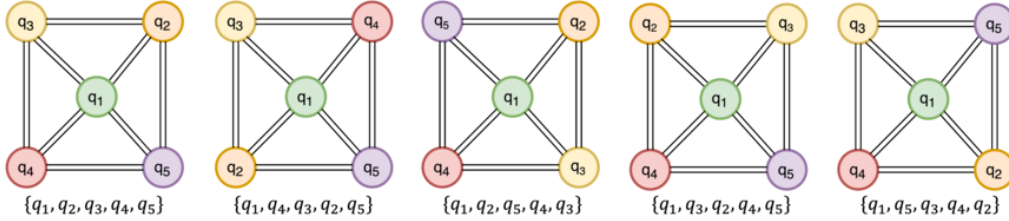


Figure 4.3: Automorphism Group of G_q

technique that can be employed to avoid this is called symmetry breaking. This makes sure that the ties are resolved at the moment they are generated, hence avoiding redundant work. An easy way to break symmetry is by having an order between the matches of data graph between the symmetric levels, which was introduced by [29]. Algorithm 2 lists the pseudocode for detecting symmetry and generating ordering for symmetry based pruning. Note, the `GetAutomorphismGroup` and `GetOrbits` routines used here are provided by NAUTY [39].

For the query graph in Figure 4.1, to avoid redundancy due to automorphism $\{q_1, q_4, q_3, q_2, q_5\}$ of G_q , it can be made sure that the data graph vertices matched at level 2 have labels less than the vertices matched at level 4. This

Algorithm 2: Symmetry breaking

Input: Template Graph, G_q
Output: Set of partial node orderings, \mathcal{P}_q

- 1 $\mathcal{A} \leftarrow \text{GetAutomorphismGroup}(G_q)$;
- 2 **while** $\mathcal{A} \neq \mathbb{I}$ **do**
- 3 $\mathcal{O}_{\mathcal{A}} \leftarrow \text{GetOrbits}(\mathcal{A})$;
- 4 Pick largest orbit $\{v_1, v_2, \dots, v_k\}$ from $\mathcal{O}_{\mathcal{A}}$;
- 5 Add $v_1 < v_2, v_1 < v_3, \dots, v_1 < v_k$ to \mathcal{P}_q ;
- 6 $\mathcal{A} \leftarrow \{\alpha \mid \alpha \in \mathcal{A}, \alpha(v_1) = v_1\}$;
- 7 **end**

is an example of *lexicographic symmetry* breaking. Another symmetry breaking order can be based on the degree of data graph nodes. A key observation is that the symmetry breaking criteria can be different for different levels but not always different for different subtrees. This will be formally established and utilized in Section 5.2.

4.2 Data Graph Preprocessing

Data graph preprocessing involves operations on the data graph to shrink the search tree during traversal. With GPU implementations, preprocessing also helps to improve coalesced memory accesses and memory utilization.

4.2.1 Peeling

Subgraph Enumeration produces isomorphic matches of the query graph in the data graph. Naturally, each match needs to have degree greater than the degree of the corresponding query node. This implies that: *Vertices in data graph with degree less than minimum query degree will never form any matches.* Using this simple observation, all vertices with degree lesser than the minimum query degree can be deleted from the data graph. This can be done iteratively as the observation holds for the resulting data graph. This process is known as *Peeling*. Empirical experiments conducted by [1] show that *Peeling* should be performed till at most 5% of nodes are deleted. Algorithm 3 shows the peeling routine.

Algorithm 3: Peeling data graph

Input: Minimum query graph degree, $d_{q,min}$, $G_d = (V_d, E_d)$:

```
1 repeat
2   PrunedVertices  $\leftarrow \{u | u \in V_d, \text{Degree}(u) < d_{q,min}\}$ ;
3   if  $|\text{PrunedVertices}| = 0$  then
4     Break;
5   end
6   PrunedEdges  $\leftarrow \{(u, v) | u \text{ or } v \in \text{PrunedVertices}\}$ ;
7    $V_d \leftarrow V_d \setminus \text{PrunedVertices}$ ;
8    $E_d \leftarrow E_d \setminus \text{PrunedEdges}$ ;
9 until  $|\text{PrunedVertices}| < 0.05 \times |V_d|$ ;
```

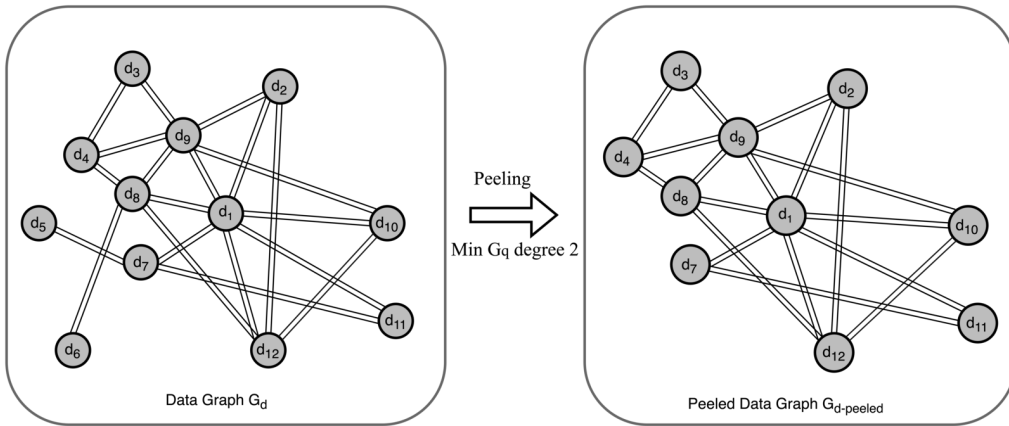


Figure 4.4: Peeling Example

4.2.2 Priority Sorting

CSR/COO storage format allows efficient storage of graphs in CPU memory. As mentioned in Section 2.2, the *Column Index* array is sorted by vertex labels to enable binary search. However, efficient symmetry breaking demands the neighbors be listed in a different order. To achieve this, we need another copy of the *Column Index* array in a *priority* sorted fashion while maintaining the *Row Pointer* boundaries. This array is referred to as *Priority Sorted Column Index* or *PSCI*. The *priority* can be any property of the vertices; in our case two properties are explored: Degree and Degeneracy. Given the size of data graph, GPUs are used for this preprocessing task. The *PSCI* array can be generated using a series of two stable sorts. Sorting can be efficiently executed on GPUs using the CUB library [40]. To start with, the *PSCI* array is taken as a copy of the original *Column Index* array. An array of triplets is

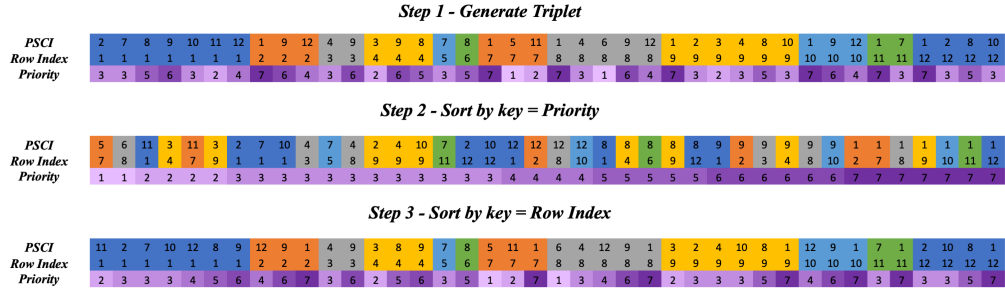


Figure 4.5: Generating Priority Sorted Column Index Array

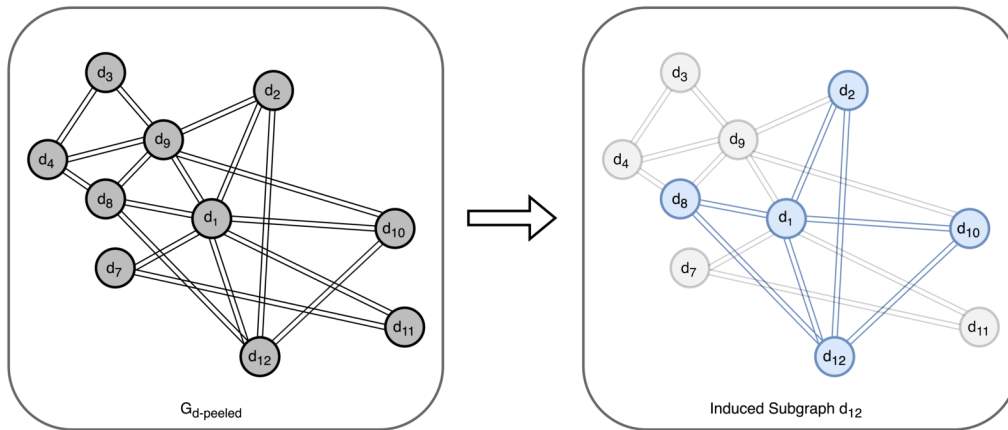


Figure 4.6: Induced Subgraph for d_{12}

created with each entry containing an element from *PSCI*, *Row Index*, and *priority*. Here, *priority* is an array populated with the entries corresponding to the required vertex property in *PSCI*. This array of triplets is subjected to two key-based stable sorts, the first sort is with the keys being entries in the *priority* array while the second sort is with the keys being entries in the *Row Index* array. Figure 4.5 shows the priority sorting performed on G_d as a series of key sorts.

4.2.3 Induced Subgraph

Search space in the subgraph enumeration problem is essentially the whole graph for a general query. This makes the problem extremely challenging to scale. However, most practical applications of subgraph isomorphism focus on dense templates. As explored in [41], restricting the search space can provide immense performance improvements. The search space can be re-

stricted by focusing on query graphs containing at least one *central node*. A central node is defined as a node in the query graph that is connected to all other nodes. Under the assumption of a central node, the search space for each match can be limited to the subgraph induced by the data graph vertex matched to the central node. Figure 4.6 shows the subgraph induced by vertex d_{12} . To save storage space and perform fast adjacency list intersections, the induced subgraph is stored in a binary adjacency matrix format. The columns of the adjacency matrix are ordered according to the *Priority Sorted Column Index* array to enable efficient symmetry breaking.

4.3 Search Tree Traversal

This section explains how the sequenced query graph G_q and processed data graph G_d are used to perform the search tree traversal for enumerating all matches.

Algorithm 4 describes the DFS-based iterative search tree traversal. The whole traversal can work within a DFS stack of size $|V_q| \times Degree_{max}(G_d)$. `GetNextNodes` function is called at each node of the search tree which generates the candidates to be visited at the next level. This function is listed in Algorithm 5. It performs an adjacency list intersection operation to generate all possible candidates for the next level (lines 2 - 5). The next loop, (lines 6-8) prunes redundant candidates which are generated due to the symmetry in G_q . Note that, the \mathcal{V}_{orient} set used in Algorithm 5 is directly available from the binary adjacency matrix generated in Section 4.2.3.

We now walk through the search tree traversal algorithm with the example query graph sequence from Figure 4.2 and the peeled data graph in Figure 4.4. Note, the pseudocode in Algorithm 4 describes DFS-based traversal while the example enumerates search tree traversal in BFS manner for illustration purposes.

Step 1: Select candidates in G_d with degree greater than or equal to q_1 . In this example, vertices $\{d_1, d_8, d_9, d_{12}\}$ will be selected. All of them are matched to q_1 and traversed separately. For conciseness, only d_1 and d_{12} are considered. The matches are shown in Figure 4.7.

Step 2: For candidates at level 2, select all neighbors of the vertex matched

Algorithm 4: DFS Traversal

Input: Sequenced Query Graph: $G_q = (V_q, E_q), S_q$
Induced Subgraph: G_{ind}^d . Initial level: l_{init}
Initial set of Nodes at l_{init} : $matches_{init}$ and size $num_matches_{init}$

Output: Set of enumerated matches and count: $final_enum, count$

```
1  $l \leftarrow l_{init}, matches[l] \leftarrow matches_{init}, num\_matches[l] \leftarrow num\_matches_{init};$ 
2  $curr\_idx \leftarrow 0, final\_enum \leftarrow \phi, count \leftarrow 0;$ 
3 while  $curr\_idx[l] \leq num\_matches[l]$  do
4   GenNextCandidates ();
5    $num\_matches[l + 1] = |matches[l + 1]|;$ 
6   if ( $num\_matches > 0$  and  $l < |V_q| - 1$ ) then
7      $l \leftarrow l + 1;$ 
8      $curr\_idx[l] \leftarrow 0;$ 
9   end
10  else if  $l == |V_q| - 1$  then
11    for  $n = l_{init}$  to  $num\_matches[l + 1]$  do
12       $match \leftarrow \phi;$ 
13      for  $k = l_{init}$  to  $l$  do
14         $match \leftarrow match \cup matches[k][curr\_idx[k]]$ 
15      end
16       $final\_enum \leftarrow final\_enum \cup match$ 
17    end
18     $count \leftarrow count + num\_matches[l + 1]$ 
19  end
20  else
21     $curr\_idx[l] \leftarrow curr\_idx[l] + 1;$ 
22    while  $curr\_idx[l] == num\_matches[l]$  and  $l > l_{init}$  do
23       $l \leftarrow l - 1;$ 
24       $currIdx[l] \leftarrow curr\_idx[l] + 1;$ 
25    end
26  end
27 end
```



Figure 4.7: Step 1

Algorithm 5: Generate Candidates for Next Level

Input: Induced Subgraph G_{ind}^d , level l
 Backward Neighbor List \mathcal{N}'
 Symmetry levels List \mathcal{S}
 Set of vertices of lower priority than given vertex $\mathcal{V}_{orient}()$
 List of matches in previous levels `matches[]`

```

1 Function GenNextCandidates()
  while (curr_idx[l] < num_matches[l]) do
2   forall  $k \in \mathcal{N}'(S_q[l+1])$  do
3      $u \leftarrow \text{matches}[k][\text{curr\_idx}[k]]$ ;
4      $\text{matches}[l+1] \leftarrow \text{matches}[l+1] \cup \mathcal{N}(G_{ind}^d(u))$ ;
5   end
6   forall  $j \in \mathcal{S}(S_q[l+1])$  do
7      $u \leftarrow \text{matches}[j]$ ;
8      $\text{matches}[l+1] \leftarrow \text{matches}[l+1] \setminus \mathcal{V}_{orient}(u)$ ;
9   end
10 end
11 Function End
  
```

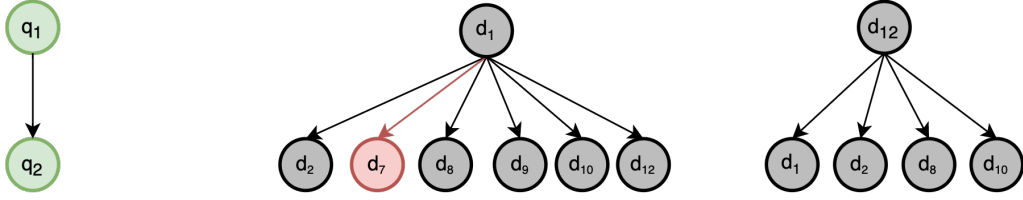


Figure 4.8: Step 2

at level 1 with degree greater than $d(q_2)$, i.e., 3. As shown in Figure 4.8, there will be 6 candidates: $\{d_2, d_7, d_8, d_9, d_{10}, d_{12}\}$ in the subtree rooted at d_1 . d_7 is pruned out of these 6 candidates since it has degree less than q_2 .

For the subtree rooted at d_{12} , there will be 4 candidates: $\{d_1, d_2, d_8, d_{10}\}$ and none can be pruned.

Step 3: Level 3 candidates are obtained by the adjacency list intersection of matches at level 2 and level 1. Since this level is symmetric to level 2, symmetry breaking needs to be performed. Decreasing degree criteria for symmetry breaking is used in this example. Figure 4.9 shows all candidates at level 3. The candidates marked in red are pruned. For example, d_9 is pruned due to symmetry breaking in

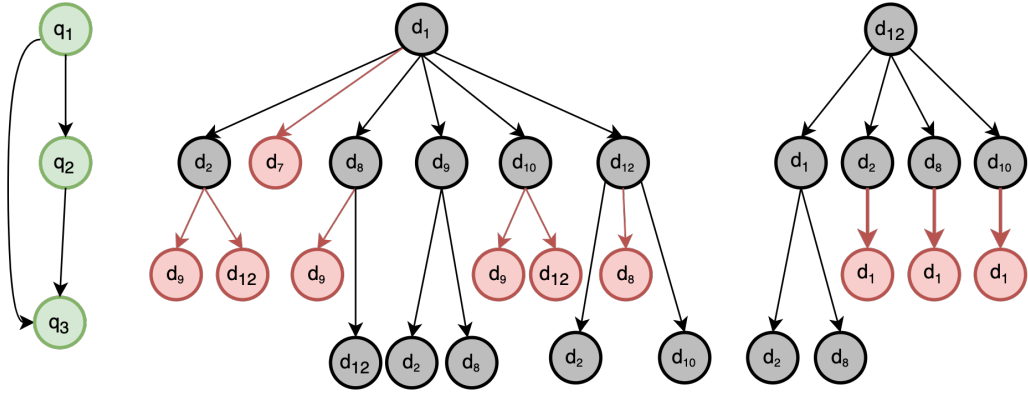


Figure 4.9: Step 3

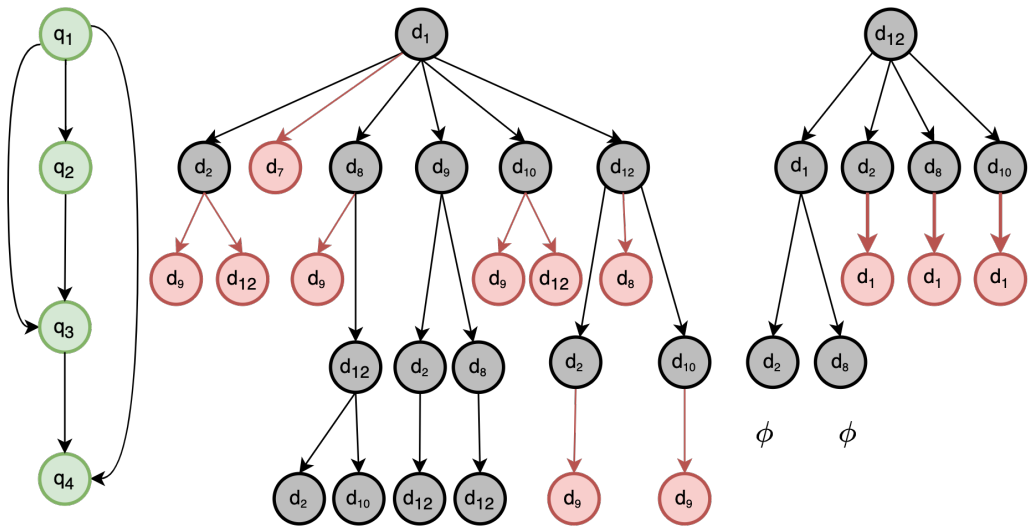


Figure 4.10: Step 4

the subtree (d_1, d_8, \dots) since $Degree(d_9) > Degree(d_8)$. Ties are resolved using the lexicographic criterion.

Step 4: For level 4 candidates, the adjacency list intersection of vertices matched at q_1 and q_3 is required. This level is also symmetric to level 2. As illustrated in Figure 4.10, the subtree rooted at d_{12} does not yield any candidates.

The other subtree gets two candidates for $(d_1, d_8, d_{12}, \dots)$. and one candidate each for the remaining subtrees. d_9 is pruned due to symmetry breaking.

Step 5: Level 5 candidates (or final matches) are generated by the adjacency list intersection of matches at q_1 , q_2 , and q_4 . This level is symmet-

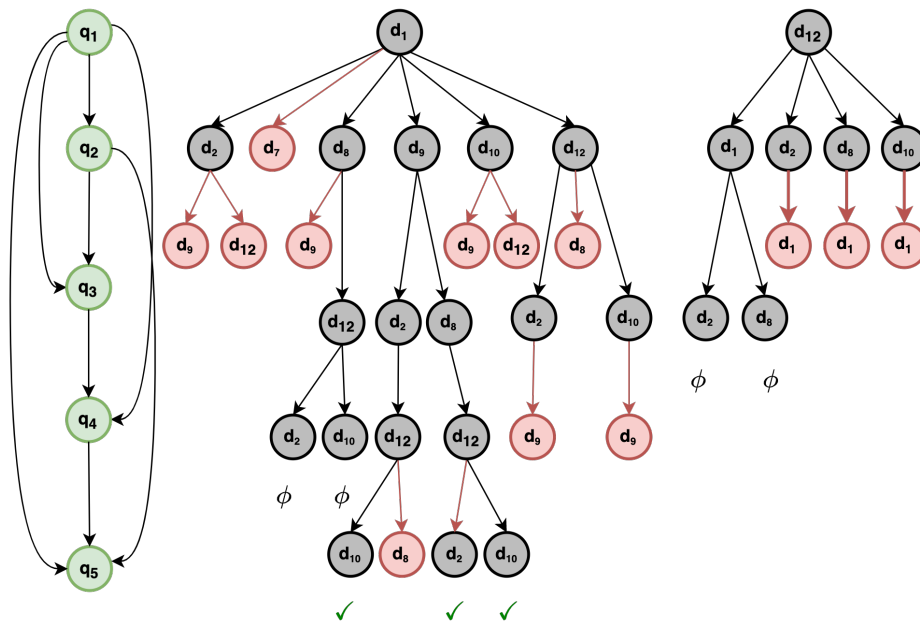


Figure 4.11: Step 5

ric to both q_2 and q_3 . The candidates generated by intersection at the first leg are $\{d_8, d_{10}\}$, symmetry breaking will eliminate d_8 as $Degree(d_8) > Degree(d_2)$. There is a tie between d_2 and d_{10} which is resolved lexicographically. Similarly, the other leg generates two more candidates for level 5.

This way the search tree traversal process terminates with final matches $\{d_1, d_9, d_2, d_{12}, d_{10}\}$, $\{d_1, d_9, d_8, d_{12}, d_2\}$, and $\{d_1, d_9, d_8, d_{12}, d_{10}\}$. There might be other matches originating from the subtrees rooted at $\{d_8, d_9\}$ which were not illustrated here.

To conclude, this chapter provides a detailed understanding of all techniques involved in subgraph enumeration. As we can see, all subtrees are independent of each other therefore they can be easily parallelized on GPUs.

One subtle observation to be made here is that the exploration is dependent on the symmetry breaking criteria. We can see from Figure 4.10 that there would be many more candidates at level 4 if symmetry breaking was performed using lexicographic or increasing degree criterion. These candidates would have ultimately pruned in lower levels, resulting in extra work. Therefore, selecting an efficient symmetry-breaking strategy is important; it is thoroughly discussed in Section 5.2.

CHAPTER 5

IMPROVEMENTS

Multiple ideas were explored to improve the performance of PARSEC [1]. This chapter describes each idea and its individual impact on overall performance. All improvements are then combined to present the final results in Chapter 6.

The criteria used for quantifying the contributions of these improvements are the number of intersections performed and the total running time. These measures are reported as speedups. Formally, we define the intersection speedup and time speedup as:

$$\text{Intersection Speedup} = \frac{\# \text{ of Intersections Baseline}}{\# \text{ of Intersections New}}$$

$$\text{Time Speedup} = \frac{\text{Run time Baseline}}{\text{Run time New}}$$

The information of data graphs and queries used to report individual performance impact is mentioned in Section 6.1.

5.1 Intersection Reuse

Adjacency list intersection operation for generating next candidates (Algorithm 5) is the most time-consuming operation in subgraph enumeration. This is a well-established fact in the literature [37], [34], [28] and it was also verified for the baseline [1]. Figure 5.1 presents the distribution of time spent by each operation in different tasks. The Intersect operation shown here is the time taken by Algorithm 5. This algorithm iterates over each element in the backward neighbors list to perform adjacency list intersection and generates possible candidates for next level (lines 2 - 5). The size of the backward neighbors list increases for larger templates, which results in even

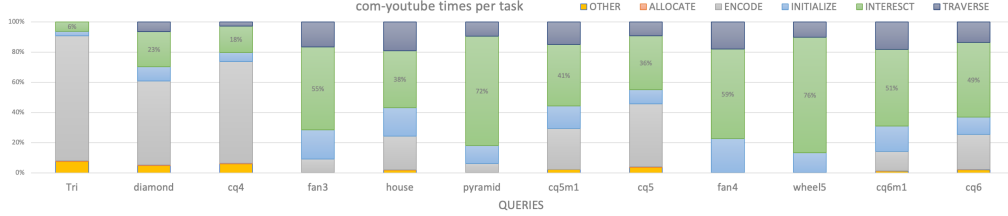


Figure 5.1: Stepwise Fraction Time Spent

more adjacency list intersections at each node.

RPS [37] reduces these operations by generating an intersection reuse plan which smartly finds the intersections that will be required by more than one node and stores them after first computation. The authors report huge scope for intersection reuse through this plan. However, most of the reuse opportunities arise due to BFS traversal. A similar extent of reuse is not possible with DFS traversal since the information is lost while backtracking. However, Some levels of the sequenced query graph have similar backward neighbor lists. For such queries, some intersections can be reused by simply storing the intermediate intersection results. To do this efficiently, one needs to match levels with similar adjacency lists. This problem is posed as a linear programming optimization problem.

Let the sequenced query graph be $G_q = (V_q, E_q)$ with $|V_q| = k$ and the match order be S_q . Let $\mathcal{N}(\cdot)$ represent the backward neighbors list of a vertex. For each vertex pair $S_q(i), S_q(j)$, let W_{ij} be a measure of the commonality between their backward neighbors. Let X_{ij} be the variable which decides if vertex i should reuse intersections from vertex j .

With these definitions the optimization problem can be modelled as:

$$\max \sum_{i=j+1}^k \sum_{j=1}^k W_{ij} X_{ij} \quad (5.1)$$

$$\text{s.t. } \sum_{j=1}^k X_{ij} \leq 1 \quad \forall i \in \{1, \dots, k\} \quad (5.2)$$

Where,

$$W_{ij} = \begin{cases} |\mathcal{N}(S_q[i]) \cap \mathcal{N}(S_q[j])| & \text{if } i > j, \mathcal{N}(S_q[i]) \supseteq \mathcal{N}(S_q[j]), \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

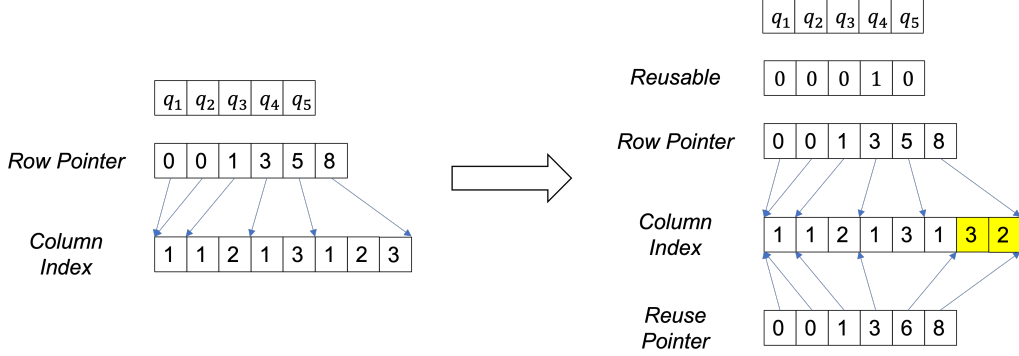


Figure 5.2: Reuse Detection Example

The maximization problem (5.1 - 5.2) is a Linear Semi-Assignment Problem (LSAP) for which the greedy solution is optimal. Hence, the solution is:

$$X_{ij} = \begin{cases} 1 & \text{if } j = \operatorname{argmax}_j(w_{ij} > 0); \\ 0 & \text{Otherwise} \end{cases}$$

Reuse detection involves a maximum operation for each level in G_q hence it is polynomial time and can be executed on the CPU for small-sized query graphs. The mappings found ($x_{ij} = 1$) imply that level i can use intersection results from level j . Hence, level j is flagged as *reusable* and its intersection results are stored along with the DFS stack. Algorithm 6 lists the reuse detection and post-processing recipe. To save constant memory footprint, the query graphs are stored in CSR format. After finding the reuse mappings, the column index of (G_q) needs to be modified for efficient search tree traversal. The modification step involves reordering the adjacency lists and generating the reuse pointer array.

We show reuse detection at work on the example query in Figure 4.1. Here, only w_{54} and w_{53} are positive with value 2. Hence, one from levels 3 or 4 can be marked reusable, let's assume level 4. The backward neighbors of level 4 are $\{1, 3\}$ and level 5 are $\{1, 2, 3\}$. A reuse pointer array, that points to the start of the remaining neighbors of level 5 is created to avoid extra checks in Algorithm 5. Once the reuse is detected and pointers are set, the data is sent to the GPU constant memory for the rest of the application. For implementing reuse, the function in Algorithm 5 needs to be changed to add conditions for checking if a level is reusable or if the reuse pointer is updated. These changes are shown as a flow chart in Figure 5.3. This way,

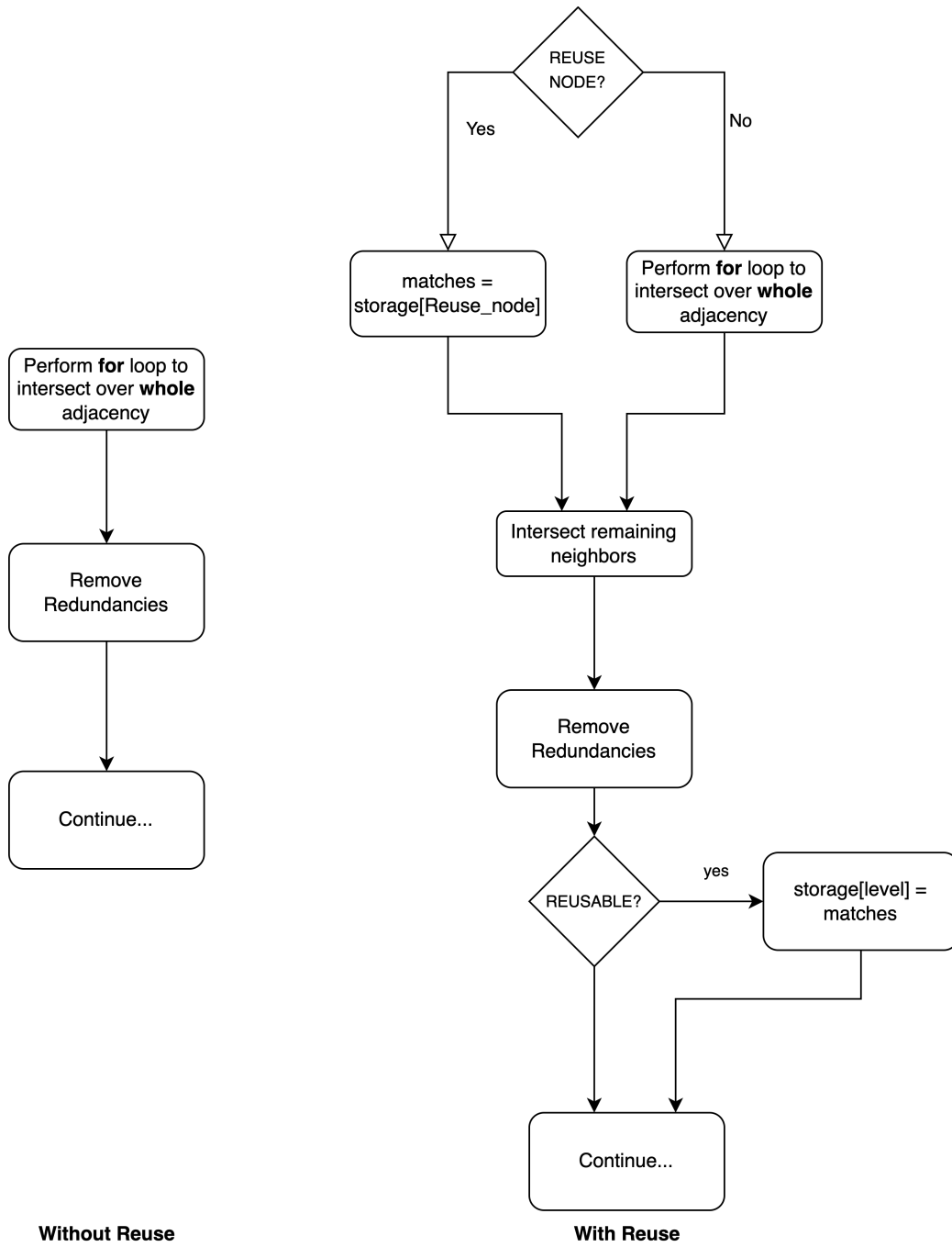


Figure 5.3: Flowchart With and Without Reuse

reuse reduces computational workload and adds to the memory workload. Since global memory writes are relatively expensive, some shared memory can be utilized to write the intermediate intersection results while longer results can be spilled to global memory.

Algorithm 6: Reuse Detection

Input: Directed Query Graph $G_q = (V_q, E_q)$
 Array of Backward Neighbor lists $\mathcal{N}(V_q)$

```

1  $W \leftarrow: [|V_q|][|V_q|];$ 
2  $X \leftarrow: [|V_q|][|V_q|];$ 
3 reusable  $\leftarrow: [|V_q|];$ 
4 forall  $j \in V_q$  do
5   | for  $i = j + 1$  to  $|V_q|$  do
6   |   | if  $\mathcal{N}(V_q)[i] \supseteq \mathcal{N}(V_q)[j]$  then  $W[i][j] \leftarrow |\mathcal{N}(V_q)[i] \cap \mathcal{N}(V_q)[j]|;$ 
7   |   | else  $W[i][j] \leftarrow 0;$ 
8   |   | end
9   | end
10 reusable[ $j$ ]  $\leftarrow (\max(W[i].) > 0);$ 
11 forall  $i \in V_q$  do
12   | for  $j = 1$  to  $i$  do
13   |   | if  $j == \operatorname{argmax}_j(W[i]. > 0)$  then 1;
14   |   | reusable[ $j$ ] = 1;
15   |   | else 0;
16   |   | end
17 end
18 // Update Reuse Pointers from  $X[i][j]$ 

```

The improvements due to intersection reuse are measured using intersection speedup. Since there are increased memory accesses and additional *if-else* conditions, the time benefit from this technique is negligible. Table 5.1 shows the improvement in the number of intersections with this technique. The intersection counts reported here also include additional read and write operations required by intersection reuse. Since the backward neighbors lists are small for sparse queries, the benefits due to intersection reuse are limited.

5.2 Hybrid Symmetry Breaking

Symmetry breaking is an important operation in the search tree traversal. [16] gives a simple ordering technique for symmetry breaking, but the order-

Table 5.1: Effect of reuse on number of intersections for com-youtube

Query	#Intersections w/o Reuse	#Intersections with Reuse	Intersection Speedup
diamond	9,722,143	9,722,143	1.00
cq4	6,028,832	6,028,832	1.00
cq5_m1	46,794,486	26,360,303	1.78
cq5	20,881,838	15,889,824	1.31
cq6_m1	130,114,516	52,225,436	2.49
cq6	49,618,729	30,194,042	1.64
cq7_m1	253,215,589	81,007,445	3.13
cq7	91,679,280	46,935,236	1.95
cq8_m1	370,669,539	102,639,103	3.61
cq8	139,279,024	62,709,660	2.22
cq9_m1	430,334,955	111,122,133	3.87
cq9	181,479,780	74,663,152	2.43

ing criteria are not well-defined. Any criterion can be used for correctness, but efficient criteria can significantly improve performance due to slimmer search tree. Handling symmetries to reduce the search space is an active area of research. The problem of finding an efficient symmetry breaking criteria was shown to be strongly NP-hard by [42]. A lot of symmetry breaking criteria have been defined since then. However, most of them have linear time complexity or need BFS traversal. In the optimization domain, this problem has also been formulated as a Constraint Satisfaction Problem (CSP) but it also works only when the tree traversal is done in BFS fashion [43]. The near-linear time complexity for each decision, constraints to perform BFS, and data sharing between subtrees make efficient symmetry breaking difficult on GPUs. To avert these, we resort to static characteristics like Degree and Degeneracy. The structure generated due to priority sorted column index array (PSCI) (discussed in Section 4.2.2) is used to further reduce global memory accesses.

With PSCI, different symmetry breaking criteria were tried and compared to the baseline [1]. The baseline criteria are based on degree for level 1 and lexicographic for level 2 onwards. The symmetry breaking performance is determined by intersection speedup. On performing these experiments, the degeneracy-based symmetry breaking seemed to perform worse than baseline across all graphs. For degree, increasing criterion perform better for some queries while decreasing criterion performed better for others. Table 5.2

shows the intersection speedups for *soc-pokec* and *cit-patents* data graphs.

Table 5.2: Intersection speedup with degree based symmetry breaking

soc-pokec	Tri	Diamond	Cq4	Cq5m1	Cq5	House	Pyramid	Fan3	Cq6m1	Cq6
Decreasing	1.22	0.53	0.66	0.89	1.19	1.49	1.88	1.76	1.06	1.41
Increasing	3.00	1.08	1.47	1.67	2.29	2.85	0.22	0.20	1.64	2.25
cit-patents	Tri	Diamond	cq4	cq5m1	cq5	house	pyramid	fan3	cq6m1	cq6
Decreasing	1.22	0.82	1.24	2.32	2.91	4.29	3.15	3.92	2.51	3.15
Increasing	3.00	1.27	1.79	3.50	4.28	6.60	1.26	1.83	2.72	3.71

Recall from 4.1.2 that the symmetry breaking criteria can be different for different levels and maybe different for different subtrees. We present the following observation to precisely find out when the criteria can be different across subtrees.

Theorem 1. *Symmetry breaking is independent across subtrees for central node queries with first symmetric level greater than two.*

Proof. Let, the data graph be $G = (V, E)$ and $\{v_1, v_2\} \in V$.

$G_{ind}(v)$ be the subgraph induced by v in G .

The backward neighbors list of a vertex is represented by $\mathcal{N}'(\cdot)$.

Consider $X := \mathcal{E}(G_{ind}(v_1)) \cap \mathcal{E}(G_{ind}(v_2))$

Assume, X to be non-empty. (proof is trivial otherwise)

\forall edges $(x, y) \in X$: Since, $(x, y) \in \mathcal{E}(G_{ind}(v_1)) \Rightarrow x, y \in \mathcal{N}(v_1)$

Similarly, $x, y \in \mathcal{N}(v_2)$

Now for Algorithm 4, since level 2 vertex (say q_2) in the query graph is not symmetric to the level 1 vertex (say q_1), candidates for level 2 will be given by adjacency list intersections of matches in $\mathcal{N}'(q_2)$.

Since, q_1 has to be a central node, $\mathcal{N}'(q_2) = q_1 \Rightarrow$ level 2 candidates for tree rooted at v_1 will be $\mathcal{N}(v_1)$ and for tree rooted at v_2 they will be $\mathcal{N}(v_2)$. Therefore, all edges in X are added to both search trees.

This implies, each subtree will enumerate the correct number of query instances independent of the symmetry breaking strategy. \square

Theorem 1 enables individual symmetry breaking criteria for subtrees. Table 5.2 shows the best strategy depends on the data graph as well as query graph. The optimal strategy may even be different for distinct subtrees in the same data graph.

To curb the variability in strategy decisions, a dynamic two-phase traversal technique is designed: The first phase is to determine if a particular subtree should perform increasing or decreasing symmetry breaking. The second phase is to use the predetermined strategy to execute the search tree traversal.

For the first phase, since the problem of finding an optimal symmetry breaking criterion is strongly NP-hard, any practical implementation can only develop Heuristics. A natural strategy to decide the criteria in phase one might seem to be based on counting the number of eligible candidates till the first symmetric level. This approach does not always work as both strategies might produce equal number of candidates at the first symmetric level. Also, fewer candidates at first symmetric level does not necessarily reflect in a smaller search tree. This hints at designing a criterion to determine the *quality* of these candidates. Since the candidates generated are sorted by priority, we use the following observations to determine their *quality*:

1. For increasing symmetry breaking, the left side of the DFS subtree has fewer candidates (vice versa for decreasing).
2. Few candidates on the left side may generate more nodes in subsequent levels, while the right side candidates (high count) might not.

So the weights should be selected based on an equalizing philosophy where higher weight is given to the left side of the tree (for increasing symmetry breaking). Increasing weights based on the index of candidates and the total number of candidates is used for this purpose.

To formalize, Let the total number of candidates at j^{th} leaf in the subtree be C_j and the number of candidates found in the previous level be S . The symmetry breaking criterion is determined based on the weighted sum (WS).

$$WS = \sum_j C_j \times W_j$$

The weightage W_j for the quality of this branch is given by:

$$W_j = \begin{cases} (S - j) & \text{for increasing and,} \\ j & \text{for decreasing.} \end{cases}$$

Table 5.3: Comparison of hybrid strategy with pure, best and worst possible strategies using intersection speedup criteria for data graph cit-patents

Query	Increasing	Decreasing	Hybrid	Best	Worst
Tri	3.00	2.20	3.00	3.00	2.20
Diamond	1.27	0.82	1.27	1.27	0.82
cq4	1.79	1.24	1.79	1.79	1.24
cq5m1	3.50	2.32	3.50	3.50	2.32
cq5	4.28	2.91	4.28	4.28	2.91
house	6.60	4.29	6.60	6.60	4.29
pyramid	1.26	2.45	2.47	2.53	1.24
fan3	1.83	3.92	3.96	4.07	1.80
cq6m1	2.72	2.51	2.72	2.72	2.51
wheel5	1.00	1.57	1.57	1.57	0.99
fan4	1.56	2.20	2.22	2.22	1.56
cq6	3.71	3.15	3.71	3.71	3.15

Table 5.4: Time speedups with hybrid Symmetry Breaking

Data Graph	Queries					
	Fan3			Pyramid		
	Time (s)		Speedup	Time (s)		Speedup
Baseline	Degree hybrid	Baseline		Degree hybrid		
soc-pokec	2.464	1.500	1.643	3.086	1.964	1.571
com-youtube	9.675	6.206	1.559	19.253	10.335	1.863
cit-patents	0.534	0.425	1.256	0.598	0.472	1.265
com-orkut	447.532	342.183	1.308	996.422	582.864	1.710
as-skitter	60.851	40.456	1.504	159.289	84.530	1.884

A criterion is picked for the subtree based on whichever weighted sum is smaller.

To find the efficiency of this strategy, we compare it with the best, worst, and pure strategies using intersection speedup. Table 5.3 shows that hybrid strategy often works better than pure strategies and is sometimes close to the best possible strategy. Table 5.4 gives the time speedups for *fan3* and *pyramid* across different data graphs. Note, these queries were specifically selected since their first symmetric level is greater than two. The reasons for intersection speedups not proportionally reflecting in time speedups are time for other operations, load imbalance, and two-phase strategy overheads.

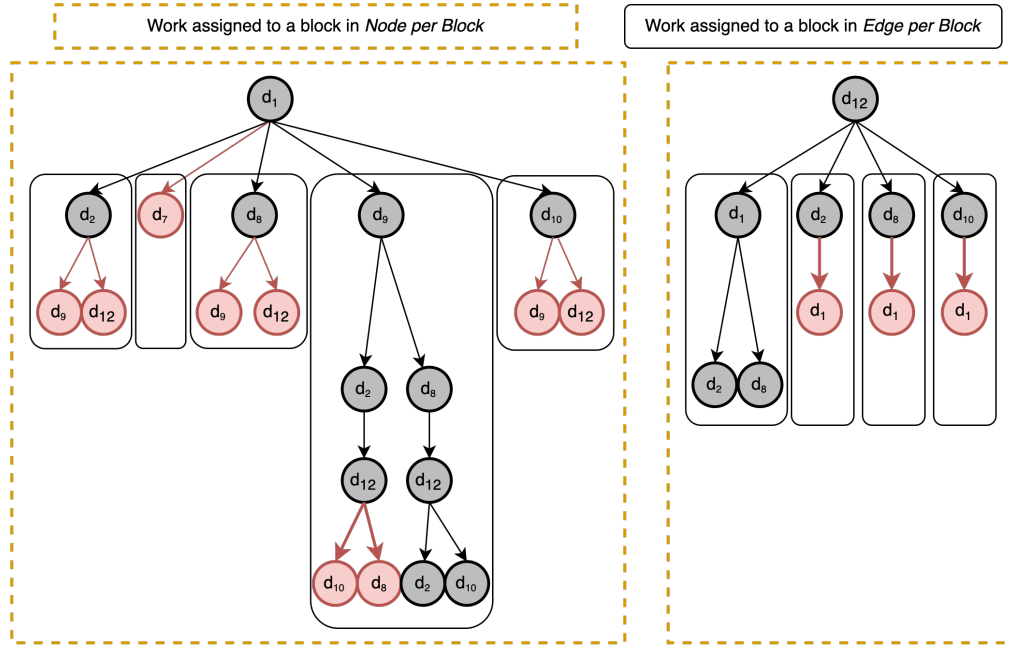


Figure 5.4: Parallelization Schemes

5.3 Hybrid Parallelism for Load Balance

As mentioned in Section 2.1.5 load balance is an important criterion while designing parallelization schemes. The baseline implementation [1] offered two types of parallelism *node per block* and *edge per block*. Figure 5.4 shows the difference between these parallelization schemes. A hybrid parallelism technique was implemented to reduce the load balance. Some insights from the baseline are presented for better understanding of this technique.

Observation 1 - Run time vs Degree

Figure 5.5 shows the relationship between degree of the root node and run time as a log-log plot. It is clear that run time is exponentially related to the root node degree. Moreover, there is huge variation in run time at the same degree.

Observation 2 - Load balance comparison

Given the trend between run time and root node degree, the node per block scheme is expected to show a huge load imbalance. The load balance is

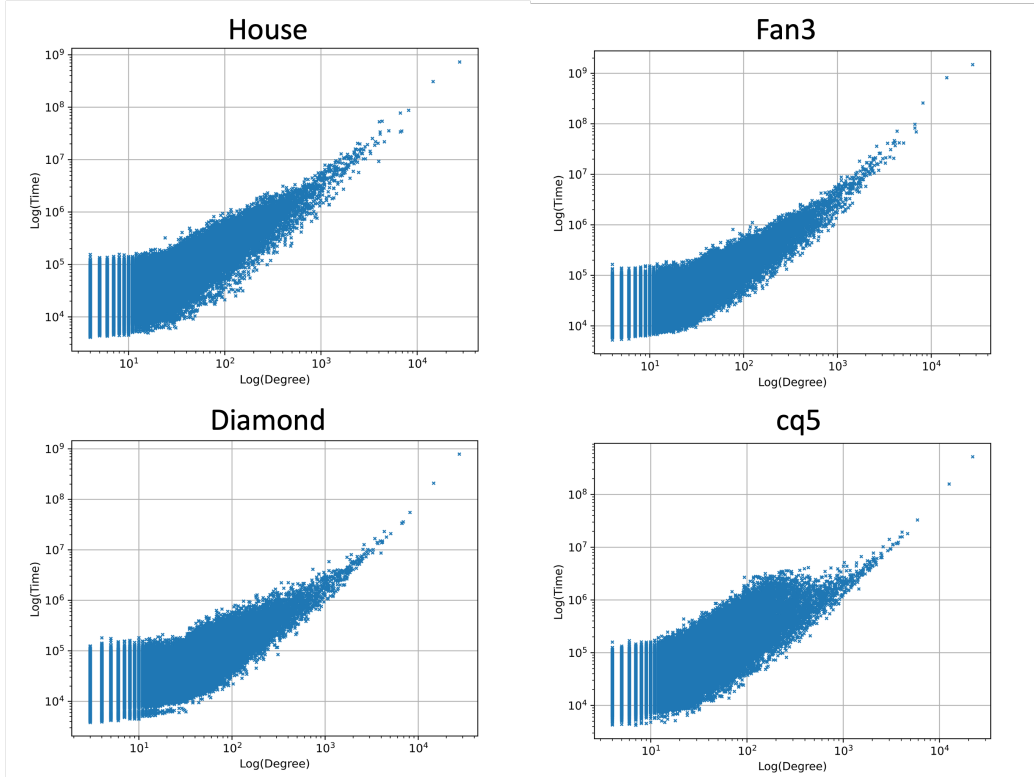


Figure 5.5: Degree vs Run Time for com-youtube

supposed to improve for the edge per block scheme as it would split the work offered by a high degree root node into several blocks. Figure 5.6 shows the load balance across SMs for various query graphs as a standard box plot. This is computed by finding the relative processing time spent by each SM during kernel lifetime. It is clear from the figure that the edge per block parallelization scheme significantly improves load balance.

Observation 3 - Load balance for low degree nodes

Since the trend between degree and run time is exponential, the subtrees originating from nodes with low degrees will likely have lesser variation in run times. Figure 5.7 shows there is negligible load imbalance when processing low degree nodes.

The claim from the authors of [1] was edge per block scheme works better than node per block for queries of size greater than five. One reason for this claim is implied by the load balance observations above. However, it does not explain the slow performance of the edge per block kernel. The reason is

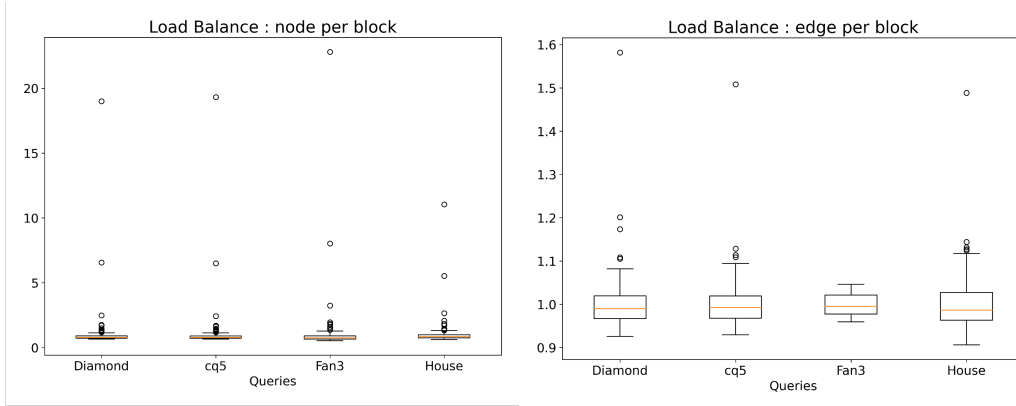


Figure 5.6: Load Balance for com-youtube

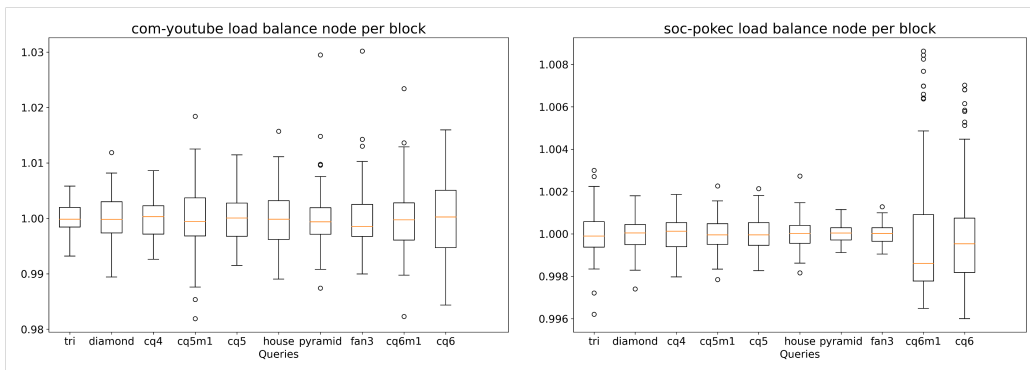


Figure 5.7: Load Balance for Nodes with Degree ≤ 256

strictly based on the implementation. For the edge per block scheme in [1], each block induces its own subgraph. This causes extra workload hence the edge per block implementation outperforms only for larger query graphs as they exhibit worse load balance.

To eliminate this redundant work, the kernel is split into two parts with different parallelization schemes. The first kernel ensures non-redundancy while inducing subgraphs by using the node per block scheme. The second kernel uses these induced subgraphs to perform search tree traversal using the edge per block scheme while ensuring better load balance. Since the kernels are split, the persistent memory technique needs to be removed. To minimize the total number of kernel launches and better resource utilization a runtime query is made to the device memory to find the number of nodes that can be simultaneously processed.

It is clear from Figure 5.7 that low degree node processing is fairly load balanced with the node per block parallelization scheme. Since splitting the ker-

nels causes poor caching performance and launch overheads, a cutoff scheme was developed. Nodes with degree less than cutoff would perform subgraph induction and search tree traversal in node per block fashion. Whereas nodes with degree higher than cutoff would perform these operations in separate kernels as mentioned above. Empirical testing was conducted to fix a cutoff value across different queries. It was found that for dense queries like *cqx* and *cqxm1* cutoff value between (1024 – 2048) is better while for sparse queries like *fans* and *wheels* cutoff value around 768 achieved best results. It was also observed that the run time is not too sensitive to cutoff. Figure 5.8 shows the relative run times across different cutoffs. Using these insights the cutoff was decided at runtime based on the input query.

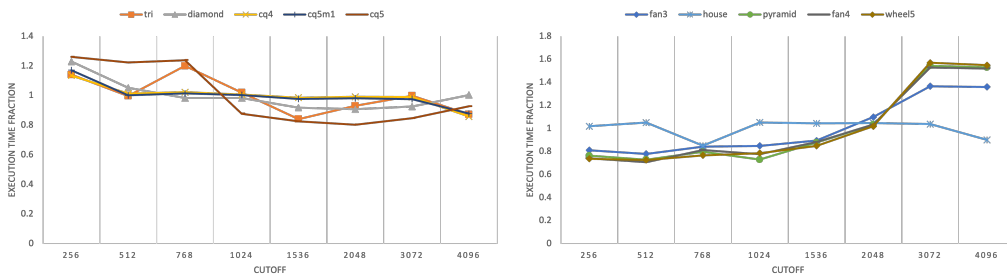


Figure 5.8: Run Time Fractions vs Cutoff for com-youtube

Figure 5.9 shows the improvement in runtimes for the com-youtube data graph across different queries.

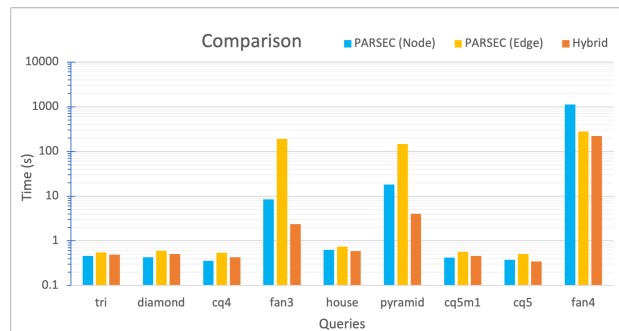


Figure 5.9: Run Times with Hybrid Parallelism for com-youtube

CHAPTER 6

RESULTS

6.1 Experimental Setup

All results presented in this chapter are generated on a supercomputing hardware cluster with a dual-socket 64-core AMD EPYC 7763 (“Milan”) CPU and NVIDIA A100 40GB HBM2 GPU. The system has 256 GB of DDR4 memory with 1.5 TB swap memory via high-performance-SSD. The program is compiled with NVCC (CUDA 11.7, driver version 515.48) SM version 80, and GCC 11.2 with the -O3 flag. For a fair comparison, the baseline is also run on the same hardware platform with equivalent compilation flags.

The data graphs used for experiments are obtained from the Stanford Network Analysis Project (SNAP) [44]. Table 6.1 summarizes the properties of these data graphs.

We use the common query graphs with central nodes that are generally used in subgraph enumeration literature. Figure 6.1 shows the query graphs used for experiments, these queries are adapted from [1]. *Clique* is a fully connected graph. As a notation in Figures and Tables, we use cqx for *clique* of size x and $cqxm1$ for size x *clique* with one edge missing. Note, removing exactly one edge from a *clique* gives a unique query due to symmetry.

Table 6.1: Data graphs used for experiments

Graph	 V 	 E 	Max Degree
as-skitter	1,696,415	11,095,298	35,455
soc-pokec	1,632,804	22,301,964	14,854
cit-patents	3,774,768	16,518,948	793
com-orkut	3,072,441	117,185,083	33,313
com-youtube	1,134,891	5,975,248	28,754

one parallelization scheme for comparison.

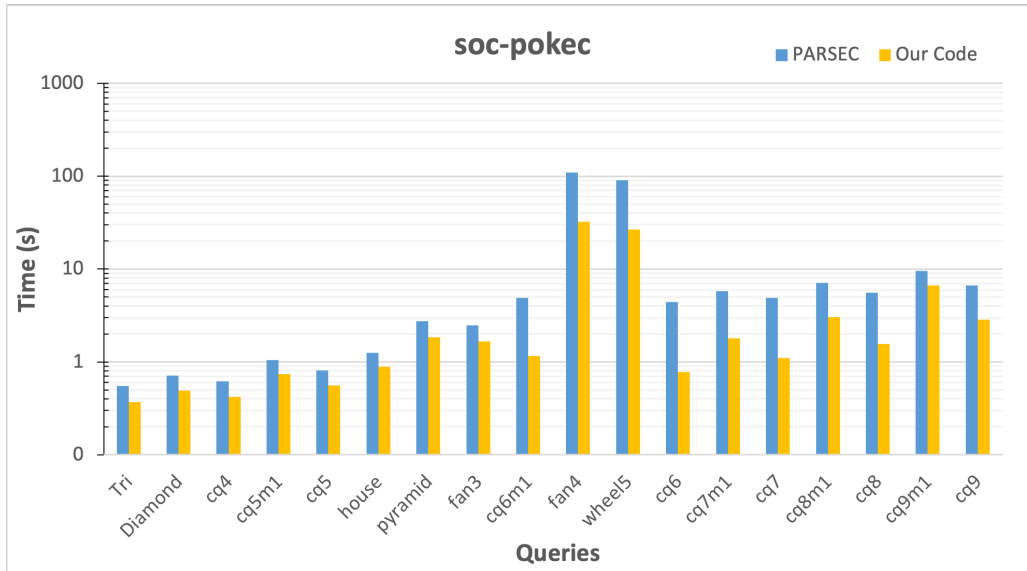
- Instances that take more than 12,000 seconds (~ 3.3 hours) are terminated, and their times are not reported. The corresponding speedup values are reported as blank in all tables.

6.3 Final Results and Analysis

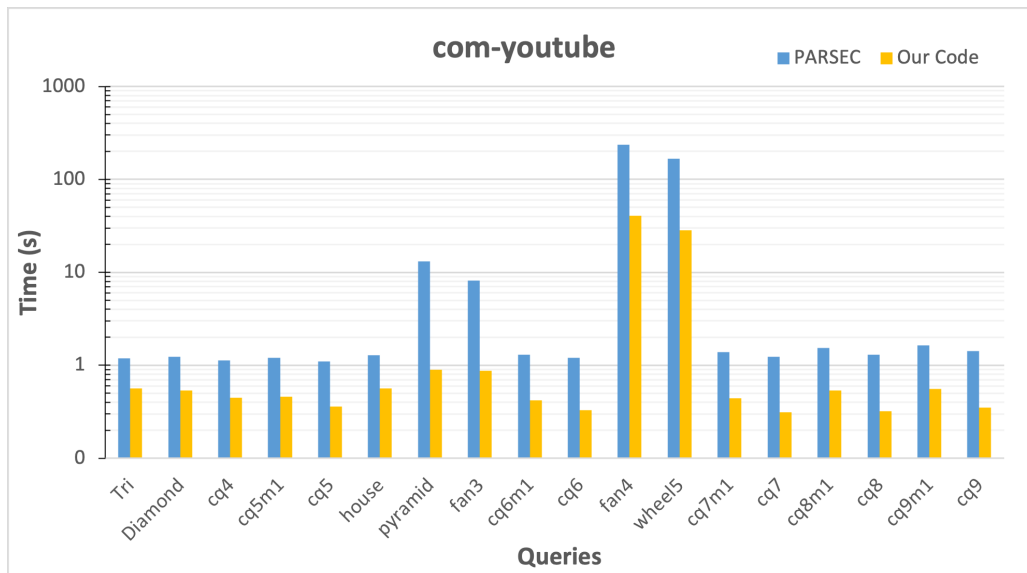
Table 6.2: Speedups across all queries and data graphs

Queries	soc-pokec	com-youtube	cit-patents	com-orkut	as-skitter	Geo Mean
Tri	1.49	2.11	1.00	0.73	1.95	1.35
Diamond	1.44	2.31	1.00	0.92	2.08	1.45
cq4	1.47	2.51	1.03	0.76	1.85	1.40
cq5m1	1.41	2.62	1.08	1.23	1.76	1.54
cq5	1.46	3.04	1.07	1.13	1.83	1.58
house	1.41	2.27	1.08	1.36	3.45	1.75
pyramid	1.50	14.65	1.15	4.5	9.75	4.06
fan3	1.50	9.38	1.26	3.59	5.89	3.27
cq6m1	4.19	3.09	5.47	2.41	1.44	3.01
fan4	3.41	5.82	2.62	–	–	3.73
wheel5	3.39	5.93	2.10	–	–	3.48
cq6	5.68	3.65	5.76	4.48	2.01	4.04
cq7m1	3.19	3.14	5.11	1.47	1.11	2.42
cq7	4.43	3.96	5.45	2.01	1.32	3.03
cq8m1	2.32	2.88	3.25	1.32	–	2.31
cq8	3.55	4.06	5.15	1.58	–	3.29
cq9m1	1.44	2.93	2.65	1.32	–	1.96
cq9	2.32	4.06	3.67	1.52	–	2.69
Geo Mean	2.25	3.74	2.20	1.61	2.29	2.33

Figure 6.2 presents a runtime comparison across all graphs and queries on a logarithmic scale. Table 6.2 provides individual speedups and geometric mean speedups across different templates and data graphs. The highest speedup among all observations is $14.65\times$, while the lowest speedup is $0.73\times$. The geometric mean speedups across different queries are separately plotted in Figure 6.3. Roughly, the speedups increase with increasing query size for queries in similar families. Queries in the family *fans*, and *wheels* have higher speedups as compared to others. This observation can be explained by Figure 5.6 since *fans* and *wheels* have higher load imbalance than other queries. Data graphs *cit-patents* and *com-orkut* perform worse than the baseline for small queries because the kernel that induces subgraph has high load imbalance. Since the processing time for these queries is fairly small, load imbalance in the subgraph induction step is significant.

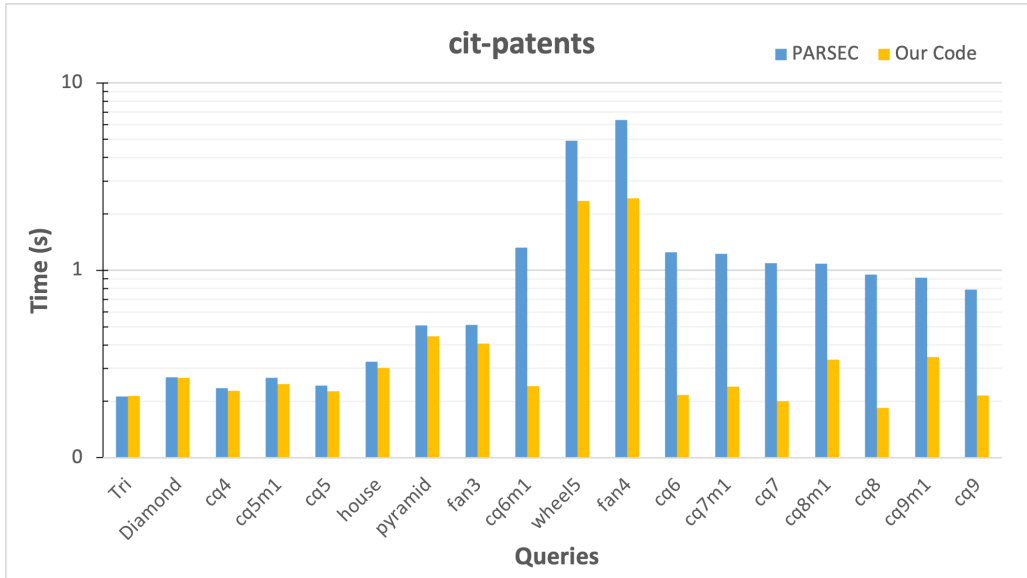


(a) Data Graph: soc-pokec

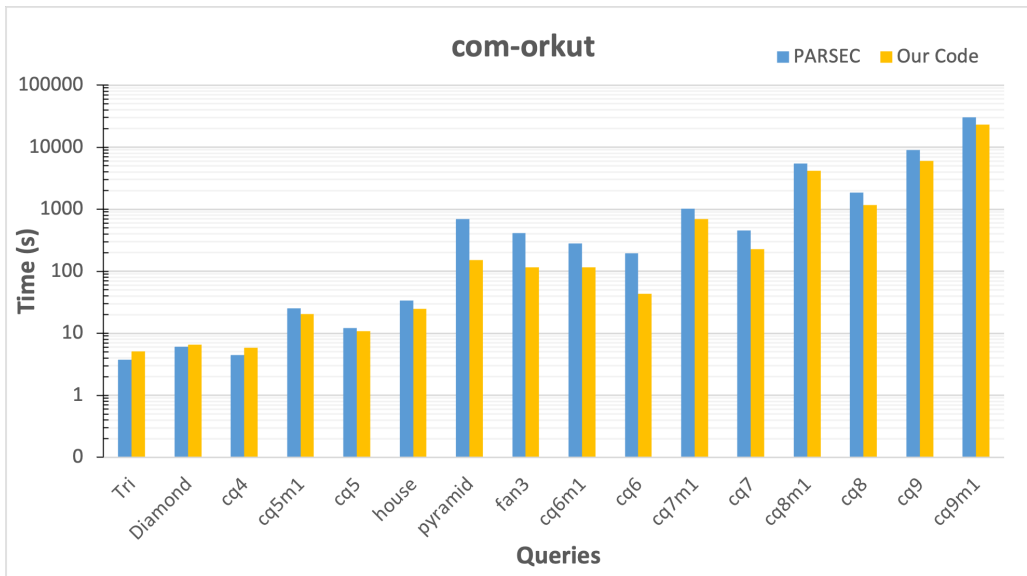


(b) Data Graph: com-youtube

Figure 6.2: Execution Times and Speedups Compared to PARSEC [1]

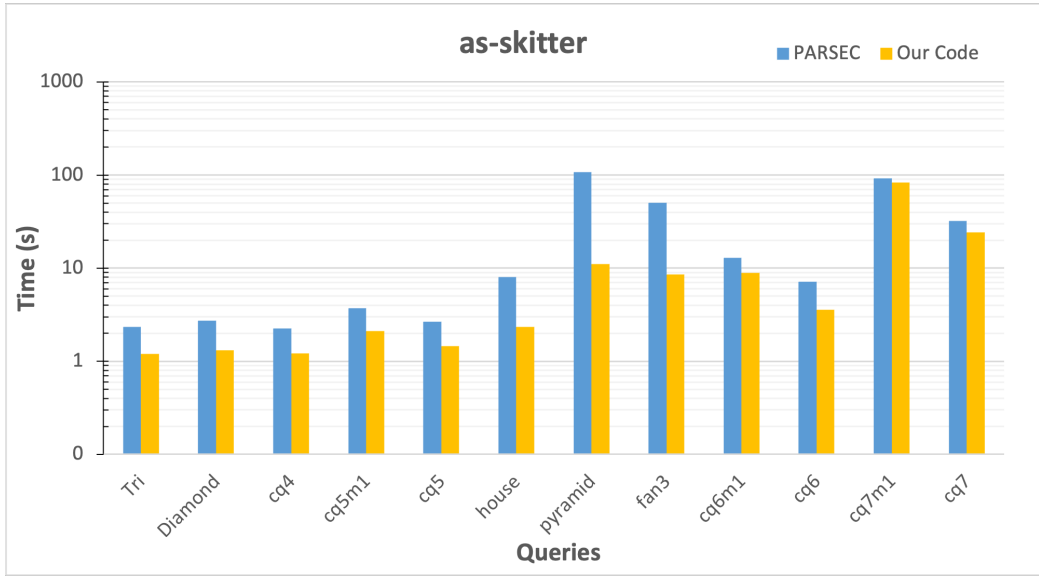


(c) Data Graph: cit-patents



(d) Data Graph: com-orkut

Figure 6.2: Execution Times and Speedups Compared to PARSEC [1]



(e) Data Graph: as-skitter

Figure 6.2: Execution Times and Speedups Compared to PARSEC [1]

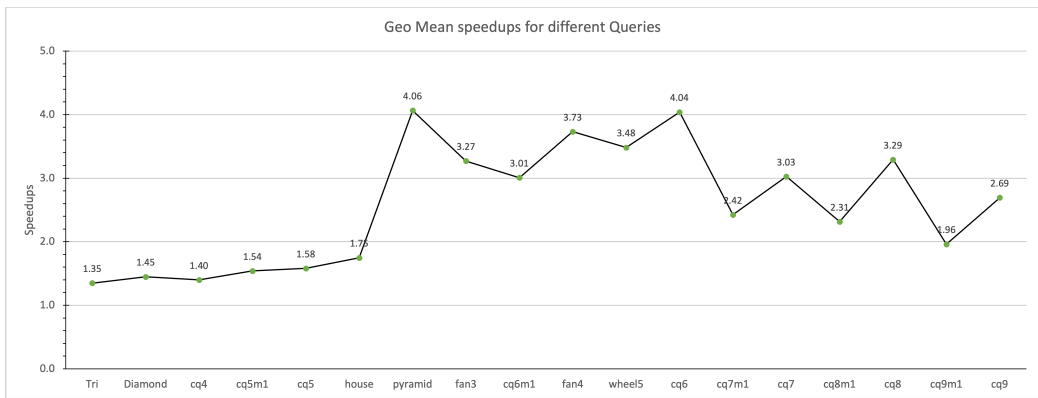


Figure 6.3: Geo Mean Speedups across Data Graphs

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this thesis, a comprehensive review of subgraph enumeration algorithms using state-space search techniques was conducted. Insights from traditional sequential algorithms and the understanding of GPU architecture were used to uncover several enhancements on the state-of-the-art subgraph enumeration application, PARSEC [1]. To reduce the number of intersections, the problem of detecting optimal reuse mappings was posed as an optimization problem and a closed form solution was found. The search space of the search tree was shrunk by using a two-phase hybrid symmetry breaking strategy which uses heuristics to dynamically decide the symmetry breaking direction. Further, an in-depth analysis of PARSEC [1] was performed to reveal the underlying load imbalance due to the parallelization scheme and static work division. The load balance was improved by using a hybrid parallelization scheme that switches after a cutoff. These different techniques help achieve modest geometric mean speedups of up to $4\times$ across different data graphs and up to $3.7\times$ across all queries. A maximum speedup of $14.6\times$ was attained. These efforts also help develop numerous computational insights which unlock opportunities for further improvements.

For future work, the load balance can be improved by using dynamic strategies like work sharing between blocks. The biggest drawback of this implementation is that it is only applicable for queries with a central node. Different algorithms need to be developed for supporting more queries. One idea might be to convert edges in query graphs to vertices to support queries with a central edge. Another could be to add dummy edges to queries to create an artificial central node.

Another drawback of this implementation is that the memory requirement is in the order of the square of undirected max degree. This restricts the implementation to be truly scalable for large dense graphs. However, real-world graphs generally have a few vertices with high degrees, these can be dealt

with using CPUs where the memory is not too restrictive. Orientation-based techniques were attempted which store induced subgraphs of the oriented graph and find remaining candidates using binary search. However, these implementations do not perform well with increasing query size and have worse load balance.

After addressing the load balance problems, this implementation can be scaled to multiple GPUs by using the CPU unified memory to store the data graph and process even bigger data graphs and larger queries in a reasonable time. Graph partition packages like METIS [45] can be utilized to scale to multiple CPUs. This will produce a truly scalable high performing subgraph enumerator unlocking new applications on much larger data graphs.

REFERENCES

- [1] V. Dodeja, M. Almasri, R. Nagi, J. E. Xiong, and W. mei Hwu, “Parsec: Parallel subgraph enumeration in cuda,” *IEEE*, 2022.
- [2] I. Wegener, *Complexity Theory - Exploring the Limits of Efficient Algorithms*. Springer Berlin, Heidelberg, 2005.
- [3] R. Alonso and S. Günnemann, “Mining contrasting quasi-clique patterns,” *CoRR*, vol. abs/1810.01836, 2018. [Online]. Available: <http://arxiv.org/abs/1810.01836>
- [4] N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S. C. Sahinalp, “Biomolecular network motif counting and discovery by color coding,” *Bioinformatics (Oxford, England)*, vol. 24, no. 13, pp. i241–i249, Jul 2008. [Online]. Available: <https://doi.org/10.1093/bioinformatics/btn163>
- [5] N. Shervashidze, S. Vishwanathan, T. Petri, K. Mehlhorn, and K. Borgwardt, “Efficient graphlet kernels for large graph comparison,” in *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, D. van Dyk and M. Welling, Eds., vol. 5. PMLR, 16–18 Apr 2009, pp. 488–495.
- [6] T. Milenković and N. Przulj, “Uncovering biological network function via graphlet degree signatures,” *Cancer informatics*, vol. 6, pp. 257–273, 2008. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/19259413>
- [7] I. Gutman, C. Rücker, and G. Rücker†, “On walks in molecular graphs,” *Journal of chemical information and computer sciences*, vol. 41, pp. 739–45, 05 2001.
- [8] J. Leskovec, A. Singh, and J. Kleinberg, “Patterns of influence in a recommendation network,” in *Proceedings of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, ser. PAKDD’06. Berlin, Heidelberg: Springer-Verlag, 2006, p. 380–389.
- [9] D. B. Kirk and W.-m. W. Hwu, *Programming Massively Parallel Processors: A Hands-on Approach*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2010.

- [10] H. Sutter and J. Larus, “Software and the concurrency revolution: Leveraging the full power of multicore processors demands new tools and new thinking from the software industry.” *Queue*, vol. 3, no. 7, p. 54–62, sep 2005. [Online]. Available: <https://doi.org/10.1145/1095408.1095421>
- [11] T. Biagini, F. Petrizzelli, M. Truglio, R. Cespa, A. Barbieri, D. Capocefalo, S. Castellana, M. F. Tevy, M. Carella, and T. Mazza, “Are gaming-enabled graphic processing unit cards convenient for molecular dynamics simulation?” *Evolutionary Bioinformatics*, vol. 15, p. 1176934319850144, 2019. [Online]. Available: <https://doi.org/10.1177/1176934319850144>
- [12] T. Dong, V. Dobrev, T. Kolev, R. Rieben, S. Tomov, and J. Dongarra, “A step towards energy efficient computing: Redesigning a hydrodynamic application on cpu-gpu,” in *Proceedings of the International Parallel and Distributed Processing Symposium, IPDPS*, 05 2014, pp. 972–981.
- [13] Business-Quant. (2022) Nvidia revenue breakdown by end-market (2017-2022). [Online]. Available: <https://businessquant.com/nvidia-revenue-by-end-market>
- [14] P. Gupta, “Cuda refresher: The cuda programming model,” NVIDIA, Tech. Rep., 2020. [Online]. Available: <https://developer.nvidia.com/blog/cuda-refresher-cuda-programming-model/>
- [15] Z. Lin, M. Alshboul, Y. Solihin, and H. Zhou, “Exploring memory persistency models for gpus,” in *2019 28th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2019, pp. 311–323.
- [16] J. R. Ullmann, “An algorithm for subgraph isomorphism,” *J. ACM*, vol. 23, no. 1, p. 31–42, jan 1976. [Online]. Available: <https://doi.org/10.1145/321921.321925>
- [17] H. Shang, Y. Zhang, X. Lin, and J. X. Yu, “Taming verification hardness: an efficient algorithm for testing subgraph isomorphism,” *Proc. VLDB Endow.*, vol. 1, pp. 364–375, 2008.
- [18] X. Ren and J. Wang, “Exploiting vertex relationships in speeding up subgraph isomorphism over large graphs,” *Proc. VLDB Endow.*, vol. 8, pp. 617–628, 2015.
- [19] F. Bi, L. Chang, X. Lin, L. Qin, and W. Zhang, “Efficient subgraph matching by postponing cartesian products,” in *Proceedings of the 2016 International Conference on Management of Data*, ser. SIGMOD ’16. New York, NY, USA: Association for Computing Machinery, 2016,

- p. 1199–1214. [Online]. Available: <https://doi.org/10.1145/2882903.2915236>
- [20] C. R. Rivero and H. M. Jamil, “Efficient and scalable labeled subgraph matching using sgmach,” *Knowledge and Information Systems*, vol. 51, pp. 61–87, 2016.
- [21] B. Archibald, F. Dunlop, R. Hoffmann, C. McCreesh, P. Prosser, and J. Trimble, “Sequential and parallel solution-biased search for subgraph algorithms,” in *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, L.-M. Rousseau and K. Stergiou, Eds. Cham: Springer International Publishing, 2019, pp. 20–38.
- [22] C. Solnon, “Alldifferent-based filtering for subgraph isomorphism,” *Artificial Intelligence*, vol. 174, no. 12, pp. 850–864, 2010. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0004370210000718>
- [23] L. Lai, L. Qin, X. Lin, and L. Chang, “Scalable subgraph enumeration in mapreduce,” *Proc. VLDB Endow.*, vol. 8, no. 10, p. 974–985, jun 2015. [Online]. Available: <https://doi.org/10.14778/2794367.2794368>
- [24] L. Lai, L. Qin, X. Lin, Y. Zhang, L. Chang, and S. Yang, “Scalable distributed subgraph enumeration,” *Proc. VLDB Endow.*, vol. 10, no. 3, p. 217–228, nov 2016. [Online]. Available: <https://doi.org/10.14778/3021924.3021937>
- [25] S. Sun and Q. Luo, “In-memory subgraph matching: An in-depth study,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Association for Computing Machinery, 6 2020, pp. 1083–1098.
- [26] P. Foggia, C. Sansone, and M. Vento, “An improved algorithm for matching large graphs,” in *Proc of the 3rd IAPR-TC-15 International Workshop on Graph-based Representation*, 01 2001.
- [27] L. Cordella, P. Foggia, C. Sansone, and M. Vento, “A (sub)graph isomorphism algorithm for matching large graphs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 10, pp. 1367–1372, 2004.
- [28] V. Carletti, P. Foggia, A. Saggese, and M. Vento, “Challenging the time complexity of exact subgraph isomorphism for huge and dense graphs with vf3,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 804–818, 2018.

- [29] J. A. Grochow and M. Kellis, “Network motif discovery using subgraph enumeration and symmetry-breaking,” in *Annual International Conference on Research in Computational Molecular Biology*. Springer, 2007, pp. 92–106.
- [30] A. for Computing Machinery. Special Interest Group on Management of Data., *Turbo-ISO*. ACM, 2013.
- [31] X. Ren and J. Wang, “Exploiting vertex relationships in speeding up subgraph isomorphism over large graphs,” *Proc. VLDB Endow.*, vol. 8, no. 5, p. 617–628, jan 2015. [Online]. Available: <https://doi.org/10.14778/2735479.2735493>
- [32] Y. Shao, B. Cui, L. Chen, L. Ma, J. Yao, and N. Xu, “Parallel subgraph listing in a large-scale graph,” in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD ’14. New York, NY, USA: Association for Computing Machinery, 2014, p. 625–636. [Online]. Available: <https://doi.org/10.1145/2588555.2588557>
- [33] B. Bhattarai, H. Liu, and H. H. Huang, “Ceci: Compact embedding cluster index for scalable subgraph matching,” in *Proceedings of the 2019 International Conference on Management of Data*, ser. SIGMOD ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1447–1462. [Online]. Available: <https://doi.org/10.1145/3299869.3300086>
- [34] S. Sun, Y. Che, L. Wang, and Q. Luo, “Efficient parallel subgraph enumeration on a single machine,” in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 2019, pp. 232–243.
- [35] D. Mawhirter, S. Reinehr, W. Han, N. Fields, M. Claver, C. Holmes, J. McClurg, T. Liu, and B. Wu, “Dryadic: Flexible and fast graph pattern matching at scale,” in *2021 30th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2021, pp. 289–303.
- [36] H.-N. Tran, J.-j. Kim, and B. He, “Fast subgraph matching on large graphs using graphics processors,” in *Database Systems for Advanced Applications*, M. Renz, C. Shahabi, X. Zhou, and M. A. Cheema, Eds. Cham: Springer International Publishing, 2015, pp. 299–315.
- [37] W. Guo, Y. Li, and L. Tan, “Exploiting reuse for gpu subgraph enumeration,” *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp. 1–1, 11 2020.
- [38] W. Guo, Y. Li, M. Sha, B. He, X. Xiao, and K. L. Tan, “Gpu-accelerated subgraph enumeration on partitioned graphs,” in *Proceedings of the*

- ACM SIGMOD International Conference on Management of Data*. Association for Computing Machinery, 6 2020, pp. 1067–1082.
- [39] B. D. McKay and A. Piperno, “Practical graph isomorphism, II,” *Journal of Symbolic Computation*, vol. 60, pp. 94–112, 2014.
- [40] D. Merril. (2016) Cub. [Online]. Available: <https://nvlabs.github.io/cub/index.html>
- [41] M. Almasri, I. E. Hajj, R. Nagi, J. Xiong, and W. mei Hwu, “Accelerating k-clique counting on GPUs,” *arXiv preprint arXiv:2104.13209*, 2021.
- [42] J. Crawford, M. Ginsberg, E. Luks, and A. Roy, “Symmetry-breaking predicates for search problems,” in *KR*, 03 1997.
- [43] S. Zampelli, Y. Deville, and P. Dupont, *Symmetry Breaking in Subgraph Pattern Matching*. John Wiley & Sons, Ltd, 2007, ch. 10, pp. 203–218. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470612309.ch10>
- [44] J. Leskovec and A. Krevl, “SNAP datasets: Stanford large network dataset collection,” 2014.
- [45] G. Karypis and V. Kumar, *METIS—A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes and Computing Fill-Reducing Ordering of Sparse Matrices*, 01 1997. [Online]. Available: <https://hdl.handle.net/11299/215346>