SISA: Set-Centric Instruction Set Architecture for Graph Mining on Processing-in-Memory Systems

Maciej Besta1, Raghavendra Kanakagiri2, Grzegorz Kwasniewski1, Rachata Ausavarungnirun3, Jakub Beránek4, Konstantinos Kanellopoulos1, Kacper Janda5, Zur Vonarburg-Shmaria1, Lukas Gianinazzi1, Ioana Stefan1, Juan Gómez-Luna1, Marcin Copik1, Lukas Kapp-Schwoerer1, Salvatore Di Girolamo1, Nils Blach1, Marek Konieczny5, Onur Mutlu1, Torsten Hoefler1

1ETH Zurich, Switzerland 2IIT Tirupati, India 3King Mongkut’s University of Technology North Bangkok, Thailand 4Technical University of Ostrava, Czech Republic 5AGH-UST, Poland

ABSTRACT
Simple graph algorithms such as PageRank have been the target of numerous hardware accelerators. Yet, there also exist much more complex graph mining algorithms for problems such as clustering or maximal clique listing. These algorithms are memory-bound and thus could be accelerated by hardware techniques such as Processing-in-Memory (PIM). However, they also come with non-straightforward parallelism and complicated memory access patterns. In this work, we address this problem with a simple yet surprisingly powerful observation: operations on sets of vertices, such as intersection or union, form a large part of many complex graph mining algorithms, and can offer rich and simple parallelism at multiple levels. This observation drives our cross-layer design, in which we (1) expose set operations using a novel programming paradigm, (2) express and execute these operations efficiently with carefully designed set-centric ISA extensions called SISA, and (3) use PIM to accelerate SISA instructions. The key design idea is to alleviate the bandwidth needs of SISA instructions by mapping set operations to two types of PIM: in-DRAM bulk bitwise computing for bitvectors representing high-degree vertices, and near-memory logic layers for integer arrays representing low-degree vertices. Set-centric SISA-enhanced algorithms are efficient and outperform hand-tuned baselines, offering more than 10× speedup over the established Bron-Kerbosch algorithm for listing maximal cliques. We deliver more than 10 SISA set-centric algorithm formulations, illustrating SISA’s wide applicability.

CCS CONCEPTS
• Hardware → Emerging architectures; Memory and dense storage; Application-specific VLSI designs; Application specific instruction set processors; • Computer systems organization → Architectures; • Theory of computation → Design and analysis of algorithms; Graph algorithms analysis; Data structures design and analysis; Parallel algorithms; • Mathematics of computing → Graph algorithms; • Information systems → Data mining; Clustering; • Computing methodologies → Parallel computing methodologies.

KEYWORDS
Graph Mining, Graph Pattern Matching, Graph Learning, Clique Mining, Clique Listing, Clique Enumeration, Subgraph Isomorphism, Parallel Graph Algorithms, Processing In Memory, Processing Near Memory, Graph Accelerators, Instruction Set Architecture

ACM Reference Format:
Maciej Besta et al.. SISA: Set-Centric Instruction Set Architecture for Graph Mining on Processing-in-Memory Systems. In MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO ’21), October 18–22, 2021, Virtual Event, Greece. ACM, New York, NY, USA. 26 pages.
https://doi.org/10.1145/3466752.3480133

1 INTRODUCTION
Graph analytics underlies many problems in machine learning, social network analysis, computational chemistry or biology, medicine, finances, and others [167]. The growing importance of these fields necessitates even more efficient large-scale graph processing. Research on graph analytics in computer architecture has mostly targeted graph algorithms based on vertex-centric formulations [6, 7, 19, 40, 109, 135, 175, 186, 218, 270, 285]. Some works also focus on edge-centric or linear algebra paradigms [140, 205, 227, 230]. The algorithms in question are usually iterative and have complexities described by low-degree polynomials [141], for example \(O(n + m)\) for Breadth-First Search (BFS) [71], \(O(\ell mL)\) for a power iteration scheme for PageRank (PR) [39], and \(O(mn)\) for Brandes’ algorithm for Betweenness Centrality (BC) [51], where \(n\) and \(m\) are numbers of vertices and edges, and \(\ell\) is a selected iteration count.

Yet, there are numerous important problems and algorithms in the area of graph mining [42, 68, 131, 210, 236] that received little or no attention in computer architecture. One large class is graph pattern matching [131], which focuses on finding certain specific subgraphs (also called motifs or graphlets). Examples of such problems are \(k\)-clique listing [78, 102], maximal clique listing [52, 58, 91, 238], \(k\)-star-clique mining [125], and many others [68]. Another class is broadly referred to as graph learning [68], with problems such as unsupervised learning or clustering [128], link prediction [9, 157, 165, 234], or vertex similarity [151]. All these problems are widely used in social sciences (e.g., studying human interactions) [91], bioinformatics (e.g., analyzing protein structures) [91], computational chemistry (e.g., designing chemical compounds) [232], medicine (e.g., drug discovery) [232], cybersecurity (e.g., identify intruder machines) [85], healthcare (e.g., identify groups of people who submit fraudulent claims) [237], web graph analysis (e.g., enhance search services) [132], entertainment (e.g.,
Thus, they often differ significantly in their performance properties from "low-complexity" problems such as BFS or PageRank.

Importantly, the established vertex-centric model, originally proposed in the Pregel graph processing system [169], does not effectively express graph mining problems. It exposes only the local graph structure: A thread executing a vertex kernel for any vertex \( v \) can only access the neighbors of \( v \). While this suffices for algorithms such as PageRank, graph mining often requires non-local knowledge of the graph structure [68]. Obtaining such knowledge in the vertex-centric paradigm is hard or infeasible. As noted by multiple researchers [144, 161, 209, 257] "(...) implementing graph algorithms efficiently on Pregel-like systems (...) can be surprisingly difficult and require careful optimizations." [209]. "It is challenging to design Pregel algorithms (...)" [257]. "Non-iterative graph algorithms might be difficult to express in the vertex-centric model which heavily relies on (...) supersteps" [135], "(...) graph algorithms, like triangle counting, are not a good fit for the vertex-centric model. [135]. Thus, most graph mining problems cannot be simply programmed with the vertex-centric software (SW) frameworks [107, 135, 164, 169, 175] and accelerated with vertex-centric hardware (HW) architectures [6, 7, 40, 109, 186, 227, 270]. Similar arguments apply to other paradigms such as GraphBLAS [140, 205] and to frameworks such as Ligra [221]. They do not support many graph mining problems, and we discuss in Table 1 and Section 4.

Several graph mining software frameworks (Peregrine [127] and others [62, 63, 83, 124, 133, 173, 174, 236, 256, 258, 273]) were proposed. Unfortunately, they focus exclusively on only a few graph pattern matching problems. Moreover, these frameworks usually do not have any formal guarantees on total work [47], and thus they do not provide time complexities and run-times competitive to those of tuned parallel graph algorithms for solving specific mining problems [42]. Overall, there is a need for a graph mining paradigm that would (1) enable expressing many graph mining problems, and (2) offer competitive theoretical work guarantees [47]. We summarize this in Table 1, which analyzes existing programming paradigms for graph processing (a total of seven). We detail this table in Section 4 – its central message is that no existing graph processing paradigm supports many graph mining problems or offers competitive time complexities.

Moreover, past works illustrated that graph mining algorithms are memory bound [65, 87, 127, 264, 271]. This is because these algorithms generate and heavily use large intermediate structures, but, similarly to algorithms such as PageRank, they are not compute-intensive [91, 127, 265]. We show this in Figure 1: When we increase the number of parallel threads, runtime decrease flattens out and stalled CPU cycle count increases. This motivates using processing-in-memory (PIM) to obtain the much needed speedups in graph mining. While PIM is not the only potential solution for hardware acceleration of graph mining, we select PIM because it (1) represents one of the most promising trends to tackle the memory bottleneck [100, 183] outperforming various other approaches [217], (2) offers well-understood designs [184], and (3) brings very large speedups in simple graph algorithms such as BFS or PageRank (see Figure 1).
more than 15 works in Table 8). Yet, graph mining algorithms are much more complex than PageRank, BFS, and similar: they employ deep recursion, create many intermediate data structures with non-trivial inter-dependencies, and have high load imbalance [91, 256]. Table 8 (shown and detailed in Section 10) extends Table 1 by analyzing specific hardware (HW) accelerators for graph processing, illustrating that no existing HW design targets broad graph mining (i.e., both graph pattern matching and graph learning), or explores PIM techniques for accelerating broad graph processing.

To address all these issues, we propose a novel design that is high-performance (empirically and theoretically), applicable to many graph mining problems, and easily amenable to PIM acceleration. We first observe that large parts of many graph mining algorithms can be expressed with simple set operations such as intersection \( \cap \) or union \( \cup \), where sets contain vertices or edges. This drives our set-centric programming paradigm, in which the developer identifies sets and set operations in a given algorithm. These set operations are then mapped to a small and simple yet expressive group of instructions, offering a rich selection of storage/performance tradeoffs. These instructions are offloaded to PIM units. We call these instructions SISA as they form “set-centric” ISA extensions that enable a simple interface between numerous graph mining algorithms and PIM hardware. Overall, our cross-layer design consists of three key elements: a new set-centric programming paradigm and formulations of graph algorithms (contribution #1), set-centric ISA extensions with its instructions, implemented set operations, and set organization (contribution #2), and PIM acceleration (contribution #3). The strength of our design comes from observing that these concepts (set algebra/notation, set representations/algorithms, PIM) fit together and only need minor HW extensions to provide an efficient architecture for graph mining.

Overall, we advocate using set algebra as a basis for the design of graph mining algorithms. Our set-centric paradigm is the first to use set operations as fundamental general building blocks for both algorithmic formulations and their execution. Using set algebra ensures that SISA set-centric algorithms are succinct, applicable to many problems, and theoretically efficient. For the in-memory acceleration of SISA, we investigate which types of PIM are beneficial for which set operations. We process sets stored as bitvectors in situ PIM [101, 184], as offered in Ambit [108, 217], ELI2IM [253], DRISA [155], or ComputeDRAM [96], for highest performance and energy efficiency (“SISA processing using memory” – SISA-PUM). In contrast, while sets stored as sparse arrays cannot be simply processed in situ with today’s technology, they can use the high throughput and low latency of near-memory PIM [101, 163, 184, 191] as offered in the 2D UPMEM architecture [106, 147] or logic layer of 3D DRAM such as Hybrid Memory Cube (HMC) [130] (“SISA processing near memory” – SISA-PNM). Here, for further speedups, we also provide a small HW controller that selects on-the-fly the best variant of a set instruction to be executed with PIM. For example, it decides on using merge or galloping set intersection, based on the properties of the processed graph, using our performance models.

Overall, our results show that graph mining algorithms, although complex and lacking straightforward parallelism, greatly benefit from PIM. For example, Bron-Kerbosch does not offer simple vertex-level parallelism known from algorithms such as PageRank: some vertices may belong to large cliques, and processing such cliques results in deep recursion trees, which take a large portion of the processing time while not offering straightforward parallelization opportunities. Our key solution is using parallelism offered by set operations and exposed with the set-centric approach. This solution harnesses parallelism at the level of bits, DRAM subarrays, and vaults. We show that SISA-enhanced algorithms are theoretically efficient (contribution #4) and empirically outperform tuned parallel baselines (contribution #5), for example offering more than 10x speedup for many real-world graphs over the established Bron-Kerbosch algorithm for listing maximal cliques [91]. Finally, for usability, we integrate SISA with the RISC-V ISA [251].

To summarize, we contribute the following:

- We propose a set-centric programming approach for a wide selection of graph mining problems, in which one exposes and exploits sets and set operations in graph algorithms.
- We develop SISA, a set-centric ISA interface between hardware and software in graph mining. We describe the syntax, semantics, and encoding of SISA.
- We provide a careful graph data layout based on dense bit vectors and sparse integer arrays, a hardware implementation of SISA that harnesses in- and near-memory processing, and associated performance models that enable automatic selection of fastest set instructions.
- We develop programming guidelines for SISA, covering details such as selecting the most beneficial SISA instructions for different set operations in graph algorithms.
- We provide an extensive theoretical analysis of SISA, analyzing the performance of graph algorithms and single set operations, for different set representations. This analysis shows that SISA offers competitive time complexities.
- We use cycle-based simulations to illustrate performance advantages of SISA over hand-tuned baselines.

## 2 NOTATION AND BACKGROUND

We first describe background and notation, see Table 2.

**Graphs** We model an undirected graph \( G \) as a tuple \((V, E)\); \( V \) and \( E \subseteq V \times V \) are sets of vertices and edges; \(|V| = n, |E| = m\). Vertices are modeled with integers \((V = \{1, \ldots, n\})\). \( G = (V, E) \) is a directed graph; \( E \) contains arcs. \( N(v) \) and \( N^+(v) \) denote the neighbors and the out-neighbors of \( v \in V \); \( d \) and \( d(v) \) denote \( G \)'s maximum degree and a degree of \( v \). In some cases, we consider labeled graphs \( G = (V, E, L) \); \( L \) is a labeling function that maps a vertex or an edge to a label.

**Set Representations** A concept used heavily in SISA is the set representation and its sparsity. Figure 4 shows example considered representations. Consider a set of \( k \) vertices \( S = \{v_1, \ldots, v_k\} \subseteq V \) (we focus on vertex sets, but SISA also works with edges). One can represent \( S \) as a simple contiguous sparse array (SA) with integers from \( S \) (“sparse” means that only non-zero elements are explicitly stored). \( S \)'s size is \( W|S| \) bits where \( W \) is the memory word size (we assume that the maximum vertex ID fits in one word). One can also represent \( S \) with a dense bitvector (DB) of size \( n \) bits: the \( i \)-th bit indicates that a vertex \( i \in S \) (“dense” means that all zero bits are explicitly stored).
We now overview SISA's cross-level design, see Figure 2. SISA's variant on-the-fly. Variants of operations, for example there is an instruction for both merge and galloping set intersection (details in Section 6). We also implement set operations. These instructions support all how and not be executed using different set algorithms. A set-centric formulation of these sets.

The set-centric approach is superior to other graph programming paradigms in that (1) it supports many graph mining problems and (2) it enables algorithms with competitive theoretical bounds on performance (we discuss (2) in Section 7; this is often a key to low runtimes [81, 141]). The analysis results for (1) are in Table 1.

To illustrate the above points, we first extensively examined the related literature to identify representative graph mining problems and important graph processing algorithms [4, 11, 59, 95, 131, 149, 151, 157, 165, 197, 199, 201, 233, 248]. For the former, we pick Ambit [217] as the implementation of SISA-PUM within a DRAM die, and logic layers for SISA-PNM, but other designs can also be used.

The SCU maintains set metadata (SM) using a dedicated in-memory SM structure. SM contains mappings between logical set IDs and set addresses, and the type of the representation as well as the cardinality of a given set. This information is used to guide SCU decisions. Finally, the SCU has a small scratchpad, the Set Metadata Buffer (SMB), to cache metadata.

4 SETS FOR SIMPLE & PROVABLY FAST GRAPH MINING

The set-centric approach is superior to other graph programming paradigms in that (1) it supports many graph mining problems and (2) it enables algorithms with competitive theoretical bounds on performance (we discuss (2) in Section 7; this is often a key to low runtimes [81, 141]). The analysis results for (1) are in Table 1.

The SCU receives SISA instructions from the CPU, and it appropriately schedules their execution on SISA-PNM and SISA-PUM. Two bitvectors are always processed with SISA-PUM, while in other scenarios SCU uses SISA-PNM. The SCU can also select the most advantageous set algorithm. For example, whenever two sets have similar sizes, it is better to intersect them using a merge-based intersection, in which input sets are streamed and they can harness high sequential bandwidth. For concreteness, in Section 8, we pick Ambit [217] as the implementation of SISA-PUM within a DRAM die, and logic layers for SISA-PNM, but other designs can also be used.

The SCU maintains set metadata (SM) using a dedicated in-memory SM structure. SM contains mappings between logical set IDs and set addresses, and the type of the representation as well as the cardinality of a given set. This information is used to guide SCU decisions. Finally, the SCU has a small scratchpad, the Set Metadata Buffer (SMB), to cache metadata.

4 SETS FOR SIMPLE & PROVABLY FAST GRAPH MINING

The set-centric approach is superior to other graph programming paradigms in that (1) it supports many graph mining problems and (2) it enables algorithms with competitive theoretical bounds on performance (we discuss (2) in Section 7; this is often a key to low runtimes [81, 141]). The analysis results for (1) are in Table 1.

To illustrate the above points, we first extensively examined the related literature to identify representative graph mining problems and important graph processing algorithms [4, 11, 59, 95, 131, 149, 151, 157, 165, 197, 199, 201, 233, 248]. For the former, we pick four problems from both graph pattern matching and graph learning areas (maximal clique listing [52], k-clique listing [66], dense subgraph discovery [105, 149], subgraph isomorphism [240], vertex similarity [151, 202], link prediction [9, 157, 165, 234], graph clustering [128, 210], verification of prediction accuracy [245]). For

| Table 2: The most important symbols and acronyms. |
|---|
| $G = (V, E)$ | An undirected graph. $V$ and $E$ are sets of vertices and edges. |
| $\overline{G} = (\overline{V}, \overline{E})$ | A directed graph. $\overline{V}$ and $\overline{E}$ are sets of vertices and directed edges. |
| $G[V]$ | A subgraph (of a graph $G$) induced on a vertex set $V$. |
| $n, m$ | The numbers of vertices and edges in $G$ ($|V| = n$, $|E| = m$). |
| $N(v), N^+(v)$ | The neighbors and the out-neighbors of a vertex $v$. |
| $d, d(v)$ | The maximum degree of $G$, the degree of $v \in V$. |
| $c$ | The graph degeneracy (a property used in theoretical analysis). |
| SA, DB | Sparse array, dense bitvector. |
| MC, BM | The latency and bandwidth of accessing DRAM. |
| $b_L$ | The bandwidth of the interconnect (e.g., QPI) between cores. |
| $l_c$ | The latency of one bulk bitwise operation run with in-situ PIM. |
| $r$ | #rows that can be processed in parallel (e.g., in a DRAM bank). |
| $R$ | The size [bits] of a single DRAM row. |
| SCU, SM | SISA Controller Unit, Set Metadata. |

Set Operations SISA uses all basic set operations: intersection $A \cap B$, union $A \cup B$, difference $A \setminus B$, cardinality $|A|$, and membership $\in A$. $A$ and $B$ usually contain vertices and sometimes edges. We use different algorithms to implement these operations (described later in the paper).

Architecture Concepts We outline the used architectural concepts in Section 1; more details are in Section 8 and 9. The architecture-related symbols are listed in Table 2 (bottom).

3 OVERVIEW & CROSS-LAYER DESIGN

We now overview SISA's cross-level design, see Figure 2. SISA's cross-layer design (see Figure 2) consists of three key elements: (a) set-centric formulations of graph algorithms, (b) the actual set-centric ISA with its instructions, implemented set operations, set organization, and a thin software layer, and (c) an example HW implementation. We support SISA with programming guidelines and a theoretical analysis.

(a) Set-Centric Formulations [Section 5 & 5.4] SISA relies on set-centric formulations of algorithms in graph mining. While some algorithms (e.g., Bron-Kerbosch [91]) by default use rich set notation, many others, such as $k$-clique listing by Danisch et al. [78], do not. In such cases, we develop such formulations. Details on deriving set-centric formulations are in Section 5.4; the key common step is to express two nested loops, commonly used to identify connections between two sets of vertices, with a single intersection of these sets.

A set can be represented in different ways, and a set operation can be executed using different set algorithms. A set-centric formulation hides these details, focusing on what a given graph algorithm does, and not how it is done.

(b.1) Set-Centric ISA (Instructions) [Section 6] Our ISA extension implements set operations. These instructions support all variants of operations, for example there is an instruction for both merge and galloping set intersection (details in Section 6). We also provide a thin software layer: iterators over sets and C-style wrappers for SISA instructions. For programmability and performance, many SISA instructions automatize selecting the best set operation variant on-the-fly.
fairness, we also consider four popular "low-complexity" problems, targeted by many past works (triangle counting, BFS, connected components, and PageRank). For the latter, we first select vertex-centric [169] and edge-centric [205], two established graph processing paradigms implemented in the Pregel and X-Stream systems. Second, we pick vertex/edge array maps from Ligra [221], an approach for developing graph algorithms based on transforming arrays of vertices or edges according to a specified map. Third, we consider GraphBLAS and its linear algebraic approach [140], where graph algorithms are expressed with linear algebra building blocks such as matrix-vector products. Moreover, we consider pattern matching frameworks [95] that usually employ some form of exploring neighbors of each vertex, combined with user-specified filtering, to search for specified graph patterns. For completeness, we also consider recent attempts at solving graph problems with novel deep learning [25] paradigms such as graph neural networks (GNN) [30, 252] and others [103], as well as joins and principles from relational databases and the associated algebra [274].

The analysis results are in Table 1. Overall, no single paradigm, except for the set-centric approach, enables efficient graph mining algorithms for the considered problems. Some paradigms, such as the vertex-centric or the edge-centric model, do not focus on such problems at all. Other paradigms, for example array maps or GNNs, address only certain problems. Finally, graph pattern matching or RDBMS can solve different graph mining problems, but they do not offer formal guarantees, as indicated by past work.

5 SET-CENTRIC GRAPH ALGORITHMS

We now present set-centric formulations of graph mining algorithms. The used set operations are in Table 3. We loosely categorize the considered problems and algorithms into graph pattern matching (§5.1) and graph learning (§5.2), based on an analysis of graph related surveys [21, 42, 71, 84, 131, 175, 194, 196, 210, 218, 236].

Notes on Listings Set operations accelerated by SISA are marked with the gray color. "[ in par]" indicates that in a given loop one can issue set operations in parallel. We ensure that the parallelization does not involve conflicting memory accesses. We use "∪=", "∩=", "\-=" to indicate that a set operation mutates its first set argument. We now focus on formulations and we discuss set representations, instructions, and parallelization later. For clarity, we exclude unrelated optimizations from the listings.

Does SISA Execute All Set Operations? SISA is used for executing set operations that benefit from hardware acceleration, but one can find certain counter-examples, i.e., it may be more beneficial to use standard implementations of set operations, or to exclude set notation and set-focused data structures completely. For example, appending a vertex $v$ to a list $L$, which can be expressed as $L ∪ \{v\}$, does not necessarily benefit from memory acceleration, if $L$ is implemented as a linked list. We provide examples of algorithms not necessarily benefiting from the memory acceleration (as offered in SISA) later in this section. We discovered that this is often the case with "low-complexity" algorithms such as Boruvka’s algorithm for solving the Minimum Spanning Tree problem.

Time Complexity of SISA Algorithms Complexities of set-centric algorithms heavily depend on many factors, such as the
We first consider graph pattern matching, an important class of problems where one searches for specific subgraphs. The main recursive function (Line 4) has three arguments that are dynamic sets containing vertices. The algorithm heavily uses different set operations. The main recursive function and SISA also provides a dedicated set-centric formulation. BK mathematics [79, 228], and computational chemistry [200].

Table 3: Overview of set-centric graph algorithms. In maximal clique listing, subgraph isomorphism, and clustering, one also uses variants of union and difference where one set is always a single-element set (i.e., $A \cup \{b\}, A \setminus \{b\}$). Bolded text indicates algorithms with set-centric formulations derived in this work.

5.1 Graph Pattern Matching

We first consider graph pattern matching, an important class of problems where one searches for specific subgraphs.

5.1.1 Triangle Counting. In the extensively researched triangle counting (TC) problem [10, 21, 211, 222], one counts the total number of 3-cycles $t_c$. This problem is somewhat similar to Bron-Kerbosch in that it is also recursive and restricts the search to a larger clique. Finding all maximal cliques, an NP-Hard problem, has many applications in social network analysis [249], bioinformatics [79, 228], and computational chemistry [200]. Algorithm 2 contains the recursive backtracking Bron-Kerbosch algorithm [52] with pivoting and degeneracy optimizations [58, 91, 238], an established and commonly used scheme for finding maximal cliques (deriving the degeneracy ordering is itself an important graph problem and SISA also provides a dedicated set-centric formulation). BK heavily uses different set operations. The main recursive function $\text{BKPivot}(R, P, X)$ has three arguments that are dynamic sets containing vertices. The algorithm recursively calls $\text{BKPivot}$ for each new candidate vertex, checks if this gives a clique, and updates accordingly $P$ and $X$. Some optimizations need more set operations, but they reduce the search space of potential cliques [238]. For example, the set of candidates (for extending a clique $c$) is $P \setminus N(u)$ instead of $P$, where $u \in P \setminus X$.

Algorithm 2: Maximal Clique Listing (Bron-Kerbosch) [52, 58].
5.1.5 Degeneracy Order and k-Core

Graphs in which the degeneracy order is high often have a large k-core. The degeneracy order of a graph is the highest order of a degenerate set of vertices. A degenerate set is a set of vertices such that every vertex in the set has a neighbor in the set. The degeneracy order is defined as the maximum degeneracy order of any degenerate set.

A k-core of a graph is a maximal subgraph in which every vertex has degree at least k. The k-core of a graph is a measure of the cliquishness of the graph. A graph with a high k-core is more cliquish than a graph with a low k-core.

5.1.6 Subgraph Isomorphism

Subgraph isomorphism (SI) is a key graph problem where one checks whether a given (usually small) graph G2 is a subgraph of a graph G1. Here, we consider an established VF2 algorithm [70]. In Algorithm 7, we first provide the most important part that recursively constructs a candidate set of vertices from G1, and verifies if it matches the pattern G2.

We use SI as an example of how SISA supports labeled graphs. In VF2 [70], for each transition between states, one first verifies if the structure of G2 matches that of G1 (Line 11). Then, label matching is verified independently (Lines 12-13). Checking if vertex labels match, i.e., if \( L(v_1) = L(v_2) \), is trivial. Yet, a graph may also contain edge labels that need to be matched. This could be done with a standard approach without set operations [70]. However, the generality of set notation also enables supporting label verification. For this, we first identify all edges in G1 where one endpoint is the newly matched vertex \( v_1 \) and the other endpoint \( v'_1 \) is already matched (i.e., \( v'_1 \in M_1(s) \)). This is done with an intersection \( N_1(v_1) \cap M_1(s) \). Then, we find the vertex with which \( v'_1 \) is matched, see the second loop in Line 17. Finally, we verify that the respective labels match (Line 18).

5.1.7 Frequent Subgraph Mining

Here, we use an established Apriori-based scheme [5], [131, Algorithm 3.1]. We show it in Algorithm 8. It first generates candidate subgraphs \( C_k \) (Line 6) and then checks their counts \( c(v) \) in the input graph (Line 8) using subgraph isomorphism (SI) as a fundamental kernel [131] (combining candidate generation and occurrence verification is a very popular FSM approach [5, 110, 145, 146], also see other references in [131]). If the count is above a certain user selected threshold \( \sigma \cdot n \), a candidate is added as a found frequent subgraph (Line 9). VF2, an SI algorithm covered in this section, was found to be an efficient kernel for FSM; all SISA operations in SI are reused. Generation of candidate subgraphs (candidate_gen) is less time-consuming than SI [131]. Still, it also benefits from set operations; for example, joining trees that represent candidates, a key operation in a kernel by Hido and Kawano [118], is done using set union [131]. These trees can be implemented with either n-bit dense bitvectors or sparse arrays, benefiting from SISA-PUM or PNM (user’s choice).

Algorithm 3: k-Clique Counting [78].

\[
k\text{-clique-stars relax the restrictive nature of cliques [125]. Algorithm 4 shows the scheme. We first find k-cliques. Then, for each k-clique, one finds additional vertices that form stars with intersections and a union. We also observe that those extra vertices that are connected to the k-clique actually form a \((k+1)\)-clique (together with this k-clique). For this, we provide another variant of k-clique-star listing, see Algorithm 5. Specifically, to find k-clique-stars, we first mine \((k+1)\)-cliques. Then, we find k-clique-stars within each \((k+1)\)-clique using set union, membership, and difference. Large cliques are expected to be rare because every vertex in a clique, regardless of the clique size, must be connected to all other vertices in this clique.
\]

Algorithm 4: k-clique-star listing [125].
Figure 3: Overview of SISA instructions and syntax at different levels of abstraction.
5.3 “Low-Complexity” Algorithms

SISA does not target the “low-complexity” algorithms, as they offer few opportunities for set-centric acceleration [37, 50, 71, 104, 177, 178, 220, 224, 226, 231, 257]. For example, in PageRank, one updates vertex ranks in two nested loops, which is not easily expressible with set operations. We analyzed many other such algorithms. This includes Dijkstra’s SSSP [224], A*-Stepping [177], Bellman-Ford [71], Betweenness Centrality schemes [226], traversals [37], Connected Components algorithms [104, 220], Low-Diameter Decomposition [178], or Borůvka’s Minimum Spanning Tree [50].

Our work is already more general than other pattern matching accelerators / frameworks, as it supports many more problems beyond simple pattern matching (e.g., vertex similarity, clustering, link prediction, complex algorithms such as Bron-Kerbosch).

As a single example, we illustrate and discuss a set-centric formulation of Breadth-First Search (BFS) [71], the basic graph traversal algorithm, see Algorithm 12. BFS is a basis of the established Graph500 benchmark [182] and a subject of extensive research in a past decade [20, 22, 39, 53, 54, 152, 182, 266]. In the bottom-up part of BFS, the key element of the set-based variant is an additional set \( \Pi \) with unvisited vertices. \( \Pi \) is represented as a dense bit vector, requiring only \( n \) bits of storage. Using \( \Pi \) and other sets enables abstracting away some branches.

5.4 Deriving a Set-Centric Formulation

One either picks a set-centric formulation of a given algorithm, or designs one. For the former, we offer more than 10 set-centric formulations. For the latter, one starts with a selected algorithm specification to be “made” set-centric. Often, algorithms use set notation, and one may simply pick operations for memory acceleration. This is the case with, for example, Jarvis-Patrick clustering. Still, one may need to apply more complex changes to “expose” set instructions. The general rule is to associate used data structures with sets, and then identify respective set operations. As an example, we compare a traditional snippet for deriving the count of all 4-cliques \( cnt \), a derived set-centric algorithmic formulation, and the corresponding SISA snippet in Table 4. The key algorithmic change is using set intersections instead of explicitly verifying if vertices are connected. For example, instead of iterating over all neighbors of \( v_1-v_3 \) (Lines 4-6, the top snippet), in SISA, we intersect neighbors of \( v_1-v_3 \) (Line 4 & 6, the middle snippet) to filter 4-cliques.

Table 4: Finding all 4-cliques: a traditional (non-set-centric) snippet, a set-centric algorithmic formulation derived in this work, and a SISA set-centric snippet. In all variants, we use directed edges (i.e., iterate over out-neighbors) to avoid counting the same clique twice.

6 SISA: DESIGN, SYNTAX, SEMANTICS

We now present the details of representing and processing sets used in set-centric formulations. This constitutes core parts of SISA’s
design. We summarize SISA in Figure 3 and we detail key SISA instructions in Table 5. We already outlined SISA’s general structure in § 3.

### 6.1 Representation of Sets

The first key question is how to represent sets: SISA’s "first-class citizens." We observe that – in each graph algorithm – there are two fundamentally different classes of data structures. One class are (1) vertex neighborhoods \( N(v) \) that maintain the structure of the input graph. There are \( n \) such sets, their total size is \( O(m) \), and each single neighborhood is static (we currently focus on static graphs) and sorted (following the established practice in graph processing [170]). Another class are (2) auxiliary structures, for example \( P \) in Bron-Kerbosch (Listing 2). These sets are used to maintain some algorithmic state. They are usually dynamic, they may be unsorted, their number (in a given algorithm) is usually a small constant, and their total size is \( O(n) \). While SISA enables using any set representation for any specific set, we offer certain recommendations to maximize performance.

SAs should be used for small neighborhoods and DBs for the large ones (in the evaluation, we vary the threshold so that 5%-30% largest neighborhoods use DBs). This approach is memory efficient. For example, for \(|N(v)| = n/2\), a DB takes only \( n \) bits while an SA uses \( 16n \) bits (for a 32-bit word size).

Auxiliary sets benefit from being stored as dense bitvectors. This is because such sets are often dynamic, and updates or removals take \( O(1) \) time. As in practice there is usually a small constant number of such sets in considered algorithms, the needed storage is not excessive, e.g., less than 3% on top of a CSR for a graph with the average degree 100 (such as orkut), assuming using 32 threads and the Bron-Kerbosch algorithm, with auxiliary sets \( P, X, \) and \( R \) (the space complexity is \( O(Tn) \) where \( T \) is #threads). We analyze and confirm it for other algorithms and datasets. For example, in SI, the storage complexity is \((TnP)\) (where \( P \) is the size of the subgraph), which is also negligible in practice as \( P \) is usually small. To control space usage, the user may pre-specify that, above a certain number of DBs, SISA starts to use SAs only.

The user controls selecting a set representation. For programmability, SISA offers a predefined graph structure, where small and large neighborhoods are automatically created (when a SISA program starts) as sparse arrays and dense bitvectors, respectively. A given neighborhood \( N(v) \) is stored as a DB whenever \(|N(v)| \geq t \cdot n \) \((t \in (0; 1))\) is a user parameter that controls a “bias” towards using DBs or SAs and it does not exceed a storage budget limit set by the user (SISA by default uses a limit of 10% of the additional storage on top of the graph size when stored only with SAs). For example, \( t = 0.5 \) indicates that each vertex connected to at least 50% of all vertices has its neighborhood stored as a DB.

Figure 4 shows an SA and a DB built from the same vertex set. Then, it illustrates an example SISA graph representation where some neighborhoods are DBs and some are SAs.

### 6.2 High-Performance Set Operations

The second key challenge in SISA is how to apply set operations for highest performance. For this, we detail the algorithmic aspects, a summary is in Table 5. HW details (used PIM and a performance model) are discussed in Section 8. An overview of the structure of SISA is in Figure 3. For each set operation acting on sets \( A \) and \( B \), we provide a number of variants of this operation, where variants differ based on how exactly sets \( A \) and \( B \) are represented.

#### 6.2.1 Set Intersection \( A \cap B \)

is a key operation in SISA, because our analysis illustrates that it is used in essentially all considered graph algorithms. We now briefly discuss the most relevant variants of \( \cap \), a summary is in Figure 4.

- **SA [sorted] \( A \cap SA \) [sorted] \( B \)** The intersection of two sorted SAs is commonly used when processing two neighborhoods. It comes in two "flavors": If \( A \) and \( B \) have similar sizes \((|A| \approx |B|)\), one prefers the **merge** scheme where one simply iterates through \( A \) and \( B \), identifying common elements \((\text{time } O(|A| + |B|))\). If one set is much smaller than the other \((|A| \ll |B|)\), it is better to use the **gallingop** scheme \([1]\), in which one iterates over the elements of a smaller set and uses a binary search to check if each element is in the bigger set \((\text{time } O(|A| \log |B|))\). SISA offers both variants, and a variant that automatically selects the best variant with a performance model (described in § 8.3).

- **SA [unsorted or sorted] \( A \cap DB \) \( B \)** Iterate over \( A \) \((O(|A|))\) and check if each element is in \( B \) \((O(1))\). This variant is often used to intersect a neighborhood with an auxiliary set represented as a bitvector, for example \( X \cap N(v) \) in Listing 2.

- **DB \( A \cap DB \) \( B \)** Apply bitwise AND over both input DBs (they both have sizes of \( n \) bits, giving \( O(n/C) \) time, where \( C \) is the maximum chunk of bits that can be processed in \( O(1) \) time using bit-level parallelism). This variant is used for example when intersecting two dense neighborhoods.

- **SA [unsorted] \( A \cap SA \) [sorted] \( B \)** Iterate over \( A \) \((O(|A|))\) and check if each element is in \( B \) \((O(1))\). This variant is used to intersect a sorted neighborhood and an auxiliary set that is implemented as an unsorted SA \((e.g., P \cap N(v) \) in Bron-Kerbosch, see Listing 2) when it is not uncommon in graph mining algorithms.

#### 6.2.2 Set Union \( A \cup B \), Set Difference \( A \setminus B \)

\( A \setminus B \) and \( A \cup B \) have variants similar to those for \( \cap \), there are also corresponding merge and gallingop variants.

#### 6.2.3 Set Membership \( x \in A \), Set Cardinality \( |A| \)

Set membership takes \( O(|A|) \) time for an unsorted SA (linear scan), \( O(\log |A|) \) time for a sorted SA (binary search), and \( O(1) \) for a DB (a single access to verify if \( x \)-th bit is set). As for set cardinality, we keep \(|A|\) for any set. This incurs only \( O(1) \) storage overhead (per set) as well \( O(1) \) time overhead needed to update the size, but it enables \( O(1) \) time to resolve any set cardinality operation. Finally, SISA provides dedicated instructions for computing cardinalities of the results of set operations, for example \( |A \cap B| \). This enables speedups as SISA avoids creating any intermediate structures needed for keeping the results of operations such as intersection.

#### 6.2.4 Adding and Removing Elements

Auxiliary sets often grow and shrink by one element. Both add and remove straightforwardly...
Table 5: Overview of SISA instructions, one row describes one specific set operation variant. Set elements are vertices (A, B ⊆ V, x ∈ V). "VIRTUAL" means "yes". "na" means "not applicable". "ins" is a proposed instruction opcode. "S (Sorted)" indicates if an instruction assumes set representations of A and B to be sorted (thus two columns).

| ins | Set op. | A and B represent. | Set algorithm | S? | Time complexity | Input size [bits] | Main form of data transfer (§ 8.3) |
|-----|---------|-------------------|---------------|----|-----------------|------------------|-------------------------------|
| 0x0 | A ⊓ B  | SA ⊓ SA           | Merge         | ⊓ | O(|A| + |B|) | W|A| + W|B| | Streaming |
| 0x1 | A ⊔ B  | SA ⊔ SA           | Galloping     | ⊔ | O(|A| log |B|) | W|A| + W|B| | Random accesses |
| 0x2 | A ⊓ B  | SA ⊓ SA           | Merge vs. gallop. | ⊓ | cf. 0x0 and 0x1 | W|A| + W|B| | cf. 0x0 and 0x1 |
| 0x3 | A ⊔ B  | SA ⊔ DB           | Bitwise AND   | ⊔ | na | O(n/(qS)) | n + n | In-situ row copies |
| 0x4 | A ⊔ B  | DB ⊔ DB           | Merge vs. gallop. | ⊔ | na | O(n/(qS)) | n + n | In-situ row copies |
| 0x5 | A∪{x}  | DB ⊔ {x}          | Set bit       | ⊔ | na | O(1) | n + W | Random access |
| 0x6 | A∪{x}  | DB ⊔ {x}          | Clear bit     | ⊔ | na | O(1) | n + W | Random access |

Figure 4: SISA representations of sets and graphs, and processing SISA sets.

Table 6: The impact of set intersection schemes (merging vs. galloping) on the runtime of graph mining algorithms. ★ means that a given SISA variant matches asymptotically the best known non-set-centric baseline, referenced in the top row. k, c, and d denote the size of the mined pattern, the graph degeneracy (a popular measure of graph sparsity) and the maximum vertex degree, respectively (other symbols are described in Section 2). Link prediction complexities are valid for the following vertex similarity measures: ˄Jaccard, Overlap, Adamic Adar, Resource Allocation, Common Neighbors; ˅Total Neighbors; ˅Preferential Attachment [151, 187].

6.3 Additional Details of SISA Design

We detail several aspects of SISA’s design; cf. Figure 3.

6.3.1 Labeled Graphs. As a baseline, we propose to use a sparse array to maintain labels, indexed by vertex IDs, similarly to other works [70]. This form benefits from SISA-PNM. The SISA user can also implement labels with a one-hot encoding and use bit vectors. This would harness SISA-PUM.

6.3.2 SISA Instructions. SISA offers instructions that package the described set operations in all the considered variants, including instructions that automatically select merge or galloping set algorithms (cf. § 6.2). Finally, SISA also provides instructions for creating and deleting sets.

6.3.3 Programming Interface (Set Iterators & Wrappers). For programmability, SISA offers a thin software layer on top of high-level
instructions that consists of abstractions and wrappers. In the former, we provide an opaque type Set that is a reference to a SISA set; this enables using C++ iterators over sets, see left side of Figure 3. In the latter, SISA provides functions that directly map to SISA set instructions. Function parameters determine an instruction variant.

### 6.3.4 Set Identification
SISA identifies sets with unique logical set IDs. These IDs are mapped by the underlying SISA HW design to any used form of physical addresses.

### 6.3.5 RISC-V Compliant Encoding
SISA can be integrated with the RISC-V ISA [251]. To enable modularity and flexibility, SISA’s new instructions are encoded using the custom opcode set [250]. We encode the opcode and functionality of custom RISC-V instructions using bits [31...25] and [6..0], see Figure 5. The former represent the different SISA instructions (up to 128). The latter are set to 0x16 to represent the custom characteristic of the instruction. Fields rs1, rs2, and rd indicate registers with IDs of input sets and the output set, respectively. In Table 5, we assign ISA codes (bits [31..25]) to respective instructions. The number of SISA instructions is less than 20, leaving space for potential new variants.

---

![Figure 5: Encoding of SISA instructions.](image)

### 7 THEORETICAL ANALYSIS

We now show that SISA-enhanced algorithms are theoretically efficient, i.e., their time complexities match those of hand-tuned graph mining algorithms. This is enabled by SISA’s ability to control used set representations and set operations, facilitates tuning performance and storage tradeoffs. To show this, we analyze how varying a used set intersection variant (merge vs. galloping) impacts the runtime of set-centric algorithms, see Table 6. We focus on intersection as it is prevalent in considered algorithms. Crucially, all set-centric variants are able to match the competitive time complexities of considered tuned graph mining algorithms.

#### 7.1 Parametrization with Degeneracy

We parametrize complexities with degeneracy $c$, a well-known measure of graph sparsity [172]. The degeneracy $c$ of a graph $G$ is the smallest number $x$ such that every subgraph in $G$ has a vertex of degree at most $x$ (i.e., every subgraph has at least one sparsely connected vertex). Different graphs have constant degeneracy, such as planar graphs [158], certain scale-free graphs [17, 88], and graphs of bounded treewidth [88]. We consider degeneracy as it is used by many recent graph mining algorithms to enhance their time complexities [78, 91, 282]. This is because several of the investigated graph algorithms orient the graph edges according to the degeneracy order [91, 282]. The purpose of this is to (1) make the graph acyclic (2) make the out-degree as small as possible. The smallest out-degree in a degeneracy ordering is the degeneracy $c$ of the graph.

By definition, the degeneracy is always less than the maximum degree: $c \leq d$. In Table 6, we express the bounds parameterized by $c$ and $d$. To get worst-case bounds that hold for all $d$ and $c$, one can replace $d$ by $n$ and $c$ by $\sqrt{m+n}$ (as $c$ also satisfies $c \leq \sqrt{2m+n} + 1$ [67, 282]). Note that the difference between the maximum degree $d$ and the degeneracy $c$ can be up to $n - 1$: For example, a star graph has maximum degree $n - 1$, but degeneracy 1.

The following observations follow directly form the definitions or the cited literature:

**Observation 7.1** ([67, 282]). For a graph $G = (V, E)$ with degeneracy $c$, $\sum_{(u,v) \in E} \min(d(u), d(v)) \leq 4cn$.

**Observation 7.2.** For every graph $G = (V, E)$, we have that $\sum_{(u,v) \in E} (d(u) + d(v)) = \sum_{i \in V} d(i)^2 \leq md$.

**Observation 7.3.** For a graph $G = (V, E)$ directed according to its degeneracy ordering,

$$\sum_{(u,v) \in E} \left( |N^+(u)| + |N^+(v)| \right) \leq mc.$$

### 7.2 Derivations of Bounds

Next, we discuss proving the bounds in Table 6.

**Triangle Counting** The algorithm iterates over all edges and performs an intersection of the out-neighbor sets of the two endpoints. If Galloping is used, the cost is

$$O \left( \sum_{(u,v) \in E} \min(|N^+(u)|, |N^+(v)|) \log c \right) = O(mc \log c),$$

by Observation 7.1. If Merging is used, the cost is

$$\sum_{(u,v) \in E} \left( |N^+(u)| + |N^+(v)| \right) = O(mc)$$

by Observation 7.3.

**k-Clique Listing** Algorithm 3 has the same cost as the Edge-Parallel algorithm when using Merging. With Galloping, intersections take an additional factor $\log c$.

**k-Clique-Star Listing** The algorithm computes all $k + 1$ cliques and then performs $O(k^2)$ work per clique. There are at most $O(m(c/3)^{k-1})$ cliques of $k + 1$ vertices (as testified by the algorithm that lists them in $O(km(c/2)^{k-1})$ time [78]).

**Jarvis-Patrick Clustering** Jarvis-Patrick Clustering (for any of the vertex-similarity measures from Corollary 9) iterates over edges and perform a set intersection on the neighbors. The difference to triangle counting is that the graphs are not oriented according to a degeneracy ordering. This changes the cost for the Galloping approach to $O(mc \log d)$ (where only the term in the logarithm chances as we may need to search sets of size up to $d$ instead of only up to $c$). The Merge approach costs $\sum_{(u,v) \in E} (d(u) + d(v)) = O(md)$, by Observation7.2. Similar discussions apply to macbLink Prediction.

**Maximal Cliques** Computing maximal cliques takes $O(cn^{3/2})$ time when both good pivoting and the degeneracy ordering are used [91]. If appropriate pivoting is used (without the degeneracy
We now discuss details of SISA hardware implementation.

8 HARDWARE IMPLEMENTATION

8.1 Processing-In-Memory for Sets

We start with how SISA uses PIM for set operations.

**SISA-PUM** First, the intersection, union, and difference of sets represented as DBs are processed with SISA-PUM that relies on in-situ DRAM bulk bitwise schemes. For concreteness, we pick Ambit [217], a recent design that enables energy-efficient bulk bitwise operations fully inside DRAM, by small extensions to the DRAM circuitry but without any changes to the DRAM interface. However, SISA is generic and other designs could also be used (e.g., ELPI2M [253], DRISA [155], ComputeDRAM [96], PCM (Pinatubo) [156]). The key extension in Ambit (for in-situ processing) is to modify a decoder for three selected DRAM rows (that share the same set of sense amplifiers) in such a way that one amplifier connects directly to three DRAM cells. This enables logical AND and OR over two of such three rows, immediately computing a result in-situ, and again one amplifier connects directly to three DRAM cells. This enables logical AND and OR over two of such three rows, immediately computing the result in the third row (NOT is provided by including a single row of dual-contact DRAM cells [217]). Importantly for SISA-PUM, only three selected designated DRAM rows (per single DRAM sub-array) are modified this way. Whenever the running code requests an in-situ memory operation, Ambit uses a recent RowClone technology [216] to copy (also in-situ) the rows that store input sets to these two designated rows, compute the result in-situ, and again use RowClone to copy the result to the destination (unmodified) DRAM row. Now, SISA-PUM uses Ambit’s execution model and interface without any modifications: set intersection and union are processed with an in-situ AND and OR, respectively. Set difference is processed using set intersection, along with the well-known set algebra rule: \( A \setminus B = A \cap B' \) [129]. Whenever needed, the negation \( B' \) can be derived with the in-situ NOT.

**SISA-PNM** A set operation with no bulk bitwise processing uses SISA-PNM that relies on high bandwidth between processing units and DRAM (as in UPMEM [147], HMC [130], or Tesseract [6]). Adding or removing an element from a set stored as a DB (\( A \cup \{x\}, A \setminus \{x\} \)) is conducted with a single DRAM access to a specific memory cell. Other set operations on SAs that employ either streaming (e.g., merge \( \cap \)) or random accesses (e.g., galloping \( \cap \)) are also executed using small in-order cores. Here, we rely on the high TSV enabled bandwidth for high performance of set operations dominated both by data streaming (merge) and random accesses (galloping).

8.2 SCU & Automatizing SISA Decisions

We use a small SISA Control Unit (SCU), cf. Section 3, to automatically decide on (1) selecting the PNM or PUM execution, and (2) merge or galloping. Once the host core decodes a SISA instruction, it passes it to the SCU. The SCU further decodes this instruction, and picks either PNM or PUM to execute the instruction. For deployment, SCU could either be added to the CPU or to the DRAM circuitry (see the feasibility discussion later in this section), or – to avoid any HW modifications – it can also be emulated by a single designated in-order logic layer core. SCU does not implement any complex logic (e.g., dynamic set modifications), it only decides on variants of schemes to execute.

**SISA-PUM & SISA-PNM** First, SCU decides whether to use SISA-PUM or SISA-PNM for given two sets. This decision is simple and is based on how sets are represented (this information is stored in a simple in-memory SM ("set metadata") structure and possibly cached in SCU’s cache).

**Variants of Set Operations** Second, SCU automatically detects if it is best to use merge or galloping, and processes input sets using the corresponding variant. This decision is guided by our performance models.

8.3 Performance Models for Set Operations

The runtime of each SISA instruction variant is dominated by either streaming or random accesses.

**Streaming** takes place when two sets \( A \) and \( B \) stored as SAs are processed using merging. We model the runtime as \( I_M + W \cdot \max(|A|, |B|) \cdot \min(b_M, b_L) \). \( I_M \) and \( b_M \) are latency and bandwidth of accessing DRAM, and \( b_L \) is bandwidth between cores. The model conservatively assumes that \( A \) and \( B \) may be located in memory locations attached to different cores (e.g., in different vaults), and thus (1) the overall bandwidth is bottlenecked by \( \min(b_M, b_L) \), and (2) we can use \( \max(|A|, |B|) \) as \( A \) and \( B \) are streamed in parallel.

To model **random accesses**, we simply count the number of performed operations and multiply it by the memory access latency. This gives \( I_M \cdot \min(|A|, |B|) \cdot \log(\max(|A|, |B|)) \) for a binary search over the larger of input sets, used when processing two SAs with galloping. Then, a specific variant is **selected automatically** to minimize the predicted runtime. To **parametrize** these models, SISA needs (1) the sizes of processed sets, (2) their representation types, and (3) \( b_M, b_L, I_M \) (1) and (2) are maintained in the metadata structure. (3) describe the execution environment and are thus identical for each set; they are stored directly in the SCU. We instantiate (3) to reflect logic layers in Tesseract [6].

8.4 Details of SISA Hardware

We now present various details on SISA HW.

**Life Cycle of a Set** Any SISA graph application is a series of standard instructions as well as SISA instructions that load, store, and process sets. A set is allocated with a standard malloc, augmented with setting the appropriate set information in the set metadata (SM) structure. Loading, processing, and storing sets is conducted by the respective existing elements such as logic layer cores; the SCU is only responsible for selecting the appropriate instruction variant to be executed. Once a set is deleted, the standard
free call is used, together with removing the respective entry from the SM structure.

**Set Metadata** SM forms a simple associative structure that holds constant amount of data per set (representation, set size). The total SM size is \( O(n) \) as there are \( n \) neighborhoods and a constant number of auxiliary sets. Thus, while we conservatively assume that SM is an in-memory structure, in practice it fits completely in cache or a small scratchpad. This is because many datasets processed by graph mining algorithms have small \( n \), in the order of hundreds or thousands [203]. These graphs pose computational challenges, but these challenges come from high computational complexities (e.g., listing maximal cliques is NP-hard) or from relatively high edge counts \( m \) (as some vertices may have high degrees [203]), but not (or to a smaller extend) from \( n \). Whenever the given SM information is not cached, there is a single additional memory access for one set operation. The SM information is used by the SISA performance models when deciding which set algorithm to execute. Yet, other information could also be stored – for example, when using other set representations such as sparse bitmaps [1, 112]. We plan on extending SISA with such schemes as future work. Each SM entry describing one set also contains the set location. Now, entries in the SM structure are indexed by set IDs. A set ID is returned by a function creating a set, cf. Figure 3. Set IDs and set creation (and destruction) calls are used by a developer analogously to pointers and malloc/free calls.

**Caching Set Metadata** Depending on how SISA HW is deployed, the SM information can be cached in either a small dedicated scratchpad or cache (if the SCU is implemented as an additional circuitry), or in the standard cache of a logic layer core (if the SCU is emulated by a such designated core).

**SISA-PNM and SISA-PUM Together** Ambit fully preserves the DRAM interface: the sets are always stored in standard DRAM rows, and moved to the designated rows only for bulk bitwise processing [217]. SISA-PNM accesses run on unmodified DRAM banks (the modifications in PNM are only related to the high bandwidth, and the SCU in SISA). Thus, SISA-PNM and -PUM are seamlessly used together.

**Harnessing Parallelism** SISA HW harnesses memory parallelism at different levels, enabling parallel execution of both a single set operation and different set operations. First, bit-level parallelism is enabled by using Ambit’s bulk bitwise operations: bits in a row are ANDed or ORed in parallel. Second, pairs of bitvectors placed in different subarrays (or, e.g., DRAM banks) can be processed in parallel. Third, processing pairs of sets stored as integer arrays in different vaults can also be parallelized. Here, SISA benefits from the same effect of bandwidth scalability as the Tesseract graph accelerator [6].

**Managing Concurrency** For simplicity, SISA relies on developers using established techniques (locks, lock-free protocols, general parallel programming principles [117] and libraries such as OpenMP [61]) to concurrently access the same set. Thus, designing a parallel graph mining algorithm that uses SISA is analogous to non-SISA based algorithms.

For **cache coherence in SISA-PUM**, we rely on mechanisms (provided by the memory controller) that flush dirty cache lines in source rows, and invalidate cache lines in destination rows. Existing schemes also rely on it, including Ambit [217], DMA accesses [69] and others [121, 216]. As in Ambit, SISA-PUM accesses are always row-wise, and thus we can also rely on Dirty-Block Index [215] and similar schemes for fast data flushing. Invalidations run in parallel with Ambit operations and thus do not incur overheads.

**Memory Layout and Storage of Sets** Advanced schemes for the layout of vertices and edges in different sets (e.g., spreading large sets across different vaults) are beyond the scope of this work. We ensure that storing SISA sets is feasible (i.e., a maximum-size neighborhood, represented as SA or DB, fits into a single vault).

### 8.5 SISA Hardware Cost and Feasibility
We also briefly discuss the hardware cost. First, the needed DRAM modifications are minimal and identical to those already discussed in Ambit. Second, as the logic to be implemented in SCU is straightforward decision making on what instruction variant to use, its costs are not prohibitive, as shown by many designs proposed in the past, for example in HyVE [122] (a hybrid vertex-edge memory hierarchy that uses ReRAM and DRAM) or in GraphH [75] (an accelerator that combines HMC with SRAM). Third, the code of all SISA instructions is also straightforward: a simple binary search (galloping), merging of two arrays (merge), or setting/clearing a DRAM cell (set element add/remove). Thus, they can be trivially deployed in in-order cores in the logic layer of 3D stacked DRAM, as shown by other designs [75].

**Integration with RocketChip** To facilitate a potential real SISA implementation, we outline the integration of SISA and RocketChip [14]. Custom SISA instructions as specified in our encoding (§ 6.3) are forwarded to the SMU and SLB tandem, which replaces the RoCC accelerator component [14] of a Rocket/Boom tile. The SLB is connected to the cache network and the SMU is connected directly to the memory bus. When the Rocket/Boom core receives an instruction whose opcode and functionality bits match those of SISA, the core forwards the instruction, using the existing RoCC interface [14], to the SMU. When the SMU completes its operation, it signals the core to continue executing the application.

### 9 EVALUATION
We illustrate example performance advantages from SISA. Due to a very large evaluation space, we provide summaries; all results follow similar patterns.

**9.1 Methodology, Setup, Parameters**
We first present our simulation setup.

**Simulation Infrastructure** We use Sniper [115] with the Pin frontend [166]. Sniper is a popular cycle-level simulator used in many works proposing various architectural extensions for both CPUs and memory subsystem [179, 242].

**SISA Implementation** We simulate the SISA HW design and the ISA, instrumenting the code so that the simulation toolchain can distinguish between SISA and non-SISA instructions. To model each component of SISA, we add the respective set instructions and simulate the SCU (a small fixed delay), the cache in SCU (with the LRU policy), the SM structure (random memory accesses whenever the SCU cache is not hit), and the execution of all used set operations by appropriate delays in the simulation execution. For operations based on streaming and random memory accesses, we use the...
To simulate SISA-PUM, we model a run-time of in-situ operations with a delay \( t_M + t_L \cdot \left\lceil \frac{n}{q} \right\rceil \); \( t_M \) is the latency to access DRAM (to initiate the operation) and \( t_L \) is the latency to execute one in-situ instruction. \( \left\lceil \frac{n}{q} \right\rceil \) models a scenario when the bitvector size \( n \) exceeds the size of all DRAM rows that can be processed in parallel. \( q \) is the count of rows within a bank that can be used in parallel and \( R \) is the size of one row.

**SISA Platform & Parameters** For concreteness, we set the platform for executing SISA instructions to match Tesseract [6] (for SISA-PNM) and Ambit [217] (for SISA-PUM). The former has simple in-order cores (1 core/vault in its logic layer) with 32 KB L1 instruction/data caches, no L2, 16 8GB HMCs (128 GB in total), 32 vaults/cube, 16 banks/vault. Each vault offers 16 GB/s of memory bandwidth to its core. Thus, we assume scalable bandwidth as proposed by Tesseract: using more vaults increases the total memory bandwidth. We set the DRAM row rank size to 8 KB, following Ambit [217]. Next, we set the parameter \( t \in [0; 1] \) (that controls the bias towards using DBs or SAs to store neighborhoods) to 0.4 (i.e., 40% of neighborhoods are stored as DBs); we also analyze other values. We ensure that the total storage used for neighborhoods does not exceed the size of the simple CSR graph storage by more than 10%. Finally, we set the size of SISA SCU’s cache to be 32 KB (matching Tesseract’s L1).

**Platform for non-SISA Instructions & Baselines** For any non-SISA instructions and baselines, we use a high-performance Out-of-Order manycore CPU. Each core has a 128-entry instruction window, a branch predictor, 32 KB L1 instruction/data caches, a 256 KB L2 cache. All cores share an 8 MB L3 cache. There is also a four-way associative 64-entry D-TLB, a 128-entry I-TLB, and a 512-entry S-TLB. For fair comparison, we also use bandwidth scalability in this configuration, i.e., we increase the memory bandwidth with the number of cores, matching it with that of SISA-PNM.

**Considered Mining Problems** The graph mining problems we consider are clustering with the Jaccard (cl- Jac), overlap (cl-ovr), and total neighbors (cl-tot) coefficients, listing \( k \)-cliques (kcc-\( k \), \( k \in \{ 4, 5, 6 \} \)), \( k \)-cliq-star (ksc-\( k \), \( k \in \{ 4, 5, 6 \} \)), maximal cliques (mc), triangles (tc), and subgraph isomorphism (si-k-\( k \) for \( k \)-stars).

**Comparison Targets: Hand-Tuned Algorithms** Our most important (the most challenging to outperform) baselines are hand-optimized parallel algorithms for each graph mining problem. Specifically, we use a tuned version from the GAP Benchmark Suite [21] for tc, Eppstein’s version of BK for mc [91], Danisch’s scheme for kcc-\( k \) [78], enhanced Jabbour’s scheme for ksc-\( k \) [125], parallel VF2 for si-\( k \)-\( k \) [70], and cl-\( k \)-jac based on counting triangles in the GAP suite [21]. All used baselines have competitive work and depth complexities, cf. Table 6. For fair comparison, all baselines benefit from the high bandwidth of PIM. We consider two classes of baselines: algorithms that do not explicitly use set algebra (denoted with _non-set) and their set-centric variants (denoted with _set-based). SISA variants are indicated with _sisa.

**Comparison Targets: Pattern Matching Frameworks** SISA and its underlying paradigm do not aim to outperform specific accelerators but complement or reinforce them, by offering a novel set-centric paradigm and building blocks. Thus, we focus on comparing to the fundamental paradigms / algebras that underlie these accelerators: neighborhood expansion for pattern matching implemented in Peregrine [127] (which represents GRAMER [265]) and relational algebra implemented in RStream [244] (which represents TrieJax [136]). We stress that, while we consider these baselines for completeness, we focus on comparing to (much faster) hand-tuned parallel algorithms for solving specific problems.

**Comparison Baselines & PIM** For fairness, all considered comparison targets take advantage from the high bandwidth of PIM setting (parametrized identically to that of SISA).

**Graphs** We select different input datasets (Table 7) from Network Repository [204], considering biological (bio-), interaction (int-), brain (bn-), economics (econ-), social (soc-), scientific-computing (sc-), discrete-math (dimacs-), and wikitext (edit-) networks. We pick graphs with different structural properties (low/high density, small/large maximum degree, low/high degree distribution skew, etc.).
9.2 Discussion of Results

We now proceed to analyze the results.

Comparison to Hand-Tuned Algorithms We first analyze run-times with all available cores, comparing SISA set-centric variants to non-set-based and set-based hand-tuned parallel baselines that all benefit from high-bandwidth storage. The results are in Figure 6. SISA is almost always the fastest by a large margin of at least 2x, often more than 10x (than non-set schemes). The differences vary depending on the processed graphs and the considered problem. Gains are usually larger on graphs with large maximum degrees, such as brain graphs, where SISA-PUM is used more often to directly process sets inside DRAM, reducing the latency. Such graphs are prevalent in many computational domains [204], and this is the case for the majority of considered datasets.

Algorithmic vs. Architectural Speedups We also observe speedups from using only set-centric formulations (over non-set-based variants). Namely, speedups of "set-based" schemes over the "non-set" ones indicate gains from purely algorithmic (set-centric) changes, while speedups of "sisa" schemes over the "set-based" indicate gains only from architectural changes (i.e., from using PIM). First, the differences between _set-based and _non-set heavily depend on the targeted mining algorithm. These speedups are particularly visible for more complex algorithms such as m, with multiple nested loops and/or recursion. Packaging different parts of such algorithms into, e.g., set intersections, and being able to control the used operation variant (e.g., merging based on streaming) helps to utilize features such as high sequential bandwidth. Contrarily, for certain simpler schemes such as clustering, the very tuned _non-set baseline outperforms _set-based (while still falling short of _sisa). Second, the difference between _set-based and _sisa depend more on the used graph. Here, in many cases, _sisa is only marginally faster than _set-based, because the graph structure (e.g., sizes of neighborhoods) favor using SAs rather than DBs, diminishing benefits from SISA-PUM (e.g., for econ-graphs) and equalizing the differences between both _set-based and _non-set take advantage from the high bandwidth setting. In other cases (e.g., bionumber-LC), more vertices have large enough degrees to benefit from DBs and low latencies of SISA-PUM.

Labels We also analyze labeled SI. Most often, labeled graphs are faster to process. Despite more memory accesses, the labels form additional constraints, which eliminates some recursive calls earlier, resulting in performance gains.

Scalability We also analyze how run-times change when varying numbers of threads T, for a fixed graph size ("strong scaling"), and when increasing T proportionally to the graph size ("weak scalability"). To fix the used graph model, we use Kronecker graphs [153] and we vary the number of edges/vertex. SISA maintains its speedups, but they become less distinctive when T is small. This is expected because fewer threads exert less pressure on the memory subsystem, and there is overall smaller potential from using PIM in SISA.

Large Graphs We execute SISA on several large graphs, including the Orkut social network with 117M edges, see Figure 8. Runtime benefits from SISA and the set-centric formulations are similar to those in smaller graphs in Figure 6. The only two graphs

Figure 6: Run-times with full parallelism. The red line cuts off of long simulation runtimes for readability (the bars reaching the line have much larger runtimes). No bar indicates the timeout of the respective baseline (>24h). The results for cl-jac (clustering based on the Jaccard coefficient) are very similar to those that use other coefficients and for link prediction as well as vertex similarity. All 32 cores are used. Acronyms are stated in “Comparison Targets: Hand-Tuned Algorithms”. Graph classes: biological (bio), brain (bn), interaction (int), informative (inf), social (soc).

where SISA and non-SISA set baselines are comparable, are sc-pwtk and soc-orkut. This is because these networks, due to their origin (social and scientific) do not have large cliques or very dense clusters (unlike, e.g., genome graphs), somewhat lowering SISA benefits.
Comparison to Other Paradigms We compare SISA set-centric algorithms to neighborhood expansion and relational algebra paradigms, representing frameworks such as Peregrine or RStream, and accelerators such as GRAMER or TriaJax. Peregrine is able to express only listing \( k \)-cliques and subgraph isomorphism, and maximal clique listing in a limited way (i.e., it does not offer a native scheme for MC and we implemented it by iterating over possible clique sizes and listing maximal cliques of each size), RStream can only find \( k \)-cliques. In each case, SISA baselines are much faster: 10-100× than Peregrine (and more than 1,000× for MC due to Peregrine’s inability to natively support MC), and more than 100× for RStream. This is because the underlying paradigms focus on programmability in the first place, sacrificing performance, while in SISA we start with tuned graph algorithms and only then restructure them with the set-centric paradigm.

Sensitivity Analysis & Design Exploration We investigate the impact from varying SISA parameters.

SCU cache Not using the SCU cache lowers performance by \( \approx 1.5 \times \) for \( T = 1 \) and \( \approx 0.05-0.1 \times \) for \( T = 32 \). The lower performance for high \( T \) is because, with more threads executing set operations, it becomes more difficult to ensure high hit ratio. Overall, the behavior of the SCU cache is similar to that of other such units such as L1, including varying cache parameters such as size.

PNM vs. PUM & Sparse/Dense Neighborhoods PNM and PUM are synergistic in SISA. PNM cores handle sparse neighborhoods and SAs well, as they offer low latency and bandwidth proportionality. PUM is well-suited for large neighborhoods stored as DBs (common in considered graphs due to their degree distribution skews). Yet, SISA-PUM adds overheads when using it for sparse sets due to low utilization of very sparse rows. Thus, it is relevant to not choose the DB bias parameter to be too high. We find that 0.4 works well for most processed graphs. We illustrate this in Figure 7b, where we analyze how the performance changes when varying the fraction of largest neighborhoods stored as DBs. Smallest and largest fractions that correspond to using only SISA-PNM or only SISA-PUM, while technically feasible, give slowest runtimes. We also vary the “galloping threshold”, i.e., the relative difference between two sets that causes the set operation to switch to the galloping variant. For example, the value of 5 indicates that galloping is used if any of the two sets is at least 5× larger than the other one. While this threshold influences performance, the general pattern stays the same.

We also analyze the impact from degree distributions of datasets, see Figure 7a. Graphs often used in graph mining, such as biological networks, that SISA focuses on, have often very heavy tails. This implies many large neighborhoods and very dense large clusters, benefiting from SISA-PUM. For example, the human genome graph has many vertices connected to more than 30% of all other vertices. Other graphs such as social networks have much lighter tails, cf. soc-orkut and sc-pwt in Figure 7a. This is because these networks, due to their origin (social, scientific) do not have large cliques or very dense clusters. Such graphs benefit less from SISA-PUM. Still, using SISA-PNM enables high performance, outperforming tuned non-set-based baselines, cf. Figure 8.

Load balancing Figure 9a illustrates total fractions of time during which each parallel thread is stalled when executing a given algorithm. SISA stall times are low because its design implicitly tackles two types of load imbalance. First, SISA’s performance models enable adaptive selection of the best variant of a set algorithm to be executed for any two sets. This minimizes load imbalance from processing two sizes that differ a lot in sizes. Second, load imbalance due to processing imbalanced pairs of sets (i.e., two very

Figure 7: Figure 7a: Differences between degree distributions in graphs used mostly in graph mining and the ones used also outside graph mining (on the right). Figure 7b: Sensitivity analysis: the percentage of neighborhoods stored as dense bitvectors vs. different thresholds for using the galloping or the merging intersection.

Figure 8: Run-times for large graphs. Graph classes: biological (bio), brain (bn), interaction (int), informative (inf), social (soc). 8 cores are used.
small and two very large sets) is alleviated by the fact that very large pairs of sets are processed with very fast SISA-PUM.

**SCU cache: shared vs. private** We also explore sharing the SCU cache among all the cores. While possibly increasing the hit rate, a single shared cache has higher access latency. This has a small (<1%) yet noticeable slowdown effect in our simulations. A potential remedy would be to include multiple SCU cache levels. To keep the core logic simple, we do not explore it further, and leave it for future work.

We also show that the reduced simulation runtimes do not artificially eliminate load imbalance. We gather traces of executed set operations in full vs. partial simulation executions, and we plot histograms of the sizes of processed sets, see Figure 9b. In both types of executions, we encounter large sets which are the primary source of load imbalance.

![Histograms of sizes of processed sets, see Figure 9b.](image)

**Figure 9: Load balancing analysis.**

**SISA Limitations** For some graphs with small maximum degrees (e.g., soc-fbMsg) in Figure 6, SISA speedups are smaller, or even (in the extreme cases) result in slowdowns. This is because the benefits from SISA-PUM, or from the automatic selection of the most beneficial set operation variant, are out-weighted by having to process too many large bitvectors (that have always size $n$ bits). This effect rare, and it can be alleviated by reducing the number of neighborhoods stored as DBs. In this case, the performance of SISA variants gradually converges towards that of standard CSR based set-centric algorithms. We plan on addressing it with advanced bitvector representations.

10 RELATED WORK

In developing an ISA extension for graph mining, we follow recent footsteps of other specialized ISAs, for example a quantum ISA [225], a neural ISA [160], or even ISAs for managing cloud resources [94, 116].

Related graph processing paradigms (Table 1) and software efforts are described in Section 1 [33, 33, 39, 167, 208]. We now briefly summarize other related areas. First, we conducted an exhaustive analysis of existing hardware accelerators as well as ISA designs for graph processing, see Table 8. The analysis indicates that SISA offers the only hardware acceleration for a broad family of problems such as maximal clique listing or clustering. The closest designs [136, 198, 265] only focus on selected pattern matching problems. Next, some works target hardware accelerated dynamic (time-evolving) graph analytics [29, 38, 55, 56, 114]. Such problems are outside the scope of this work. Moreover, several analyses illustrate how to efficiently use existing hardware for graph analytics, but purely from a software development perspective [82, 89, 272].

**Such works are orthogonal to SISA.** Several works focus on external memory graph processing in the context of hardware acceleration [82, 134, 171]. One could possibly use these designs as other SISA backends for external memory set instructions; we leave this for future work.

While in the current SISA version we focus on implementing and executing set operations in set-centric algorithm formulations using PIM, SISA could be extended into different directions. This includes parallel and distributed execution of set operations, and implementing them using high-performance techniques such as Remote Direct Memory Access [32, 34, 99, 212]. One could also enable more efficient execution of set-centric graph mining algorithms in the context of modern complex heterogeneous architectures that may host massively parallel on-chip networks [31], NUMA and systems with locality effects [214, 235], or FPGAs [28, 40, 80]. One could also incorporate various forms of graph compression and summarization [35, 41, 43, 162].

**Graph Algorithms and Set Algebra** Sets are used in different graph algorithms, to simplify operations on selected data structures [39, 52, 142, 177, 194, 211, 219]. For example, the BFS frontier can be modeled as a set. Similarly, inserting and removing an element from the frontier was often modeled with inserting and removing an element from a set [20]. The only work (that we know of) which focuses on graph processing and sets is due to Han et al. [112] and Aberger et al. [1], where the authors accelerate set intersections for graph analytics. Contrarily, SISA is the first attempt to accelerate general graph mining by identifying different set operations used in these algorithms, formulating these operations as an ISA extension, and supporting these instructions with in-situ and near-memory acceleration. Here, SISA's main contribution is not to simply use set notation. Instead, from the algorithmic perspective, SISA is the first design that (1) uses set operations as the primary building blocks, which break down complex graph mining algorithms into simple units of parallel execution, and (2) identifies the appropriate set operations (i.e., operations that are easily accelerated with PIM) and reformulates selected algorithms so that they use such operations, cf. Table 3.

**Set Programming** Some works propose to use sets as a basis for general programming to enhance coding productivity, with use cases in software prototyping. These works include SETL [139, 213], ISETL [86], and CLAIRE [57]. These efforts do not focus on graph processing or improving performance.

**SISA vs. AutoMine** AutoMine [174] uses set operations to express finding graph patterns. It focuses on automatic compilation of set schedules into efficient code. This part is orthogonal to our work and AutoMine could easily be combined with SISA to,
| Reference / Accelerator | Prob. | Key memory mechanism | Pattern M. Learning | “Low-c” | ia | si | vs | lp | el | av | bf | pr | ce |
|-------------------------|-------|----------------------|---------------------|------------------|-----|----|----|----|----|----|----|----|----|----|
| [P1] GauX-X [60]        | SpMV  | [e] CAM/MAC          | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P1] GraphSAR [76]      | ver-c | [e] ReRAM           | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P1] GraphDelta [12]    | low-c | [e] DRAM            | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P1] GraphMa [154]      | edge-c| [e] DRAM            | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] GraphVine [24]     | ver-c | [e] 3D DRAM         | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] ReGraph [159]      | BFS   | [e] ReRAM           | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] Sparsa [275]       | ver-c | [e] ReRAM           | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] GraphQ [285]       | ver-c | [e] HMC             | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] GraphS [13]        | low-c | [e] SOT-MRAM        | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] RAGra [123]        | ver-c | [e] 3D ReRAM        | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] GRAM [276]         | ver-c | [e] ReRAM           | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] Messagefusion [23] | ver-c | [e] HMC             | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] Missaayhi et al. [160] | low-c | [e] HMC             | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] RPFBS [111]        | BFS   | [e] ReRAM           | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] GraphD [227]       | SpMV  | [e] ReRAM           | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] GraphSP [270]      | ver-c | [e] HMC             | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] Tesseract [6]      | low-c | [e] HMC             | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] PIM Enabled [7]    | low-c | [e] HMC             | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] Gu et al. [97]     | low-c | 3D DRAM             | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |
| [P2] LAM [283, 284]     |       | SpMSpM              | × × × × × × × × × × |     |    |    |    |    |    |    |    |    |    |

Table 8: Comparison of SISA to graph-related accelerators, focusing on supported graph mining problems and offered architecture elements. **”**: Support / significant focus. **”**: Partial support / some focus. **“**: no support / no focus. Addressed problems: see Table 1 for details; ver-c: vertex-centric, edge-c: edge-centric, low-c: general low-complexity problems, SpMV: sparse matrix-vector products, SpMSpM: sparse matrix-sparse matrix products, GCN: graph convolution networks, GNN: graph neural networks. Graph problems and algorithms: as in Table 1. Considered architecture and stack elements: "is" is an ISA, or its extensions, "xl" is a cross-layer design, "ab" is a programming paradigm Classes of accelerators: **[P1]**: in-situ PIM, **[P2]**: near memory PIM (e.g., logic layers), **[A]**: ASIC, **[M]**: focus on memory hierarchy enhancements, **[F]**: FPGA, **[e]** focus on extensions and modifications to the established (already proposed) HW technology, *Applies to some works in a given group. Note that the generality of SISA comes from harnessing all basic set algebra operations.
We develop the first hardware acceleration for broad graph mining. First, we offer a set-centric programming paradigm, where one identifies and exposes set operations in graph mining algorithms, resulting in “set-centric” algorithmic formulations. This enables competitive time complexities and succinct formulations. We support labeled graphs and non anti-monotonic set operations [136, 174, 198, 265].

Second, the set-centric algorithms are mapped to SISA, a small yet expressive family of instructions that form a "set-centric" ISA extension for graph mining. SISA could be extended into multiple directions, for example, with CISC-style set instructions that accept multiple arguments (e.g., to intersect multiple sets in a single instruction \(A_1 \cap \ldots \cap A_p\)) to facilitate optimizations such as vectorization with loop unrolling. Due to the generality of set algebra, we predict that SISA can be used for problems beyond graph mining and general static graph computations, for example dynamic (time-evolving) graph processing, or data mining beyond graphs.

Third, we pick in-situ and logic layer PIM for hardware acceleration, and offer automated selection of the most beneficial instruction variants, maximizing speedups over hand-tuned baselines of parallel graph mining algorithms. However, the interface based on set algebra could use other hardware backends for SISA instructions. For example, one could use a GPU backend for fast SIMD-based set intersections [112], implement set operations on FPGAs [40], execute set operations in caches [3, 185], or use ReRAM [227] for efficient in-memory analog matrix-vector multiplications, which can also be used to implement some instances of set intersection.

Finally, our cross-layer architecture could also be extended in other directions, for example by providing compiler support for generating SISA programs from set-centric formulations. Here, one could use, e.g., AutoMine’s [174] compiler generated schedules as input to some SISA programs.

ACKNOWLEDGMENTS

We thank Mark Klein, Hussein Harake, Colin McMurtrie, Angelo Mangili, and the whole CSCS team granting access to the Ault and Daint machines, and for their excellent technical support. We thank Timo Schneider for his immense help with computing infrastructure at SPCL. This project received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 programme (grant agreement DAPP, No. 678880).

REFERENCES

[1] Christopher R Aberger, Andrew Lamb, Susan Tu, Andres Nötzli, Kunle Olukotun, and Christopher Ré. 2017. Emptyheaded: A relational engine for graph processing. ACM Transactions on Database Systems (TODS) 42, 4 (2017), 1–44.
[2] Abraham Addisie, Hiwot Kassa, Opeoluwa Matthews, and Valeria Bertacco. 2018. Heterogeneous memory subsystem for natural graph analytics. In 2018 IEEE International Symposium on Workload Characterization (ISWC). IEEE, 134–145.
[3] Shaizeen Aga, Supreet Jeloka, Arun Subramaniam, Satish Narayanasamy, David Blazew, and Reetuparna Das. 2017. Compute caches. In 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 481–492.
[4] Charu C Aggarwal and Hanxin Wang. 2010. Managing and mining graph data. Vol. 40. Springer.
[5] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. Citeseer, 487–499.
[6] Junwan Ahn, Sungjoo Hong, Sungjoo Yoo, Onur Mutlu, and Kiyoung Choi. 2015. A scalable processing-in-memory accelerator for parallel graph processing. In ISCA.
[7] Junwan Ahn, Sungjoo Yoo, Onur Mutlu, and Kiyoung Choi. 2015. PIM-enabled instructions: a low-overhead, locality-aware processing-in-memory architecture. In Computer Architecture (ISCA), 2015 ACM/IEEE 42nd Annual International Symposium on. IEEE, 336–348.
[8] Sam Ainsworth and Timothy M Jones. 2018. An event-triggered programmable prefetcher for irregular workloads. ACM SIGPLAN Notices 53, 2 (2018), 578–592.
[9] Mohammad Al Hasan, Vaneet Chaujo, Saeed Salem, and Mohammed Zaki. 2006. Link prediction using supervised learning. In SDM06: workshop on link analysis, counter-terrorism and security.
[10] Mohammad Al Hasan and Vachik S Dave. 2018. Triangle counting in large networks: a review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8, 2 (2018), e1226.
[11] Mohammad Al Hasan and Mohammed J Zaki. 2011. A survey of link prediction in social networks. In Social network data analytics. Springer, 243–275.
[12] Shaashin Angizi and Deliang Fan. 2019. GraphIDE: A graph processing accelerator leveraging in-dram-computing. In Proceedings of the 2019 on Great Lakes Symposium on VLSI ‘19–50.
[13] Shaashin Angizi, Jiao Sun, Wei Zhang, and Deliang Fan. 2019. GraphS: A graph processing accelerator leveraging SOT-MRAM. In 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 378–383.
[14] Krste Asanovic, Rimas Avizienis, Jonathan Bachrach, Scott Beamer, David Bianchi, Christopher Celio, Henry Cook, Daniel Dabbelt, John Hauser, Adam Israelevitz, et al. 2016. The rocket chip generator. EECs Department, University of California, Berkeley, Tech. Rep. UCB/Eecs-2016-17 (2016).
[15] Osama G Attia, Tyler Johnson, Kevin Townsend, Philip Jones, and Joseph Zambreno. 2014. CyGraph: A Reconfigurable Architecture for Parallel Breadth-First Search. In 2014 IEEE International Parallel & Distributed Processing Symposium Workshops (IPDPSW). IEEE, 228–235.
[16] Jonathan W Bahb, Matthew Frank, and Anant Agarwal. 1996. Solving graph problems with dynamic computation structures. In High-Speed Computing, Digital Signal Processing, and Filtering Using Reconfigurable Logic, Vol. 2914. International Society for Optics and Photonics, 225–237.
[17] Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. Science 286, 5439 (1999), 509–512. https://doi.org/10.1126/science.286.5439.509.arXiv https://science.sciencemag.org/content/286/5439/509.full.pdf
[18] Akshay Basak, Shuangchen Li, Xing Hu, Sang Min Oh, Xinfeng Xie, Li Zhao, Xiaowei Jiang, and Yuan Xie. 2019. Analysis and optimization of the memory hierarchy for graph processing workloads. In 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 373–386.
[19] Omar Batari, Radwa El Shewi, Ayman G Fayouni, Reza Nouri, Ahmed Barnawi, and Sherif Sakr. 2015. Large scale graph processing systems: survey and an experimental evaluation. Cluster Computing 18, 3 (2015), 1189–1213.
[20] Scott Beamer, Krste Asanović, and David Patterson. 2013. Direction-optimizing breadth-first search. Scientific Programming 21, 3–4 (2013), 137–148.
[21] Scott Beamer, Krste Asanović, and David Patterson. 2015. The GAP benchmark suite. arXiv preprint arXiv:1508.03619 (2015).
[22] Scott Beamer, Aydin Buluc, Krste Asanovic, and David Patterson. 2013. Distributed memory breadth-first search revisited: Enabling bottom-up search. In 2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum. IEEE, 1618–1627.
[23] Leul Belayneh, Abraham Addisie, and Valeria Bertacco. 2019. Messagefusion: On-path message coalescing for energy efficient and scalable graph analytics. In 2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED). IEEE, 1–6.
[24] Leul Belayneh, Abraham Addisie, and Valeria Bertacco. 2020. GraphVine: exploiting multiscalar for scalable graph analytics. In 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 762–767.
[25] Tal Ben-Nun, Maciej Besta, Simon Huber, Alexandros Nikolaou Zogias, Daniel Peter, and Torsten Hoefler. 2019. A modular benchmarking infrastructure for high-performance and reproducible deep learning. In 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). IEEE, 66–77.
Yossi Shiloach and Uri Vishkin. 1982. An O (log) parallel connectivity algorithm. *Journal of Algorithms* 3, 1 (1982), 57–67.

Julian Shun and Guy E Blelloch. 2013. Ligra: a lightweight graph processing framework for shared memory. In *ACM SIGPLAN Notices*, Vol. 48. 135–146.

Julian Shun and Kanat Tangwongsan. 2015. Multiscope triangle computations without tuning. In *Data Engineering (ICDE)*, 2015 IEEE 31st International Conference on. IEEE, 149–159.

Shreyas G Singapura, Ajijeth Sriavastava, Rajgopal Kannan, and Viktor K Prasanna. 2017. OSCAR: Optimizing SCAttach reuse for graph processing. In *2017 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 1–7.

S Skiena. 1990. Dijkstra’s algorithm. Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica. Reading, MA: Addison-Wesley (1990), 225–227.

Robert S Smith, Michael J Curtis, and William J Zeng. 2016. A practical quantum instruction set architecture. *arXiv preprint arXiv:1608.03535* (2016).

Edgar Solomonik, Maciej Besta, Flavio Vella, and Torsten Hoefler. 2017. Scaling betweeness centrality using communication-efficient sparse matrix multiplication. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 47.

Linghao Song, Youwei Zhuo, Xuehui Qian, Hai Li, and Yiran Chen. 2018. GraphR: Accelerating graph processing using ReRAM. In *High Performance Computing Computer Architecture (HPCA), 2018 IEEE International Symposium on*. IEEE, 531–543.

Victor Spirin and Leonid A Mirny. 2003. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences* 100, 21 (2003), 12123–12128.

Narayanan Sundaram, Nadafithur Sathar, Md Mostofa Ali Patwary, Subramanya R Dullor, Michael J Anderson, Satya Goutam Vadlamudi, Dipankar Das, and Pradeep Dubey. 2015. Graphmat: High performance graph analytics made productive. *Proceedings of the VLDB Endowment* 8, 11 (2015), 1214–1225.

Michael Sutton, Tai Ben-Nun, and Amnon Barak. [n. d.]. Optimizing Parallel Graph Connectivity Computation via Subgraph Sampling. [n. d.]

Ichigaku Takigawa and Hiroshi Mamitsuka. 2013. Graph mining: procedure, modules in molecular networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 11 (2013), 50–57.

Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller. 2004. Link prediction in relational data. In *Proceedings of the 21st international conference on Machine Learning*. ACM, 406–413.

Adrian Tate, Amir Kamil, Anshu Dubey, Armin Grollinger, Brad Chamberlain, Brice Goglin, Carter Edwards, Chris J Newburn, David Padua, Didem Unat, et al. 2014. Programming abstractions for data locality. *FADL Workshop*, April 24–29, Swiss National Supercomputing Center ...

Carlos HC Teixeira, Alexandre J Fonseca, Marco Serafini, Georgos Siganos, Brice Goglin, Carter Edwards, Chris J Newburn, David Padua, Didem Unat, et al. 2016. *Graph Connectivity Computation via Subgraph Sampling*. [n. d.]

Mingyu Yan, Lei Deng, Xing Hu, Ling Liang, Yujing Feng, Xiaochun Ye, Zhimin Zhang, Dongrui Fan, and Yuan Xie. 2020. Hyagen: a gen accelerator with hybrid architecture. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 15–29.

Mingyu Yan, Xin Liu, Zhong Chen, Li, Aniir Maks, Hau Lin, Xia Mian, Itir Akgun, Yujing Feng, Peng Gu, and Lei Deng. 2019. Alleviating irregularity in graph analytics acceleration: A hardware/software co-design approach. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, 655–672.

Chengbo Yang. 2018. An Efficient Digger for Large Scale Graph Processing on an OpenCL-based FPGA. In *arXiv preprint arXiv:1806.11509* (2018).

Tianyang Yang, Zhaozhi Deng, Zhiwei Liu, Shousui Yin, Jiachen Li, and Leibo Liu. 2020. *GraphABC*: Scaaling Out Graph Analytics with Asynchronous Block Coordinate Descent. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 419–432.

Pengcheng Yao, Long Zhang, Zhen Zeng, Yu Huang, Chuangyi Gui, Xiaofei Liao, Hai Jin, and Jingling Xue. [n. d.]. A Locality-Aware Energy-Efficient Accelerator for Graph Mining Applications. [n. d.]

Fengcheng Yao, Long Zhang, Zhen Zeng, Yu Huang, Chuangyi Gui, Xiaofei Liao, Hais Jin, and Jingling Xue. 2020. A Locality-Aware Energy-Efficient Accelerator for Graph Mining Applications. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 895–907.

Andy Yoo, Edmund Chow, Keith Henderson, William McLendon, Bruce Hendrickson, and Umey Catalyurek. 2005. A scalable distributed parallel breadth-first search algorithm on the BlueGene/L. In *Proceedings of the 2005 ACM/IEEE conference on Supercomputing*. IEEE Computer Society, 25.

Xiangyao Yu, Christopher J Hughes, Nadathur Satish, and Srinivas Devadas. 2015. IMP: Indirect memory prefetcher. In *Proceedings of the 48th International Symposium on Microarchitecture*. IEEE, 178–190.

Jiali Zhang, Soroosh Khoram, and Jing Li. 2017. Boosting the Performance of FPGA-based Graph Processor Using Hybrid Memory Cube: A Case for Breadth First Search. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA ’17)*. ACM, New York, NY, USA, 207–216. https://doi.org/10.1145/3020078.3021737
