aDFS: An Almost Depth-First-Search Distributed Graph-Querying System

Vasileios Trigonakis, Jean-Pierre Lozi, Tomáš Faltín, Nicholas P Roth, Iraklis Psaroudakis, Arnaud Delamare, Vlad Haprian, Călin Iorgulescu, Petr Koupy, Jinsoo Lee, et al.

aDFS: An Almost Depth-First-Search Distributed Graph-Querying System

Vasileios Trigonakis,[1] Jean-Pierre Lozi,[1] Tomáš Faltín,[1,2] Nicholas P. Roth,[3] Iraklis Psaroudakis,[1] Arnaud Delamare,[1] Vlad Haprian,[1] Călin Iorgulescu,[1] Petr Koupy,[1] Jinsoo Lee,[1] Sungpack Hong,[1] and Hassan Chafi[1]

[1]Oracle Labs, firstname.lastname@oracle.com [2]Charles University
[3]KUNGFU.AI (work done while at Oracle Labs), nicholas.roth@kungfu.ai

Abstract

Graph processing is an invaluable tool for data analytics. In particular, pattern-matching queries enable flexible graph exploration and analysis, similar to what SQL provides for relational databases. Graph queries focus on following connections in the data; they are a challenging workload because even seemingly trivial queries can easily produce billions of intermediate results and irregular data access patterns.

In this paper, we introduce aDFS: A distributed graph-querying system that can process practically any query fully in memory, while maintaining bounded runtime memory consumption. To achieve this behavior, aDFS relies on (i) almost depth-first (aDFS) graph exploration with some breadth-first characteristics for performance, and (ii) non-blocking dispatching of intermediate results to remote edges. We evaluate aDFS against state-of-the-art graph-querying (Neo4J and GraphFrames for Apache Spark), graph-mining (G-Miner, Fractal, and Peregrine), as well as dataflow joins (BiGJoin), and show that aDFS significantly outperforms prior work on a diverse selection of workloads.

1 Introduction

Graph processing is a very active area of research, with a plethora of prior work focusing on classic graph algorithms [35, 36, 38, 48, 53, 58, 61, 62, 82, 85], graph mining [27, 29, 32, 34, 42, 54, 69, 74, 77], as well as graph querying [1, 3, 8, 10, 12, 18, 31, 86] and graph-query languages [4, 5, 11, 14, 17]. Graph algorithms (such as PageRank [57]) are typically used in batch computations, while graph mining is used to extract structural properties and compute cumulative statistics of a graph by exploring its subgraph structures.

Graph queries are a key tool for graph analysis, as indicated by the large number of existing systems and graph-query languages. Graph queries provide an expressive interface for interactive graph exploration with rich dynamic projection and filtering support that is analogous to SQL for relational databases (see Section 5 for further details). They focus on data connections, i.e., edges, allowing users to submit queries with any pattern, filter, or projection. For instance, the following PGQL [14] query:

```
SELECT a1.name, a2.name, a1.country = a2.country,
       ABS(a1.salary - a2.salary) AS salary_diff
MATCH (a1:author)-[:likes]->(a2:author),
      (a2)-[:likes]->(a1)
WHERE ABS(a1.age - a2.age) <= 10
ORDER BY salary_diff DESC
```

enumerates the `author`s of similar age that `like` each other. Answering such a query requires finding all homomorphic [45] matches of the query pattern in the target graph, while enforcing filters (e.g., `a1 IS author`) and projecting the requested output (e.g., whether `a1.country = a2.country`).

The dynamic user-defined patterns, filters, and projections, the focus on edges, and the homomorphic matching make graph query execution a challenging workload that needs to handle very large intermediate and final result sets, with a combinatorial explosion effect. For example, on the well-researched Twitter graph [47], the single-edge query `(a)->(b)` matches the whole graph, amounting to 1.4 billion results, and the two-edge query `(a)->(b)->(c)` amounts to 9.3 trillion matches. This means matching the `(a)->(b)->(c)->(a)` cycle needs to consider 9.3 trillion intermediate results. Compared to relational queries, graph queries can exhibit extremely irregular access patterns [51, 63] and lack of spatial locality, while calling for low-latency data access.

High-performance graph-querying systems ideally need to (i) keep the computation in main memory to guarantee low latency, (ii) scale out to multiple machines in a distributed manner to handle graphs and queries that exceed the capacity of a single machine, and (iii) control their memory usage at the machine level. Controlling memory consumption during query execution becomes paramount for cloud graph-processing services, in which multiple users submit queries that produce results of unpredictable size. Allowing a single query to monopolize memory would hinder service quality for users running other queries.

Query execution on graphs is typically based on one of the two classic graph-traversal strategies: Breadth-first search (BFS) or depth-first search (DFS). Both BFS and DFS have major advantages and drawbacks for distributed graph queries:

BFS traversals are easier to parallelize but, as with distributed joins, suffer from explosion in the size of intermediate results, cannot be easily pipelined, and stress the network bandwidth to shuffle data across levels of pattern matching. DFS traversals reduce the size of intermediate results, but are challenging to parallelize and result in random data access patterns, wasting locality when iterating over neighbors.

In this paper, we introduce aDFS (almost-DFS): A novel distributed graph-querying system that brings the best of both DFS/BFS worlds. aDFS extends the graph-processing capabilities of PGX.D [39] with queries and processes graphs partitioned across multiple machines *fully in memory*, combining BFS and DFS traversals to *bound the maximum amount of memory* required for query execution, while achieving a *high degree of parallelism*. DFS, together with a distributed flow-control mechanism, guarantee that the amount of runtime memory remains within limits, while the BFS exploration allows for better locality and parallelization during execution.

Worker threads in aDFS mainly prioritize DFS execution for completing—and thus freeing—intermediate results. The execution switches to BFS when matching a remote edge (i.e., an edge pointing to a remote machine) or when the runtime detects that the query contains limited parallelism (i.e., a small set of intermediate results). To elaborate, for local edges, worker threads perform DFS, unless aDFS detects that there is a limited amount of available work on the local machine, in which case they switch to per-thread BFS exploration until there is enough parallelism. For remote edges, threads buffer the matched intermediate results and continue with matching the next edge in a BFS manner (i.e., the next edge is possibly at the same depth as the current one). Once a buffer is full, the worker thread sends its contents to the target machine, unless it is blocked by the flow-control mechanism, which enforces target memory limits. Section 3 expands on the design and implementation of aDFS.

Section 4 thoroughly evaluates aDFS and shows that it is capable of executing trillion-scale queries, with a 10GB per-machine runtime memory cap. When running our largest query, aDFS computes a 9.3 trillion count pattern on the Twitter graph with a rate of 7.3 billion matches per second. We compare aDFS to two graph systems (i.e., Apache Spark GraphFrames [31] and Neo4j [10]) and two relational databases (i.e., MonetDB [9] and PostgreSQL [15]) using the LDBC graph and query suite [68]. aDFS completes the set of queries 43 and 53 times faster than GraphFrames and Neo4j,[1] respectively, and 8 and 26 times faster than MonetDB and PostgreSQL, respectively (as Section 4.2 shows, LDBC is "relational-friendly"). We also compare aDFS to these four systems with schema-less graphs and show that either aDFS is 16 to 9,200 times faster than the rest, or the other systems simply fail to complete the queries. Finally, we compare aDFS with (i) three state-of-the-art graph-mining systems: G-Miner [27], Fractal [32] and Peregrine [42], as well as (ii) BiGJoin [19], a dataflow join system. We show that aDFS is up to 12, 625, and 18 times faster than G-Miner, Fractal and Peregrine, respectively, and performs comparably to BiGJoin on mining-oriented workloads. We discuss related work further in Section 5.

The main contributions of this paper are the following:

- aDFS, which is, to the best of our knowledge, the first graph-querying system that strictly bounds runtime memory while operating with fully-distributed computations over partitioned graphs;
- The novel combination of DFS (for eager completion of intermediate results), BFS (for performance), and flow control (for controlling the size of the intermediate state) to achieve performance and scalability while capping memory usage; and
- The evaluation of aDFS, which shows that aDFS significantly outperforms the state of the art and is capable of executing queries with trillions of matches.

2 Background and Motivation

Representing data as a graph is becoming increasingly popular. The main advantage of graphs is that they focus on modeling fine-grained relationships between entities. In contrast, the relational model concentrates more on rows and relies on the heavyweight primary-key foreign-key (PK-FK) and join mechanisms to link entities. However, when using graphs, different models, data representations, and ways of exploring them have a major effect on processing performance.

2.1 The Property Graph (PG) Model

Property graphs represent the graph topology as vertices and edges, and store *properties* and *labels* separately. Properties can be associated to any vertex or edge and take the form of typed key-value pairs. Labels are key-only and represent types or categories, e.g., *person* or *animal*. Separating the topology from properties avoids the proliferation of edges and allows for quick traversals of the graph over its real structure.

2.2 Graph Pattern-Matching Queries

Several languages for graph querying exist, such as PGQL [14], SPARQL [17], Gremlin [5], and Cypher [11]. In its simplest form, graph querying makes it possible to find patterns in graphs, with filters and projections. aDFS uses PGQL, which is modeled after SQL: Projection and aggregation operations are the same as in SQL, including `GROUP BY` and `ORDER BY`. PGQL adds support for graph patterns and vertex and edge labels. For example, the query presented in Section 1 adds the `MATCH` clause to an otherwise valid SQL query. It matches patterns that are *homomorphic* [45] to the `(a1)->(a2)->(a1)` cycle while enforcing filters (e.g., `a1` has label `author`), and it projects or aggregates the requested data—including even arbitrary expressions—out of the matched vertices and edges.

Graph queries require homomorphic matching of the pattern, as data is projected out of all matches, even if they are permutations of each other. In contrast, graph-mining systems

[1] Using Neo4j Community Edition (benchmarks not audited by Neo4j).

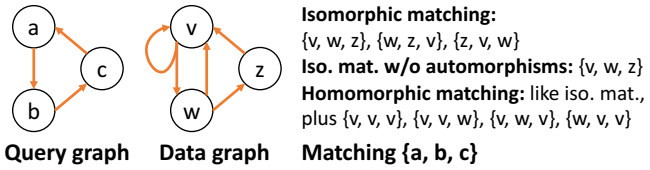

Figure 1: Homomorphic vs. isomorphic matching of a pattern.

focus on *isomorphic matching* [34] and often use automorphism elimination to prune the search space further. Figure 1 highlights the differences between isomorphic and homomorphic matching of a query pattern. In graph queries, isomorphic matching and automorphism elimination can be simulated with filters and/or with query specifiers, such as `GROUP BY`.

In this work, we focus on the backbone of graph pattern-matching queries. Accordingly, aDFS only supports a subset of PGQL 1.1 [14]; in particular, two following features are missing: Regular path expressions and subqueries. Nevertheless, we design aDFS with these features in mind and we intend to work on them in future work.

2.3 Graphs vs. Relational Joins and RDF

In PG graph systems, edges are stored explicitly and can be traversed directly. In contrast, in relational databases, relationships are represented with PK-FK. Following any relationship means joining two tables—or doing a self-join if the keys belong to the same table—and producing the intermediate result. Therefore, while matching multiple-hop paths is a relatively cheap operation in graph systems, doing the same in SQL requires a chain of multiple expensive join operations that materialize intermediate results. Thus, graph systems can be much more efficient than relational databases when it comes to matching graph patterns (see Section 4 for a comparison).

Another alternative model is the Resource Description Framework (RDF) that uses {*subject*, *predicate*, *object*} triples to represent graphs, which can be queried with languages such as SPARQL [17]. RDF became popular with the semantic web [21] and has been the model of choice for many graph databases starting in the early 2000s [37, 80]. A number of works have focused on distributed RDF graphs [40, 64].

Although the RDF model is equivalent to PG in terms of expressiveness, there are differences: (i) RDF adds links for every graph data piece, including constant literals, (ii) it does not have explicit vertices/edges—yet it can be viewed as representing graphs, and (iii) and it does not store properties separately. The de-facto implementation of RDF triples results in similar PK-FK behavior as aforementioned for relational databases. Triples force RDF systems to process and join a much larger number of intermediate results using e.g., a key-value-style storage, and lose the graph structure, resulting in slower neighbor lookup. To address these drawbacks, some RDF systems use asynchronous processing [37], or compute graph indices, using e.g., the CSR representation, to mock the graph structure [80]. aDFS focuses on PGQL queries and the PG model, avoiding complex and expensive joins. We note that the pattern-matching part of query execution is largely orthogonal to the graph model and aDFS's techniques could be used for RDF graphs.

2.4 DFS, BFS, and Intersections for Graph Exploration

DFS can expand one intermediate result at a time, starting from the first variable in the pattern and continuing to the next ones until the whole pattern is matched. However, this behavior results in totally random accesses and is impractical for distributed graph traversals: The only way to continue with strict DFS is to directly send the intermediate result to the remote machine and wait until it is picked up and completed.

Thus, graph exploration is traditionally done using BFS: For each query edge (hop), the entire result set is computed, and only then does the exploration of the next hop start. This approach has two main advantages: (i) it is easy to implement, as work is naturally divided into simple steps (hops), and (ii) it is relatively easy to parallelize, as the entire input is known before processing a hop (of course, skewed vertex degrees still pose a problem). However, BFS has one major shortcoming: Because the intermediate result set is produced between stages, an intermediate result-set explosion can quickly occur.

Figure 2 illustrates this issue showing the average total per-machine memory usage and execution time when matching cycles of various lengths using aDFS and BFS (implemented in our runtime) on a small graph [16] (875K vertices and 5.1M edges). While both approaches are able to match cycles of length one to four with similar performance, the memory consumption of BFS explodes for five-hop cycles at approximately 60GB on each of the eight machines in the experiment, and BFS crashes with six-hop cycles after 96 minutes when one machine runs out of memory (~768GB). Meanwhile, the memory consumption of aDFS is almost constant.

Recent graph-mining and graph-querying systems [22, 42] adopt a pattern-matching approach that relies on intersecting neighborhood lists. Instead of being vertex-centric (i.e., starting from vertices and following edges), the intersection approach focuses on edges. The benefit of the intersection-based model is that it takes $O(|V|)$ steps since it allows intersecting multiple incoming edges at a time, as compared to the vertex-based approaches that are $O(|E|)$. However, intersections require complete subgraph parts to operate. This necessitates pulling/gathering possibly large amounts of data from remote machines. To make things worse, queries enumerate all automorphisms (i.e., the exploration space could locally explode) and offer arbitrary user filters and projections, meaning that in an intersection-based model, one would need to pull not only the vertex/edge data, but also all the properties required by the query. Therefore, we use a vertex-centric approach in aDFS that builds mini-frontiers based on the first query vertex and enables aDFS to operate on fully partitioned graphs with limited memory.

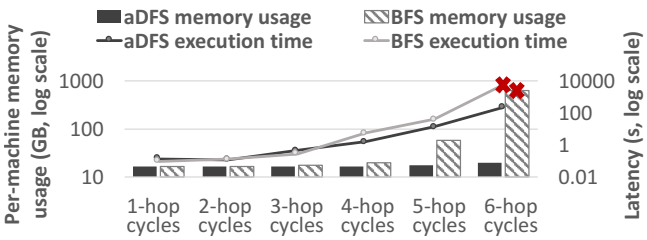

Figure 2: Matching cycles using aDFS vs. BFS.

3 aDFS: A Pattern Matching and Querying System for Distributed Graphs

The main design goals of aDFS are (i) enabling fast, fully in-memory distributed queries of any size, while (ii) allowing for limited, controllable memory consumption during execution. The rationale for these two goals is as follows. First, high-performance graph queries demand in-memory execution and the ever-increasing size of data calls for distribution. Second, server systems, especially in cloud deployments, are shared by multiple concurrent users, hence no single query can be permitted to saturate the system memory. aDFS achieves these two goals through the following design principles.

1. **§3.3: DFS-first and asynchronous communication.** The eager match completion of DFS gives aDFS fine-grained control on the size of intermediate results during query execution, but strict DFS would be inefficient when matching a remote edge, i.e., an edge that leads to a remote machine. For that reason, worker threads do not block when encountering a remote edge, but place the intermediate result in a message buffer and continue with other local work instead. Buffers batch intermediate results: once full, a buffer's contents are asynchronously sent to the remote machine for further processing. Threads only need to block if flow control dictates so. This buffering results in essentially BFS exploration of the remote edges of a vertex.
2. **§3.4: Flow control.** Cross-machine communication is controlled through a flow-control mechanism that caps the number of in-flight intermediate result buffers. The finite nature of these message buffers allows strictly configuring the amount of runtime memory that aDFS requires, while the flow-control mechanism guarantees query termination and deadlock freedom.
3. **§3.5: Dynamic local DFS/BFS.** Besides the BFS style of buffering for remote edges, aDFS includes a dynamic approach for deciding whether to go DFS or expand with BFS for local matches in order to improve parallelism, locality, and work sharing across threads.

Before diving into these design principles, we first present the architecture of aDFS from a high-level point of view (Section 3.1) and describe how aDFS generates execution plans for graph queries (Section 3.2).

3.1 High-Level aDFS Architecture

Figure 4 shows the high-level architecture of aDFS. Graphs are kept in memory and are partitioned across machines based on simple random vertex partitioning. Random partitioning achieves cross-machine balance and does not overfit to the workload. Of course, intelligent partitioning schemes could bring performance benefits and are left for future work. aDFS's approach is orthogonal to partitioning, i.e., it can work with any partitioning approach.

For efficient traversals, graphs are stored in the classic CSR (Compressed Sparse Row) graph format. Due to graph partitioning, messaging is necessary for moving intermediate results to the machine which holds the target vertex. aDFS maintains two threads on dedicated cores on each machine for messaging; a sender and a receiver. Consequently, worker threads in aDFS place their messages in software queues, from where they are picked up by the sender.

3.2 Distributed Query Execution Planning

Users submit declarative PGQL queries [14] to aDFS. As Figure 3 illustrates, each query goes through three transformation steps (marked i through iii) before being executed in step iv.

Step i: Logical query planner. The first step translates the PGQL query into a logical query plan, which consists of the logical operators of Table 1. Similar to relational query planning, a given query can be executed by multiple logical query plans. In the example of Figure 3, an alternative plan could rewrite the query as `(a)-[e]-(c)->(a)->(b)`. This first step directly translates the query to an admissible plan, which is then optimized in the following steps.

Step ii: Distributed query planner and optimizer. This step specializes the logical query plan by taking into account the specific characteristics of aDFS's runtime. The query planner rewrites the logical plan in terms of *stages* and transitions from one stage to another (called *hops*). A stage is responsible for matching or accessing exactly one vertex and contains all the information necessary for matching the corresponding vertex and for transitioning to the next vertex with a hop. In the example of Figure 3, the topmost stage "a" matches the first vertex `a` of the query, while the next one matches `b`. An out-neighbor hop takes the execution from `a` to `b`.

aDFS supports four types of hops that specialize for distributed execution: *neighbor match*, *edge match*, *output*, and *inspection*. Neighbor and edge hops have the same behavior as the corresponding logical operators in Table 1. An *output* hop produces a final match using the current intermediate result and is always used in the last stage of a match.

Inspection hops are specific to distributed processing: They bring the current intermediate result back to an already matched vertex in order to continue query evaluation. In the example of Figure 3, after matching `a` and `b` of `(a)->(b)`, the query again needs the neighbor list of the already matched vertex `a` in order to continue with matching `(a)<-(c)`. Since the matched vertex `b` might be in a different machine than `a`,

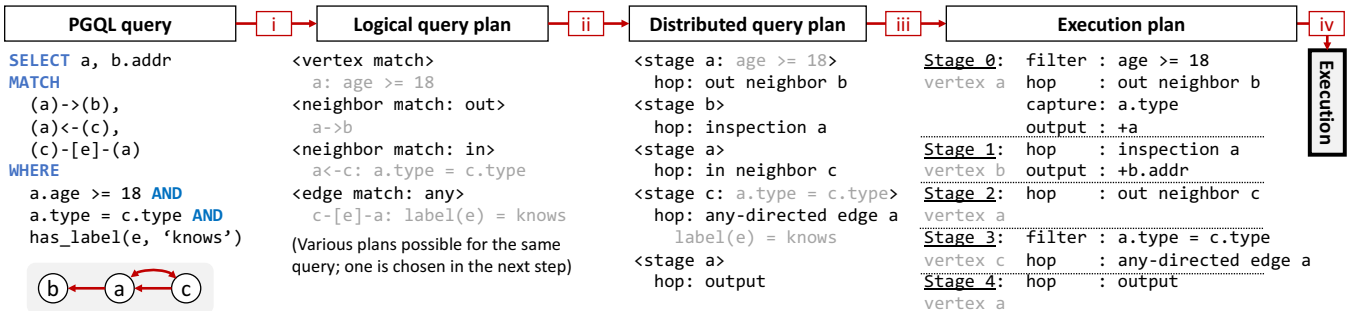

Figure 3: From a PGQL query to aDFS execution. Three transformation steps before execution.

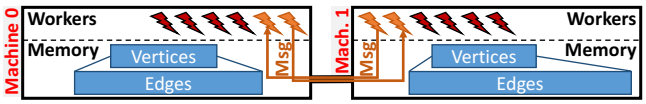

Figure 4: High-level architecture of aDFS.

the query planner introduces an inspection step to "link" this disconnected pattern and bring back the context to the machine of `a`. If `a` resides in the current machine, an inspection hop is essentially a no-op.

In this step, aDFS rewrites the logical query plan with a cost-based optimizer, implemented using dynamic programming, that is based on the following heuristics: (i) heavily filtered vertices are preferred for the earlier stages of the plan, (ii) inspection hops are not free and increase the plan's cost, and (iii) the cost of an edge hop is approximately *log* of the cost of a neighbor hop, as it can be implemented with a binary search in the neighbor list of the source vertex. The optimizer further detects whether a query has a single starting vertex, by extracting ID equality filters (e.g., `ID(person) = 123`). In the example of Figure 3, the optimizer rewrites the query as `(a)-[e]-(c)->(a)->(b)` because it avoids an inspection hop and `a` and `e` are more filtered as compared to `b`.

Steps iii–iv: Execution plan and execution. Finally, aDFS generates a concrete execution plan. Apart from stages and hops, the execution plan contains filters (on vertices and edges), as well as information on what data should be included in the intermediate results in order to execute filters of later stages and produce the final output. For example, in the query of Figure 3, Stage 0 must collect `a.type`, since it is required by the filter of Stage 3. Similarly, Stage 0 must put vertex `a` in the intermediate result as it is part of the projection of the query. Overall, each stage builds up the intermediate result such that another thread, local or remote, can pick it up and continue the computation. The resulting execution plan is then submitted to the aDFS runtime, on which we focus next.

Op.	Example	Short description
Vertex match	`(x)`	Match vertices of the graph (without following any edge)
Neighbor match	`(x)->(y)`	Having matched the left vertex `x`, match its neighbors `y`; can be in-, out-, or any-directional
Edge match	`(x)->...(y)->(x)`	The vertex `x` is known (already visited)—test whether `x` exists in the neighbor list of the left vertex `y`; can be in-, out-, or any-directional

Table 1: Graph operators used in the logical query plan.

3.3 aDFS's Depth-First Runtime

The runtime of aDFS is based on the *stage* and *hop* constructs described above. aDFS initiates query execution by applying Stage 0 (matching of the first vertex variable of the execution plan) to each vertex of the graph. This bootstrapping process happens (i) across machines, i.e., each machine starts from the locally-stored vertices, and (ii) concurrently within each machine, i.e., each worker thread handles a distinct set of vertices and performs the bootstrapping process on these vertices one after the other. Hops that follow remote edges send the intermediate match (batched) to the destination remote machine where they are picked up and taken over by a local thread.

Bootstrapping a match. Figure 5 includes a high-level activity diagram of the aDFS runtime. Completing the execution of this diagram from Stage 0 to the last query stage implements the complete matching starting from a single vertex of the graph. We explain these steps using the example of Figure 6. Text in the *blue italic face* represents the activities in Figure 5. The aDFS runtime assigns vertex `Joe` (the dark gray rounded rectangle of Figure 6) to a worker thread `t`, which tries to generate new matches. The thread first tries to match `Joe` with Stage 0's `p1` using *apply stage*. If the filter `p1.name = "Joe"` returned `FALSE`, the thread would try to *backtrack* to a previous stage and, because there is none, it would simply complete this invocation. If there were more top-level vertices to explore, `t` would start again with a different vertex.

In the example of Figure 6, we assume that the execution plan matches vertex `p1` as Stage 0. `p1` matches `Joe` and `t` continues with the *hop: follow next edge* operation, starting from edge ①. Since the `:friend` label filter is satisfied and the edge is local, `t` proceeds via *DFS next stage* to Stage 1 where `p2` is matched with the vertex that has `age = 20`. At this point, since the filter `p2.age < 35` is satisfied and there is no next stage, `t` produces a query output row and *backtracks*

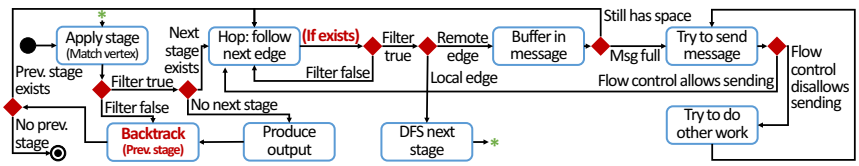

Figure 5: Matching operations starting from a given vertex. The **backtrack** activity represents returning to the previous DFS stage. Similarly, if the conditional in the red bold font returns false, the execution backtracks to the previous stage (if any).

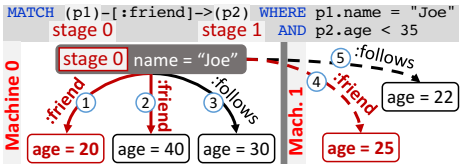

Figure 6: Example graph query execution. Rounded rectangles represent vertices, red vertices and edges are matched.

to Stage 0 to continue with the next edge. At this point, edge ② with label `:friend` is matched but since it leads to the vertex with `age = 40`, the filter on `p2` is not satisfied. Backtracking to Stage 0 brings us to edge ③ with label `:follows`, which is not matched.

Thread `t` is now done with local edges and starts processing the remote ones (aDFS does not necessarily match all local edges first). The first one, edge ④, has label `:friend`, thus `t` places the current intermediate result in a messaging buffer targeting Machine 1 (*buffer in message*). Once the buffer is filled up, `t` *tries to send a message* with the contents of the buffer to the destination. As Section 3.4 describes in more detail, flow control might temporarily block `t` from sending the message; in that case, `t` *tries to do some other work* (e.g., handling an incoming message). Once the thread returns from performing these other tasks, it retries sending the blocked message. Finally, `t` attempts to match the last remote edge ⑤, which does not match because of its label. With all the edges of vertex `Joe` explored and no previous stage to backtrack to, `t` completes the invocation.

Handling incoming messages (intermediate results). Workers eagerly try to receive and process remote messages, always prioritizing the latest stage with available work. Threads try to process messages: (i) before starting new work, i.e., before *apply stage* at Stage 0 (new top-level vertex), (ii) when flow control (temporarily) disallows message sending—in that case, the impacted thread picks up a new message to process while waiting for flow control to release the blocked message, and (iii) once the matching operations (see Figure 5) have completed on all local vertices—at that point, workers continuously wait for incoming messages to complete any pending work from remote machines.

3.4 Flow Control

aDFS allows specifying the total memory size `M` of the messaging buffers that hold the intermediate results in any machine, making it possible to cap runtime memory utilization. Besides these buffers, aDFS only needs a small per-thread, per-stage, additional memory allocation to hold the current ongoing local match and metadata for thread blocking.

In order to enforce this memory cap, aDFS employs a simple flow-control protocol. aDFS partitions the buffers that hold intermediate results across the query stages, such that no stage can consume all buffers (required to prevent starvation). When a buffer with intermediate results is full, the corresponding worker requests permission to send the contents of the buffer to the target machine. The flow-control protocol keeps track of the amount of data `D` that has been sent to that machine but not yet processed. If `D` is above a threshold (computed based on the memory cap `M`; a machine does not accept more than `M / #Machines` worth of intermediate results from any other), flow control blocks the message transmission (controlled per stage, not for the whole query) and the thread continues with some other work before retrying to send the message. Once a message has been processed, the handling thread informs the source machine that its chunk of intermediate results has been completed and makes the corresponding memory available for another message. Note that this simple protocol *strictly bounds memory consumption*, i.e., no pattern can violate the memory configuration of aDFS.

Flow-control performance. We evaluate the performance overheads of flow control; see Section 4 for details on the detailed experimental setup. Figure 7 compares the query execution latency without flow control (i.e., all messages of intermediate results are sent as produced) and with different per-machine flow-control limits in aDFS. In this experiment, we use a buffer size of 256KB and eight machines. The per-machine limit N is the total number of outgoing buffers that the query execution is allowed to have, therefore it also dictates the maximum amount of memory M that a machine can use during the execution of the query. Since all intermediate results could be targeting a single machine at some point during execution, $M = N \times (\text{size of one buffer}) \times (\#\text{ machines})$.

We execute simple `SELECT COUNT(*)` queries that include basic patterns such as `(a)->(b)->(a)` (Q1 and Q2) and `(a)->(b)->(c)` (Q3 to Q6), with different filters. The results show that aDFS is not very sensitive to the flow-control limit, unless the limit is very low, e.g., 512 messaging buffers. In that case, the flow control only allows a single outstanding message per worker, per stage, per machine.

Figure 8 gives more insights in the execution of Q3 with Livejournal: `SELECT COUNT(*) MATCH (a)->(b)->(c).` The figure shows the maximum number of incoming and outgoing messages for the busiest stage on any of the eight machines, as well as the number of cases in which the flow-control limits were reached. For very low limits ($N = 512$ messages) the

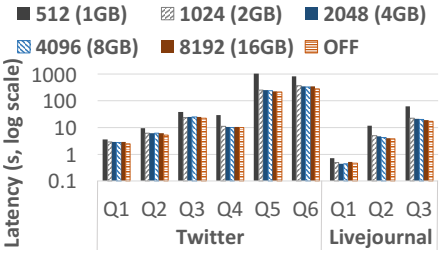

Figure 7: Performance of simple queries (8 machines) with different flow-control limits. In parentheses: Total per-machine maximum memory consumption.

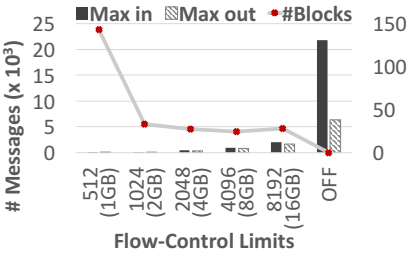

Figure 8: Messaging and blocking statistics on Q3/Livejournal with different flow-control limits. In parentheses: Total per-machine max. memory consumption.

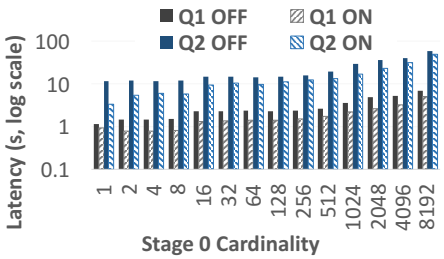

Figure 9: aDFS with dynamic local-edge BFS "ON" or "OFF" for two queries while varying the number of intermediate results of the first query stage.

amount of blocking is very high, which penalizes performance (more than 3× higher latency). Still, the overhead for switching stages due to flow control is generally low: Setting N to 8,192 results in only ~10% performance loss as compared to no flow control (`OFF`), while reaching 10× fewer maximum incoming messages (2,087 vs. 21,793) and 4× fewer outgoing messages (1,636 vs. 6,430).

3.5 Dynamic BFS for Local Edges in aDFS

For remote edges, aDFS essentially does (per-thread) BFS: A thread matching a remote edge simply buffers the intermediate result and continues exploring and matching the same stage, which might produce new intermediate results.

While local processing could happen in pure DFS, doing so can result in artificially limited parallelism for queries that produce small sets of intermediate results. A characteristic example is queries with a very narrow starting Stage 0, such as `MATCH (a)->... WHERE ID(a) = X`; this narrow-start behavior appears in several real-life queries (e.g., the LDBC queries of Section 4). In such a query, the whole Stage 0 might produce a single intermediate result, giving limited opportunities for parallelism. For these workloads, DFS can significantly delay the expansion of intermediate results that are produced in the system (both locally and through messages).

In aDFS, we solve this DFS limitation by dynamically switching from depth-first exploration to per-thread breadth-first for local edges. aDFS maintains per-stage counts of the number of buffers with intermediate results that are ready to be taken care of by worker threads. A low number of intermediate results means that the stage has not expanded enough, hence some threads could end up not having sufficient work to perform. When threads in aDFS are processing a local edge, they use this information to decide whether to go for BFS, i.e., buffer the intermediate result in a local buffer and continue at the same stage.

In practice, we keep these local buffers small, i.e., up to a few kilobytes, in order to promote quick local work creation. We further use a *DFS threshold* to decide when to work depth-first: When the sum of the number of local buffers (produced by the breadth-first expansion) plus the number of message buffers from remote machines is greater than 4× the number of threads, threads switch to DFS. Having a low threshold plus small local buffers allows aDFS to keep the maximum additional memory consumption limited: If the DFS threshold is set to n, the maximum number of threads is t, the size of local buffers is b, and the query contains s stages, the maximum additional memory in a machine is $(n+t)*(s-1)*b$. In the configuration used for our experiments ($t=28$, $n=4t=128$, $s\leq 11$, and $b=8{,}192$), local buffers consume less than 12MB additional memory.

Controlled Experiment. Figure 9 illustrates the benefits of this local-match BFS mode on 8 machines (see Section 4 for detailed experimental settings) with the following two queries:

```
1: (a)->()->() WHERE ID(a) < $i
2: (a)->()-[e]->()->() WHERE e.cost < 0.5
                       AND ID(a) < $i
```

using the Twitter graph extended with a uniform random edge property with values in $[0.0, 100.0)$. In both queries, the `ID(a) < $i` filter determines the cardinality of the first query stage and is used to narrow the starting point. In Q2, the edge filter also guarantees that the third stage includes a small number of intermediate results. The dynamicity of aDFS brings significant performance benefits, especially for queries with very narrow starting points. For example, for Q1 with `$i = 1`, Machine 0 hosts the match for Stage 0; without the breadth-first mode ("OFF"), a single thread handles all the 55K local edges which lead to Stage 1. In contrast, enabling dynamic local BFS ("ON") generates more work early on and allows splitting the work among local threads, each of which operates on approximately 2,000 vertices for Stage 1.

LDBC Q20. We also briefly analyze the BFS-mode benefits on LDBC Q20 (see Section 4 for more details):

```
MATCH (tC:tagClass)<-[:subClassOf]-(:tagClass)
  <-[:hasType]-(:tag)<-[:hasTag]-(:post|comment)
WHERE tC.name IN ('Politics', 'Art', 'Country')
```

In this query, the first two stages match `tagClass`es and Stage 0 results in only three intermediate results due to the filter. The local BFS optimization brings 32% latency benefits (8 vs. 5.5 seconds), by better paralellizing the work across threads. Without the optimization, the most busy thread, i.e., the one that "gets stuck" in performing local DFS work the

most, spends 4 seconds in these local explorations: It matches about 1,000 vertices in Stage 1, which result in 5.2 million local matches in Stage 2 and 5 million in Stage 3. In comparison, with the optimization, the most busy thread spends only 1.6 seconds in DFS work: It handles 4 million local edges in Stage 2, which it successfully distributes to other threads with approximately 500 local BFS buffers. Overall, enabling dynamic local BFS provides significant speedup on realistic workloads, while incurring at most a 5% slowdown.

4 Evaluation

The goals of our evaluation are (i) to understand how well aDFS performs as compared to other systems (graph, relational, mining and dataflow join systems) that could be used in similar use cases, (ii) explain how different parts of aDFS contribute to performance and memory, and (iii) show how aDFS scales as we increase the number of machines.

4.1 Experimental Settings

Hardware details. We use a cluster of eight nodes, each having two Intel Xeon E-2690 v4 2.60GHz CPUs with 14 cores (hyperthreads disabled/DVFS enabled), for 28 cores in total. Each processor contains 756GB of DDR4-2400 memory and LSI MegaRAID SAS-3 3108 storage. Each node includes a Mellanox Connect-X InfiniBand card, all connected to an EDR 100Gbit/s InfiniBand network.

Graphs and queries. Unless specified otherwise, our experiments use the five graphs of Table 2. As we mention in Section 2, the scope of this paper covers user-provided fixed-pattern queries, thus aDFS implements only a subset of PGQL 1.1. Accordingly, we use the 12 LDBC Business Intelligence (BI) standard queries [68] supported by PGQL 1.1 (later PGQL versions support the remaining LDBC queries). Out of these 12 queries, four represent simple path patterns (i.e., Q4, Q17, Q23, Q24) and are directly supported in aDFS. The remaining ones either include regular path queries (e.g., `... MATCH (a)-/:knows*/->(b)`), or include sub-queries in projection or filters (e.g., `SELECT ... FROM (SELECT ...) ...`). We devise a simplified variant of these queries in order to support the benchmark specification as closely as possible. For example, the original Q6 is:

```
SELECT id(person),
SUM((SELECT COUNT(*) MATCH (m)<-[:replyOf]-(:cmt))) AS rN
SUM((SELECT COUNT(*) MATCH (:prsn)-[:likes]->(m))) AS lN,
COUNT(*) AS msgN
MATCH (tag:tag) <-[:hasTag]- (m:post|comment)
        -[:hasCreator]-> (person:prsn)
WHERE tag.name = ?
GROUP BY person, tag
ORDER BY msgN + (2 * rN) + (10 * lN) DESC, id(person)
```

We simplify the query by removing the two `COUNT` subqueries in projections and from `ORDER BY`. We plan to extend the PGQL support of aDFS in future work.

Note that the queries include patterns of varying complexity, e.g., the one in Q6 above is rather simple, while Q17 matches the following complex pattern:

```
(x:person)-[:livesIn]->(c1:city)-[:partOf]->(cy:country),
(y:person)-[:livesIn]->(c2:city)-[:partOf]->(cy),
(z:person)-[:livesIn]->(c3:city)-[:partOf]->(cy),
(x)-[:knows]-(y)-[:knows]-(z)
```

Methodology. We perform 15 runs of each query and report the median latency (in Figure 10, the error bars cover all runs). For each experiment set, we execute the queries in a per-graph round-robin fashion in order to reduce caching effects (e.g., data in the LLC or instruction caches). We use eight machines for aDFS, GraphFrames, G-Miner, Fractal as well as BiGJoin, and make sure all systems are configured to use InfiniBand. The four other systems are single machine.

Engines and their configurations. We configure aDFS to use up to 4,096 messaging buffers of 256KB per machine for messaging. This setting translates to approximately 1GB of intermediate results that can be produced per machine and limits the worst-case maximum memory consumption of a single machine to approximately 8GB (1GB outgoing, plus 7GB incoming). We further use the configuration of Section 3.5 for the local-edge dynamic BFS, resulting in up to a few MBs of extra memory per machine. Altogether, the aDFS runtime consumes approximately 10GB per machine. Of course, the graph (that resides in memory) and the final query results consume extra memory than these 10GB. We use such a low-memory configuration because (i) aDFS is designed for server deployments and we want to evaluate the performance at a realistic setting, where a single query cannot monopolize memory, and (ii) as we show in Figure 7, this configuration is already sufficient for aDFS to perform well.

We first compare aDFS to two graph systems and two relational systems which we describe below. In Section 4.5, we further compare aDFS to three graph-mining systems and a dataflow join system.

GraphFrames [31] is a distributed graph-querying system built on top of Apache Spark [2, 79]; we use version 0.7 on top of Spark 2.4.1 with 600GB executor memory per machine. ***Neo4j*** [10] is a single-machine graph database, which stores its data on disk but uses an in-memory cache for performance (caching effects are obvious in the first run of each query). We use Neo4j Community Edition 3.5.3 and allow it to manage the full memory of the machine. ***MonetDB*** [9, 26] is an in-memory column-store relational database. Its distributed support is rather rudimentary, resulting in worse than single-machine performance for our join-heavy workloads. Therefore, we use MonetDB 11.31.13 on a single machine, configured to use the whole 756GB of memory. ***PostgreSQL*** [15] is a relational database. We use version 11.2, tuned for a single connection with memory cache size of 564GB and 198GB of

Graph	#V	#E	Schema	Description
Livejournal [20]	484K	68.9M	No	Users and friendships
URandom	100M	1B	No	Uniform random edges
Twitter [47]	42.6M	1.47B	No	Tweets and followers
LDBC(100) [68]	283M	1.78B	Yes	LDBC social
Webgraph-UK [25]	77.7M	2.97B	No	2006 .uk domains

Table 2: The set of graphs we use in the evaluation.

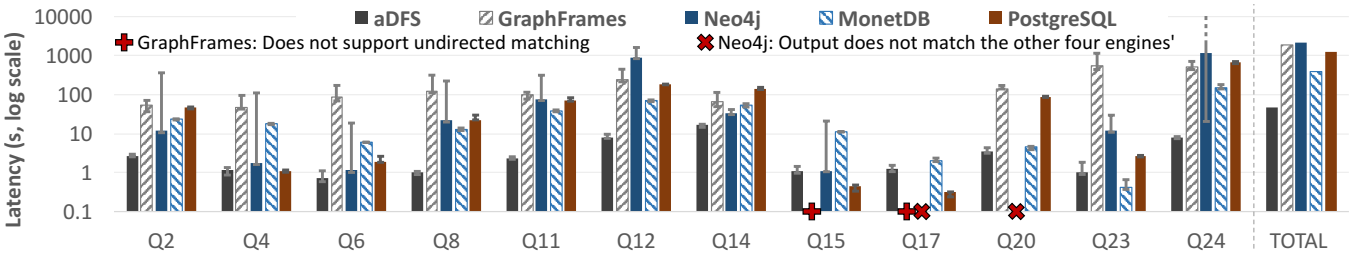

Figure 10: aDFS vs. other graph/relational systems on LDBC. Q*N* is the *N*th LDBC BI query. Error bars show min/max latencies. TOTAL is the sum of all latencies—i.e., the time to complete a single run of all 12 (10 for GraphFrames and Neo4j) queries.

shared buffers. For both MonetDB and PostgreSQL, we use the optimized schema/indices designed for the original LDBC evaluation paper [68]. We choose these four systems as they cover a broad spectrum of data processing: Distributed graph dataframes, single-machine graph databases, and in-memory or traditional relational databases.

4.2 aDFS vs. Other Engines: LDBC

Experiment. We perform an end-to-end comparison of aDFS to the four aforementioned systems. We use the LDBC graph and BI queries which constitute an unfavorable workload for aDFS and GraphFrames: the LDBC graph has a relational schema, carefully partitioned in tables, such as `person` and `post`. For relational databases (as well as Neo4j), this schema enables the exploration of small parts of the graph for most queries. For example, the pattern `(:post)-[:hasCreator]->(:person)` (taken from an actual query) needs to only access the tables `post` and `person`, which are a relatively small part for the graph. In contrast, aDFS and GraphFrames operate on the original graph model, where the whole dataset is a single graph. The end result is that these two systems perform more broad exploration even on queries that are very narrow in terms of schema accesses.

Optimizing for relational schemas is outside of the scope of this work. Still, we choose LDBC with BI queries for our end-to-end comparison as it shows how aDFS performs on queries that can be expressed well both in graph and relational systems. The next sections focus on schema-less graphs.

Results. Figure 10 depicts the query latencies of the five systems. For most queries, aDFS is one to two orders of magnitude faster than GraphFrames. aDFS delivers 102× average speedup and takes 43× less total time than GraphFrames to complete the 10 out of 12 supported queries. GraphFrames translates graph queries into dataframe joins, offered by Apache Spark, which are significantly slower than aDFS's graph traversals. Additionally, GraphFrames is memory hungry, consuming hundreds of gigabytes of memory in comparison to the small footprint of aDFS. Furthermore, aDFS completes the 10 supported queries 53× faster than Neo4j, with a 35× average speedup, even though Neo4j leverages the graph schema, as well as the large amount of available memory with its graph cache. With Neo4j, the whole graph

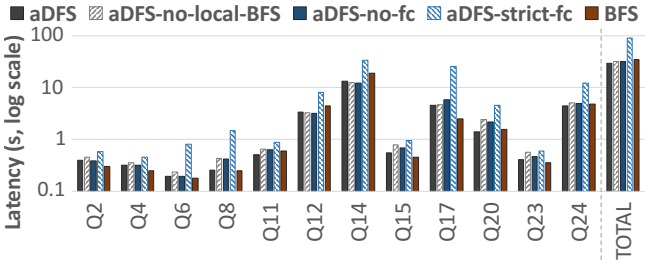

Figure 11: aDFS with various configurations on LDBC.

resides in memory after the first run: the error bars clearly show the effects of the first slow run.

Comparing aDFS to the relational systems, MonetDB and PostgreSQL, shows two different behaviors depending on the query size. For large queries, such as Q12 and Q24, which expand to large parts of the graph with long paths, aDFS is significantly faster. On the contrary, for small, very relational queries, such as Q15, Q17, and Q23, the relational systems can be faster than aDFS. This is expected given that just the distributed bootstrapping and coordination overheads in aDFS account for several tens of milliseconds. These different queries highlight the tradeoff between the relational table-focused joins and the graph exploration approach of aDFS. Overall, aDFS completes the whole set of queries 8.4 and 26 times faster than MonetDB and PostgreSQL, respectively. The average speedups are 10× and 25× against MonetDB and PostgreSQL, respectively. Conversely, MonetDB is 2.4× faster than aDFS on Q23, while PostgreSQL is on average 2.6× faster for Q4, Q15, and Q17.

In conclusion, aDFS achieves better overall performance than the four other systems while consuming lower/capped runtime memory.

4.3 Dissecting aDFS with LDBC

Experiment. We again use the LDBC benchmark to show how different design characteristics of aDFS contribute to performance and memory usage. In particular, we compare the pattern-matching-only latency of the default *aDFS* (as used in Section 4.2) to *aDFS-no-local-BFS* (we disable the machine-local dynamic BFS), *aDFS-no-fc* (we further disable flow

control), *aDFS-strict-fc* (we make flow control very strict), and *BFS* (we use the BFS implementation of Section 2.4).

Results. Figure 11 includes the results for these configurations. All in all, aDFS is the fastest. With the "dynamic BFS for local edges" option, aDFS is 31% faster on average than *aDFS-no-local-BFS* for 10 queries, while incurring 4% overhead for the remaining two queries. As we described earlier, queries often have some very "narrow" execution points with a handful of intermediate results, which leads to poor parallelization with strict local DFS. In terms of memory consumption, aDFS consumes slightly more memory than *aDFS-no-local-BFS*, not only due to the local buffers, but also thanks to better parallelization, which results in more parallel message traffic.

Disabling flow control on top of *aDFS-no-local-BFS* can bring some benefits as shown by *aDFS-no-fc*. However, the performance gains are low, as *aDFS-no-fc* hits almost no flow-control limits for this workload—i.e., local DFS and prioritizing messaging buffers from later stages of the queries result in an efficient execution flow, since none of the stages "explodes" in terms of memory. Still, *aDFS-no-fc* exhibits a 20% speedup with up to 5× higher memory consumption.

aDFS-strict-fc represents the closest realistic configuration to DFS. Processing one intermediate result at a time would naturally perform poorly, hence, we instead disable dynamic local BFS and configure each stage to have exactly one outgoing buffer to the next stage per target machine. The results show that excessive flow control reduces performance. In particular, *aDFS-strict-fc* is up to 6× slower than aDFS, while consuming up to 4× less memory.

Finally, as a reference, *BFS* implements a basic BFS-only runtime. As expected, *BFS* performs better than aDFS for certain queries, as it better leverages locality and parallelization. Still, aDFS executes the total workload 16% faster than *BFS*, while *BFS* consumes up to 6× more memory.

In conclusion, aDFS includes a set of design characteristics that when put together achieve great performance with low and controlled memory consumption.

4.4 aDFS vs. Other Engines: Large Schema-Less Queries

Experiment. The classic property graph model is schema-less, which enables users to easily query the whole dataset (unlike the relational model which requires several joins and unions of results). Therefore, we now compare aDFS to the other four systems with the schema-less graphs of Table 2: this workload shows the full power of aDFS in handling very large queries. For the relational systems, the graphs consist of two tables: One for vertices and another one for edges. Regarding queries, we use two simple patterns, a cycle `(a)->(b)->(a)` as Q1 and a two-hop path `(a)->(b)->(c)` as Q2, combined with aggregations in the `SELECT` clause (variant "a" performs a `COUNT(*)` and variant "b" `AVG` aggregations on a random vertex property). The conclusions remain the same for other patterns and projections (not shown). Note that it is impossible to evaluate more elaborate patterns, as the competing systems can barely handle the simple patterns that we use.

Results. Figure 12 depicts the results. In most cases, aDFS is about 2 orders of magnitude faster than the other systems. For the large queries and graphs, we also see that the other systems are either not able to complete the queries within eight hours, or crash. In particular, GraphFrames crashes after having consumed its 600GB of executor memory.

The speedups of aDFS over the other systems (for the completed queries where there is no timeout) are: 16 to 62× for GraphFrames, 1,105 to 9,200× for Neo4j, 20 to 169× for MonetDB, and 60 to 190× for PostgreSQL. Neither the join-based systems (GraphFrames, MonetDB, and PostgreSQL) nor Neo4j are able to handle well these immense graph explorations, although they have access to hundreds of gigabytes of memory. In particular, Neo4j spills to disk, hence the extreme performance difference compared to aDFS. Clearly, for graphs and queries at this scale, a fast graph-optimized solution such as aDFS, which easily handles these queries, is required. With the largest query (Q2a on Twitter) aDFS performs a 9.3T `COUNT` in 1,286 seconds, resulting in 7.3B matches per second, while consuming less than 10GB per-machine memory for intermediate results.

4.5 aDFS vs. Graph Mining, Dataflow Joins

Experiment. We compare aDFS to (i) three graph-mining systems,[2] namely ***G-Miner*** [27], ***Fractal*** [32], and ***Peregrine*** [42], as well as a dataflow join system, ***BiGJoin*** [19]. We use workloads from the G-Miner paper [27]: TC, i.e., Triangle Counting, and counting instances of a more complex pattern referred to as the P-pattern, with the four graphs that are used to evaluate these operations in the paper. All systems are distributed apart from Peregrine. For BiGJoin, we only perform the evaluation on TC as it does not support filters, and tune the batch size for performance (10^8). For aDFS, we express both triangles and the P-pattern as graph queries.

Results. Figure 13 includes the performance of the four systems. Triangle counting (TC) highlights the difference between matching and not matching automorphisms: For the three graph-mining systems, the search for "unique" triangles is baked in the pattern-matching algorithm, whereas in aDFS, we implement isomorphism with automorphism elimination using dynamic filtering (i.e., `(a)->(b)->(c)->(a) WHERE ID(a) < ID(b) AND ID(b) < ID(c)`). This results in expensive filtering and heavier cross-machine communication than with the other systems. Still, aDFS is faster than G-Miner and Fractal for all graphs by up to 14× for G-Miner and by up to several orders of magnitude for Fractal. Peregrine outperforms all other graph-mining systems including aDFS on three out of the four graphs, as it is able to intersect adjacency lists to quickly find common neighbors, an optimization that

[2] We requested the artifact of Automine [54] for evaluation, but the authors were not able to provide us with it.

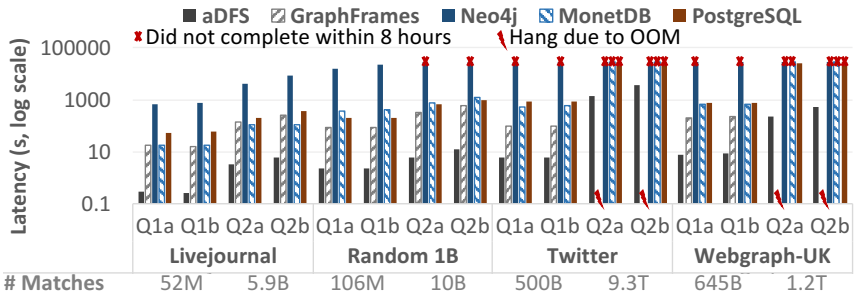

Figure 12: aDFS vs. other graph and relational systems on simple-pattern queries.

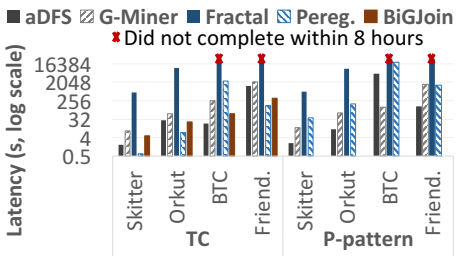

Figure 13: aDFS vs. graph-mining and dataflow join systems on triangles, patterns.

performs particularly well for triangles and which can be implemented in a straightforward manner on a single machine, where the whole graph is accessible. There is no clear winner between aDFS and BiGJoin on TC, with each system outperforming the other on two graphs. By intersecting local edges, BiGJoin's approach allows for reduced communication and better performance on the two graphs with the highest average degrees (Orkut and Friendster).

The P-pattern does not require automorphism checks, as its vertices are differentiated by labels. We express it as:

```
(c:c)->(b1:b)->(:a)->(c)->(b2:b)->(:d)
WHERE b1 <> b2
```

in PGQL. When matching the P-Pattern, aDFS significantly outperforms all other systems for all but one datapoint (G-Miner on BTC); it is on average 12 and 366× faster than Peregrine and Fractal, respectively, and 8× faster than G-Miner on three graphs. G-Miner achieves the best performance on BTC mainly because it replicates the target vertex label with each edge, which increases locality and reduces communication traffic. Such an optimization is not practical in a real-world system in which vertices can have many labels and properties of various types: Replicating these for each edge can have unacceptable memory overhead.

Overall, although aDFS is designed for different workloads, i.e., expressive graph queries, it is still very competitive with state-of-the-art graph-mining systems and a dataflow join system on triangle counting and/or a mining-oriented workload.

4.6 aDFS Scalability

Experiment and results. We use the LDBC workload to illustrate the scalability of aDFS as we vary the number of machines. Figure 14 includes the speedups, normalized to the latency of a single machine. Overall, aDFS exhibits very good scalability: The average speedup is 1.6× from one to two machines, 2.5× from one to four machines, and 5.4× from one to eight machines. These numbers include various distributed coordination and query compilation overheads, as well as additional fixed costs. The core runtime of aDFS actually has even better scalability: looking at pure pattern-matching execution time, without coordination overheads, `GROUP BY`, and `ORDER BY`, the speedup improves to 1.7×, 2.6×, and 6× from one to two, four, and eight machines respectively (not shown). aDFS is designed to scale: More machines translate to more compute resources, more buffers for intermediate results, and often more BFS exploration and higher network utilization, as the percentage of remote edges increases with the number of machines.

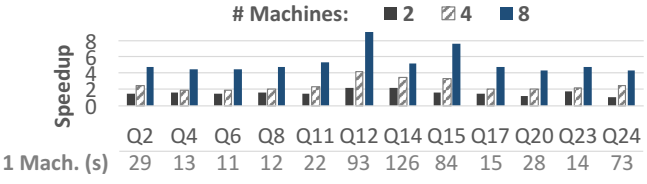

Figure 14: Scalability of aDFS vs. using a single machine.

5 Related Work

Database Management Systems (DBMSs). DBMSs offer graph support via a multi-model premise, but focus on SQL-like rather than pattern-matching querying [7, 12, 13, 52]. Kalinsky et al. [43] acknowledge that using DBMS joins for graph pattern matching is suboptimal, and propose hardware support to alleviate the issue. In contrast to DBMSs, aDFS is an efficient in-memory distributed graph-querying system that considers graph storage and queries as first-class citizens and focuses on analytical rather than transactional workloads.

Graph Algorithms. There is a plethora of related work for executing graph algorithms (such as PageRank [57]). Single-machine solutions focus on various topics such as proposing DSLs [38] or programming models [55] for graph algorithms, performance optimizations [65, 67], leveraging hardware features such as NUMA [81] and GPUs [56, 84], or supporting out-of-core computing [83]. Distributed solutions focus on topics such as asynchronous processing and performance [35, 50], efficient partitioning [78, 82, 85], leveraging hardware features such as RDMA [75], support for secondary storage [61], distributing sequential algorithms [33], approximate computing [70], alternative programming paradigms [76], or fault tolerance [30, 71]. aDFS focuses on graph queries rather than algorithms, but it shares features with some of these distributed solutions, such as the use of asynchronous processing or (random) graph partitioning.

Graph Querying. A number of single-node graph-querying systems were proposed by academia: Sun et al. [66] and Lin et al. [49] build relational and transactional systems, Graphflow [44] is an active graph database that supports evaluating one-time and continuous subgraph queries, TurboFlux [46] optimizes fast continuous subgraph matching over a fast graph update stream, and CECI [22] uses multiple embedding clusters and intersections of neighborhood lists to optimize subgraph matching (CECI can be distributed through graph replication/sharing, not graph distribution as with aDFS, due to the challenges mentioned in Section 2.4). In earlier work, we prototyped simple distributing DFS exploration [60].

There are numerous industrial graph-querying solutions. Neo4j [10] is single-machine and supports Cypher [11] queries. Amazon Neptune [1] is built for the Amazon cloud. Facebook Dragon [3] builds indices on updates for accessing data. Microsoft Graph Engine [8] is an in-memory data processing system based on Trinity [63], and TigerGraph [18] distributes GSQL [4] queries based on the source vertex data for a given query hop. Furthermore, there are also open-source distributed solutions. JanusGraph [6] uses distributed graph storage but does not distribute computation. GraphFrames [31] implements graph pattern matching with Spark using joins of dataframes. Wukong [64, 73] is a distributed graph-based RDF store that leverages hardware features, such as RDMA and GPUs, which we do not focus on. To the best of our knowledge, aDFS is the first truly distributed graph-querying system that works on fully-partitioned graphs and strictly bounds memory while maintaining great performance.

Graph-Mining systems. Graph-mining focuses on extracting structural properties and computing complex aggregate statistics [34, 74] of a graph by exploring its subgraph structures. Examples include triangle counting, maximal clique finding, community detection, and graph matching [27, 54, 59]. Graph-querying systems typically employ a vertex/edge-centric processing approach: A state is maintained per vertex and communicated to its neighbors [54, 69]. Graph-mining systems typically follow a subgraph-centric (often undirected and schema-less) processing approach: They attach information to a large amount of intermediate results composed of subgraphs [54] rather than specific vertices. Additionally, graph-mining systems typically leverage automorphism elimination [29, 32, 42], while while graph-querying engines generate homomorphisms to answer user graph queries.

Recent single-machine systems include RStream [72], AutoMine [54], and Peregrine [42]. Distributed systems include Arabesque [69], NScale [59], G-thinker [77], BiGJoin [19], G-Miner [27], ASAP [41], and Fractal [32]. aDFS shares features with some of these systems. For example, forms of asynchronous computations are used in G-Miner [27] (with a "task-pipeline" to hide communication overheads) and BiGJoin [19] (with data-parallel dataflow computations that pick up dynamically joined columns with the least matches). Techniques to reduce memory consumption are used by G-Thinker [77] (buffering excess subgraph-tasks in a disk-based priority queue), BiGJoin [19] (primarily using batching to limit memory consumption but not for intermediate results as with aDFS) and Fractal [32]. Fractal combines a DFS strategy with a "from-scratch processing" paradigm which leads to re-computation overheads (absent in aDFS), as well as imbalances across workers that are mitigated by work stealing: workers break the DFS strategy to steal enumerations, which can be at any level of the matched graph pattern, from other workers. aDFS uses asynchronous DFS-based graph traversals together with flow control to strictly bound memory consumption, and can switch to BFS, in the same graph pattern-matching level, to generate more local work and to buffer remote edges (see Section 3). Our in-depth evaluation shows that the performance of aDFS for graph pattern-matching is competitive with that of state-of-the-art graph-mining systems.

BFS/DFS. The BFS/DFS tradeoff has been explored in the context of single-machine parallel task-scheduling runtimes. Typically, DFS is used to schedule a task graph in order to curtail memory [28], and BFS is used opportunistically (often called "work stealing") to maximize parallelism [23, 24]. aDFS leverages these insights in the context of distributed graph query processing.

6 Concluding Remarks

Conclusions. We have introduced aDFS: A system that uses an efficient, almost-DFS approach to execute pattern-matching queries on distributed graphs. aDFS is able to execute virtually any query on any in-memory graph using at most a fixed, configurable amount of memory. aDFS is also very fast and scalable. We compared aDFS to eight state-of-the-art systems with diverse characteristics—graph or relational/join-based, distributed or single machine, in-memory or disk-based—and showed that aDFS is up to orders of magnitude faster than them.

Limitations and future work. aDFS uses simple algorithms for query optimization and graph partitioning, as this paper focused on runtime support for distributed graph querying. In the future, we will improve query planning and optimization, together with graph partitioning and caching. We will also consider query optimization opportunities to enable pruning of the traversal space when the underlying data has a relational-style schema, as described in Section 4.2.

Acknowledgments. Tomáš Faltín was partially supported by the Charles University, project GA UK No. 396721. We would like to thank our anonymous reviewers, as well as our shepherd, Keval Vora, for their feedback.

References

[1] Amazon Neptune – Fast, reliable graph database built for the cloud. https://aws.amazon.com/neptune/.

[2] Apache Spark – Unified analytics engine for big data. https://spark.apache.org.

[3] Facebook Dragon – A distributed graph query engine. https://code.fb.com/data-infrastructure/dragon-a-distributed-graph-query-engine/.

[4] GQL Standard – Graph Query Language. https://www.gqlstandards.org.

[5] Gremlin – A graph traversal language. https://github.com/tinkerpop/gremlin/wiki.

[6] JanusGraph – Distributed, open source, massively scalable graph database. https://janusgraph.org.

[7] Microsoft Azure Cosmos DB – Fast NoSQL database with open APIs for any scale. https://azure.microsoft.com/en-gb/services/cosmos-db/.

[8] Microsoft Graph Engine – Serving big graphs in real-time. https://www.graphengine.io.

[9] MonetDB – An open-source database system. https://www.monetdb.org.

[10] Neo4j – Graph database platform. https://neo4j.com.

[11] Neo4j Cypher Query Language – Developer guides. https://neo4j.com/developer/cypher/.

[12] OQGRAPH – The Open Query GRAPH engine for MariaDB. https://openquery.com.au/products/graph-engine.

[13] OrientDB Community Edition. https://orientdb.org.

[14] PGQL 1.1 Specification – Property Graph Query Language. https://pgql-lang.org/spec/1.1/.

[15] PostgreSQL – The world's most advanced open source database. https://www.postgresql.org.

[16] SNAP: Network Datasets – Google web graph. https://snap.stanford.edu/data/web-Google.html.

[17] SPARQL Query Language for RDF – SPARQL Protocol and RDF Query Language. https://www.w3.org/TR/rdf-sparql-query/.

[18] TigerGraph Distributed Query Mode – Documentation. https://docs.tigergraph.com/dev/gsql-ref/querying/distributed-query-mode.

[19] Khaled Ammar, Frank McSherry, Semih Salihoglu, and Manas Joglekar. Distributed Evaluation of Subgraph Queries Using Worst-Case Optimal Low-Memory Dataflows. *PVLDB*, 2018.

[20] Lars Backstrom, Daniel P. Huttenlocher, Jon M. Kleinberg, and Xiangyang Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *SIGKDD*, 2006.

[21] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5), 2001.

[22] Bibek Bhattarai, Hang Liu, and H. Howie Huang. CECI: Compact Embedding Cluster Index for Scalable Subgraph Matching. In *SIGMOD*, 2019.

[23] Guy E. Blelloch, Phillip B. Gibbons, and Yossi Matias. Provably Efficient Scheduling for Languages with Fine-Grained Parallelism. *J. ACM*, pages 281–321, 1999.

[24] Robert D. Blumofe, Christopher F. Joerg, Bradley C. Kuszmaul, Charles E. Leiserson, Keith H. Randall, and Yuli Zhou. Cilk: An Efficient Multithreaded Runtime System. In *PPOPP*, 1995.

[25] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. A Large Time-Aware Web Graph. *SIGIR Forum*, 42(2), 2008.

[26] Peter A. Boncz, Martin L. Kersten, and Stefan Manegold. Breaking The Memory Wall in MonetDB. *Commun. ACM*, 51(12), 2008.

[27] Hongzhi Chen, Miao Liu, Yunjian Zhao, Xiao Yan, Da Yan, and James Cheng. G-Miner: An Efficient Task-Oriented Graph Mining System. In *EuroSys*, 2018.

[28] Shimin Chen, Todd C. Mowry, Chris Wilkerson, Phillip B. Gibbons, Michael Kozuch, Vasileios Liaskovitis, Anastassia Ailamaki, Guy E. Blelloch, Babak Falsafi, Limor Fix, and Nikos Hardavellas. Scheduling Threads for Constructive Cache Sharing on CMPs. In *SPAA*, 2007.

[29] Soumyava Das and Sharma Chakravarthy. Duplicate Reduction in Graph Mining: Approaches, Analysis, and Evaluation. *IEEE KDD*, 30(8), 2018.

[30] Roshan Dathathri, Gurbinder Gill, Loc Hoang, and Keshav Pingali. Phoenix: A Substrate for Resilient Distributed Graph Analytics. In *ASPLOS*, 2019.

[31] Ankur Dave, Alekh Jindal, Li Erran Li, Reynold Xin, Joseph Gonzalez, and Matei Zaharia. GraphFrames: An Integrated API for Mixing Graph and Relational Queries. In *GRADES*, 2016.

[32] Vinicius Dias, Carlos H. C. Teixeira, Dorgival Guedes, Wagner Meira, and Srinivasan Parthasarathy. Fractal: A General-Purpose Graph Pattern Mining System. In *SIGMOD*, 2019.

[33] Wenfei Fan, Jingbo Xu, Yinghui Wu, Wenyuan Yu, Jiaxin Jiang, Zeyu Zheng, Bohan Zhang, Yang Cao, and Chao Tian. Parallelizing Sequential Graph Computations. In *SIGMOD*, 2017.

[34] Brian Gallagher. Matching Structure and Semantics: A Survey on Graph-Based Pattern Matching. In *AAAI Fall Symposium*, 2006.

[35] Joseph E. Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, and Carlos Guestrin. PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs. In *OSDI*, 2012.

[36] Joseph E. Gonzalez, Reynold S. Xin, Ankur Dave, Daniel Crankshaw, Michael J. Franklin, and Ion Stoica. GraphX: Graph Processing in A Distributed Dataflow Framework. In *OSDI*, 2014.

[37] Sairam Gurajada, Stephan Seufert, Iris Miliaraki, and Martin Theobald. TriAD: A Distributed Shared-Nothing RDF Engine Based on Asynchronous Message Passing. In *SIGMOD*, 2014.

[38] Sungpack Hong, Hassan Chafi, Edic Sedlar, and Kunle Olukotun. Green-Marl: A DSL for Easy and Efficient Graph Analysis. In *ASPLOS*, 2012.

[39] Sungpack Hong, Siegfried Depner, Thomas Manhardt, Jan Lugt, Merijn Verstraaten, and Hassan Chafi. PGX.D: A Fast Distributed Graph Processing Engine. In *SC*, 2015.

[40] Jiewen Huang, Daniel J. Abadi, and Kun Ren. Scalable SPARQL Querying of Large RDF Graphs. *PVLDB*, 4(11), 2011.

[41] Anand Padmanabha Iyer, Zaoxing Liu, Xin Jin, Shivaram Venkataraman, Vladimir Braverman, and Ion Stoica. ASAP: Fast, Approximate Graph Pattern Mining at Scale. In *OSDI*, 2018.

[42] Kasra Jamshidi, Rakesh Mahadasa, and Keval Vora. Peregrine: A Pattern-Aware Graph Mining System. In *EuroSys*, 2020.

[43] Oren Kalinsky, Benny Kimelfeld, and Yoav Etsion. The TrieJax Architecture: Accelerating Graph Operations Through Relational Joins. In *ASPLOS*, 2020.

[44] Chathura Kankanamge, Siddhartha Sahu, Amine Mhedbhi, Jeremy Chen, and Semih Salihoglu. Graphflow: An Active Graph Database. In *SIGMOD*, 2017.

[45] Jinha Kim, Hyungyu Shin, Wook-Shin Han, Sungpack Hong, and Hassan Chafi. Taming Subgraph Isomorphism for RDF Query Processing. *Proc. VLDB Endow.*, 8(11), 2015.

[46] Kyoungmin Kim, In Seo, Wook-Shin Han, Jeong-Hoon Lee, Sungpack Hong, Hassan Chafi, Hyungyu Shin, and Geonhwa Jeong. TurboFlux: A Fast Continuous Subgraph Matching System for Streaming Graph Data. In *SIGMOD*, 2018.

[47] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue B. Moon. What Is Twitter, A Social Network Or A News Media? In *WWW*, 2010.

[48] Aapo Kyrola, Guy E. Blelloch, and Carlos Guestrin. GraphChi: Large-Scale Graph Computation on Just a PC. In *OSDI*, 2012.

[49] Chunbin Lin, Benjamin Mandel, Yannis Papakonstantinou, and Matthias Springer. Fast In-Memory SQL Analytics on Typed Graphs. *PVLDB*, 10(3), 2016.

[50] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J.M. Hellerstein. Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. *PVLDB*, 5(8), 2012.

[51] Andrew Lumsdaine, Douglas P. Gregor, Bruce Hendrickson, and Jonathan W. Berry. Challenges in Parallel Graph Processing. *Parallel Processing Letters*, 17(1), 2007.

[52] Hongbin Ma, Bin Shao, Yanghua Xiao, Liang Jeff Chen, and Haixun Wang. G-SQL: Fast Query Processing via Graph Exploration. *PVLDB*, 9(12), 2016.

[53] Grzegorz Malewicz, Matthew H. Austern, Aart J.C Bik, James C. Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. Pregel: A System for Large-scale Graph Processing. In *SIGMOD*, 2010.

[54] Daniel Mawhirter and Bo Wu. AutoMine: Harmonizing High-Level Abstraction and High Performance for Graph Mining. In *SOSP*, 2019.

[55] Donald Nguyen, Andrew Lenharth, and Keshav Pingali. A Lightweight Infrastructure for Graph Analytics. In *SOSP*, 2013.

[56] Amir Hossein Nodehi Sabet, Junqiao Qiu, and Zhijia Zhao. Tigr: Transforming Irregular Graphs for GPU-Friendly Graph Processing. In *ASPLOS*, 2018.

[57] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.

[58] Vijayan Prabhakaran, Ming Wu, Xuetian Weng, Frank McSherry, Lidong Zhou, and Maya Haradasan. Managing Large Graphs on Multi-Cores with Graph Awareness. In *USENIX ATC*, 2012.

[59] Abdul Quamar, Amol Deshpande, and Jimmy Lin. NScale: Neighborhood-Centric Large-Scale Graph Analytics in the Cloud. *The VLDB Journal*, 25(2), 2016.

[60] Nicholas P. Roth, Vasileios Trigonakis, Sungpack Hong, Hassan Chafi, Anthony Potter, Boris Motik, and Ian Horrocks. PGX.D/Async: A Scalable Distributed Graph Pattern Matching Engine. In *GRADES Workshop*, 2017.

[61] Amitabha Roy, Laurent Bindschaedler, Jasmina Malicevic, and Willy Zwaenepoel. Chaos: Scale-Out Graph Processing From Secondary Storage. In *SOSP*, 2015.

[62] Amitabha Roy, Ivo Mihailovic, and Willy Zwaenepoel. X-Stream: Edge-Centric Graph Processing Using Streaming Partitions. In *SOSP*, 2013.

[63] Bin Shao, Haixun Wang, and Yatao Li. Trinity: A Distributed Graph Engine on a Memory Cloud. In *SIGMOD*, 2013.

[64] Jiaxin Shi, Youyang Yao, Rong Chen, Haibo Chen, and Feifei Li. Fast and Concurrent RDF Queries with RDMA-Based Distributed Graph Exploration. In *OSDI*, 2016.

[65] Julian Shun and Guy E. Blelloch. Ligra: A Lightweight Graph Processing Framework for Shared Memory. In *PPoPP*, 2013.

[66] Wen Sun, Achille Fokoue, Kavitha Srinivas, Anastasios Kementsietsidis, Gang Hu, and Guo Tong Xie. SQLGraph: An Efficient Relational-Based Property Graph Store. In *SIGMOD*, 2015.

[67] Narayanan Sundaram, Nadathur Satish, Md. Mostofa Ali Patwary, Subramanya Dulloor, Michael J. Anderson, Satya Gautam Vadlamudi, Dipankar Das, and Pradeep Dubey. GraphMat: High Performance Graph Analytics Made Productive. *PVLDB*, 8(11), 2015.

[68] Gábor Szárnyas, Arnau Prat-Pérez, Alex Averbuch, József Marton, Marcus Paradies, Moritz Kaufmann, Orri Erling, Peter A. Boncz, Vlad Haprian, and János Benjamin Antal. An Early Look at The LDBC Social Network Benchmark's Business Intelligence Workload. In *GRADES Workshop*, 2018.

[69] Carlos H. C. Teixeira, Alexandre J. Fonseca, Marco Serafini, Georgos Siganos, Mohammed J. Zaki, and Ashraf Aboulnaga. Arabesque: A System for Distributed Graph Mining. In *SOSP*, 2015.

[70] Keval Vora, Rajiv Gupta, and Guoqing Xu. KickStarter: Fast and Accurate Computations on Streaming Graphs via Trimmed Approximations. In *ASPLOS*, 2017.

[71] Keval Vora, Chen Tian, Rajiv Gupta, and Ziang Hu. CoRAL: Confined Recovery in Distributed Asynchronous Graph Processing. In *ASPLOS*, 2017.

[72] Kai Wang, Zhiqiang Zuo, John Thorpe, Tien Quang Nguyen, and Guoqing Harry Xu. RStream: Marrying Relational Algebra with Streaming for Efficient Graph Mining on A Single Machine. In *OSDI*, 2018.

[73] Siyuan Wang, Chang Lou, Rong Chen, and Haibo Chen. Fast and Concurrent RDF Queries Using RDMA-Assisted GPU Graph Exploration. In *USENIX ATC*, 2018.

[74] Takashi Washio and Hiroshi Motoda. State of the Art of Graph-Based Data Mining. *ACM SIGKDD Explorations Newsletter*, 5(1), July 2003.

[75] Ming Wu, Fan Yang, Jilong Xue, Wencong Xiao, Youshan Miao, Lan Wei, Haoxiang Lin, Yafei Dai, and Lidong Zhou. GraM: Scaling Graph Computation to the Trillions. In *SoCC*, 2015.

[76] Chengshuo Xu, Keval Vora, and Rajiv Gupta. PnP: Pruning and Prediction for Point-To-Point Iterative Graph Analytics. In *ASPLOS*, 2019.

[77] Da Yan, Hongzhi Chen, James Cheng, M. Tamer Özsu, Qizhen Zhang, and John C. S. Lui. G-thinker: Big Graph Mining Made Easier and Faster. *CoRR*, abs/1709.03110, 2017.

[78] Da Yan, James Cheng, Yi Lu, and Wilfred Ng. Blogel: A Block-Centric Framework for Distributed Computation on Real-World Graphs. *PVLDB*, 7(14), 2014.

[79] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauly, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. In *NSDI*, 2012.

[80] Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, and Zhongyuan Wang. A Distributed Graph Engine for Web Scale RDF Data. *PVLDB*, 6(4), 2013.

[81] Kaiyuan Zhang, Rong Chen, and Haibo Chen. NUMA-Aware Graph-Structured Analytics. In *PPoPP*, 2015.

[82] Mingxing Zhang, Yongwei Wu, Kang Chen, Xuehai Qian, Xue Li, and Weimin Zheng. Exploring the Hidden Dimension in Graph Processing. In *OSDI*, 2016.

[83] Mingxing Zhang, Yongwei Wu, Youwei Zhuo, Xuehai Qian, Chengying Huan, and Kang Chen. Wonderland: A Novel Abstraction-Based Out-Of-Core Graph Processing System. In *ASPLOS*, 2018.

[84] Yu Zhang, Xiaofei Liao, Hai Jin, Bingsheng He, Haikun Liu, and Lin Gu. DiGraph: An Efficient Path-Based Iterative Directed Graph Processing System on Multiple GPUs. In *ASPLOS*, 2019.

[85] Xiaowei Zhu, Wenguang Chen, Weimin Zheng, and Xiaosong Ma. Gemini: A Computation-Centric Distributed Graph Processing System. In *OSDI*, 2016.

[86] Lei Zou, M. Tamer Özsu, Lei Chen, Xuchuan Shen, Ruizhe Huang, and Dongyan Zhao. gStore: A Graph-Based SPARQL Query Engine. *The VLDB Journal*, 23(4), 2014.