Single-Purpose Algorithms vs. a Generic Graph Summarizer for Computing $k$-Bisimulations on Large Graphs

Jannik Rau  
University of Ulm  
Germany  
jannik.rau@uni-ulm.de

David Richerby  
University of Essex  
United Kingdom  
david.richerby@essex.ac.uk

Ansgar Scherp  
University of Ulm  
Germany  
ansgar.scherp@uni-ulm.de

ABSTRACT
Summarizing graphs w.r.t. structural features is important in large graph applications, to reduce the graph’s size and make tasks like indexing, querying, and visualization feasible. The generic parallel BRS algorithm can efficiently summarize large graphs w.r.t. a custom equivalence relation — defined on the graph’s vertices $V$. Moreover, $\sim$ can be chained $k \geq 1$ times, so the equivalence class of a vertex depends on the structure of the graph up to distance $k$. This allows the computation of stratified $k$-bisimulation, a popular concept in structural graph summarization.

In this work, we investigate whether the generic BRS approach can outperform an existing single-purpose parallel algorithm for bisimulation. Furthermore, we investigate whether an existing sequential bisimulation algorithm can effectively be computed by the parallel BRS algorithm and how such generic implementations compete against a parallel variant. To give a fair comparison, we have reimplemented the original algorithms in the same framework as was used for the generic BRS algorithm. We evaluate the performance of the two native implementations against the implementations in the BRS algorithm for $k$-bisimulation with $k = 1, \ldots, 10$, using five real-world and synthetic graph datasets containing between 100 million and two billion edges.

Our results show that our generic BRS algorithm outperforms the respective native bisimulation algorithms for any value of $k$. The generic algorithm has no disadvantage over the native parallel algorithm. Furthermore, the execution times of the generic BRS algorithm for the native parallel and native sequential bisimulation variants are very similar. This shows that the bisimulation variant computed by the native sequential algorithm can be effectively computed in a parallel by our BRS algorithm. These insights open a new path for efficiently computing bisimulations on large graphs.

KEYWORDS
Semi structured data, $k$-bisimulation, Structural Graph Summarization, FLUID, Parallel Computation

1 INTRODUCTION
Large-scale graphs, for example in the form of directed, edge-labeled graphs, are an increasingly popular way to model data [18], often in the form of RDF graphs [13]. These contain information about (real-word) entities, denoted by the graph’s vertices, and their relationships, represented by the graph’s edges. An RDF graph can be expressed as a set of triples $(s, p, o)$, each of which formulates a subject-predicate-object expression.\footnote{https://www.w3.org/TR/2014/REC-n-triples-20140225/Overview.html} A triple $(s, p, o)$ corresponds to an edge with label $p$ going from vertex $s$ to vertex $o$.

Figure 1a shows an example RDF graph,\footnote{For brevity, URIs are replaced by short identifiers.} which contains information about a student and the university he works at. Such knowledge graphs, have emerged and grown in both the open-source and enterprise domains [18]. Applications range from search engines, recommendation systems, fraud detection and business intelligence, to simple information holders.

An equivalent formulation is the property graph [18]. Property graphs include vertex labels, as well as property-value pairs for vertices and edges. Figure 1b shows an example property graph, which contains similar information to the graph in Figure 1a.

Storing, indexing, querying, understanding, and visualizing large graphs is difficult [8]. One way to mitigate this is graph summarization [10]. Graphs can be summarized with respect to selected structural features (e.g., incoming/outgoing paths), statistical measures (e.g., occurrences of specific vertices) or frequent patterns found in the graph [10]. This produces a summary graph, which is usually smaller than the original but contains an approximation of or exactly the same information as the original graph w.r.t. the selected features. Thus, tasks that were to be performed on the original graph can be performed on the summary, instead.

Figure 1: Graphs showing a student–university relationship.
This work focuses on structural graph summaries, i.e., summarization w.r.t. selected structural features. More precisely, we consider structural summaries based on equivalence relations, which form a lossless partition of the input graph’s vertices. Consequently, the summary graph contains precise structural information about selected features of the input graph [10]. Other forms of graph summarization include those based on frequent patterns or statistical measures [10], which result in approximate (lossy) summaries. Different approaches to lossless summarization have been reported in the literature, capturing different aspects of the input graph [7, 10]. Consequently, many single-purpose algorithms exist, each computing one specific summary.

Motivated by this observation, a generic structural summarization approach has been proposed [6, 7]. The BRS algorithm summarizes an input graph w.r.t. a custom equivalence relation specified in the formal language FLUID [7]. This allows graphs to be summarized w.r.t. different structural features using one generic approach, allowing the user to select the specific structural features that will appear in the summary graph. Researchers and practitioners in the graph domain benefit, as they do not have to rely on there being a pre-existing summary tailored to their specific needs. Rather, they can define their custom equivalence relation and summarize graphs using the BRS algorithm. The FLUID language allows the various combination of features found in structural summaries reported in the literature. The first group of features is comprised of a vertex’s local information, e.g., the label set of a vertex, the direct neighbors of a vertex, or the set of incoming or outgoing edge labels. The second group is comprised of a vertex’s global information at distance \( k \geq 1 \). This includes, e.g., local information about reachable vertices up to distance \( k \) or information about incoming or outgoing paths of length up to \( k \). Summarizing a graph w.r.t. global information can be achieved by applying variants of stratified bisimulation (formally introduced in Section 3.2). Informally, a stratified \( k \)-bisimulation groups together vertices that have equivalent structural neighborhoods up to distance \( k \). Several existing structural graph summarization approaches use stratified bisimulation to incorporate global information into the summary graph [7, 10]. To accommodate this, the FLUID-based BRS algorithm can chain any definable equivalence relation \( k \) times, such that the resulting equivalence classes are determined by global information up to distance \( k > 1 \). This can be used to compute stratified \( k \)-bisimulations. Prior work has shown that the BRS algorithm can effectively compute \( k \)-bisimulations for large values of \( k \) on real-world and synthetic graphs with up to one billion edges [5].

However, it is not known if the general approach of the BRS algorithm sacrifices performance: could it be outperformed by efficient, parallel single-purpose \( k \)-bisimulation algorithms such as those of Schätzle, Neu, Lausen, and Przajcziel-Zablocki [33]? Furthermore, there are sequential algorithms for bisimulation such as the one by Kaushik, Shenoy, Bohannon, and Gudes [20]. Being a sequential algorithm, it is naturally disadvantaged against parallel algorithms. However, we show in this work that the FLUID language for the BRS algorithm can declaratively define the bisimulation of Kaushik et al. Thus, our generic approach provides a parallelizes that of Kaushik et al. “for free”. We evaluate the performance of the parallelized computation of the Kaushik et al. graph summary model and compare it with the sequential native algorithm. We also compare both Kaushik et al. variants, the native and generic, with the parallel native algorithm of Schätzle et al. [33] and a BRS implementation of Schätzle et al. from [5].

In total, we have four \( k \)-bisimulation algorithms, two native algorithms by Kaushik et al. and Schätzle et al., and two generic algorithms implemented using our formal language FLUID and the BRS algorithm. For a fair comparison, the existing native algorithms have been reimplemented using the same underlying graph processing framework as the BRS algorithm. We execute the four algorithms on five graph datasets—two synthetic and three real-world—of different sizes, ranging from a few ten million of edges to billions of edges. The real-world datasets’ sizes are around 100M, 150M, and 2B triples, whereas the synthetic datasets are 100M and 1B triples. We evaluate the algorithms’ performance for computing \( k \)-bisimulation to a distance of \( k = 1, \ldots, 10 \). We measure running time (per the \( k \) bisimulation iterations) and the maximum memory consumption during the computation. All algorithms perform computations in-memory only, to eliminate side effects. We leave considering solutions with external memory for future work.

The questions we address are:

RQ 1 Do the native bisimulation algorithms have an advantage over a generic solution?
RQ 2 How well do the native and generic algorithms scale on large real-world and synthetic graphs?
RQ 3 Is it possible to effectively scale a sequential algorithm by turning it into a parallel variant by using a general formal language and algorithms for graph summaries.

The following sections are structured as follows. Section 2 discusses related work in the domain of graph summarization. Section 3 defines preliminaries, while the algorithms are introduced in Section 4. Section 5 outlines the experimental apparatus including datasets, experimental procedure, implementations, and the applied measures. Section 6 describes the results obtained from the experiments, which are discussed and interpreted in Section 7. Finally, Section 8 concludes the work.

2 RELATED WORK

Summary graphs can be constructed in different ways. A classification of existing methods into four categories has been proposed by Čekirči et al. [10], who classify techniques into Structural, Pattern-mining, Statistical, and Hybrid approaches. Structural approaches can be used for improving query evaluation on a graph \( G \) by using the resulting summary graph \( SG \) as and index for \( G \) [10].

**Quotient Summaries.** Two common techniques can be distinguished in structural approaches. The first, producing quotient summaries, summarizes a graph \( G \) w.r.t. an equivalence relation \( \sim \subseteq V \times V \) defined on the vertices \( V \) of \( G \) [8, 10]. The resulting summary graph \( SG \) consists of vertices \( V_S \), which correspond to the equivalence classes \( A \) of the equivalence relation \( \sim \). Tran et al. [38] construct a summary graph \( SG \) using \( k \)-bisimulation as the equivalence relation. Informally, \( k \)-bisimulation groups together vertices, that have an equivalent structural neighborhood up to distance \( k \). A more detailed definition of \( k \)-bisimulation is given in Section 3.2. Tran et al. use the bisimulation-based summary graph to evaluate data partitioning and different query processing with \( SG \) serving as
Non-quotient Summaries. The second structural summarization technique, producing non-quotient summaries, does not use equivalence relations to summarize a graph. Rather, the summary graph is comprised of vertex summaries \(v_s\), which group together vertices \(v\) of the original graph \(G\) according to certain criteria [10]. Hence, the main difference between quotient summaries discussed above and the non-quotient summaries is that in the latter case a vertex \(v\) can belong to zero, one, or multiple vertex summaries \(v_s\). In contrast, in quotient summaries every vertex \(v\) has exactly one corresponding vertex summary \(v_S\), which is the equivalence class of \(v\) under \(\sim\) [10].

Early work of non-quotient summarization includes Goldman and Widom [15], who created a vertex summary \(v_s\) for every labeled path in the original graph \(G\). A vertex \(v\) of \(G\) is associated with a vertex summary \(v_s\), if it is reachable by the corresponding label path. The summary graph \(SG\) is used as a path-index, as well as a tool for understanding the schema structure in semi-structured databases and hence finds application in query formulation and query optimization. Revisiting the summarization tool SchemEX [24], its first layer – the RDF class layer – consists of vertex summaries \(v_s\), representing all the classes \(c\) present in the input RDF graph \(G\). A vertex \(v\) of \(G\) is associated with a vertex summary \(v_s\), if \(v\) is of the corresponding type \(c\). Since a vertex \(v\) can have multiple types \(c_1, c_2, \ldots\), it is possible that \(v\) is associated with several vertex summaries \(v_{s_1}, v_{s_2}, \ldots\) and therefore the index’s RDF class layer is considered a non-quotient summary. Kellou-Menouer and Kedad [21] provide a schema extraction approach based on clustering. It utilizes density-based clustering to establish a partition of the vertices based on type profiles. For each type \(T\), a type profile \(TP_T = \{(\text{label}_1, \alpha_1), (\text{label}_2, \alpha_2), \ldots\}\) is constructed, consisting of tuples of edge labels for outgoing edges \((v, w)\) and incoming edges \((w, v)\), with \(v \in T_j\). The associated probabilities \(\alpha_i\) denote how likely it is that a vertex \(v \in T_j\) has an edge with the respective \(\text{label}_i\). If a type profile \(TP_T\) contains all entries \((\text{label}_1, \alpha_1)\) of another type profile \(TP_k\) and every \(\alpha_i\) is greater than a certain threshold \(\theta\) (e.g., \(\theta = 0.6\)), then the vertices in \(T_k\) are added to \(T_j\) to create overlapping classes. Clustering can be found in more structural non-quotient approaches [22, 25, 29, 39].

Besides structural graph summarizing, the aforementioned classification is comprised of three more categories. Pattern-mining approaches utilize algorithms to identify frequent patterns in the original graph \(G\), which are then used to construct the summary graph \(SG\) [10]. Song et al. [34] construct \(d\)-summaries to summarize a knowledge graph \(G\). A summary \(P\), which is a graph pattern found in \(G\), is considered a \(d\)-summary, iff all the summary vertices \(u \in P\) are \(d\)-similar \((R_d)\) to all their respective original vertices \(v \in V\). Informally, \(uR_d v\) iff \((1) u\ and v\ share\ the\ same\ label\ and\ (2)\ for\ every\ neighbor\ \(u' \in P\ of\ u\ connected\ over\ an\ edge\ with\ label\ \(\ell\)\ there\ exists\ a\ respective\ neighbor\ \(v' \in V\ connected\ via\ the\ same\ edge\ label\ and\ \(u' R_{d-1} v'\)\).

Their definition of \(d\)-similarity is very similar to \(k\)-bismulation (Section 3.2) and mainly differs in the domain on which it is defined, namely summary vertices and original vertices. Statistical approaches construct summary graphs by considering quantitative properties of the input graph \(G\) [10]. The summarization operation \(k\)-SNAP [37] minimizes a function based on occurrences of user selected edge labels to produce a summary graph \(SG\), which contains exactly \(k\) vertex summaries. In its top-down approach, it starts by partitioning the graph based on user selected vertex attributes. Afterwards, the algorithm splits elements (vertex summaries) of the partition based on the aforementioned function, until the partition’s size is \(k\). Combining the first step, partitioning vertices by label, and the second step, minimizing a function which considers edge labels, \(k\)-SNAP can be considered a hybrid approach, combining structural and statistical concepts.

3 PRELIMINARIES

We define the data structure on which the bisimulation algorithms operate in Section 3.1. Section 3.2 formally defines \(k\)-bisimulation as an equivalence relation on graph vertices. Based on this, Section 3.3 introduces the bisimulation models of Schätzle et al. [33] and Kaushik et al. [20]. Finally, we briefly introduce our approach of a generic graph summarization model in Section 3.4, and show how the bisimulation models of Kaushik et al. and Schätzle et al. can be declaratively expressed in a formal language.

3.1 Data Structures

The algorithms operate on multi-relational, labeled graphs \(G\).\(^4\)

**Definition 1.** A multi-relational labeled graph \(G\) is defined as \(G = (V, E, l_V, l_E)\), where \(V = \{v_1, v_2, \ldots, v_n\}\) is a set of vertices and \(E \subseteq V \times V\) is a set of directed edges between the vertices in \(V\). Furthermore, each vertex \(v \in V\) and each edge \((u, v) \in E\) has zero or more associated labels from finite sets \(\Sigma_V\) and \(\Sigma_E\), respectively. The

\(^4\)We could, instead, use multi-relational property graphs, in which vertices and edges can also be labeled with key–value pairs. However, the bisimulation variants we study do not consider these key–value pairs, so multi-relational, labeled graphs suffice.
mapping of vertex and edge labels is done via the functions $l_V : V \rightarrow P(\Sigma_V)$ and $l_E : E \rightarrow P(\Sigma_E)$, respectively.

Figure 2 depicts an example graph $G = (V, E, \Sigma_V, \Sigma_E)$. The graph represents two universities and three employees.

Example. Figure 2 depicts an example graph $G = (V, E, l_V, l_E)$ with vertices

$$V = \{\text{asc}, \text{dri}, \text{jra}, \text{uess}, \text{uulm}\},$$

"Ansgar Scherp", "David Richerby", "Jannik Rau", "University of Essex", "University of Ulm"),

and edges

$$E = \{(\text{asc}, "Ansgar Scherp"), (\text{asc}, \text{uulm}),$$

(dri, "David Richerby"), (dri, uess),

(jra, "Jannik Rau"), (jra, uulm),

(uess, "University of Essex"), (uulm, "University of Ulm")\}

and respective vertex and edge label sets

$$\Sigma_V = \{\text{Lecturer, Literal, Organization, Professor, Student}\},$$

$$\Sigma_E = \{\text{name, worksAt}\}.$$

The graph represents two universities and three employees. For example, the edges (asc, uulm) and (jra, uulm) labeled with worksAt together with edges (asc, "Ansgar Scherp"), (jra, "Jannik Rau") and (uulm, "University Ulm") labeled with name and vertex labels Professor (asc), Student (jra) and Organization (uulm) state that the professor Ansgar Scherp and the student Jannik Rau both work at the organization University of Ulm.

In a graph $G = (V, E, l_V, l_E)$ the in-neighbors of a vertex $v \in V$ are the set of vertices $N^+(v) = \{u \mid (u, v) \in E\}$ from which $v$ receives an edge. Similarly, $v$'s out-neighbors are the set $N^-(v) = \{w \mid (v, w) \in E\}$ to which it sends edges. In- and out-neighbors are also referred to as predecessors and successors, respectively.

### 3.2 Bisimulation

A bisimulation is an equivalence relation defined on the vertices of a directed graph $G$ that groups together vertices with equivalent structural neighborhoods. Considering the outgoing edges to determine equivalence results in a forward bisimulation; considering the incoming edges gives backward bisimulation. Vertices $u$ and $v$ are forward- (respectively, backward-) bisimilar if, for every out- (resp., in-) neighbor $u'$ of $u$ there is a corresponding out- (resp., in-) neighbor $v'$ of $v$, and vice versa. Furthermore, the two neighbors $u'$ and $v'$ must be bisimilar [20, 38]. It is important to note that this definition corresponds to a complete bisimulation. For the neighbors $u'$ and $v'$ to be bisimilar, their in-/out-neighbors must be bisimilar as well, and so on. Thus, a change to an edge in a graph could cause vertices at an arbitrary distance from that edge to become bisimilar, or to cease to be bisimilar. A $k$-bisimulation is a bisimulation on $G$ that considers features a distance at most $k$ from a vertex when deciding whether that vertex is equivalent to some other.

**Definition 2.** The forward $k$-bisimulation $\approx_{fw}^k \subseteq V \times V$ with $k \in \mathbb{N}$ is defined as follows:

- $u \approx_{fw}^0 v$ for all $u, v \in V$,
- $u \approx_{fw}^{k+1} v$ if $u \approx_{fw}^k v$ and, for every successor $u'$ of $u$, there exists a successor $v'$ of $v$ such that $u' \approx_{fw}^k v'$.

Table 1 shows the forward bisimulation partitions of the graph in Figure 2. All vertices are 0-bisimilar. The 1-bisimulation partition has two blocks: vertices that have a successor, and vertices that do not. Finally, the 2-bisimulation partition contains three blocks: vertices with non-literal and literal successors, vertices with only literal successors, and vertices with no successors.

**Definition 3.** The backward $k$-bisimulation $\approx_{bw}^k \subseteq V \times V$ with $k \in \mathbb{N}$ is defined as follows:

- $u \approx_{bw}^0 v$ for all $u, v \in V$,
we modify Definition 2 as follows:

\[ \forall v \in V, \text{ predecessors of } v \text{ have predecessors.} \]

Table 3: Edge-labeled forward 2-bisimulation partition of the example graph according to Definition 4.

| \( k \) | Partition blocks |
|--------|-----------------|
| 0      | \( V \)         |
| 1      | \{asc, dri, jra\}, \{uess, uulm\}, literals |
| 2      | \{asc, dri, jra\}, \{uess, uulm, "Jannik Rau", "David Richerby", "Ansgar Scherp"\}, \{"University of Essex", "University of Ulm"\} |

Table 4: Vertex-labeled backward 2-bisimulation partition of the example graph according to Definition 5.

| \( k \) | Partition blocks |
|--------|-----------------|
| 0      | \{asc\}, \{dri\}, \{jra\}, \{uess, uulm\}, literals |
| 1      | \{asc\}, \{dri\}, \{jra\}, \{uess\}, \{uulm\}, \{"Jannik Rau"\}, \{"David Richerby"\}, \{"Ansgar Scherp"\}, \{"University of Essex"\}, \{"University of Ulm"\} |

Kaushik et al.’s Backward Bisimulation. This uses vertex labels so we modify Definition 3 to:

Definition 5. The vertex-labeled backward 2-bisimulation \( \cong_{bw}^{k} \subseteq V \times V \) with \( k \in \mathbb{N} \) is defined as follows:

- \( u \cong_{bw}^{0} v \iff I_{bw}(u) = I_{bw}(v) \),
- \( u \cong_{bw}^{k+1} v \iff u \cong_{bw}^{k} v \) and, for every successor \( u' \) of \( u \) there exists a predecessor \( v' \) of \( v \) such that \( u' \cong_{bw}^{k} v' \).

Table 4 shows the vertex-labeled backward 2-bisimulation partitions of the graph in Figure 2. Initially, the vertices are partitioned by label, giving the 0-bisimulation partition. Then, uess, uulm and the literal vertices are split according to the labels of their parents. In the final 2-bisimulation partition, no vertex is 2-bisimilar to any other: the partition only contains singletons.

Henceforth, if not further specified, \( \cong_{bw}^{k} \) and \( \cong_{bw}^{k} \) will correspond to the edge-labeled forward 2-bisimulation and the vertex-labeled backward 2-bisimulation.

3.4 Graph Summary Model

The BRS algorithm summarizes graphs with respect to a graph summary model (GSM) [7].

Definition 6. A graph summary model (GSM) is a mapping from graphs \( G = (V, E) \) to equivalence relations \( \sim \subseteq V \times V \).

The equivalence classes of \( \sim \) define a partition of the graph \( G \). A simple example of a GSM would be label equality, i.e., two vertices are equivalent iff they have the same label. Depending on the application, one might want to summarize a graph w.r.t. different GSMS. Therefore, the algorithm works with GSMS defined in the formal language FLUID [7]. FLUID’s building blocks are simple and complex schema elements, along with six parameterizations, of which we use two. The chaining parameterization enables summarization similar to k-bisimulation. The direction parameterization determines whether incoming or outgoing edges are used. The other four parameterizations described in [7] are not needed here.

The definitions of simple and complex schema elements, and parameterizations are based on [5, 7].

Definition 7. The three simple schema elements (SSEs) are:

1. Object cluster (OC) compares types and vertex identifiers of all neighboring vertices: two vertices \( v \) and \( v' \) are equivalent iff \( I_{v}(v) = I_{v'}(v') \) and \( N^{+}(v) = N^{+}(v') \).
(2) Predicate cluster (PC) compares labels: \(v\) and \(v'\) are equivalent if (i) their vertex label sets are both empty or both non-empty\(^5\) and (ii) they have the same labels on their outgoing edges: specifically, \(\{E(v, w) \mid w \in N^+(v)\} = \{E(v', w') \mid w' \in N^+(v')\}\).

(3) Predicate–object cluster (POC) combines PC and OC: \(v\) and \(v'\) are equivalent if, \(I_5(v, w) = I_5(v', w')\), \(N^*(v) = N^*(v')\), and \(I_2(v, w) = I_2(v', w')\) for all \(w \in N^+(v)\).

The SSEs consider local information about a vertex, i.e., the directly connected neighbors (“ego network”) and edge labels. To define equivalence based on vertices at distance up to \(k\), FLUID provides complex schema elements [7].

**Definition 8.** A complex schema element (CSE) consists of three equivalence relations and is defined as \(CSE := (∼_s, ∼_p, ∼_o)\). Vertices \(v\) and \(v'\) are equivalent, iff

1. \(v ∼_s v'\),
2. for all \(w \in N^+(v)\) there is a \(w' \in N^+(v')\) with \(I_2(v, w) ∼_p I_2(v', w')\) and \(w ∼_o w'\), and vice versa.

Introducing the identity relation \(id = \{(v, v) \mid v \in V\}\) and tautology relation \(Π = V \times V\), one is able to represent the three SSEs as CSEs, i.e., \(OC = (T, id, T)\), \(PC = (id, T, T)\), and \(POC = (T, id, id)\). An example of a CSE that takes more than local information into account is given by \((T, id, PC) = (T, id, T, T)\). This considers vertices as equal iff they have the same outgoing edge labels leading to neighbors with the same outgoing edge labels.

The chaining parameterization is of special interest to us, as it enables computing \(k\)-bisimulations by increasing the considered neighborhood for determining vertex equivalence [7]. The chaining parameterization is defined by nesting CSEs.

**Definition 9.** The chaining parameterization \(cp(CSE, k)\) takes as input a complex schema element \(CSE := (∼_s, ∼_p, ∼_o)\) and a chaining parameter \(k \in \mathbb{N}_{>0}\) and returns an equivalence relation \(CSE^k\) that corresponds to recursively applying CSE to a distance of \(k\) hops. \(CSE^k\) is defined inductively as follows:

\[
CSE^1 = (∼_s, ∼_p, ∼_o),
\]

\[
CSE^{k+1} = (∼_s, ∼_p, CSE^k) \quad \text{for } k > 1.
\]

To model backward \(k\)-bisimulations, we need to work with incoming edges, whereas the simple and complex schema elements consider only outgoing edges. In FLUID, this is done with the direction parameterization [7] but, here, we simplify notation.

**Definition 10.** For any simple or complex schema element \(SE\), the inverse schema element \(SE^{-1}\) is defined analogously to \(SE\) but using the relation \(E^{-1} = \{(y, x) \mid (x, y) \in E\}\) in place of the graph’s edge relation \(E\). In particular, \((∼_s, ∼_p, ∼_o)^{-1}\) denotes the inverse of the CSE \((∼_o, ∼_p, ∼_s)\).

Using the chaining parameterization, the edge-labeled forward \(k\) -bisimulation of Schätzle et al. ([33] and Definition 4) can be defined. Using chaining and the inverse schema element definition allows us to define the vertex-labeled backward \(k\)-bisimulation of Kaushik et al. ([20] and Definition 5).

\[
CSE_{Sch} := cp((T, id, T), k)
\]

\[
CSE_{Kau} := cp((OC_{type}, T, OC_{type})^{-1}, k).
\]

Vertices are equivalent under \(OC_{type}\) iff they have the same labels [7].

## 4 ALGORITHMS

We introduce the single-purpose algorithms of Schätzle, Neu, Lau- sen, and Przajaciel-Zablocki [33] in Section 4.1 and Kaushik, Shenoy, Bohannon, and Gudes [20] in Section 4.2. Finally, we introduce the generic BRS algorithm in Section 4.3.

The first two algorithms compute summaries in a fundamentally different way to the BRS algorithm. At the beginning of the execution, BRS considers every vertex to be in its own equivalence class. During execution, vertices with the same vertex summary are merged. Therefore, BRS can be seen as a bottom-up approach. In contrast, Schätzle et al. consider all vertices to be equivalent at the beginning, and Kaushik et al. initially consider all vertices with the same label to be equivalent. The partition is then successively refined. These two algorithms can be seen as top-down approaches.

In a partition \(P_1 = \{B_{i1}, B_{i2}, \ldots\}\), the sets \(B_{ij}\) are known as blocks.

**Definition 11.** Let \(P_1 = \{B_{i1}, B_{i2}, \ldots\}\) and \(P_j = \{B_{j1}, B_{j2}, \ldots\}\) be two partitions of a finite set \(V\). \(P_i\) is a refinement of \(P_j\) if every block \(B_{ik}\) of \(P_i\) is contained in a block \(B_{jll}\) of \(P_j\).

In Table 5, the considered algorithms are compared concerning their paradigm (e.g., parallel vs. sequential), origin (which work they are based on), worst-case complexity, and the computed graph summary model. The following subsections describe the algorithms’ origins and procedure in more detail.

### 4.1 Native Schätzle et al. Algorithm

The approach taken by Schätzle et al. [33] is distributed and uses the MapReduce paradigm. It is based on an implementation of Blom and Orzan [4] (see Table 5) for reducing labeled transition systems (LTS), a form of a labeled directed graph, modulo strong bisimulation. Blom and Orzan’s implementation, in turn, is a distributed version of the “naive” method established by Kanellakis and Smolka [19]. Kanellakis and Smolka examine the problem of testing observational equivalence on Calculus of Communicating Systems (CCS) expressions, which can be represented as labeled directed graphs [19]. Their definition of \(k\)-limited observational equivalence corresponds to a stratified \(k\)-bisimulation.

Two differences between the work of Schätzle et al. and Blom and Orzan is that Schätzle et al. use RDF graphs and the Apache Hadoop Framework.\(^4\) This is a more modern framework than that of Blom and Orzan, who use the Message Passing Interface\(^5\) to compute \(k\)-bisimulations of labeled transition systems.

The approach of Schätzle et al. is detailed in this section, though the concepts are nearly the same for Blom and Orzan’s work. We

\(^5\)https://www.mpi-forum.org/index.html

\(^6\)https://hadoop.apache.org/
have adapted the definitions and pseudocode to multi-relational labeled property graphs (Definition 1).

Two fundamental concepts are the signature and the ID of a vertex $v$ with respect to the current iteration’s partition $P_i$.

**Definition 12.** The signature of a vertex $v$ with respect to a partition $P_i = \{B_{i1}, B_{i2}, \ldots\}$ of $V$ is given by

$$\text{sig}_{P_i}(v) = \{(\ell, B_{ij}) \mid (v, w) \in E \text{ with } I_E(v, w) = \ell \text{ and } w \in B_{ij}\}.$$ 

Hence, the signature of a vertex $v$ with respect to the current iteration’s partition $P_i$ denotes the set of outgoing edge labels to blocks of $P_i$ [33]. By Definition 4, vertices $u$ and $v$ are $(k+1)$-bisimilar, $u \equiv_t v$, iff $\text{sig}_{P_{k+1}}(u) = \text{sig}_{P_{k+1}}(v)$. Therefore the signature can be used to identify the block of a vertex and to represent the current bisimulation partition [33]. Thus, the signature of a vertex $v$ w.r.t. $P_i$ can be represented as

$$\text{sig}_{P_{k+1}}(v) = \{(l, \text{sig}_{P_{l+1}}(w)) \mid (v, w) \in E \text{ with } I_E(v, w) = \ell\}.$$ 

We now introduce the ID function. The nested structure of vertex signatures means that they may become very large\(^8\) so we use an identifier function to map a vertex to its block [33].

**Definition 13.** The ID of a vertex $v$ with respect to a partition $P_i$ is given by the function $ID_{P_i} : V \rightarrow \mathbb{N}$ defined by

$$ID_{P_i}(v) = \text{hash}(\text{sig}_{P_i}(v)),$$

which computes a hash value for $\text{sig}_{P_i}(v)$.

Now the signature of a vertex $v$ w.r.t. the current partition $P_i$ can be represented as

$$\text{sig}_{P_{k+1}}(v) = \{(l, ID_{P_{l+1}}(w)) \mid (v, w) \in E \text{ with } I_E(v, w) = \ell\}.$$ 

With $\text{sig}_{P_i}(v)$ and $ID_{P_i}(v)$, the procedure for computing an edge-labeled forward $k$-bisimulation partition is sketched in Algorithm 1. The initial partition is just $V$ (Line 2), as every vertex is $0$-bisimilar to every other vertex. Next, the algorithm performs $k$ iterations (lines 3–8). In each iteration $i \in \{1, \ldots, k\}$, the information needed to construct a vertex’s signature $\text{sig}_{P_i}(v)$ is sent to every vertex $v$ (Line 5). This information is the edge label $I_E(v, w) = p \in \Sigma_E$ and the block identifier $ID_{P_{l+1}}(w)$ for every $w \in N^+(v)$. The signature $\text{sig}_{P_i}(v)$ is then constructed using the received information, and the identifiers $ID_{P_i}(v)$ are updated for all $v$ (Line 6). At the end of each iteration, the algorithm checks if any vertex ID was updated, by comparing the number of distinct values in $ID_{P_i}$ and $ID_{P_{i-1}}$ (Line 7).\(^9\) If no vertex ID was updated, we have reached full bisimulation [4, 33] and hence can stop execution early. At the end, the resulting $k$-bisimulation partition $P_k$ is constructed by putting vertices $v$ in one block if they share the same identifier value $ID_{P_k}(v)$ (line 9).

| Algorithm       | Paradigm          | Based on                        | Worst-Case Complexity | Graph Summary Model |
|-----------------|-------------------|---------------------------------|------------------------|---------------------|
| Set Union Approaches (Bottom-Up) |                  |                                 |                        |                     |
| BRS [6]         | Parallel – Signal/Collect | Set Union Problem [36]       | $O(n + m \cdot a(m + n))$ | Generic, Section 3.4 |

Set Refinement Approaches (Top-Down)

| Schätzle et al. [33] | Distributed – MapReduce          | Blom and Orzan [4], a distributed version of Kanellakis and Smolka [19] | $O(m \cdot n + n^2)$ | Specific, Definition 4, Equation (1) |
| Kaushik et al. [20]  | Sequential                       | Paige and Tarjan [31]         | $O(k \cdot m)$        | Specific, Definition 5, Equation (2) |

**Table 5: Comparison of the considered algorithms.**

---

4.2 Native Kaushik et al. Algorithm

In contrast to the Schätzle et al. algorithm, the native Kaushik et al. [20] algorithm provides a sequential implementation and computes vertex-labeled backward $k$-bisimulation. It is an adapted version of Paige and Tarjan’s “naïve method” [31] for solving the relational coarsest partition problem. The notation used by Kaushik et al. and Paige and Tarjan are clarified such that they are uniform for both, which is not the case in the original papers. The details can be found in Table 10 in Appendix A.

For a set $S \subseteq V$, let $N^+(S) = \bigcup_{v \in S} N^+(v)$ be the set of successors of $S$, i.e., the vertices that have incoming edges from some vertex in $S$. The following definitions are due to Paige and Tarjan [31], but

---

\(^8\)Since blocks can be split but not merged, it suffices to count the blocks.
we have modified them to deal with backward instead of forward bisimulation.

Definition 14 (Stable subset). A subset $B \subseteq V$ is stable with respect to another subset $S \subseteq V$ if either $B \subseteq N^*(S)$ or $B \cap N^*(S) = \emptyset$.

That is, vertices in $B$ cannot be distinguished by their relation to $S$: either all vertices in $B$ get at least one edge from $S$, or none do.

Definition 15 (Stable partition). A partition $P_i$ of $V$ is stable with respect to a subset $S \subseteq V$ if every block $B_{ij} \in P_i$ is stable w.r.t. $S$. $P_i$ is stable if it is stable w.r.t. each of its blocks $B_{ij}$.

Thus, a partition $P_i$ is stable if none of its blocks $B_{ij}$ can be split into a set of vertices that receive edges from some $B_{ik}$ and a set of vertices that do not. This corresponds to the endpoint of a bisimulation computation: no further distinctions can be made.

Definition 16. In the relational coarsest partition problem, we are given a graph $G = (V, E, I_V, I_E)$ and an initial partition $P_0$ of $V$. We must find the coarsest stable refinement of $P_0$, the unique stable partition $P$ such that $P$ is a refinement of $P_0$ and every other stable partition is a refinement of $P$.

Note that, among all stable refinements of $P_0$, the coarsest stable refinement has the fewest blocks. The naïve method of Paige and Tarjan for solving the relational coarsest partition problem is defined in Algorithm 2. It uses the split() function. This function’s arguments are a partition $P$ and a vertex set $S$, which we call a “splitter”. split$(P, S)$ replaces every block $B$ of $P$ that is not stable w.r.t. $S$ with the two sub-blocks consisting, respectively, of vertices in $B$ that have successors in $S$, and those that do not.

Definition 17 (Split function). For any partition $P_i$ and set $S \subseteq V$, let split$(S, P_i)$ be the refinement of $P_i$ obtained by replacing each block $B_{ij} \in P_i$ that is not stable w.r.t. $S$, by the blocks $B_{ij} \cap N^*(S)$ and $B_{ij} - N^*(S)$. We refer to the set $S$ as a splitter.

This leads to Algorithm 2, which repeatedly refines the initial partition by using its own blocks or unions of them as splitters. When no more splitters exist, the partition is stable [31]. The resulting stable partition is equivalent to the full backward bisimulation of the initial partition $P_0$ [19].

Algorithm 2: Naïve algorithm for relational coarsest partition of initial partition $P_0$ of $V$

```
function naïveApproach(P_0)
    i ← 0;
    while ∃ a splitter $S$, which is a union of blocks $B_{ij} \in P_i$
        do
            i ← i + 1;
            $P_i$ ← split($S, P_{i-1}$);
    return $P_i$;
```

The algorithm of Kaushik et al., Algorithm 3, adapts Algorithm 2. The first difference is that in each iteration $i \in \{1, \ldots, k\}$ the partition is stabilized with respect to each of its own blocks (lines 7–16). This ensures that, after iteration $i$, the algorithm has computed the $i$-bisimulation [20], which is not the case in Algorithm 2. Second, blocks are split as in Definition 17, (lines 9–15). As a result, Algorithm 3 computes the $k$-backward bisimulation. To check if $P$ is the relational coarsest partition of $P_0$ (i.e., if full bisimulation has been reached), the algorithm uses the Boolean variable wasSplit (lines 6 and 15). If this is false at the end of an iteration $i$, the algorithm stops early (line 17). Moreover, Algorithm 3 tracks which sets have been used as splitters (line 3), to avoid checking for stability against sets w.r.t. which the partition is already known to be stable. The algorithm provided by Kaushik et al. does not include this. If a partition $P$ is stable w.r.t. a block $B$, each refinement of $P$ is also stable w.r.t. $B$ [31]. So after the partition $P$ is stabilized w.r.t. a block copy $B^\text{copy}$ (line 7 to line 16), we can add $B^\text{copy}$ to the usedSplitters set and not consider it in subsequent iterations.

Algorithm 3: Bisimulation Algorithm by Kaushik et al. [20]

```
function bisimKaushik(G = (V, E, I_V, I_E), k \in \mathbb{N})
    /* $P := \{B_1, B_2, \ldots, B_k\}$ */
    P ← partition V by label;
    usedSplitters ← \emptyset;
    for $i \leftarrow 1$ to $k$
        pcopy ← $P_i$;
        wasSplit ← false;
        for $B_j^\text{copy} \in pcopy - usedSplitters$
            do
                /* Use blocks of copy partition to stabilize blocks of original partition */
                for $B_j \in P$
                    succ ← $B_j \cap N^*(B^\text{copy}_j)$;
                    nonSucc ← $B_j - N^*(B^\text{copy}_j)$;
                    /* Split non-stable blocks */
                    if succ ≠ \emptyset and nonSucc ≠ \emptyset then
                        P.add(succ);
                        P.add(nonSucc);
                        P.delete($B_j$);
                        wasSplit ← true;
                        usedSplitters.add($B^\text{copy}_j$);
                    break;
                return $P$;
```

4.3 Generic BRS Algorithm

The parallel BRS algorithm is not specifically an implementation of $k$-bisimulation. Rather, one can define a custom graph summary model ~ in the formal language FLUID (see Section 3.4). The BRS algorithm then summarizes a graph w.r.t. ~.

However, the $k$-bisimulation models of Schätzle et al. and Kaushik et al. can be expressed in FLUID through the chaining parameterization $cp(C\text{SE}, k)$ (Definition 9) as demonstrated in Section 3.4. Thus, the BRS algorithm can compute $k$-bisimulation partitions. In other words, $k$-bisimulation can be incorporated into any graph summary model defined in FLUID. The algorithm is executed in parallel and
uses the Signal/Collect paradigm [6]. In Signal/Collect, vertices collect information from their neighbors, sent over the edges as signals. Using this paradigm, the algorithm is implemented as a set union approach, starting with every vertex in its own singleton set. Before outlining the algorithm, we give a necessary definition.

**Definition 18.** Suppose we have a graph summary $GS$ of $G = (V, E, ℓ, v, i)$ w.r.t. some GSM $∼$. For each $v ∈ V$, the vertex summary $vs$ is the subgraph of $GS$ that defines $v$’s equivalence class w.r.t. $∼$.

Our work is based on an implementation of the BRS algorithm by Blume et al. [5]. This algorithm uses hash-based messages during the parallel computation of the graph summary and has been shown to be memory efficient [5]. We give the pseudocode of the BRS algorithm in Algorithm 4 and briefly describe it below. An a step-by-step example run of the algorithm is provided in [5].

**Initialization (lines 3–5).** We compute the local schema information for each vertex $v ∈ V$ w.r.t. $∼_v$ and $∼_o$ of the provided graph summary model $(∼_v, ∼_o)$. Hashes of these are stored as the identifiers $i_d_v$ (line 4) and $i_d_o$ (line 5).

At the end of the initialization, every vertex has an identifier $i_d_v$ and $i_d_o$. Two vertices $v, v'$ are equivalent w.r.t. $(∼_v, ∼_o)$, iff $v.i_d_v = v'.i_d_v$. Thus, this initialization step can be seen as iteration $k = 0$ of bisimulation.

**Case of $k = 1$ bisimulation (lines 6–11).** Every vertex $v$ sends, to each in-neighbor $w$, its $i_d_v$ value and the edge schema $ℓ(w, v)$ (line 9), i.e., the set of labels of the edge $(w, v)$. Correspondingly, each vertex receives a set of schema $(L, i_d_o)$ pairs from its out-neighbors, which are collated into the set $M_o$ (line 10) and merged using an order-independent hash to give $v$’s new $i_d_o$ (line 11).

**Case of $k > 1$ bisimulation (lines 12–32).** In the first iteration (lines 13–19), every vertex $v$ sends a message to each of its out-neighbors $w$. The message contains $v.i_d_v$ and $i_d_o$ values, and the edge schema $ℓ(w, v)$ (line 15). Subsequently, the incoming messages of the vertex are merged into a set of tuples with the received information $(ℓ(w, v), i_d_v)$ and $(ℓ(w, v), i_d_o)$ (lines 16 and 17). Finally, the identifiers $i_d_v$ and $i_d_o$ of $v$ are updated by hashing the corresponding set (lines 18 and 19). Note that, whenever an update of an identifier value $i_d_v$ of vertex $v$ is performed, the algorithm combines the old $v.i_d_v$ with the new hash value, indicated by $@login$.

In the remaining iterations, the algorithm performs the same steps (lines 20–27), with the exception of identifier $i_d_v$. In merging messages for $i_d_v$ (line 25), when merging the messages in $i_d_v$, it is not necessary to consider $ℓ(w, v)$, as the iterations $2$ to $k − 1$ it is only needed to update the $i_d_v$ values using the hash function as described above. This is possible since $i_d_v$ by definition already contain the information about the edge schema $ℓ(w, v)$, which is computed in the first iteration.

In the final iteration, a last update of the identifiers w.r.t. the final messages is made (lines 28–32). The final messages received by the vertices are the values stored in the out-neighbors’ $i_d_o$ values. To this end, each vertex signals its $i_d_v$ value to its in-neighbors (line 30). Again, the messages a vertex receives are merged (line 31) and hashed to update the final $i_d_v$ value (line 32).

**Algorithm 4: Parallel BRS algorithm [5]**

```plaintext
1 function PARALLELSUMMARIZE(G, SG, (∼_v, ∼_o))
2 returns graph summary SG
3 /* Initialization */
4 for all v ∈ V do in parallel
5     v.id_v ← hash(VERTEXSCHEMA(v, E, ∼_v));
6     v.id_o ← hash(VERTEXSCHEMA(v, E, ∼_o));
7 /* If k = 1, only signal edge labels and v.id_o */
8 if k = 1 then
9     for all v ∈ V do in parallel
10         for all w ∈ V −{v} do
11             SendMsgs((ℓ(w, v), 0, v.id_o));
12             M_v ← {⟨L, i_d_o⟩ | (L, i_d_o) was received};
13             v.id_o ← v.id_o @ MergeAndHash(M_v);
14 else /* Signal initial messages. Update v.id_v and v.id_o */
15     for all v ∈ V do in parallel
16         for all w ∈ V −{v} do
17             SendMsgs((ℓ(w, v), v.id_v, v.id_o));
18             /* Collect all incoming messages of v */
19             M_v ← {⟨L, i_d_o⟩ | (L, i_d_o) was received};
20             M_o ← {⟨L, i_d_o⟩ | (L, i_d_v, i_d_o) was received};
21             /* Update identifiers by hashing the messages */
22             v.id_v ← v.id_v @ MergeAndHash(M_v);
23             v.id_o ← v.id_o @ MergeAndHash(M_o);
24 /* Signal messages k − 2 times. As above, but we do not include L when updating v.id_v. (See text.) */
25     for i ← 2 to k − 1 do
26         for all v ∈ V do in parallel
27             for all w ∈ V −{v} do
28                 SendMsgs((ℓ(w, v), v.id_v, v.id_o));
29                 M_v ← {⟨L, i_d_v⟩ | (L, i_d_v) was received};
30                 M_o ← {⟨L, i_d_v⟩ | (L, i_d_v, i_d_o) was received};
31                 v.id_v ← v.id_v @ MergeAndHash(M_v);
32                 v.id_o ← v.id_o @ MergeAndHash(M_o);
33 /* Signal final messages. Update v.id_v */
34     for all v ∈ V do in parallel
35         for all w ∈ V −{v} do
36             SendMsgs((∅, 0, v.id_o));
37             M_v ← {⟨L, i_d_v⟩ | (L, i_d_v) was received};
38             v.id_v ← v.id_v @ MergeAndHash(M_v);
39             SG ← FINDANDMERGE(SG, V);
40 return SG;
```

After this update, equivalence between any two vertices $v$ and $v'$ can be determined. Vertices with the same $i_d_v$ value are merged (line 33), which ends the $k$-bisimulation.
5 EXPERIMENTAL APPARATUS

5.1 Datasets

It is important to evaluate graph summarization algorithms on real-world datasets, as synthetic datasets do not capture the characteristics of real-world graphs [6, 26]. Thus, three real-world datasets and two synthetic datasets were chosen for the experiments.

We experiment with smaller and larger graphs. Table 6 lists statistics of these datasets, where \( |\text{rng}(I_V)| = |\{I_V(v) \mid v \in V\}| \) is the number of different label sets, \( |I_V(v)| \) is the number of labels of a vertex \( v \in V \), \( d(G) \) the average degree of vertices in \( G \), \( d_G(v) \) the degree of a vertex \( v \in V \), and \( \Delta(G) \) the maximum degree in \( G \).

5.1.1 Real-World Datasets. Three real-world datasets were chosen.

The Laundromat100M dataset contains 100 million edges of the LOD Laundromat service.\(^{11}\) LOD Laundromat automatically cleans existing linked datasets and provides the cleaned version on a publicly accessible Website [1]. The smaller dataset consists of about 29.87 million vertices and 87.85 million edges. Compared to the other two smaller datasets (BTC150M and BSBM100M), Laundromat100M contains a huge number of different labels (33, 431) and label sets (7, 373). However, on average a vertex’s label set \( I_V(v) \) only contains 0.93 labels. Laundromat100M has the smallest average and maximum values for both degree and in-degree.

The BTC150M and BTC2B datasets contain, respectively, approximately 150 million and 1.9 billion edges of the Billion Triple Challenge 2019 (BTC2019) dataset.\(^{12}\) Herrera, Hogan and Käfer statistically analyzed the entire BTC2019 dataset. They found that 93% of the total edges (-2 billion) originate from Wikidata.\(^{13}\) BTC150M is a chunk of those 2 billion edges. Moreover, investigation of used vocabularies, predicates, and classes, led Herrera et al. to the conclusion that BTC2019 is a “highly diverse dataset” [17]. The BTC150M dataset consists of about 5 million vertices and 145 million edges. Compared to the other two smaller datasets, it contains the least amount of different labels (69) and label sets (137). However, on average a vertex has 12.38 different labels, which is about 12 times more than the other two. Furthermore, BTC150M contains the greatest number of different edge labels (10, 750) and has the highest average and maximum degrees and the highest maximum in-degree. BTC2B consists of about 79.65 million vertices and 1.92 billion edges. It contains 113, 365 different label sets and 38, 136 different edge labels. On average, a vertex has 10.40 different labels, which is the largest value among the larger datasets.

5.1.2 Synthetic Datasets. Two versions of the Berlin SPARQL Benchmark (BSBM) [3] were used. Originally developed for “comparing the SPARQL query performance of native RDF stores with that of SPARQL-to-SQL rewriters” [3], the benchmark consists of a data generator, used here, and a test driver. The data generator produces RDF datasets which consist of products, vendors, offers, and reviews, simulating an e-commerce use case. Datasets of arbitrary sizes can be generated by specifying the number of products represented.

5.2 Procedure

An experiment consists of the algorithm to run, the dataset to summarize, and the bisimulation degree \( k \). In case of the BRS algorithm, it additionally consists of the graph summary model to use. We have two different graph summary models defined by Schätzle et al. [33] and Kaushik et al. [20]. Each model comes in two implementations, one native implementation as defined by the original authors and a generic implementation through our BRS algorithm. We use the terms BRS-Schätzle and BRS-Kaushik, respectively, to refer to our implementations of the GsMs of Schätzle et al. and Kaushik et al. using the BRS algorithm; we refer to our single-purpose implementations of these two GsMs as native Schätzle and native Kaushik.

The four implemented algorithms are applied on each of the five datasets, for a total of 20 experiments. Each experiment is executed with a bisimulation degree of \( k = 10 \), using the following procedure. We run the algorithm six times in a row with the specific configuration. We use the first run (run 0) as a warm-up run and hence do not account it for measurement of the dependent variables. The following five runs (runs 1–5) are used to measure the variables. Thus, the last five runs of all experiments follow the same environment, i.e., none of the experimental results are influenced by any processes that ran before the experiment started.

5.3 Implementation

All algorithms, i.e., the algorithms of Schätzle et al. and Kaushik et al. implemented as single-purpose algorithms and implemented using the BRS algorithm, are implemented in Scala upon the Apache Spark Framework. This API offers flexible support for parallel computation and message passing. For the sequential native Kaushik algorithm (Algorithm 3), the Apache Spark API was used to implement the graph data structure, the partition data structures, and to parse and initialize these data structures.

The stabilization routine of the native Kaushik algorithm (lines 4–18) is implemented sequentially and has limited scope for parallelization. The first inner loop (line 7) has the following problem: In each iteration of the second inner loop at line 8, there is a chance...
that new blocks are added to the current partition \( P_i \). This leads to two new blocks which must be checked for instability with respect to all the other block-copies in the copy partition.

For our experiments, we use an Ubuntu 20 system with 32 cores and 2 TB RAM, with exclusive access to avoid interference from other processes. The Apache Spark contexts were given the full resources, i.e., 32 cores and 1.94 TB heap space. Time and memory measurements were taken using the Apache Spark Monitoring API.

5.4 Measures

To evaluate the algorithms’ performance, we use their running time and memory consumption. For every run of an experiment we report the total run time, run time of each of the \( k \) iterations, and the maximum JVM on-heap memory during execution.

6 RESULTS

Table 7 lists the average total run time (minutes) for each experiment, i.e., each combination of the algorithms and datasets, for a \( k = 10 \) bisimulation. As the BRS algorithm takes an additional initialization step (see Algorithm 4, line 13 to line 19), Table 8 shows the breakdown of the running time of the algorithms in initialization and computing the actual \( k \) bisimulation iterations. It is separated into average run time for initialization and computation of all ten iterations. Native Kaushik also has an initialization step, where the vertices are partitioned based on their label. This partitioning is done in parallel and took less than a second, so we do not account for it separately. Table 9 reports the maximum JVM on-heap memory in GB used during the execution for each experiment.

6.1 Smaller Datasets (100M+ Edges)

Figure 3 shows plots of the average run time (minutes) for each of the ten iterations on the smaller datasets. As the iterations in the native Kaushik experiments take much longer than in the other experiments (Figures 3a, 3c and 3e), additional plots without the native Kaushik results are provided (Figures 3b, 3d and 3f), allowing easier comparison between the other algorithms.

| Graph               | \(|V|\) | \(|E|\) | \(|\Sigma_V|\) | \(|\mu(V)|\) | \(|\Sigma_E|\) |
|---------------------|--------|--------|----------------|----------------|----------------|
| Laundromat100M      | 30 M   | 88 M   | 33,431         | 7,373          | 0.93 \(\pm\) 44 | 5,630          |
| BSBM100M            | 18 M   | 90 M   | 1,289          | 2,274          | 1.02 \(\pm\) 13 | 39             |
| BTC150M             | 5 M    | 145 M  | 69             | 137            | 1.04 \(\pm\) 26 | 10,750         |
| BSBM1B              | 172 M  | 941 M  | 6,153          | 27,306         | 1.03 \(\pm\) 18 | 39             |
| BTC2B               | 80 M   | 1,919 M | 113,365       | 576,265        | 0.95 \(\pm\) 1.82 | 38,136         |

Table 6: Statistics of the datasets.

| Schätzle et al. | Kaushik et al. |
|-----------------|-----------------|
| BRS             | Native          |
| BRS             | Native          |
| Laundromat100M  | 5.56            | 5.60           | 586.66         |
| BSBM100M        | 6.46            | 4.20           | 77.84          |
| BTC150M         | 4.08            | 6.14           | 78.02          |
| BSBM1B          | 54.44           | 85.98          | OOM            |
| BTC2B           | 61.96           | 83.74          | canceled       |

Table 7: Average total run time (minutes) for the \( k=10 \) bisimulation. Numbers are averaged over 5 runs.

| Schätzle et al. | Kaushik et al. |
|-----------------|-----------------|
| BRS-Schätzle et al. | BRS-Kaushik et al. |
| Init. | Iterations | Init. | Iterations |
| Laundromat100M  | 1.10           | 4.46           | 1.10           | 4.50           |
| BSBM100M        | 1.00           | 5.46           | 1.05           | 4.15           |
| BTC150M         | 1.47           | 2.61           | 1.44           | 3.10           |
| BSBM1B          | 10.20          | 44.24          | 10.43          | 53.57          |
| BTC2B           | 16.22          | 45.74          | 16.29          | 69.63          |

Table 8: Initialization and bisimulation iteration (minutes) of average total run time for BRS.

| Schätzle et al. | Kaushik et al. |
|-----------------|-----------------|
| BRS             | Native          |
| BRS             | Native          |
| Laundromat100M  | 211.5           | 147.9          | 210.5          | 335.1          |
| BSBM100M        | 248.6           | 113.1          | 172.0          | 327.0          |
| BTC150M         | 140.6           | 107.7          | 130.1          | 181.7          |
| BSBM1B          | 1,248.2         | 1,335.4        | 1,249.2        | OOM            |
| BTC2B           | 1,249.1         | 1,249.7        | 1,249.3        | canceled       |

Table 9: Maximum JVM on-heap memory (GB) for the \( k=10 \) bisimulation over 5 runs.

BRS-Schätzle shows a relatively constant iteration time on all smaller datasets. On Laundromat100M (Figure 3b), the algorithm...
Figure 3: Average Iteration Times (minutes) of Experiments on Smaller Datasets.
computes an iteration in about 0.4 to 0.6 minutes. On BTC150M (Figure 3d) an iteration takes approximately 0.2 to 0.4 minutes and on BSBM100M (Figure 3f) 0.5 to 0.8 minutes. The total run time (Table 7) is lowest on BTC150M, taking about 4.08 minutes on average. On Laundromat100M, the average run time is about 5.56 minutes and on BSBM100M, the algorithm needs 6.46 minutes on average. On all smaller datasets the initialization step takes about 1 minute on average (Table 8).

Native Schätzle shows a similar behavior, with a relatively constant iteration time across the datasets. On Laundromat100M (Figure 3b), the algorithm computes an iteration in about 0.7 to 0.95 minutes. On BTC150M (Figure 3d) an iteration takes approximately 0.55 to 0.8 minutes. Finally, on BSBM100M (Figure 3f) the average iteration time ranges from 0.9 to 1.15 minutes. The average run time (Table 7) is fastest on BTC150M with 6.14 minutes on average. This is followed by an average run time of 7.72 minutes on Laundromat100M and 9.40 minutes on BSBM100M. Comparing the run times of the two implementations, BRS is slightly faster on all smaller datasets, but uses more memory (Table 9).

BRS-Kaushik shows similar results to BRS-Schätzle. Again iteration time is relatively constant on all datasets taking about 0.3 to 0.6 minutes on Laundromat100M (Figure 3b) and 0.3 to 0.5 minutes on BTC150M and BSBM100M (Figures 3d and 3f). As before, it is fastest on BTC150M with, an average run time of 4.54 minutes. On BSBM100M and Laundromat100M, the average run times are about 5.20 and 5.60 minutes, respectively. Again, initialization takes about 1 minute on all smaller datasets (Table 8).

Native Kaushik shows a different behavior across the datasets. It does not have constant iteration time on any dataset: the iteration time increases in early iterations until it reaches a maximum value, and then decreases. The only exception to this behavior occurs on BTC150M, where the iteration time decreases from iteration one to two before following the described pattern for the remaining iterations. On Laundromat100M (Figure 3a) an iteration takes about 10 to 100 minutes. The maximum iteration time is reached in iteration six. On BTC150M (Figure 3c) the iteration time ranges from approximately 1 to 17 minutes. Here, the maximum iteration time is reached in iteration five. Finally, on BSBM100M (Figure 3e) the algorithm needs about 0.6 to 30 minutes to compute an iteration. The maximum iteration time is iteration three on this dataset. Furthermore, the full bisimulation w.r.t. Definition 5 is reached and detected by the algorithm at iteration seven. Native Kaushik runs fastest on BSBM100M, taking about 77.84 minutes on average. This is followed by an average run time of 78.02 minutes on BTC150M and 586.66 minutes on Laundromat100M. The BRS implementation is significantly faster on all smaller datasets (Table 7) and uses slightly less memory (Table 9).

In summary, BRS-Schätzle computes the 10-bisimulation the fastest on BTC150M and Laundromat100M, taking 4.08 and 5.56 minutes on average. On BSBM100M, BRS-Kaushik runs the fastest, taking 5.20 minutes on average. Native Schätzle consumes the least memory on all smaller datasets.

6.2 Larger Datasets (1B+ Edges)

Figure 4 shows plots of the average run time (minutes) for each of the ten iterations on the larger datasets. BRS-Schätzle computes the 10-bisimulation on BTC2B and BSBM1B with a relatively constant iteration time. On BTC2B (Figure 4a) an iteration takes approximately 4 to 6 minutes and on BSBM100M (Figure 4b) 4 to 5 minutes. The total run time (Tables 7 and 8), is lowest on BSBM1B taking 54.44 minutes on average, with 10.20 minutes for initialization and 44.24 minutes for the 10 iterations. On BTC2B the algorithm needs 16.22 minutes to initialize the graph and 45.74 minutes to compute the 10 iterations. This results in a total run time of 61.96 minutes on average. Native Schätzler also has nearly constant running time across iterations. On BTC2B (Figure 4a) an iteration takes approximately 8 minutes. On BSBM1B (Figure 3f) the average iteration time ranges from 8 to 9.5 minutes. The average running time for all ten iterations is lowest on BTC2B with 83.74 minutes on average, compared to 85.98 minutes on BSBM1B. Comparing the run times of the two implementations, BRS is slightly faster on the BSBM1B
dataset and noticeably faster on BTC2B. Both algorithms consume about the same amount of memory during execution on BTC2B. On BSBM1B, native Schätzle uses slightly more memory than BRS.

Native Kaushik did not complete one iteration on BTC2B in 24 hours, so execution was canceled. The algorithm ran out of memory on BSBM1B (Table 9). Hence, the plots in Figure 4 only provide results for BRS and native Schätzle.

Finally, we consider BRS-Kaushik. It has a relatively constant iteration time on BSBM1B (Figure 4b), ranging from about 4 to 6 minutes. On BTC2B it shows a slightly different behavior (Figure 4a). In iterations one to three the time ranges from 2 to 4 minutes. From iterations four to nine the algorithm consistently takes about 8 minutes per iteration. In the final iteration ten, the running time increases slightly to about 9.5 minutes. BRS-Kaushik runs fastest on BSBM1B with 64 minutes on average, taking 10.43 minutes for initialization and 53.57 minutes for the 10 iterations. This is followed by 85.92 minutes on BTC2B with 16.29 minutes for initialization and 69.93 minutes for the 10 iterations.

In summary, on BTC2B and BSBM1B BRS-Schätzle computes the ten iterations the fastest taking 61.96 and 54.44 minutes on average respectively. Native Schätzle consumes slightly more memory on BSBM1B, whereas on BTC2B BRS and native Schätzle consume about the same amount of memory.

### 7 DISCUSSION

#### 7.1 Main Results

All the algorithms can compute 10-bisimulations on the small synthetic and real-world graphs. The parallel algorithms BRS and native Schätzle compute the 10-bisimulations in about 4.08 to 9.40 minutes on the smaller datasets. The sequential native Kaushik algorithm takes one to two orders of magnitude longer and was unable to compute a 10-bisimulation at all for the two largest datasets.

The parallel algorithms BRS and native Schätzle operate on the graph’s vertices and edges. As the numbers of vertices and edges do not change during execution, the algorithms perform the same number of steps in each iteration, taking a roughly constant amount of time. The peaks in certain iterations could be due to the elimination of duplicates when merging incoming messages, which can vary between iterations. Furthermore, both parallel algorithms perform significantly faster than the sequential Kaushik et al. algorithm. BRS takes 4.08 to 5.60 and native Schätzle 6.14 to 9.40 minutes on the smaller datasets, whereas the sequential algorithm takes 77.84 to 586.66 minutes. Hence, even for smaller graphs of 100M to 150M triples, we observe that our parallel algorithm is much faster than the sequential alternative.

The parallel algorithms perform well on both the synthetic BSBM1B graph and the real-world BTC2B graph. The results indicate that the native algorithm does not have any advantage over the generic one, as BRS outperforms native Schätzle on every dataset. Finally, we showed that BRS-Kaushik effectively parallelizes the sequential bisimulation algorithm of Kaushik et al.

Both BRS and native Schätzle seem to scale linearly with the input graph’s size. BTC150M contains approximately 5M vertices and 145M edges. BTC2B is comprised of approximately 80M vertices and 28 edges, which is a factor of about 18 and 14 respectively. BRS-Schätzle takes 4.08 minutes, BRS-Kaushik 4.54 minutes and native Schätzle 6.14 minutes on the smaller version. On the larger version, they take 15 times, 19 times and 14 times longer. BSBM1B contains about 10 times as many vertices and edges as BSBM100M. The experiments on BSBM1B took about 10 times longer for BRS-Schätzle, 12 times for BRS-Kaushik, and 9 times for native Schätzle.

Experimental results for native Kaushik indicate that the algorithm does not scale linearly with the input graph’s size. The worst-case complexity of Kaushik et al. is stated by the authors to be $O(k \cdot m)$ [20], where $m$ is the number of edges in the input graph. As the algorithm operates on blocks, we suggest that one should rather express the worst-case complexity in terms of the maximum number of blocks. In the worst case, the initial partition $P_0$ contains $n = |V|$ blocks, which corresponds to the scenario where each vertex has a different label set. Then, the algorithm checks for each of the $n$ blocks, if it is stable to each of the $n$ block-copies. This would be the only iteration, as in a partition of size $n$, no block can be split. Hence, the worst-case complexity of Algorithm 3 is $O(n^2)$, where $n$ is the number of vertices in the input graph. Rather, in practice, the worst-case complexity may be given by $O(b_{\text{max}} \cdot |P_i| \cdot k)$, where $b_{\text{max}}$ corresponds to the maximum number of blocks present in some iteration $i \in \{1, \ldots, k\}$ that have not been used as a splitter before (i.e., $|\text{Prop}_{\text{prop}} \cup \text{usedSplitters}|$ in Algorithm 3, line 7).

A reason for the long running time of native Kaushik on Laundromat100M could be the size of the initial partition $P_0$. It depends on the number of different label sets present in the graph’s vertices (Definition 5). Laundromat100M contains 33,431 different label sets, which is much higher than for BTC150M with 69 and BSBM100M with 1,289 (Table 6). As a consequence, the algorithm has to perform stability checks and (potential) splits on blocks more often than on the other datasets. This does not contradict the similar run times on BTC150M and BSBM100M (69 label sets vs. 1,289 label sets). First, BSBM100M reaches full bisimulation w.r.t. Definition 5 in iteration seven and hence execution is stopped early. Second, as can be seen in Figure 3e, the iteration time starts to decrease rapidly from iteration four to five on BSBM100M, going down from about 20 minutes to about 1 minute. This indicates that the partition hardly changes any further and therefore only a small number of stability checks and splits are performed in iterations five to seven.

Native Kaushik operates on the blocks of the graph’s current partition. In each iteration $i$, the algorithm produces a partition $P_i$, which is stable w.r.t. all the blocks in $P_{i-1}$ [20]. Consequently, if a block is split into two new blocks, these must also be checked for stability in that iteration (see also discussion in Section 5.3). Hence, the more a partition increases in an iteration, the more steps must be performed by the algorithm. In addition, bisimulation relationships between the vertices change far more often in early iterations [33]. This explains the specific shape of the run time curve (Figures 3a, 3c and 3e) for the native Kaushik et al. algorithm. The exception is BTC150M (Figure 3c), where the iteration time first decreases from iteration one to two, indicating that the 1- and 2-bisimulations of this graph are very similar.
7.2 Threats to Validity
All algorithms are implemented in Scala, in the same framework. Optimizations from the Spark Tuning Guide\textsuperscript{16} were used equally to implement the algorithms such as arrays of objects (tuples) and primitives (integers). Each algorithm was executed using the same procedure on the same machine with exclusive access during the experiments. Hence, the experiments for each of the algorithms were performed in the same environment.

Each experimental configuration was run six times, to account for possible side effects from the hardware caching data and other machine properties. The first run is discarded in the evaluations to address these potential side effects, leaving the reported averages over five independent measurements.

Our implementation of the native Schätzle et al. algorithm follows the original work [33], using sets of tuples for sending and merging messages. This could be optimized by using arrays of tuples instead. Another optimization could be to remove the check at each iteration whether a full bisimulation has been reached (line 7 in Algorithm 1). To do this, Schätzle et al. count the distinct block identifiers of the current iteration. They evaluated their algorithm experimentally on different versions of three synthetic benchmarks and on one real world graph [33]. On the synthetic datasets, full bisimulation was reached in iterations four (SP\textsuperscript{3}Bench), seven (LUBM), and twelve to fourteen (SIB). The real-world graph (DBPedia version 3.7) reached full bisimulation at iteration 37. Hence, together with the results provided here, where no graph reaches full bisimulation w.r.t. Definition 4 in the first 10 iterations, this step could be considered as unnecessary computing for the stratified bisimulations. However, we implement the algorithms as originally defined; optimizations are left for future work.

7.3 Generalization
In our experiments, synthetic and real-world graphs were considered, the latter being especially important to analyze the practical application of an algorithm [6, 26]. Two different GSMs were used for evaluation of the BRS algorithm. The GSM of Schätzle et al. computes a stratified forward bisimulation, based on edge labels (Definition 4). The GSM of Kaushik et al. computes a stratified backward bisimulation based on vertex labels (Definition 5). Hence, the two GSMs consider different structural features for determining vertex equivalence. Regardless, for both GSMs, BRS scales linearly with the input graph’s size and computes the 10-bisimulations the fastest on every dataset.

7.4 Future Work
The provided results can be used as a foundation for several future works. First, it would be interesting to see the performance of the BRS algorithm for more GSMs that can be found in the literature. Blume et al. [7] provide a large overview of existing GSMs that could be evaluated for values of $k$.

Further, the concept of operating on blocks during execution, rather than on vertices and edges, which is applied by the sequential Kaushik et al. algorithm, could be evaluated for a parallel implementation. Rajasekaran and Lee [32] provide a parallel version of Paige and Tarjan’s [31] optimized algorithm for computing relational

---

\textsuperscript{16}https://spark.apache.org/docs/latest/tuning.html
This research is the result of the Bachelor’s thesis of the first author at the University of Ulm in 2021. We thank Till Blume for very helpful discussions, his support in implementing the k-bisimulation algorithms in FLUID, and his graph summarization framework [6].

REFERENCES

1. Wouter Beek, Laurens Rietveld, Hamid R. Bazoobandi, Jan Wielemaker, and Abraham Bernstein, Chris Welty, Craig A. Knoblock, Denny Vrandecic, Paul De Bra, and Andreas Schep. 2013. 2013. External memory K-bisimulation reduction of big graphs. In GRADES 2013, co-located with SIGMOD/PODS 2013, New York, NY, USA, June 22-26, 2013. Proceedings, Part II (Lecture Notes in Computer Science, Vol. 7936). Springer, 227–240. https://doi.org/10.1007/978-3-319-64244-9_19

2. Fabio Benedetti, Sonia Bergamaschi, and Laura Po. 2015. Exposing the Underlying Software Stack (Scala and Apache Spark API). Also, a server machine with exclusive access during experimentation. For very large graphs, an external memory solution is desirable.

3. Mathias Konrath, Thomas Gottron, Steffen Staab, and Ansgar Scherp. 2012. Time and Memory Efficient Parallel Algorithm for Structural Graph Summaries and two Ex- tensions to Incremental Summarization and k-Bisimulation for Long k-Chaining. CoRR abs/1211.12495 (2012). arXiv:1211.12495 https://arxiv.org/abs/1211.12495

4. Mathias Konrath, Thomas Gottron, Steffen Staab, and Ansgar Scherp. 2018. Incremental and Parallel Computation of Structural Graph Summaries for Evolving Graphs. In CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, Mathieu d’Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 75–84. https://doi.org/10.1145/3343031.3411878

5. Angela Bonifati, Stefania Dumbrava, and Haridimos Kondylakis. 2020. Graph Summarization. CoRR abs/2004.14794 (2020). arXiv:2004.14794 https://arxiv.org/abs/2004.14794

6. Stéphane Campinas, Thomas Perry, Diego Ceccarelli, Renaul Delbo, and Giovanni Tummaselli. 2012. Introducing RDF Graph Summary with Application to Assisted SPARQL Formulation. In 29th International Workshop on Database and Expert Systems Applications, DEBS 2012, Vienna, Austria, September 3-7, 2012. Abdellah Hamed, Amin Tjoa, and Roland W. Wagner (Eds.). IEEE Computer Society, 261–266. https://doi.org/10.1109/DEBS.2012.38

7. Şelal Cebi, Ertuğrul Kaş, and Ali Aykıncıcu. 2018. Graph summarization: a survey. VLDB J. 26, 3 (2017), 295–327. https://doi.org/10.1007/s00778-018-0528-3

8. Marek Cigan, Kjetil Nørvaag, and Ladislav Hluchý. 2012. The SemSets model for graph summarization. VLDB J. 21, 4 (2012), 631–651. https://doi.org/10.1007/s00778-011-0274-x

9. Roy Goldman and Jennifer Widom. 1997. DataGuides: Enabling Query Formulation in Semistructured Databases. In VLDB’97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece. Matthias Jarke, Michael J. Carey, Klaus R. Dittrich, Frederich H. Lochovsky, Pericles Loucopoulos, and Manfred A. Jeusfeld (Eds.). Morgan Kaufmann Publishers Inc., 436–445. http://www.computer.org/csdl/proceedings/vldb/1997/P436.PDF

10. Thomas Gottron, Ansgar Scherp, Bastian Krayer, and Arne Peters. 2013. LODatio: using a schema-level index to support users infinding relevant sources of linked data. In Proceedings of the 7th International Conference on Knowledge Capture, K-CAP 2013, Banff, Canada, June 29-26, 2013, V. Richard Benjamins, Mathias d’Aquin, and Andrew Gordon (Eds.). ACM, 105–108. https://doi.org/10.1145/2479832.2479841

11. José-Miguel Herrera, Aidan Hogan, and Tobias Käfer. 2019. BTC-2019: The 2019 Billion Triple Challenge Dataset. In The Semantic Web – ISWC 2019 – 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11779). Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon (Eds.). Springer, 163–180. https://doi.org/10.1007/978-3-030-30796-7_11

12. Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard de Melo, Claudio Gutiérrez, José Emilio Labra Gayo, Sabrina Kirran, Sebastian Neumaier, Axel Polleres, Roberto Naviglio, Axel-Cyrille Ngonga Ngomo, Subway M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2020. Knowledge Graphs. CoRR abs/2003.02320 (2020). arXiv:2003.02320 https://arxiv.org/abs/2003.02320

13. Paris C. Kanelakis and Scott A. Smolka. 1990. CCS Expressions, Finite State Processes, and Three Problems of Equivalence. Inf. Comput. 86, 1 (1990), 43–68. https://doi.org/10.1016/0019-0571(90)90025-D

14. Raghav Kausik, Pradeep Shenoy, Philip Bohannon, and Elud Gudes. 2002. Ex- ploring Local Similarity for Indexing Paths in Graph-Structured Data. In Proceed- ings of the 18th International Conference on Data Engineering, San Jose, CA, USA, February 26 - March 1, 2002, Rakesh Agrawal and Klaus R. Dittrich (Eds.). IEEE Computer Society, 129–140. https://doi.org/10.1109/ICDE.2002.994703

15. Kzenon Kellou-Menouer and Zoubida Kedad. 2015. Schema Discovery in RDF Data Sources. In Conceptual Modeling - 34th International Conference, ER 2015, Stockholm, Sweden, October 19-22, 2015, Proceedings (Lecture Notes in Computer Science, Vol. 9381). Paul Johannesson, Mong-Li Lee, Stephen W. Liddle, Andreas L. Opdahl, and Oscar Pastor López (Eds.). Springer, 287–298. https://doi.org/10.1007/978-3-319-64244-9_19

16. Mathias Kounath, Thomas Gottron, Steffen Staab, and Ansgar Scherp. 2012. SchemaEX - Efficient construction of a data catalogue by stream-based indexing of linked data. J. Web Semant. 16 (2012), 52–58. https://doi.org/10.1016/j.websem.2012.06.002

17. Kristen LeFevre and Evimaria Terzi. 2010. GraSS: Graph Structure Summarization. In Proceedings of the SIAM International Conference on Data Mining, SDM 2010, Austin - May 1, 2010, Columbus, Ohio, USA. SIAM, 454–465. https://doi.org/10.1137/100797418

18. Yongming Luo, George H. L. Fletcher, Jan Hidders, Paul De Bra, and Yiqun Wu. 2013. Regularities and dynamics in bisimulation reductions of big graphs. In First International Workshop on Graph Data Management Experiences and Systems, GRADES 2013, co-located with SIGMOD/PODS 2013, New York, NY, USA, June 24, 2013, Peter A. Boncz and Thomas Neumann (Eds.). CWI/ACM, 13. https://doi.org/10.4135/978-1-4625-24548-8

19. Tomas Milo and Dan Suciu. 1999. Index Structures for Path Expressions. In Database Theory - ICDT ’99. 7th International Conference, Jerusalem, Israel, January 10-12, 1999, Proceedings (Lecture Notes in Computer Science, Vol. 1540). Catell Breali and Peter Buneman (Eds.). Springer, 277–295. https://doi.org/10.1007/3-540-49257-7_18

20. Saket Navlakha, Rajeev Rastogi, and Nisheeth Shrivastava. 2008. Graph summarization with bounded error. In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008. Jason Tsong-Li Wang (Ed.). ACM, 419–432. https://doi.org/10.1145/1376616.1376661
Single-Purpose Algorithms vs. a Generic Graph Summarizer for Computing $k$-Bisimulations on Large Graphs

[30] Thomas Neumann and Guido Moerkotte. 2011. Characteristic sets: Accurate cardinality estimation for RDF queries with multiple joins. In Proceedings of the 27th International Conference on Data Engineering, ICDE 2011, April 11-16, 2011, Hannover, Germany, Serge Abiteboul, Klemens Böhm, Christoph Koch, and Kian-Lee Tan (Eds.). IEEE Computer Society, 984–994. https://doi.org/10.1109/ICDE.2011.5767868

[31] Robert Paige and Robert Endre Tarjan. 1987. Three Partition Refinement Algorithms. SIAM J. Comput. 16, 6 (1987), 973–989. https://doi.org/10.1137/0216062

[32] Sanguthevar Rajasekaran and Insup Lee. 1998. Parallel Algorithms for Relational Coarsest Partition Problems. IEEE Trans. Parallel Distributed Syst. 9, 7 (1998), 687–699. https://doi.org/10.1109/71.707548

[33] Alexander Schätzle, Antony Neu, Georg Lausen, and Martin Przyjaciel-Zablocki. 2013. Large-scale bisimulation of RDF graphs. In Proceedings of the Fifth Workshop on Semantic Web Information Management, SWIM@SIGMOD Conference 2013, New York, NY, USA, June 23, 2013, Roberto De Virgilio, Fausto Giunchiglia, and Letizia Tanca (Eds.). ACM, 1–8. https://doi.org/10.1145/2484712.2484713

[34] Qi Song, Yinghui Wu, and Xin Luna Dong. 2016. Mining Summaries for Knowledge Graph Search. In IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain, Francesco Bonchi, Josep Domingo-Ferrer, Ricardo Baeza-Yates, Zhi-Hua Zhou, and Xindong Wu (Eds.). IEEE Computer Society, 1215–1220. https://doi.org/10.1109/ICDM.2016.0162

[35] Blerina Spahiu, Riccardo Porrini, Matteo Palfonari, Anisa Rula, and Andrea Maurino. 2016. ABSTAT: Ontology-driven Linked Data Summaries with Pattern Minimalization. In Proceedings of the 2nd International Workshop on Summarizing and Presenting Entities and Ontologies (SumPre 2016) co-located with the 13th Extended Semantic Web Conference (ESWC 2016), Anissaras, Greece, May 30, 2016 (CEUR Workshop Proceedings, Vol. 1605), Andreas Thalhammer, Gong Cheng, and Kalpa Gunaratna (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-1605/paper3.pdf

[36] Robert Endre Tarjan and Jan van Leeuwen. 1984. Worst-case Analysis of Set Union Algorithms. J. ACM 31, 2 (1984), 245–251. https://doi.org/10.1145/62.2160

[37] Yuanxuay Tian, Richard A. Hankins, and Jignesh M. Patel. 2008. Efficient aggregation for graph summarization. In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008, Jason Tsong-Li Wang (Ed.). ACM, 567–580. https://doi.org/10.1145/1376616.1376675

[38] Thanh Tran, Gunter Ladwig, and Sebastian Rudolph. 2013. Managing Structured and Semistructured RDF Data Using Structure Indexes. IEEE Trans. Knowl. Data Eng. 25, 9 (2013), 2076–2089. https://doi.org/10.1109/TKDE.2012.134

[39] Octavian Udrea, Andrea Pugliese, and V. S. Subrahmanian. 2007. GRIN: A Graph Based RDF Index. In Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada. AAAI Press, 1465–1470. http://www.aaai.org/Library/AAAI/2007/aaai07-232.php

APPENDIX

A MAPPINGS OF THE NOTATIONS USED BY KAUSHIK ET AL. AND PAIGE AND TARJAN

In Table 10 the notation used by Kaushik et al., and by Paige and Tarjan is mapped to a single scheme.
| Our notation | Meaning | Paige and Tarjan | Kaushik et al. |
|--------------|---------|-----------------|----------------|
| $V$          | Finite set on which the relational coarsest partition will be computed (interpreted as a set of vertices in this work) | $U$ | $V_G$ |
| $E \subseteq V \times V$ | Binary relation over the finite set $V$ (interpreted as a set of directed edges between the vertices in this work) | $E$ | $E_G$ |
| $P_i$ and $P_i^{copy}$ | Partition of $V$ and its copy in iteration $i \in \{0, 1, 2, \ldots\}$, where $P_0$ is the initial Partition | $P$ and $Q$ | $Q$ and $X$ |
| $S \subseteq V$ | Any subset of the finite set $V$ | $S$ | $A$ and $B$ |
| $N^+(S) = \{ y \in V \mid \exists x \in S \text{ with } xEy \}$ | The set of elements, which are related to any element $x \in S$ (if one interprets $V$ as vertices and $E$ as edges, then this refers to the successors of the vertices in $S$) | $E(S)$ | $\text{Succ}(A)$ |
| $N^-(S) = \{ x \in V \mid \exists y \in S \text{ with } xEy \}$ | The set of elements, which have any related element $y \in S$ (if one interprets $V$ as vertices and $E$ as edges, then this refers to the predecessors of the vertices in $S$) | $E^{-1}(S)$ | Not used |
| $B_{ij}$ | Block $j$ of Partition $P_i$ | $B$ | $Q$ and $X$ |

Table 10: Notation used in this work for the approach of Kaushik et al.