6-1-2011

Labeling Recursive Workflow Executions On-the-Fly

Zhuowei Bao  
*University of Pennsylvania*, zhuowei@cis.upenn.edu

Susan B. Davidson  
*University of Pennsylvania*, susan@cis.upenn.edu

Tova Milo  
*Tel Aviv University*, milo@post.tau.ac.il

Follow this and additional works at: [http://repository.upenn.edu/cis_reports](http://repository.upenn.edu/cis_reports)  
Part of the [Databases and Information Systems Commons](http://repository.upenn.edu/cis_reports)

Recommended Citation


This paper is posted at ScholarlyCommons. [http://repository.upenn.edu/cis_reports/962](http://repository.upenn.edu/cis_reports/962)  
For more information, please contact repository@pobox.upenn.edu.
Labeling Recursive Workflow Executions On-the-Fly

Abstract
This paper presents a compact labeling scheme for answering reachability queries over workflow executions. In contrast to previous work, our scheme allows nodes (processes and data) in the execution graph to be labeled on-the-fly, i.e., in a dynamic fashion. In this way, reachability queries can be answered as soon as the relevant data is produced. We first show that, in general, for workflows that contain recursion, dynamic labeling of executions requires long (linear-size) labels. Fortunately, most real-life scientific workflows are linear recursive, and for this natural class we show that dynamic, yet compact (logarithmic-size) labeling is possible. Moreover, our scheme labels the executions in linear time, and answers any reachability query in constant time. We also show that linear recursive workflows are, in some sense, the largest class of workflows that allow compact, dynamic labeling schemes. Interestingly, the empirical evaluation, performed over both real and synthetic workflows, shows that our proposed dynamic scheme outperforms the state-of-the-art static scheme for large executions, and creates labels that are shorter by a factor of almost 3.

Keywords
Dynamic Reachability Labeling, Recursive Workflow

Disciplines
Databases and Information Systems

Comments

This technical report is available at ScholarlyCommons: http://repository.upenn.edu/cis_reports/962
Labeling Recursive Workflow Executions On-the-Fly

Zhuowei Bao  
Department of Computer and Information Science  
University of Pennsylvania  
Philadelphia, PA 19104, USA  
zhuowei@cis.upenn.edu

Susan B. Davidson  
Department of Computer and Information Science  
University of Pennsylvania  
Philadelphia, PA 19104, USA  
susan@cis.upenn.edu

Tova Milo  
School of Computer Science  
Tel Aviv University  
Tel Aviv, Israel  
milo@post.tau.ac.il

ABSTRACT
This paper presents a compact labeling scheme for answering reachability queries over workflow executions. In contrast to previous work, our scheme allows nodes (processes and data) in the execution graph to be labeled on-the-fly, i.e., in a dynamic fashion. In this way, reachability queries can be answered as soon as the relevant data is produced. We first show that, in general, for workflows that contain recursion, dynamic labeling of executions requires long (linear-size) labels. Fortunately, most real-life scientific workflows are linear recursive, and for this natural class we show that dynamic, yet compact (logarithmic-size) labeling is possible. Moreover, our scheme labels the executions in linear time, and answers any reachability query in constant time. We also show that linear recursive workflows are, in some sense, the largest class of workflows that allow compact, dynamic labeling schemes. Interestingly, the empirical evaluation, performed over both real and synthetic workflows, shows that our proposed dynamic scheme outperforms the state-of-the-art static scheme for large executions, and creates labels that are shorter by a factor of almost 3.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications—scientific databases

General Terms
Algorithms, Performance, Theory

1. INTRODUCTION
Scientific workflow systems are now becoming “provenance aware” by automatically recording data and module dependency during execution (e.g., Taverna [14], VisTrails [7] and Kepler [5]). By using such information, provenance queries such as “Was data item A (or Module M) used to produce data item B, either directly or indirectly?” are enabled. Answering such queries entails evaluating reachability queries over large, graph-structured data, which can be expensive [8].

Reachability labels are an important tool for efficiently processing reachability queries on large graphs. The main idea is to assign each vertex a label such that, using only the label of any two vertices, we can quickly decide if one can reach the other. However, the effectiveness of this approach crucially depends on the ability to develop compact and efficient labeling schemes that take small storage space and allow fast query processing. More precisely, we say that a labeling scheme is compact if it uses logarithmic-size labels ($O(\log n)$ bits for any graph with $n$ vertices), and efficient if it answers any reachability query in constant time ¹. This is indeed the best one can hope for, as even assigning unique ids for $n$ vertices requires labels of $\log n$ bits.

An important observation in the context of workflow systems is that the execution graph (or run) from which provenance information is obtained is not arbitrary, but is derived from a workflow specification. Workflow specifications are commonly modeled as directed graphs whose vertices denote modules and whose edges denote data flow; furthermore, they are typically fairly small graphs (10s of vertices). A specification can be executed many times, using different data inputs or parameter settings, and generating multiple runs. A run is modeled as a directed, acyclic graph (DAG) in which vertices represent module executions and whose edges carry the data output by the source and input by the sink. Workflow runs can be much larger (1000s of vertices) and structurally more complex than the specification due to repeated execution of sub-workflows, e.g., sequentially (loops), in parallel (forked executions) or through recursion.

Much research has been devoted recently to develop compact and efficient labeling schemes for workflow runs [6, 13] and graphs in general [24, 16, 17, 15, 2, 9]. A significant shortcoming, however, of the existing schemes is that they all need to examine the entire graph before labeling is performed. This may not be realistic in our setting since scientific workflows can take a long time to execute and users may want to ask provenance queries over partial executions. Labeling must therefore be done on-the-fly. That is, we must label modules as soon as they are executed and data as soon as it is produced, and cannot modify the labels subsequently.

Our goal is thus to develop a dynamic labeling scheme for workflow runs. Dynamic labeling has been previously considered in the context of XML trees [10, 20, 23], but workflow

¹We follow the standard assumption that any operation on two words ($\log n$ bits) can be done in constant time [6].
runs can have an arbitrarily more complex DAG structure. Although there have been efficient dynamic algorithms for maintaining the transitive closure of DAGs, they all produce a linear-size index per vertex, which is unacceptable for large graphs. Nevertheless, we will show in this paper that the knowledge of the specification can be exploited to obtain a compact (logarithmic size) and efficient (constant query time) dynamic labeling scheme for runs.

We next give a brief summary of prior work on reachability labeling and highlight the contributions of this paper.

**Prior Work.** Reachability labeling has been studied for different classes of graphs in both static and dynamic settings. The main goal is to bound the maximum length of labels. Clearly, the more general the class of graphs is, the more difficult it is to obtain compact labeling schemes; dynamic labeling is also harder than static labeling. The maximum label lengths for different classes of graphs are summarized in Figure 1, and the main results of this paper are shaded.

![Figure 1: A Comparison of Maximum Label Length](image)

#### Static. (Trees) The earliest work for labeling static trees [22] proposed an interval-based scheme that uses labels of $2 \log n$ bits, where $n$ is the number of nodes in the tree. Considerable effort [1, 4, 18, 12] was devoted to reduce the constant factor (2). The best known scheme [4] uses labels of $\log n + O(\sqrt{\log n})$ bits, which is still separated from the known lower bound of $\log n + O(\log \log n)$ bits [3]. Motivated by the fact that XML trees are not deep, recent work [12] developed a scheme that uses labels of $\log n + 2 \log d + O(1)$ bits, where $d$ is the depth of the tree.

#### (Workflow Runs) Workflow runs are modeled as DAGs derived from a given specification. [6] proposed a compact static scheme for labeling runs that uses labels of $3 \log n + O(1)$ bits. However, it can only be applied to non-recursive workflows (with only loops and forks). [13] also proposed a static scheme for labeling runs by transforming the graph into a tree and then applying the interval-based scheme. Since the size of the new tree can be exponential in the size of the original graph, it results in linear-size labels.

#### (General DAGs) In contrast to the above results, compact labeling is impossible for general directed acyclic graphs (DAGs), since a known lower bound on the maximum label length is $\Omega(n)$ bits. This triggered several alternative approaches for efficiently answering reachability queries over large DAGs: Chain Decomposition [15], Tree Cover [2] and 2-Hop [9]. Other recent work includes Path-Tree [17] and 3-Hop [16] that combine the previous three approaches, and GRAIL [24] that is based on randomized interval labeling.

#### Dynamic. (Trees) The dynamic problem is harder than the static case; it was shown in [10] that labeling dynamic trees requires labels of $\Omega(n)$ bits. [10] also proposed a prefix-based scheme, which provides a matching upper bound of $O(n)$ bits, and if the depth of the dynamic tree is bounded by a constant, it produces labels of $O(\log n)$ bits. Other variant prefix-based schemes with similar bounds were also studied, e.g., ORDPATH [20], implemented in Microsoft SQL Server, supports frequent inserts in XML documents, and DDE [23] is tailored for both static and dynamic XML documents.

#### (Workflow Runs and General DAGs) To our knowledge, the present work is the first to study dynamic labeling of workflow runs (and more generally of DAGs). The main contributions of this paper are summarized as follows.

- We propose a formal model, based on graph grammars, that captures a rich class of workflows with recursion, loops and forks. Based on this model we define execution-based and derivation-based dynamic labeling problems for workflow runs (Section 2).
- To get a handle on the difficulty of the two problems, we first provide tight lower and upper bounds of $\Theta(n)$ bits on the maximum label length. As a side effect, we also give tight bounds of $n - 1$ bits for the general problem of labeling dynamic DAGs (Section 3).
- Nevertheless, we identify a common class of workflows with linear recursion, and show that dynamic, yet compact ($O(\log n)$ bits) labeling is possible for linear recursive workflows. Moreover, our scheme labels a dynamic run in linear time, and answers any reachability query in constant time (Sections 4 and 5).
- We also show that linear recursive workflows are, in some sense, the largest class of workflows that allow compact dynamic labeling schemes (Section 6).
- Finally, we empirically evaluate the proposed dynamic labeling scheme over both real and synthetic workflows. Interestingly, our dynamic scheme creates even shorter labels for large runs than the state-of-the-art static scheme [6] by a factor of almost 3 (Section 7).

## 2. MODEL AND PROBLEM STATEMENT

We start with notations and basic definitions over graphs in Section 2.1. An informal description of our workflow model is given in Section 2.2, followed by a formalization based on graph grammars in Section 2.3. Finally, Section 2.4 formulates the dynamic workflow labeling problems.

### 2.1 Preliminaries

Throughout the paper, the term graphs refers to directed acyclic graphs with no self-loops or multi-edges. Every vertex of a graph can be associated with two kinds of labels.

#### The one, given in the graph, denotes a module name, and the other, created by our algorithm, is used for answering reachability queries. To distinguish the two labels, we will use two-terminal graphs.
and a single target sink, denoted by $t(g)$. Given a finite set $\Sigma$ of names, the set of all two-terminal graphs whose vertices are labeled by names chosen from $\Sigma$ is denoted by $G_\Sigma$.

Next, we introduce four graph operations, namely series composition, parallel composition, vertex insertion and vertex replacement. The first two operations are used to formalize loop and fork executions in Section 2.3. The last two operations are used to formalize execution-based and derivation-based dynamic workflow runs in Section 2.4.

**Definition 1.** A series composition of two-terminal graphs $g_1, g_2, \ldots, g_n$ forms a new two-terminal graph, denoted as $S(g_1, g_2, \ldots, g_n)$, by taking the union of their vertex sets and edges sets, and adding $(t(g_i), s(g_{i+1}))$ for all $1 \leq i \leq n - 1$.

**Definition 2.** A parallel composition of two-terminal graphs $g_1, g_2, \ldots, g_n$ forms a new graph, denoted as $P(g_1, g_2, \ldots, g_n)$, by simply taking the union of their vertex sets and edge sets.

**Definition 3.** An insertion of a vertex $v$ to a graph $g$, with respect to a subset $C$ of vertices of $g$, forms a new graph, denoted as $g + \{v, C\}$, by adding $v$ and $(v', v)$ for all $v' \in C$.

**Definition 4.** A replacement of a vertex $u$ of a graph $g$ with another graph $h$ forms a new graph, denoted as $g[u/h]$, by deleting $u$ and all edges incident to $u$, and adding $h$ and $(v, s)$ for all predecessors $v$ of $u$ and all sources $s$ of $h$ and $(t, v)$ for all successors $v$ of $u$ and all sinks $t$ of $h$.

### 2.2 Workflow Model

Our workflow model has two components: workflow specification and workflow run. A workflow specification describes a particular execution of the given specification.

**Workflow Specification.** A workflow specification defines the control and data flow between a set of modules by means of a DAG. In this graph, each vertex represents a module, which takes a set of data items as input and produces a set of data items as output, and is labeled with a module name. Each directed edge represents the data flow between two modules (i.e., data items that are produced by one module and consumed by the other). We also assume that each workflow has a single source (i.e., with no incoming edges), which sends out all initial data and starts the execution, and a single sink (i.e., with no outgoing edges), which collects all final results and stops the execution.

The modules are either atomic or composite. Atomic modules are treated as “black boxes”, since their internal structure is hidden. In contrast, composite modules, treated as “white boxes”, are known to be implemented by other sub-workflows. Intuitively, we can open a white box by replacing the composite module with the corresponding sub-workflow. Some composite modules are allowed to be repeatedly executed in series or in parallel. We call them loop and fork modules respectively. Note that a composite module can be implemented by a sub-workflow that contains other composite modules (including itself). This may lead to recursion.

**Example 1.** Our running example of a workflow specification is shown in Figure 2, where the uppercase letters (i.e., $L, F, A, B, C$) are the names of composite modules, and the lowercase letters (i.e., $s_0, \ldots, s_k, t_0, \ldots, t_k$) are the names of atomic modules. In particular, $L$ and $F$ are the names of loop and fork modules respectively; $g_0$ is a start graph. The thick arrows describe the possible implementations of each composite module (e.g., $A$ has two possible implementations $h_3$ and $h_4$). Also observe that $A$ and $C$ form a recursion.

**Workflow Run.** A workflow specification is repeatedly executed using different data input and parameter settings. A valid workflow run begins with the start graph, and selects one possible implementation to execute for each composite module (“or” semantics). For a loop or fork module, the selected implementation is repeatedly executed one or more times in series or in parallel, respectively. Moreover, it must execute all the modules in the start graph and the selected implementation graphs (“and” semantics). Since all composite modules are expanded during the execution, the resulting workflow run consists only of atomic modules.

**Example 2.** One possible run derived from the specification in Figure 2 is shown in Figure 3, where $v_1, \ldots, v_{20}$ are unique identifiers for atomic modules. In this run, $h_1$ (the implementation of a loop module) is replicated twice in series. In the first copy of $h_1$, $h_2$ (the implementation of a fork module) is replicated twice in parallel. For purposes of illustration, we show only the detailed execution for one copy of $h_2$. Observe that, due to the recursion over $A$ and $C$, $h_3$ and $h_4$ may be repeatedly executed until $h_k$ is selected.

### 2.3 Workflow Grammar

We next present a formalization of our workflow model based on graph grammars. A graph grammar is similar in spirit to the well-known string grammars, such as context-free grammars. It defines a set of graph-based productions (i.e., rules), and uses them to generate a set of graphs as its language. More precisely, we consider graph grammars...
based on vertex replacement, that is, every production defined by the grammar replaces a single vertex (i.e., the head of the rule) with a graph (i.e., the body of the rule). Our idea is to map every specification to a graph grammar in such a way that the set of possible runs, derived from this specification, corresponds to exactly its graph language. To capture loop and fork executions, our grammar may have an infinite (but controlled) number of productions.

**Definition 5.** A workflow specification is defined as a system $S = (\Sigma, \Delta, \Delta_{c}, \Delta_{r}, I, g_{0})$, where
- $\Sigma$ is a finite nonempty set of names;
- $\Delta$ is a nonempty subset of $\Sigma$, called the set of atomic names, and $\Sigma \setminus \Delta$ is called the set of composite names;
- $\Delta_{c}$ and $\Delta_{r}$ are two disjoint subsets of $\Delta \setminus \Delta_{c}$, called the sets of loop names and fork names respectively;
- $I$ is a finite set of pairs $(A, h)$, where $h \in G_{\Sigma}$ is called an implementation graph of $A \in \Sigma \setminus \Delta$; and
- $g_{0} \in G_{\Sigma}$ is called a start graph.

A vertex labeled with an atomic name is said to be an atomic vertex. Similarly, we can define composite vertex, loop vertex and fork vertex. In the rest of this paper, we loosely follow the convention that $v$, $v'$ and $v_{i}$ are used for atomic vertices; and $u$, $u'$ and $u_{i}$ for composite vertices.

**Example 3.** The specification in Figure 2 is written as $(\Sigma, \Delta, \Delta_{c}, \Delta_{r}, I, g_{0})$, where $\Sigma = \{s_{0}, \ldots, s_{6}, t_{0}, \ldots, t_{6}, L, F, A, B, C\}$, $\Delta = \{s_{0}, \ldots, s_{6}, t_{0}, \ldots, t_{6}\}$, $\Delta_{c} = \{L\}$, $\Delta_{r} = \{F\}$ and $I = \{(L, h_{1}), (F, h_{2}), (A, h_{3}), (A, h_{4}), (B, h_{5}), (C, h_{6})\}$.

**Definition 6.** Given a workflow specification $S = (\Sigma, \Delta, \Delta_{c}, \Delta_{r}, I, g_{0})$, the workflow grammar of $S$ is defined as a system $G = (\Sigma, \Delta, g_{0}, P)$, where $\Sigma$, $\Delta$, and $g_{0}$ are as in $S$, and $P$ is the possibly infinite set of productions below:

$P = \{A := h | (A, h) \in I\}$

$\cup \{A := S(h, h, \ldots, h) \mid (A, h) \in I, A \in \Delta_{c}, i > 1\}$

$\cup \{A := P(h, h, \ldots, h) \mid (A, h) \in I, A \in \Delta_{r}, i > 1\}$

Let $G = (\Sigma, \Delta, g_{0}, P)$ be a workflow grammar. We say that a graph $g_{2}$ is directly derived from a graph $g_{1}$ (with respect to $G$), denoted by $g_{1} \Rightarrow_{G} g_{2}$, if there is a production $A := h \in P$ such that $g_{2} = g_{1}[u/h]$, where $u$ is a composite vertex of $g_{1}$ with $\text{name}(u) = A$. Let $\Rightarrow_{G}^{*}$ be the reflexive and transitive closure of $\Rightarrow_{G}$, then $g_{2}$ is derived from $g_{1}$ (with respect to $G$), if $g_{1} \Rightarrow_{G}^{*} g_{2}$. The graph language of $G$, denoted by $L(G)$, is defined as the set of all two-terminal graphs which can be derived from the start graph and consist only of atomic vertices. Formally,

$L(G) = \{ g \in G_{\Delta} \mid g_{0} \Rightarrow_{G}^{*} g \}$

**Definition 7.** Given a workflow specification $S$, the set of workflow runs, with respect to $S$, is defined as $L(G)$, where $G$ is the workflow grammar of $S$.

**Example 5.** The workflow run in Figure 3 can be derived from the start graph using the workflow grammar in Figure 4. A graph derivation is sketched in Figure 5, where $u_{1}, \ldots, u_{8}$ are unique identifiers for composite vertices.

### 2.4 Dynamic Workflow Labeling Problems

The classical (static) graph reachability labeling problem is defined as follows. Given a graph $g$, assign each vertex of $g$ a reachability label such that, using only the labels of any two vertices of $g$, we can decide if one can reach the other.

This paper studies the problem of labeling dynamic workflow runs. It differs from the above problem in two aspects. Firstly, the input graph is a workflow run derived from a given specification. Formally, $g \in L(G)$, where $G$ is a given workflow grammar. Secondly, rather than taking the entire graph as input, we get a sequence of “updates” that leads to a graph $g \in L(G)$. We do not know the update sequence in advance, but receive them online. We must label all new vertices introduced by one update before the next update is applied, and cannot modify their reachability labels subsequently. Moreover, these labels can be used to determine reachability in any intermediate graph.

Based on different models of updates, we introduce two related dynamic workflow labeling problems. The first one describes the real life setting where run steps are reported and logged one by one, and the second one is used as an auxiliary tool for exploring the structure of workflow runs.

**Execution-Based Problem.** The first problem defines the update as a vertex insertion. Recall from Definition 3 that every insertion creates a new vertex along with a set of directed edges from existing vertices to this vertex. We begin with an empty graph $g_{0}$, and get a sequence of insertions that leads to a graph $g \in L(G)$, called a graph execution.
Definition 8. An execution-based dynamic reachability labeling scheme for a workflow grammar $G$ is a pair $(\phi, \pi)$, where $\phi$ is a labeling function and $\pi$ is a binary predicate. The input is an execution of a graph $g \in L(G)$, denoted by

$$g_0 \xrightarrow{+\phi(v_1, C_1)} g_1 \xrightarrow{+\phi(v_2, C_2)} g_2 \xrightarrow{+\phi(v_3, C_3)} \ldots \xrightarrow{+\phi(v_n, C_n)} g_n = g$$

where $g_0$ is an empty graph and $g_i = g_{i-1} + (v_i, C_i)$, for all $1 \leq i \leq n$. In the $i$th step of the graph execution, $\phi$ assigns a reachability label $\phi(v)$ for the new vertex $v_i$. Note that by that time we can see only the first replacements. $\phi$ and $\pi$ are such that for any execution of a graph $g \in L(G)$, any intermediate graph $g_i$ ($1 \leq i \leq n$), and any two vertices $v$ and $v'$ of $g_i$, $\pi(\phi(v), \phi(v')) = \text{true}$ if and only if $v \sim g_i v'$.

Derivation-Based Problem. The other problem defines the update as a vertex replacement. Recall from Definition 4 that every replacement substitutes an existing vertex for a new subgraph. We begin with the start graph $g_0$ (defined by the given workflow), and get a sequence of replacements that leads to a graph $g \in L(G)$, called a graph derivation.

Definition 9. A derivation-based dynamic reachability labeling scheme for a workflow grammar $G$ is a pair $(\phi, \pi)$, where $\phi$ is a labeling function and $\pi$ is a binary predicate. The input is a derivation of a graph $g \in L(G)$, denoted by

$$g_0 \xrightarrow{n_1(h_1)} g_1 \xrightarrow{n_2(h_2)} g_2 \xrightarrow{n_3(h_3)} \ldots \xrightarrow{n_k(h_k)} g_k = g$$

where $g_0$ is the start graph and $g_i = g_{i-1} + (h_i)$, for all $1 \leq i \leq k$. Initially, $\phi$ assigns a reachability label $\phi(v)$ for each vertex $v$ of $g_0$. In the $i$th step of the graph derivation, $\phi$ assigns a reachability label $\phi(v)$ for each vertex $v$ of $h_i$. Again, by that time we can see only the first $i$ replacements. $\phi$ and $\pi$ (as above) preserve the reachability between any pair of existing vertices. In fact, this is a necessary condition to allow persistent reachability labels. So, to prove correctness for both the execution-based and derivation-based schemes, we only need to ensure that for any two vertices $v$ and $v'$ of the final graph $g$, $\pi(\phi(v), \phi(v')) = \text{true}$ if and only if $v \sim g v'$.

At a first glance, the above two problems differ significantly from each other. The former receives and labels vertices one by one, while the latter by group. On the one hand, the execution-based model is more realistic, since it captures how runs advance; atomic modules of a workflow are executed in some topological ordering, due to data dependencies. On the other hand, the derivation-based model is more informative, since each step of a graph derivation describes exactly how a composite module is executed (e.g., which sub-workflow is used or how many times a loop is repeated). However, our study in this paper reveals a tight relation between the two problems. We will show in Section 5.3 that a derivation-based scheme can be converted to an execution-based scheme, which creates the same reachability labels. A further study in Section 6 shows that in general, the execution-based problem allows shorter labels.

3. COMPACTNESS RESULTS

The effectiveness of reachability labeling crucially depends on the ability to design a compact labeling scheme that allows fast query processing. As mentioned before, a labeling scheme is said to be compact if it creates labels of $O(\log n)$ bits for any input graph with $n$ vertices. Clearly, a compact labeling scheme creates the shortest possible labels up to a constant factor. In Section 5, we present a compact dynamic labeling scheme for a restricted class of workflows. Unfortunately, it is impossible to design a compact one for arbitrary workflows. In this section, we provide matching lower and upper bounds on the maximum label length for both execution-based and derivation-based problems.

3.1 Lower Bounds

To establish the lower bounds, we first show in Theorem 1 that for some fixed workflow grammar, any possible dynamic labeling scheme requires linear-size reachability labels.

**THEOREM 1.** There is a workflow grammar $G$ such that for any execution-based (resp. derivation-based) dynamic labeling scheme $(\phi, \pi)$ for $G$, there is an execution (resp. derivation) of a graph $g \in L(G)$ with $n$ vertices such that $\phi$ assigns a reachability label of $\Omega(n)$ bits for some vertex of $g$.

**PROOF.** We first consider the execution-based problem. Let $G$ be the workflow grammar shown in Figure 6, where $A$ is a composite name and all the others are atomic names. Given an execution-based dynamic labeling scheme $D = (\phi, \pi)$ for $G$, for all $k \geq 0$, we define $L_k(G)$ to be the set of all graphs $g \in L(G)$ that are derived from $g_0$ by applying the production $A := h_1 k$ times; and $S(D, k)$ to be the set of all reachability labels $\phi(v)$ that are assigned to a vertex $v$ with $\text{Name}(v) = a$ of a graph $g \in L_k(G)$. Finally, $N(k)$ is the minimum of $|S(D, k)|$ over all possible schemes $D$.

![Figure 6: A workflow grammar that requires linear-size reachability labels (proof of Theorem 1).](image)

At a first glance, the above two problems differ significantly from each other. The former receives and labels vertices one by one, while the latter by group. On the one hand, the execution-based model is more realistic, since it captures how runs advance; atomic modules of a workflow are executed in some topological ordering, due to data dependencies. On the other hand, the derivation-based model is more informative, since each step of a graph derivation describes exactly how a composite module is executed (e.g., which sub-workflow is used or how many times a loop is repeated). However, our study in this paper reveals a tight relation between the two problems. We will show in Section 5.3 that a derivation-based scheme can be converted to an execution-based scheme, which creates the same reachability labels. A further study in Section 6 shows that in general, the execution-based problem allows shorter labels.

![Figure 7: A graph $g \in L_{k+1}(G)$ that is derived from $g_0$ by applying the production $A := h_1 k$ times.](image)
that is derived from $A$ by applying the production $A := h_1 k$
times, $|S(D, k+1)| \geq |S_1| + |S_2| + 1 \geq 2N(k) + 1$. This holds for all possible schemes $D$, hence $N(k + 1) \geq 2N(k) + 1$.

Since $N(0) = 0$ and $N(1) = 1$, we can prove by induction that $N(k) > 2^{k/2}$ for all $k \geq 2$. Therefore, we must assign a reachability label of at least $k/2$ bits for some vertex of a graph $g \in L_k(G)$. Finally, observe that $g$ is derived from $g_0$ by applying the production $A := h_1 k$ times and the other production $A := h_2 k + 1$ times. Let $n$ be the number of vertices of $g$, then $n = 3 + 4k + (k + 1) = 5k + 4$. So $k/2 = (n - 4)/10 = \Omega(n)$. The theorem follows.

We can use a similar proof for the derivation-based problem. The only modification is that, rather than inserting three vertices, we apply only one step of a graph derivation to obtain the intermediate graph shown in Figure 7.  \qed

3.2 Matching Upper Bounds

To match the above lower bounds, we first present a simple execution-based dynamic labeling scheme $(\phi, \pi)$, which uses linear-size reachability labels. Given an execution of a graph $g$ with $n$ vertices, let $v_i$ be the $i$th vertex to be inserted, then $\phi(v_i)$ is a binary string of $i - 1$ bits. It simply encodes the reachability with respect to the previous $i - 1$ vertices already inserted to the graph. Formally, for all $1 \leq i \leq n$ and $1 \leq j \leq i - 1$, let $\phi(v_i)[j]$ be the $j$th bit of $\phi(v_i)$, then

$$\phi(v_i)[j] = \begin{cases} 1 & \text{if } v_i \rightarrow v_j, \\ 0 & \text{otherwise} \end{cases}$$

To decide if $v \rightarrow v'$, we first compute the index of $v$ and $v'$ by the length of $\phi(v)$ and $\phi(v')$. Let $i = |\phi(v)| + 1$ and $i' = |\phi(v')| + 1$. Then $v \rightarrow v'$ can be decided by being

$$\pi(\phi(v), \phi(v')) = \begin{cases} \text{true} & \text{if } i < i' \text{ and } \phi(v')[i] = 1 \\ \text{false} & \text{otherwise} \end{cases}$$

The maximum length of labels used by this scheme is $n - 1$ bits, which matches the lower bound of $\Omega(n)$ bits in Theorem 1. In fact, this scheme can be used to label executions of arbitrary DAGs. It was shown in [10] that even labeling dynamic trees with $n$ nodes requires labels of $n - 1$ bits. Hence, we provide as a side benefit tight lower and upper bounds of $n - 1$ bits on the maximum label length for the general problem of labeling (execution-based) dynamic DAGs.

However, the above execution-based scheme does not work for the derivation-based problem, because a graph derivation may not introduce new vertices in a topological ordering. In Section 5.2, we will present a compact derivation-based dynamic labeling scheme, which creates logarithmic-size reachability labels for a restricted class of workflows. If we use that scheme to label arbitrary workflows, it guarantees linear-size labels. Details are deferred to Section 6.

4. QUERYING DYNAMIC WORKFLOWS WITH LINEAR RECURSION

Linear-size reachability labels do not scale to large graphs. As demonstrated in the proof of Theorem 1, such large labels are required when workflows have unrestricted recursion. Luckily, workflows that one encounters in practice typically have a more restricted, linear form of recursion (to be formally defined below), which does allow for compact dynamic labeling. Indeed, we will see in Section 6 that the class of linear recursive workflows is the largest for which compact derivation-based dynamic labeling is possible.

The rest of this section is organized as follows. Section 4.1 defines the class of linear recursive workflows. To develop the labeling schemes, we first introduce a tree representation for linear recursive workflows, called the explicit parse tree, in Section 4.2, and then describe how to efficiently answer reachability queries using explicit parse trees in Section 4.3.

4.1 Linear Recursive Workflows

Let $G = (\Sigma, \Delta, g_0, P)$ be a workflow grammar. We say that a name $A$ directly induces a name $B$ (in $G$), denoted by $A \rightarrow_C B$, if there is a production $A := h \in P$ such that $h$ has a vertex $v$ with Name$(v) = B$. Let $\rightarrow_C^*$ be the reflexive and transitive closure of $\rightarrow_C$. We say that $A$ induces $B$ (in $G$), if $A \rightarrow_C^* B$. Given a production $A := h \in P$, a vertex $v$ of $h$ is said to be recursive, if Name$(u)$ induces $A$.

Example 6. Consider the workflow grammar in Figure 4. $A$ directly induces $B$ and $C$, due to the presence of $A := h_3$. Moreover, in this production, the vertex labeled with $C$ is recursive, since $C$ directly induces $A$ by $C := h_6$.

Definition 10. A workflow grammar is said to be linear recursive, if any production has at most one recursive vertex.

Example 7. It can be verified that the workflow grammar in Figure 4 is linear recursive. Observe that $A := h_3$ has only one recursive vertex (labeled with $C$). In contrast, the workflow grammar in Figure 6 is not linear recursive, since $A := h_1$ has two recursive vertices (both labeled with $A$).

4.2 Explicit Parse Tree

We start by considering an arbitrary workflow grammar $G$. The derivation of a graph $g \in L(G)$ can be naturally captured by a canonical parse tree $t$, whose nodes represent nested subgraphs and edges represent composite vertices created during the graph derivation. The root of $t$ corresponds to the start graph $g_0$. A subgraph $h_1$ is the parent of a subgraph $h_2$ if the graph derivation replaces a composite vertex $v$ of $h_1$ with $h_2$, and the edge from $h_1$ to $h_2$ represents $v$.

Example 8. The canonical parse tree for the graph in Figure 3 is shown in Figure 8, whose nodes (denoted by dashed boxes) and edges (denoted by bold lines) are annotated with nested subgraphs and composite vertices that they represent. Note that $x_3, x_4$ are unique identifiers for nodes of the tree; $u_1, u_4, u_5$ and $v_1, v_2, v_3, v_4$ are unique identifiers for composite and atomic vertices of the graph respectively. Consider the edge $(x_3, x_4)$ annotated with $u_4$. It implies that the production $L := S(h_1, h_1)$ is applied to replace the loop vertex $u_1$ of $h_0$ with a series composition of two $h_1$'s.

For linear recursive workflow grammars, we can convert a canonical parse tree to an explicit parse tree by inserting three kinds of special nodes: $L$ (loop) nodes, $F$ (fork) nodes and $R$ (recursive) nodes. The children of an $L$ or an $F$ node represent one or more copies of the same loop or fork subgraph, which are combined in series or in parallel respectively; and the children of a $R$ node represent a sequence of nested subgraphs, which form a linear recursion.

Example 9. The explicit parse tree for the graph in Figure 3 is shown in Figure 9, where $x_0, ..., x_{15}$ are unique
identifiers for nodes of the new tree. Comparing with the canonical parse tree in Figure 8, we can see that $x_1$ is split into two nodes $x_2$ and $x_{13}$ rooted at a special $L$ node $x_1$. Moreover, $x_1, x'_5, x'_6$ are moved to be the children $x_6, x_8, x_9$ of a special $R$ node $x_7$. Note that $x_6, x_8, x_9$ are linked by dashed edges annotated with recursive vertices $u_5$ and $u_6$.

It is important to note that the canonical parse tree for graphs generated by a fixed grammar may have unbounded depth due to recursion. However, in the explicit parse tree, the sequence of nested subgraphs in a linear recursion are flattened to be the children of a $R$ node. Hence, for linear recursive grammars, the depth of the explicit parse tree is bounded by a constant that depends only on the grammar.

**Lemma 4.1.** Given a linear recursive workflow grammar $G = (\Sigma, \Delta, g_0, \mathcal{F})$, let $t$ be an explicit parse tree for a graph $g \in \mathcal{L}(G)$ and $d$ be the depth of $t$, then $d \leq 2|\Sigma \setminus \Delta|$.

**Proof.** Let $r$ be the root of $t$. Consider a leaf node $x$ at the deepest level of $t$. Let $k_1$ and $k_2$ be the number of special and non-special nodes on the path from $r$ to $x$ respectively. Since the parent of a special node must be a non-special node, and both $r$ and $x$ are non-special nodes, we have $k_1 \leq k_2 - 1$. Hence, $d = k_1 + k_2 - 1 \leq 2k_2 - 2$. On the other hand, since each outgoing edge of a non-special node is annotated with a composite vertex, there is totally $k_2 - 1$ vertices annotated on the path from $r$ to $x$. Moreover, these vertices must have distinct composite names, since all recursive vertices are annotated on dashed edges. Hence, $k_2 - 1 \leq |\Sigma \setminus \Delta|$. It follows that $d \leq 2k_2 - 2 \leq 2|\Sigma \setminus \Delta|$.

4.3 **Answering Reachability Queries**

Let $t$ be an explicit parse tree for a graph $g$. To avoid confusion, we use $x$ and $y$ to refer to nodes of $t$ and $u$ and $v$ to vertices of $g$. Note that in this section, we abuse the notation slightly by using $u$ and $v$ to refer to both composite and atomic vertices of $g$. $\text{Annt}(x)$ denotes the subgraph annotated on a non-special node $x$, and $\text{Annt}(x, y)$ denotes the composite vertex annotated on an edge $(x, y)$. Recall that a graph $g_2$ is said to be derived from a graph $g_1$ if $g_2$ can be obtained from $g_1$ by applying a sequence of productions. We extend this notion to vertices as follows: A vertex $v$ is *directly derived from* a vertex $u$, denoted by $u \Rightarrow v$, if a production $\text{Name}(u) := h$ is applied to replace $u$ with a graph $h$, and $v$ is a vertex of $h$. Let $\Rightarrow^*$ be the reflexive and transitive closure of $\Rightarrow$, then $v$ is derived from $u$, if $u \Rightarrow^* v$.

To efficiently answer reachability queries using explicit parse trees, we introduce the notions of context and origin.

**Definition 11.** A non-special node $x$ of $t$ is said to be the **context of a vertex** $v$ of $g$, if $v$ is a vertex of $\text{Annt}(x)$.

**Definition 12.** A vertex $u$ of $g$ is said to be the **origin** of a vertex $v$ of $g$ with respect to a non-special node $y$ of $t$, if $v$ is derived from $u$, and $y$ is the context of $u$.

Note that the context and origin are always unique and can be defined for both atomic and composite vertices.

**Example 10.** Consider the explicit parse tree in Figure 9. The context of $v_5$ (bottom left) is $x_7$, since $v_5$ is an atomic vertex of $\text{Annt}(x_7)$. The origin of $v_5$ with respect to $x_2$ (top left) is $u_2$, since $u_2$ is a composite vertex of $\text{Annt}(x_2)$ from which $v_5$ is derived. Similarly, the origin of $u_5$ (bottom right) with respect to $x_6$ (bottom left) is $u_5$ (on the dashed edge).

The main idea is as follows. To decide if $v$ can reach $v'$ in $g$, we find the least common ancestor of their context $x$ and $x'$, denoted by $\text{LCA}(x, x')$. If $\text{LCA}(x, x')$ is a special $L$ or $F$ node, we can immediately answer “yes” or “no” by showing that $v$ and $v'$ belong to two distinct copies of the same loop (reachable) or fork (unreachable); otherwise (if $\text{LCA}(x, x')$ is a special $R$ node or a non-special node), we show that the original query for $v$ and $v'$ over $g$ can be reduced to a simple query for their origins $u$ and $u'$ with respect to a small subgraph $h$. The details are given in Lemma 4.2.
LEMMA 4.2. Let $t$ be an explicit parse tree for a graph $g$. Given any two vertices $v$ and $v'$ of $g$, let $x$ (resp. $x'$) be the context of $v$ (resp. $v'$) in $t$, then

- if $\text{LCA}(x, x')$ is a special node, let $y$ (resp. $y'$) be a child of $\text{LCA}(x, x')$ who is an ancestor of $x$ (resp. $x'$). Assume w.l.o.g. that $y$ is on the left of $y'$.
  - if $\text{LCA}(x, x')$ is an $\mathcal{L}$ node, then $v \sim_y v'$;
  - if $\text{LCA}(x, x')$ is an $\mathcal{F}$ node, then $v \sim_{y'} v'$, $v' \sim_{y'} v$;
  - if $\text{LCA}(x, x')$ is a $\mathcal{R}$ node, let $u$ (resp. $u'$) be the origin of $v$ (resp. $v'$) with respect to $y$ (note that $u' = \text{Annt}(y, z)$, where $z$ is the right sibling of $y$), and $h = \text{Annt}(y)$, then $v \sim_y v'$ iff $u \sim_h u'$.
- if $\text{LCA}(x, x')$ is a non-special node, let $u$ (resp. $u'$) be the origin of $v$ (resp. $v'$) with respect to $\text{LCA}(x, x')$, and $h = \text{Annt}(\text{LCA}(x, x'))$, then $v \sim_y v'$ iff $u \sim_h u'$.

Proof. First of all, we claim the following lemma, which easily follows from Definition 4. The proof is omitted.

LEMMA 4.3. Suppose that a graph $g_2$ is derived from a graph $g_1$. Let $v$ and $v'$ be two vertices of $g_2$, and $u$ and $u'$ be two vertices of $g_1$, such that $v$ (resp. $v'$) is derived from $u$ (resp. $u'$). Then $v \sim_{u'} v'$ if and only if $u \sim_{u_1} u'$.

We prove Lemma 4.2 by four cases. (1) If $\text{LCA}(x, x')$ is an $\mathcal{L}$ node, let $y_1, y_2, \ldots, y_k$ be all children of $\text{LCA}(x, x')$ (including $y$ and $y'$), and $h = S(\text{Annt}(y_1), \text{Annt}(y_2), \ldots, \text{Annt}(y_k))$. Let $u$ (resp. $u'$) be the origin of $v$ (resp. $v'$) with respect to $y$ (resp. $y'$). Since $y$ is on the left of $y'$, by Definition 1, $u \sim_h u'$. By Lemma 4.3, $v \sim_{u'} v'$; (2) If $\text{LCA}(x, x')$ is an $\mathcal{F}$ node, the lemma can be proved similarly by Definition 2; (3) If $\text{LCA}(x, x')$ is a $\mathcal{R}$ node, let $u' = \text{Annt}(y, z)$, where $z$ is the right sibling of $y$, and let $w$ be the origin of $v'$ with respect to $y$. Since $y$ is on the left of $y'$, $w$ is derived from $u'$. Since $u'$ is derived from $u$, $v'$ is also derived from $u'$. Hence, $u'$ is the origin of $v'$ with respect to $y$. By Lemma 4.3, $v \sim_{u'} v'$ if and only if $u \sim_{u_1} u'$; and (4) If $\text{LCA}(x, x')$ is a non-special node, by Lemma 4.2, $v \sim_{u'} v'$ if and only if $u \sim_h u'$.

Example 11. We demonstrate the above four rules using the running example. First, consider $v_5$ and $v_{16}$ (top right) in Figure 9. The least common ancestor of their context $x_7$ and $x_{12}$ is an $\mathcal{L}$ node $x_{11}$. By Lemma 4.2, $v_5 \sim_{v_{16}} v_{11}$, which is confirmed by Figure 3. Similarly, consider $v_5$ and $v_{13}$ (middle right). The least common ancestor of their context $x_{7}$ and $x_{10}$ is an $\mathcal{F}$ node $x_{3}$. Hence, $v_5 \sim_{v_{13}} v_{13}$ and $v_{13} \sim_{v_{5}} v_5$. Next, consider $v_8$ and $v_8$. The least common ancestor of their context $x_2$ and $x_5$ is a $\mathcal{R}$ node $x_3$, note that the dashed edges are ignored. Moreover, $u_4$ and $u_5$ are the origin of $v_5$ and $v_5$ with respect to $x_6$ (note that $u_5$ is annotated on the dashed edge $(x_6, x_8)$). Since $u_4 \sim_{v_{11}} u_5$, by Lemma 4.2, $v_{5} \sim_{v_{13}} v_{5}$. Finally, consider $v_5$ and $v_{11}$ (bottom left). The least common ancestor of their context $x_7$ and $x_{10}$ is a non-special node $x_6$. $u_4$ and $v_5$ are the origin of $v_5$ and $v_{11}$ with respect to $x_6$. Hence, $u_4 \sim_{v_{11}} v_{11}$ implies that $v_5 \sim_{v_5} v_{11}$.

5. LABELING DYNAMIC WORKFLOWS WITH LINEAR RECURSION

Our dynamic schemes are built upon a skeleton-based labeling workflow grammar [6]. As a preprocessing step, we label the workflow specification using any static reachability scheme, and then extend the reachability labels on the specification, called the skeleton labels, to label workflow runs on-the-fly.

We start by discussing how to label the specification in Section 5.1. Section 5.2 presents a compact derivation-based dynamic labeling scheme for linear recursive workflows; we sketch how to adapt it to an execution-based scheme in Section 5.3. Finally, Section 5.4 proves the correctness and analyze the quality of our proposed dynamic schemes.

5.1 Labeling Workflow Specifications

Given a workflow specification $S = (\Sigma, \Delta, \Delta_c, \Delta_x, \mathcal{I}, g_0)$, we want to label the start graph and all implementation graphs in $S$. Formally, the set of graphs to be labeled is $G(S) = \{g_0\} \cup \{h \mid (A, h) \in \mathcal{I}\}$

It is important to note that all graphs in $G(S)$ are small compared with runs derived from the specification. In practice, the largest real-life workflow that we have collected has fewer than 30 vertices, while a realistic run may repeatedly execute a loop or fork module (sub-workflow) hundreds of times. Therefore, we claim that any static scheme is scalable to label the specification. Our experiments in Section 7 show that even linear-size skeleton labels, created by the scheme described in Section 3.2, can take negligible storage overhead.

5.2 Derivation-Based Dynamic Scheme

Given a labeled specification, we next explain the design of reachability labels for its runs. Let $G$ be a linear recursive workflow grammar, and $t$ be an explicit parse tree for a graph $g$ in $L(G)$. Recall from Lemma 4.2 that to decide if $v$ can reach $v'$ in $g$ we only need to (1) find the least common ancestor $\text{LCA}(x, x')$ of their context $x$ and $x'$ in $t$; and (2) if $\text{LCA}(x, x')$ is a special $\mathcal{R}$ node or a non-special node answer an equivalent query for their origin $u$ and $u'$ with respect to a small subgraph $h$. To encode Step (1), we use a prefix-based scheme [18] for to label $t$. To encode Step (2), we enrich a prefix-based label with skeleton labels as well as other necessary information (e.g., node types).

The formal design of a reachability label is given below. We use $(\phi_G, \pi_G)$ to denote the static labeling scheme for the specification, and $(\phi_F, \pi_F)$ to denote our proposed dynamic labeling scheme for runs. Recall that $\phi_G$ and $\phi_F$ are labeling functions, and $\pi_G$ and $\pi_F$ are reachability predicates. To label a vertex $v$ of $g$, we consider a path in $t$

$$x_0 \xrightarrow{u_0} x_1 \xrightarrow{u_1} x_2 \xrightarrow{u_2} \ldots \xrightarrow{u_{k-1}} x_k$$

where $x_0$ is the root of $t$, $x_k$ is the context of $v$, and for all $0 \leq i \leq k-1$, $x_i$ is the parent of $x_{i+1}$ and $u_i = \text{Annt}(x_i, x_{i+1})$ is the composite vertex annotated on the edge $(x_i, x_{i+1})$. Note that $u_i = \text{null}$ if $x_i$ is a special node, otherwise, $u_i$ is the origin of $v$ with respect to $x_i$. To unify the notation, let $u_0 = v$. Then $\phi_F(v)$ consists of a list of entries $\phi_F(v) = \{\text{Entry}(x_0, u_0), \text{Entry}(x_1, u_1), \ldots, \text{Entry}(x_k, u_k)\}$ where $\text{Entry}(x_i, u_i) = (\text{index}, \text{type}, \text{skl}, rcc_1, rcc_2)$ is a tuple obtained from a pair $(x_i, u_i)$ by Algorithm 1.

The details of Algorithm 1 are explained as follows. The index $\text{index}$ of $x$ is a positive integer if $x$ is the $i$th child of its parent (Line 1). Note that the index of the root of $t$ is set
Algorithm 1 Entry Construction

Input: $(x, u)$ is a pair of a node of $t$ and a vertex of $g$ $(\phi, \pi_G)$ is a static scheme for labeling specification

Output: Entry$(x, u)$ is an entry for $(x, u)$

1: $\text{index} \gets$ the index of $x$
2: $\text{type} \gets$ the type of $x$
3: $\text{skl}_l, \text{rec}_l, \text{rec}_2 \gets$ null
4: if $x$ is a non-special node then
5:   $\text{skl} \gets \phi(u)$
6: if $\text{Annt}(x)$ has one recursive vertex then
7:   /* the parent of $x$ must be a special $\mathcal{R}$ node */
8:   $w \gets$ the recursive vertex of $\text{Annt}(x)$
9:   $\text{rec}_l \gets \pi_G(\phi_G(w), \phi_G(u))$
10:  $\text{rec}_2 \gets \pi_G(\phi_G(w), \phi_G(u))$
11:  $\text{end if}$
12:  $\text{end if}$
13:  $\text{return}$ $(\text{index}, \text{type}, \text{skl}, \text{rec}_1, \text{rec}_2)$

Example 12. We label the running example using the explicit parse tree in Figure 9. For example,

\[
\phi_9(v_5) = \{\text{Entry}(x_0, u_1), \text{Entry}(x_1, \text{null}), \text{Entry}(x_2, u_2), \text{Entry}(x_3, \text{null}), \text{Entry}(x_4, u_3), \text{Entry}(x_5, \text{null}), \text{Entry}(x_6, u_4), \text{Entry}(x_7, v_5)\}
\]

where

\[
\begin{align*}
\text{Entry}(x_0, u_1) &= (0, \mathcal{N}, \phi_G(u_1), \text{null}, \text{null}) \\
\text{Entry}(x_1, \text{null}) &= (1, \mathcal{L}, \text{null}, \text{null}, \text{null}) \\
\text{Entry}(x_2, u_2) &= (2, \mathcal{R}, \phi_G(u_2), \phi_G(u_3), \text{true}, \text{false}) \\
\text{Entry}(x_3, \text{null}) &= (1, \mathcal{L}, \phi_G(u_2), \phi_G(v_5), \text{null}, \text{null})
\end{align*}
\]

Since $u_5$ is the recursive vertex of $h_3$, by Algorithm 1,

\[
\begin{align*}
\text{Entry}(x_0, u_1).\text{rec}_1 &= \pi_G(\phi_G(u_4), \phi_G(u_3)) = \text{true} \\
\text{Entry}(x_0, u_1).\text{rec}_2 &= \pi_G(\phi_G(u_1), \phi_G(u_4)) = \text{false}
\end{align*}
\]

Similarly,

\[
\phi_9(v_{16}) = \{\text{Entry}(x_0, u_1), \text{Entry}(x_1, \text{null}), \text{Entry}(x_{12}, v_{16})\}
\]

where the first two entries are defined above, and

\[
\text{Entry}(x_{12}, v_{16}) = (2, \mathcal{N}, \phi_G(v_{16}), \text{null}, \text{null})
\]

The dynamic labeling algorithm $\phi$ can be divided into two interleaved steps. First, we generate the explicit parse tree in a top-down fashion by Algorithm 2. During this process, we also label all new vertices introduced in each step by Algorithm 3. Details are explained as follows.

\[\text{Algorithm 2 Dynamic Generation of Explicit Parse Tree}\]

Input: $g_0 \Rightarrow g_1 \Rightarrow g_2 \Rightarrow \ldots \Rightarrow g_k = g$

is a derivation of a graph $g \in L(G)$

Output: $t$ is an explicit parse tree for $g$

1: Create a node $r$ annotated with $g_0$
2: Insert $r$ as the root of $t$
3: for all $i := 1$ to $k$ do
4:   $y \gets$ the context of $u_i$ in $t$
5:   if $u_i$ is not recursive then
6:     if Name$(u_i)$ is a loop or a fork name then
7:       /* $h_i$ has no recursive vertices */
8:       Let $h_i = S(h_i, h_i, \ldots, h_i)$ or $P(h_i, h_i, \ldots, h_i)$
9:      Create a special $\mathcal{L}$ or $\mathcal{F}$ node $x$
10:     for all $j := 1$ to $l$ do
11:       Create a node $x_j$ annotated with $h_i$
12:       Insert $x_j$ as the $j$th child of $x$
13:     end for
14:   else
15:     if $h_i$ has one recursive vertex then
16:       Create a special $\mathcal{R}$ node $x$
17:       Create a node $x'$ annotated with $h_i$
18:       Insert $x'$ as the single child of $x$
19:     else
20:       Create a node $x$ annotated with $h_i$
21:       end if
22:     end if
23:     Insert $x$ as the next child of $y$
24:     Annotate the edge $(y, x)$ with $u_i$
25:   else
26:     Create a node $x$ annotated with $h_i$
27:     /* the parent of $y$ must be a special $\mathcal{R}$ node */
28:     Insert $x$ as the right sibling of $y$
29:     Create a dashed edge $(y, x)$ annotated with $u_i$
30:   end if
31: end for
32: return $t$

Since the skeleton labels are shared by multiple runs, skl is implemented as a pointer to the label, rather than the label itself.
The following lemma ensures that the three cases (1a), (1b) and (1c) in Algorithm 2 are indeed disjoint.

**Lemma 5.1.** Let $G = (\Sigma, \Delta, g_0, P)$ be a linear recursive workflow grammar. If a production $A := h \in P$ has a recursive vertex $u$ in $h$, then (1) $A$ is not a loop or a fork name; and (2) Name$(u)$ is not a loop or a fork name.

Proof. We first prove the part (1). Suppose $A$ is a loop or a fork name, by Definition 6, either $A := S(h, h) \in P$ or $A := P(h, h) \in P$. But both productions have at least two recursive vertices, which contradicts Definition 10.

We next prove the part (2). Suppose Name$(u)$ is a loop or a fork name. Consider two cases. (a) If Name$(u) = A$, then $A$ is a loop or a fork name, which contradicts (1). (b) If Name$(u) \neq A$, since $u$ is recursive, Name$(u)$ induces $A$. So there is a production Name$(u) := h' \in P$ such that for some vertex $u'$ of $h'$, Name$(u')$ induces $A$. Since $A$ directly induces Name$(u)$, Name$(u')$ induces Name$(u)$. Hence, Name$(u) := h'$ has a recursive vertex $u'$, which again contradicts (1).

**Algorithm 3:** During the dynamic generation of the explicit parse tree (Algorithm 2), we also perform the following labeling. For a non-special node $x$, we create a label $\phi_g(v)$ for each vertex $v$ of $A_{\text{nt}}(x)$. For a special node $x$, we also create a temporary label. By abusing the notation, we denote this label by $\phi(x)$. Note that to obtain a new label for a node $x$, we take an existing label from its parent $y$, and append only one new entry built by Algorithm 1 (Line 13, 16, 21).

**Algorithm 3 Labeling Function $\phi_g$**

Input: A derivation of a graph $g \in L(G)$

$(\phi_G, \pi_G)$ is a static scheme for $G$

Output: $t$ is an explicit parse tree for $g$

$\phi_g$ is a labeling function for $g$

1: Initially, create a root $r$ of $t$ (by Algorithm 2)

2: for each vertex $v$ of $A_{\text{nt}}(r)$ do

3: $\phi_g(v) \leftarrow \{\text{Entry}(r, v)\}$

4: end for

5: for each derivation step do

6: Update $t$ top-down (by Algorithm 2)

7: for each newly inserted node $x$ do

8: $y \leftarrow$ the parent of $x$

9: if $y$ is a non-special node then

10: $u \leftarrow A_{\text{nt}}(y, x)$

11: if $x$ is a non-special node then

12: for each vertex $v$ of $A_{\text{nt}}(x)$ do

13: $\phi_g(v) \leftarrow \text{append Entry}(x, v)$ to $\phi_g(u)$

14: end for

15: else

16: $\phi_g(x) \leftarrow \text{append Entry}(x, \text{null})$ to $\phi_g(u)$

17: end if

18: else

19: $x$ must be a non-special node */

20: for each vertex $v$ of $A_{\text{nt}}(x)$ do

21: $\phi_g(v) \leftarrow \text{append Entry}(x, v)$ to $\phi_g(y)$

22: end for

23: end if

24: end for

25: end for

26: return $t, \phi_g$

---

**Algorithm 4 Binary Predicate $\pi_G$**

Input: $\phi_g(v)$ is a reachability label for $v$

$\phi_g(v')$ is a reachability label for $v'$

$\pi_G$ is a binary predicate for skeleton labels

Output: $\pi_g(\phi_g(v), \phi_g(v')) = \text{true}$ iff $v \sim_G v'$

1: $i \leftarrow \min\{j \mid \phi_g(v_0^{\text{kl}})(j).\text{index} = \phi_g(v')(j).\text{index} \text{ and } \phi_g(v)(j+[1].\text{index} \neq \phi_g(v')(j+[1].\text{index}\}$

2: if $\phi_g(v)(i).\text{type} = L$ then

3: return $\phi_g(v)(i+[1].\text{index} < \phi_g(v')(i+[1].\text{index}\}$

4: else if $\phi_g(v)(i).\text{type} = F$ then

5: return false

6: else if $\phi_g(v)(i).\text{type} = R$ then

7: return $\phi_g(v)(i+[1].\text{index} < \phi_g(v')(i+[1].\text{index}\}$

8: return $\phi_g(v)(i+[1].\text{rec}1$

9: else

10: return $\phi_g(v')(i+[1].\text{rec}2$

11: end if

12: else $\{\phi_g(v)(i).\text{type} = N\}$

13: return $\pi_G(\phi_g(v)(i).\text{skl}, \phi_g(v')(i).\text{skl})$

14: end if

---

To decide if $v \sim_G v'$, we compare $\phi_g(v)$ and $\phi_g(v')$ using the binary predicate $\pi_G$ described in Algorithm 4, where $\phi_g(v)(i)$ denotes the $i$th entry of $\phi_g(v)$.

**Example 13.** Returning to the running example, consider $v_5$ and $v_{16}$ in Figure 9. Since $\phi_g(v_5)$ and $\phi_g(v_{16})$, shown in Example 12, share the first two common entries, moreover, $\phi_g(v_5)[2].\text{type} = L$

(1)

$\phi_g(v_5)[3].\text{index} = 1 < \phi_g(v_{16})[3].\text{index} = 2$

(2)

by Algorithm 4, $\pi_g(\phi_g(v_5), \phi_g(v_{16})) = \text{true}$. Note that (1) implies that the least common ancestor of their context $x_7$ and $x_{12}$ is a special $L$ node, and (2) implies that $x_2$ is on the left of $x_{12}$. By Lemma 4.2, $v_5 \sim_G v_{16}$. The other cases ($(v_5, v_{13})$, $(v_5, v_8)$ and $(v_5, v_{11})$) can be verified similarly.

5.3 Execution-Based Dynamic Scheme

The above derivation-based scheme can be converted to an execution-based scheme, which creates exactly the same reachability labels. The main challenge is how to dynamically build the explicit parse tree and figure out the context and origin of a newly inserted vertex, given only an execution of a graph. We first give a solution based on a natural restriction on the workflow specification, and then discuss how to remove the restriction by using execution logs.

Let $G(S)$ be the set of the start graph and all implementation graphs of a workflow specification $S$, as defined in Section 5.1. We assume that for all graphs $G \in G(S)$,

1. All vertices of $h$ have distinct names; and

2. $s(h)$ and $t(h)$ have unique atomic names, i.e., their names do not occur in any other graph in $G(S)$.

Recall that $s(h)$ and $t(h)$ denote the source and the sink of a sub-workflow $h$ that only distribute and collect the data. We call them the dummy modules. In fact, any specification can be modified to satisfy the above two conditions by renaming module names and introducing new dummy modules.

The execution-based labeling algorithm is sketched as follows. For a newly inserted vertex, we can decide if it is a
terminal or a non-terminal, by checking its module name (Condition 2). If it is a source, then we can infer one new step of the graph derivation, and update the explicit parse tree as before; otherwise, the context of this new vertex can be determined by any of its immediate predecessors, and the origin can be decided by again checking its module name (Condition 1). If it is a sink, then we know that the current step of the graph derivation is completed.

Example 14. Consider an execution of the graph in Figure 3, which inserts an atomic vertex \( v_1 \) in the 4th step. We start with an empty graph, and get the first vertex \( v_1 \) inserted. Since \( \text{Name}(v_1) = s_0 \), by checking the specification in Figure 2, we know that the start graph \( g_0 \) is being executed. So by Algorithm 2, we create the root of the explicit parse tree in Figure 9 that corresponds to \( g_0 \), and assign the reachability label \( \phi_0(v_1) \) according to Algorithm 3. Next, suppose \( v_2 \) is inserted. Again, by checking \( \text{Name}(v_2) = s_1 \), we know that the first copy of the loop subgraph \( h_1 \) is being executed. So we update the explicit parse tree in Figure 9 by inserting a special \( \mathcal{L} \) node \( x_1 \) along with its first child \( x_2 \). Note that although the original derivation-based scheme creates all the children of a special \( \mathcal{L} \) or \( \mathcal{F} \) node in a single step, the labeling function \( \phi_0 \) given by Algorithm 3 can be done on a node-by-node basis. Hence, we can assign the reachability label \( \phi_0(v_2) \) without seeing other copies of \( h_1 \). The remaining vertices can be labeled in a similar manner.

During the above labeling process, when the first vertex of a new subgraph is inserted, we can predict future insertions for other atomic vertices of this subgraph. Moreover, their reachability labels can be created at this point, though we do not give out these labels until they are actually inserted. In principle, we are allowed to modify these unassigned labels based on the upcoming insertions. We will see in Section 6 that this relaxation provides execution-based schemes with the potential to create shorter reachability labels.

To remove the restrictions, the only extra information we need is a mapping from vertices of the run to vertices of the specification. Note that in the above algorithm, this is done by comparing module names. In reality, most scientific workflow systems record this mapping in execution logs, by assigning a unique id for each module in the specification.

### 5.4 Correctness and Quality Analysis

**Theorem 2. (Correctness)** Let \( (\phi_0, \pi_0) \) be our dynamic labeling scheme for a linear recursive workflow grammar \( G \). For any graph \( g \in L(G) \) and any two vertices \( v \) and \( v' \) of \( g \), \( \pi_g(\phi_0(v), \phi_0(v')) = \text{true if and only if } v \sim_g v' \).

**Proof.** Let \( t \) be an explicit parse tree for \( g \). Let \( x \) (resp. \( x' \)) be the context of \( v \) (resp. \( v' \)). Let \( \text{LCA}(x, x') \) the least common ancestor of \( x \) and \( x' \). Let \( \phi_0(v)[i] \) be the \( i \)th entry of \( \phi_0(v) \). Recall that \( \phi_0(v)[i] = \text{Entry}(x, u_i) \) is obtained from a pair \((x, u_i)\) by Algorithm 1, where \( x_i \) is the ancestor of \( x \) at the \( i \)th level, and \( u_i = \text{null} \) if \( x_i \) is a special node, otherwise, \( u_i \) is the origin of \( v \) with respect to \( x_i \).

We prove the correctness of Algorithm 4. Line 1 computes the maximum common prefix of entries with the same index, say the first \( i \) entries. Then \( \phi_0(v)[i] = \text{Entry}LCA(x, x'), \text{null} \) and \( \phi_0(v')[i] = \text{Entry}LCA(x, x'), \text{null} \), where \( \text{null} \) is the origin of \( v \) or \( v' \) with respect to \( \text{LCA}(x, x') \). Consider four cases. (1) If \( \text{LCA}(x, x') \) is a special \( \mathcal{L} \) node (Line 2), then let \( \phi_0(v)[i+1] = \text{Entry}(y, \text{null}) \) and \( \phi_0(v')[i+1] = \text{Entry}(y', \text{null}) \), where \( y \) and \( y' \) are two distinct children of \( \text{LCA}(x, x') \). Hence, by Lemma 4.2, \( v \sim_g v' \) if and only if \( y \) is on the left of \( y' \). Note that the ordering of \( y \) and \( y' \) can be decided by their indexes (Line 3); (2) If \( \text{LCA}(x, x') \) is a special \( \mathcal{F} \) node (Line 4), then by Lemma 4.2, \( v \sim_g v' \) if and only if \( y \) is a sink, then we know that the current step of the graph derivation is completed.

Let \( \phi_0(v)[i] = \text{Entry}(y, \text{null}) \) and \( \phi_0(v')[i] = \text{Entry}(y', \text{null}) \), where \( y \) and \( y' \) are two distinct children of \( \text{LCA}(x, x') \). We assume without loss of generality that \( y \) is on the left of \( y' \). The other case can be handled in the same way. Let \( u \) (resp. \( u' \)) be the origin of \( v \) (resp. \( v' \)) with respect to \( y \). Note that \( \phi_0(v)[i+1] = \text{Entry}(y, u) \). Moreover, \( u' = \text{Annt}(y, z) \) is a recursive vertex of \( h = \text{Annt}(y) \) annotated on the dashed edge \((y, z)\), where \( z \) is the right sibling of \( y \). By Lemma 4.2, \( v \sim_g v' \) if and only if \( u \sim_h u' \). Hence, by Algorithm 1, \( v \sim_g v' \) if and only if \( \phi_0(v)[i+1] = \text{true} \) and (4) \( \text{LCA}(x, x') \) is a non-special node (Line 12), then \( \phi_0(v)[i] = \text{Entry}(LCA(x, x'), v) \) and \( \phi_0(v')[i] = \text{Entry}(LCA(x, x'), u) \), where \( u \) (resp. \( u' \)) is the origin of \( v \) (resp. \( v' \)) with respect to \( \text{LCA}(x, x') \). Let \( h = \text{Annt}(LCA(x, x')) \). By Lemma 4.2, \( v \sim_g v' \) if and only if \( u \sim_h u' \). Hence, by Algorithm 1, \( v \sim_g v' \) if and only if \( \phi_0(v)[i].\text{skl}, \phi_0(v')[i].\text{skl} = \text{true} \).

The quality of a labeling scheme \((\phi, \pi)\) is measured by label length, construction time (i.e., the time to compute \( \phi \)) and query time (i.e., the time to evaluate \( \pi \)). Among them, label length is the main factor. Since the derivation-based and execution-based schemes create same labels, they differ only in the construction time. All parameters for quality analysis are listed in Table 1, where \( G \) is a linear recursive grammar, \( t \) is an explicit parse tree for a graph \( g \in L(G) \) and \( h \) is a subgraph of \( g \). The size of a graph refers to the number of vertices. Note that for a fixed \( G, n_G \) and \( t_G \) are constants. By Lemma 4.1, \( d_i \) is also bounded by a constant.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_\phi )</td>
<td>the size of ( g )</td>
</tr>
<tr>
<td>( n_h )</td>
<td>the size of ( h )</td>
</tr>
<tr>
<td>( n_t )</td>
<td>the size of ( t )</td>
</tr>
<tr>
<td>( n_d )</td>
<td>the depth of ( t )</td>
</tr>
<tr>
<td>( n_G )</td>
<td>the max size of a specification graph</td>
</tr>
<tr>
<td>( t_G )</td>
<td>the time to compute skeleton labels</td>
</tr>
</tbody>
</table>

**Theorem 3. (Quality Analysis)** Let \( G \) be a linear recursive workflow grammar. For any graph \( g \in L(G) \), our dynamic labeling scheme \((\phi_0, \pi_0)\) guarantees

1. logarithmic label length: for any vertex \( v \) of \( g \), \( |\phi_0(v)| = O(\log n_\phi) \) bits
2. linear total construction time: computing \( \phi_0(v) \) for each vertex \( v \) of \( g \) takes a total of \( O(n_\phi) \) time.
3. constant query time: for any two vertices \( v \) and \( v' \) of \( g \), computing \( \pi_0(\phi_0(v), \phi_0(v')) \) takes \( O(1) \) time.
Proof. First, we prove the logarithmic label length. Let $\phi_g(v)[i]$ be the $i$th entry of $\phi_g(v)$. By Algorithm 1, 

$$|\phi_g(v)[i]| \leq \log \theta_i + 2 + 2 \log n_G + 1 + 1 \text{ bits}$$

Recall that we use only a pointer to each skeleton label, rather than the label itself. So it takes only $\log n_G$ bits. By Algorithm 3, $\phi_g(v)$ has at most $d_i$ entries, and $\theta_i \leq n_i \leq n_g$. 

$$|\phi_g(v)| \leq d_i (\log \theta_i + \log n_G + 4) = O(\log n_G) \text{ bits}$$

Next, we prove the linear total construction time. For the derivation-based scheme, Algorithm 3 has two steps: (1) update $t$ by inserting a new subtree $t'$ that corresponds to $h$ using Algorithm 2. This step can be done in $O(n'_t) = O(n_h)$ time, where $n'_t$ is the number of nodes of $t'$ and $n'_t \leq n_h$; and (2) create $\phi_g(v)$ for each vertex $v$ of $h$. By Algorithm 1, it may involve comparing two skeleton labels, and thus takes $t_G$ time. So the total construction time is $O(n_h * t_G) = O(n_h)$. For the execution-based scheme, the only extra computation is to decide if the newly inserted vertex is a terminal by comparing its module name, which can be done in $O(1)$ time. So the total construction time remains linear. Finally, we prove the constant query time. Since $\phi_g(v)$ and $\phi_g(v')$ have at most $d_i$ entries, by Algorithm 4, finding the maximum common prefix of entries with same index (Line 1) takes $O(d_i)$ time. The rest of computation may involve comparing two skeleton labels (Line 13), which takes $t_G$ time. So the query time is $O(d_i) + t_G = O(1)$. □

6. LABELING DYNAMIC WORKFLOWS WITH NONLINEAR RECURSION

Although our dynamic labeling scheme is for linear recursive workflows, it can be adapted to label nonlinear recursive workflows. The only modification is to create a simplified explicit parse tree without special $R$ nodes by treating all vertices in a non-recursive way. A further optimization can be achieved by compressing at most one recursive vertex using a special $R$ node, while treating other recursive vertices (if they exist) in a non-recursive way. However, the depth of the modified explicit parse tree is no longer bounded by a constant, but is proportional to the depth of recursion. This dynamic scheme may therefore create linear-size reachability labels, matching the lower bound in Theorem 1.

The remaining question is whether any nonlinear recursive workflow allows compact dynamic labeling. Theorem 4 shows that the answer is “no” for the derivation-based problem. It gives a stronger result than Theorem 1, showing that there is no compact derivation-based dynamic labeling scheme for any given nonlinear recursive workflow. Combining Theorems 3 and 4, we conclude that linear recursive workflows are the largest class of workflows that allow compact derivation-based dynamic labeling schemes.

**Theorem 4.** For any nonlinear recursive workflow grammar $G$ and any derivation-based dynamic labeling scheme $(\phi, \pi)$ for $G$, there is a derivation of a graph $g \in L(G)$ with $n$ vertices such that $\phi$ assigns a reachability label of $\Omega(n)$ bits for some vertex of $g$.

**Proof.** Since $G$ is nonlinear recursive, by Definition 10, there is a production $A := h$ with at least two recursive vertices. By applying a sequence of productions, we can obtain from $A := h$, a new production $A := h'$ with two recursive vertices $u_1$ and $u_2$ both named $A$.

![Figure 10: A new production $A := h'$ constructed from $A := h^*$ with two parallel recursive vertices $u_1$ and $u_2$.](image1)

![Figure 11: A new production $A := h^*$ constructed from $A := h'$ with two series recursive vertices $u_1$ and $u_2$.](image2)

![Figure 12: A nonlinear recursive workflow grammar that allows a compact execution-based dynamic scheme.](image3)

Next, we want to find a differential vertex $w$ that reaches exactly one of $u_1$ and $u_2$. Consider two cases. (1) If $u_1$ and $u_2$ are not reachable from each other in $h'$ (see Figure 10), then we replace $u_1$ with a new copy of $h'$, and obtain a new production $A := h^*$. Let $u'_1$ be the new copy of $u_1$, and $w$ be the source of the new copy of $h'$, then $w$ reaches exactly one of $u'_1$ and $u_2$; and (2) If one of $u_1$ and $u_2$ can reach the other in $h'$ (let $u_1 \sim_h u_2$, see Figure 11), then again we replace $u_1$ with a new copy of $h'$, and obtain a new production $A := h^*$. Let $u'_1$ be the new copy of $u_1$, and $w$ be the sink of the new copy of $h'$, then $w$ reaches exactly one of $u'_1$ and $u_2$.

Recall the production $A := h_1$ in Figure 6, where the vertex named $a$ reaches exactly one of the two vertices named $A$. Using the new production $A := h^*$, we can prove the theorem using the technique of Theorem 1. □

We next turn to the execution-based problem. To apply the same proof, we need to ensure that the differential vertex $w$ precedes both recursive vertices $u_1$ and $u_2$ in the given insertion sequence, so that $\phi(w)$ can divide all reachability labels that will be later assigned to the subgraphs derived from $u_1$ and $u_2$ into two disjoint sets. E.g., the proof of Theorem 1 relies on the fact that the differential vertex (named $a$) precedes two recursive vertices (named $A$) in the given insertion sequence (see Figure 6). Unfortunately, Case (2) in the proof of Theorem 4 (see Figure 11) violates the above condition. The following example inspired by Case (2) shows that some nonlinear recursive workflow indeed allows compact execution-based dynamic labeling schemes.

![Example 15. Consider the workflow grammar $G$ shown in Figure 12. Since any graph $g \in L(G)$ is a simple path, we simply label the $i$th vertex by an index of $i$. This naive dynamic scheme creates logarithmic-size reachability labels. However, Case (1) (see Figure 10) respects the condition. We thus get a similar result to Theorem 4 for a subclass of nonlinear workflows, called the parallel recursive workflows.](image4)

![Definition 13. A workflow grammar is said to be parallel recursive, if there is a production with two recursive vertices that are not reachable from each other (in parallel).](image5)
Theorem 5. For any parallel recursive workflow grammar $G$ and any execution-based dynamic labeling scheme $(\phi, \pi)$ for $G$, there is an execution of a graph $g \in L(G)$ with $n$ vertices such that $\phi$ assigns a reachability label of $\Omega(n^d)$ bits for some vertex of $g$.

Proof. (Sketch) We follow the same proof as Theorem 4. The correctness follows from Definition 13, which ensures that the new production $A := h^*$ must fall into Case (1). □

This paper leaves open the problem of whether non-parallel recursive workflows (with only series recursive vertices) allow compact execution-based dynamic labeling schemes.

7. EXPERIMENTAL EVALUATION

We empirically evaluated the proposed dynamic labeling scheme in terms of label length, construction time, query time and preprocessing overhead. We performed three sets of experiments: The first uses a collected, real-life scientific workflow (Section 7.2). The second measures a variety of synthetic workflows with different characteristics (Section 7.3). The last compares our dynamic scheme against the state-of-the-art static scheme [6] (Section 7.4).

7.1 Experimental Setup

All labeling schemes are implemented in Java 6. Our experiments were performed on a local PC with Intel Pentium 2.80GHz CPU and 2GB memory running Windows XP.

Real-Life and Synthetic Datasets. The real-life workflow, called BioAID, was taken from the myExperiment workflow repository [21]. To focus on the specific factors that affect the labeling performance, we also created a family of synthetic workflows using the simple topology shown in Figure 13. Due to the lack of realistic workflow runs, we simulate the execution by repeating loops, forks and recursion a random number of times. For each specification, we vary the size of runs from $1K$ to $32K$ by a factor of 2, and randomly select one derivation and one execution for each run as dynamic inputs. All data are stored in XML files.

Labeling Methodology. We compare two schemes for labeling workflow runs: (1) the one presented in this paper, which is denoted by DRL, for (D)ynamic scheme for (R)ecursive workflows; and (2) the state-of-the-art static scheme [6], which is also skeleton-based, and is denoted by SKL for (S)keleton-based scheme. To obtain skeleton labels for both schemes, we apply two simple schemes for labeling the specification: (1) TCL denotes the one given in Section 3.2. It precomputes the (T)ransitive (C)losure for all vertices, and can be used to label either a static graph or an execution-based dynamic graph; and (2) BFS does not perform any labeling, but answers a reachability query by a (B)readth (F)irst (S)earch over the graph. Since a skeleton-based scheme for labeling runs is parameterized by the scheme for labeling the specification, we denote by DRL(TCL) and DRL(BFS) (resp. SKL(TCL) and SKL(BFS)) the corresponding combinations of the two.

Evaluation Methodology. The result for label length and construction time is an average over $10^5$ sample runs, and the one for query time is an average over $10^5$ sample queries.

7.2 Labeling Real-Life Workflows

In the first set of experiments, we evaluate DRL using BioAID. It consists of 11 sub-workflows with an average size of 10.5 and a nesting depth $^5$ of 2. There are 2 loop modules, 4 fork modules and one linear recursion of length 2. Note that the derivation-based and execution-based dynamic schemes differ only in the construction time, and the scheme used to label the specification affects only the query time and the preprocessing overhead.

Figure 14 reports the maximum and average label length. As expected, both increase logarithmically with the size of the run (note that the x-axis is log scale). The average length is always shorter than the maximum length by a small constant (about 6 bits). More interestingly, both lines are almost parallel to the asymptotic line $f(n) = \log(n) + 13$. Hence, they are bounded by $c \log(n) + O(1)$, where $c$ is a small constant factor close to 1.

Figure 15 reports the total construction time for both derivation-based and execution-based schemes. Observe that they increase linearly with the size of the run. On average, we label a new vertex on the fly in less than 5 $\mu$s, which is comparable to the time of updating the graph itself (about 6 $\mu$s). Moreover, the derivation-based scheme is faster than the execution-based scheme. This is because the latter needs to find the context and origin of the newly inserted vertex.

Figure 16 reports the query time for DRL(TCL) and DRL(BFS). Recall that TCL allows constant query time, but uses linear-size labels; in contrast, BFS does not use any labels, but has linear query time. However, since the specification graphs are small and fixed, DRL answers reachability queries in almost constant time, when combined with either (Figure 16). But DRL(TCL) is slightly faster than DRL(BFS) by about 2 $\mu$s. We also measured the preprocessing overhead for DRL(TCL), and found that the overhead is negligible: the skeleton labels take totally 160 bits and are built in less than 0.05 ms.

Conclusions: The experimental results confirm the theoretical quality analysis of DRL in Theorem 3. Moreover, DRL is scalable for large dynamic runs, even when combined with simple skeleton schemes like TCL and BFS. Due to the small size of the specification graphs, the benefit of using more sophisticated schemes to label the specification is limited.

7.3 Labeling Synthetic Workflows

In the second set of experiments, we evaluate DRL for a variety of synthetic workflows created from the specification in Figure 13. It consists of a chain of nested sub-workflows with one loop module $L$, one fork module $F$ and one recursive module $R$. Note that the recursive sub-workflow $h^n_L$ may in general contain several $R$ modules. All sub-workflows are random two-terminal graphs of some fixed size. The parameters are: (a) the size of sub-workflows; (b) the nesting depth of sub-workflows; and (c) if the workflow is linear recursive (i.e., if $h^n_L$ has more than one $R$ modules). Due to space constraints, we report only the main factor, label length.

First, we generate a set of linear recursive workflows by varying the size of sub-workflows from 10 to 160 by a factor of 2, and fixing the nesting depth of sub-workflows to be 5. Figure 17 reports the maximum label length for dynamic runs with $5K$ vertices. As we can see, the maximum label length increases almost logarithmically with the size of sub-workflows. To explain the result, recall that a tighter upper
The bound of label length, given in the proof of Theorem 3, is
\[
|\phi_\ell(v)| \leq d_0 + (\log \theta + \log n_c + 4) \tag{3}
\]
where all parameters for quality analysis are defined in Table 1. In this experiment, \(d_0\) is fixed, and \(n_c\) increases by a factor of 2. We now estimate \(\theta\). Since \(n_c \cdot n_t\) is roughly the size of the run (a fixed constant of 5\(K\)), \(n_t\) decreases by a factor of 2. Note that \(t\) is a balanced tree with fixed depth. In general, \(\theta_t\) decreases much more slowly than \(n_t\). It follows that the increase of \(\log n_c \cdot \log t\) dominates the decrease of \(\log \theta_t\) in (3). This confirms the result in Figure 17.

Next, we generate a set of linear recursive workflows by varying the nesting depth of sub-workflows from 5 to 25 by a constant of 5, and fixing the size of sub-workflows to be 20. Figure 18 reports the maximum label length for dynamic runs with 5\(K\) vertices. Observe that the maximum label length increases linearly with the nesting depth of sub-workflows. This is again confirmed by (3), where \(n_c\) and \(\theta_t\) are fixed, and \(d_0\) is proportional to the nesting depth.

Finally, we generate a nonlinear recursive workflow with two \(R\) modules in \(h'_d\) (see Figure 13) and a linear recursive one with only one \(R\) module in \(h'_d\). For both workflows, the size of sub-workflows is 20, and the nesting depth is 5. Figure 19 reports the maximum label length. Not surprisingly, the nonlinear recursive workflow produces longer labels than the linear recursive one. Although DRL creates linear-size labels for nonlinear recursive workflows in the worst case, Figure 19 shows that it performs reasonably well in practice: the maximum label length for a run with 32\(K\) vertices is less than 120 bits. Note that if we use TCL to label the run dynamically, it gives a label of exactly 32\(K\) – 1 bits.

Conclusions: The main factor that affects the performance of DRL is the nesting depth of sub-workflows. However, we observe from our experience that most real-life workflows are linear recursive, and have a nesting depth of less than 5. DRL is also effective to label nonlinear recursive workflows.

7.4 DRL (Dynamic) vs SKL (Static)

In the last set of experiments, we compare DRL and SKL. The limitations of SKL are: (1) SKL is a static scheme, which takes the entire run graph as input; (2) SKL supports only non-recursive workflows (with loops and forks); and (3) SKL entails skeleton labels over a global specification graph, in which all composite modules are replaced with corresponding sub-workflows. We show only results for the real-life workflow. To achieve a fair comparison, we remove the recursion. The results for synthetic workflows are similar.

Figure 20 reports the maximum label length. Observe that DRL creates shorter labels than SKL when the run size is larger than 1.5\(K\). This is because DRL uses a prefix-based scheme [18] to label the explicit parse tree, while SKL uses an interval-based scheme [22]. The former performs better on
balanced trees with relatively high degrees and low depth. This is exactly the case when the run becomes large. More precisely, the upper bound of the label length for SKL is
\[ |\phi_g(v)| \leq 3 \cdot \log n_c + O(\log n_c) \] (4)
where \( n_c = O(n_b) \) and \( n_c = O(1) \). So the logarithmic label length for SKL has a factor of 3. Recall from Figure 14 that the factor for DRL is close to 1. Hence, for large runs, DRL creates shorter labels than SKL by a factor of almost 3. This is confirmed by the slopes of the two lines in Figure 20.

Figure 21 reports the total construction time. Since SKL builds simpler (but larger) labels than DRL, execution-based and derivation-based SKL builds simpler (but larger) labels than DRL by a factor of 2 and 4 respectively. However, unlike DRL, SKL cannot start labeling until the entire run is completed.

Figure 22 reports the query time for all four combinations. BFS performs a linear-time graph search over the specification. Consequently, when combined with BFS, the cost of comparing skeleton labels is the dominant factor. Note that SKL searches over a global specification graph with 106 vertices, while DRL searches over an individual sub-workflow with only 10.5 vertices on average. Hence, SKL(BFS) is slower than DRL(BFS) by one order of magnitude. In contrast, when combined with TCL, the cost of comparing skeleton labels is negligible. Given that SKL enables simple decoding which compares only three indexes and one skeleton label, SKL(TCL) is slightly faster than DRL(TCL). However, such efficiency is traded by high preprocessing overhead reported in Table 2.

Conclusions: DRL creates shorter labels than SKL, and is more robust to the scheme for labeling the specification.

8. CONCLUSIONS

This paper studies derivation-based and execution-based dynamic reachability labeling problems for recursive workflows with loops and forks. We provide tight lower and upper bounds of \( \Theta(n) \) bits on the maximum label length, and present a compact dynamic labeling scheme for linear recursive workflows which uses labels of \( \log(n) \) bits. The evaluation, performed over both real and synthetic workflows, shows that our dynamic scheme creates shorter labels than the start-of-the-art static scheme [6] by a factor of almost 3.

This paper also shows an interesting characterization: A workflow allows a compact derivation-based dynamic scheme if and only if it is linear recursive. However, finding an execution-based characterization is still an open problem.

9. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments. This work was supported in part by the US National Science Foundation grants IIS-0803524 and IIS-0629846, by the Israel Science Foundation, by the US-Israel Binational Science Foundation, and by the EU grant MANCOOSI.

10. REFERENCES


