Distributed query evaluation with performance guarantees

Citation for published version:

Digital Object Identifier (DOI):
10.1145/1247480.1247537

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, June 12-14, 2007

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract

Partial evaluation has recently proven an effective technique for evaluating Boolean XPath queries over a fragmented tree that is distributed over a number of sites. What left open is whether or not the technique is applicable to generic data-selecting XPath queries. In contrast to Boolean queries that return a single truth value, a generic XPath query returns a set of elements, and its evaluation introduces difficulties to avoiding excessive data shipping. This paper settles this question in positive by providing evaluation algorithms and optimizations for generic XPath queries in the same distributed and fragmented setting. These algorithms explore parallelism and retain the performance guarantees of their counterpart for Boolean queries, regardless of how the tree is fragmented and distributed. First, each site is visited at most three times, and down to at most twice when optimizations are in place. Second, the network traffic is determined by the final answer of the query, rather than the size of the tree, without incurring unnecessary data shipping. Third, the total computation is comparable to that of centralized algorithms on the tree stored in a single site. We show both analytically and experimentally that our algorithms and optimizations are scalable and efficient on large trees and complex XPath queries.

Categories and Subject Descriptors
H.2.4 [Database Management]: Systems—Query Processing

General Terms
Algorithm, Performance

Keywords
Distributed XML documents, XPath Queries, Parallel Query Processing

1. Introduction

Partial evaluation has recently found an effective application in the evaluation of Boolean XPath queries over an XML tree that is fragmented, both horizontally and vertically, and is distributed over a number of sites [5]. The rough idea of partial evaluation (see [18] for a survey) is the following. Consider a function $f(x_1, x_2)$. Assumes that we are given part of its input, say only argument $x_1$. Then, partial evaluation specializes function $f$ with respect to the known argument $x_1$, without waiting for the other argument $x_2$. That is, it performs the part of $f$'s computation that depends only on $x_1$, and generates a partial answer, i.e., a residual function $f'$ that depends on the as yet unavailable argument $x_2$. To illustrate its application in Boolean XPath evaluation, let us consider an example.

Consider the XML tree $T$ shown in Fig. 1, which represents the clientele of an investment company. For each client, the company stores her name and the country where she resides. Furthermore, it stores the broker(s) whom the client is using and the market(s) in which the client trades, through the broker(s). For each such market, the company stores the code, buying price buy, and quantity $q$ of stock(s) that the client owns. In practice such trees are often decomposed into a number of sub-trees, or fragments, and are distributed over the Internet for geographical or administrative reasons [3], a setting commonly found in e-commerce, Web services, or while managing large-scale network directories [16]. In Fig. 1, we use dashed lines to show one possible fragmentation. The fragment marked as $F_0$, which includes the root of the tree, might be stored in the investment company's local US server (site $S_0$). However, for tax reasons, trade data for Canadian customers might have to be stored in a Canada-based server. Therefore, the fragment of the tree marked as $F_3$ is stored in a remote Canadian site $S_3$. Similarly, the NASDAQ market might require that all its own trade data are only remotely accessed and only through recognized brokers for security concerns. Therefore fragments $F_2$ and $F_3$ again need to be stored in site $S_3$ outside our investment company. The fragmentation and distribution of the tree $T$ are depicted in Fig. 2. In spite of the reasons that lead to fragmentation, conceptually this is still a single XML tree over which we would like to pose queries.

Now consider $Q = //stock/code/text() = "GOOG"]$, a Boolean XPath query that is posed at site $S_0$. When being evaluated at the root of the tree $T$, the query returns a
single truth value, true if and only if there is a client trading GOOG stock. A naive way to evaluate the query is by first shipping fragments \( F_1, F_2, F_3, F_4 \) to site \( S_0 \), assembling them with \( F_0 \) into a tree, and then evaluating \( Q \) on the tree. This incurs excessive overhead of data shipping, and worse still, may not be doable since data at some sites is not allowed to be shipped to \( S_0 \) for privacy or security reasons. Another approach is to employ a sophisticated algorithm for evaluating XPath queries in a centralized system [20]. This would require a single depth-first traversal of \( T \), visiting each node exactly once, something we cannot expect to improve upon. However, this seemingly “optimal” approach would visit fragments \( F_0, F_1, F_2, F_3, F_4, F_0, F_0, F_3, F_0 \) in that order, visiting \( F_0 \) four times and \( F_1 \) twice.

An algorithm, called ParBoX, was developed in [5] based on partial evaluation, which evaluates Boolean XML queries over a fragmented tree that is distributed over a number of different sites. The algorithm partially evaluates the whole query \( Q \), in parallel, over each fragment of the tree. Since each fragment contains only part of the tree, this partial evaluation of \( Q \) over each fragment results in a partial answer to the query, which is a Boolean expression with variables. The partial answers are all collected to a single coordinator site and are composed resulting in the final answer to \( Q \).

The ParBoX algorithm has a number of desirable properties, namely: (a) Each site is visited only once, irrespectively of the number of fragments stored there. (b) The communication cost is bounded by the size of the query and the number of fragments, and is independent of the size of the XML document. (c) The total amount of computation performed at all sites holding a fragment is comparable to the computation of the optimal centralized algorithm over the whole tree. (d) The algorithm does not impose any condition on how the XML documents are fragmented, what the sizes of these fragments are, or how they are assigned to sites.

A question left open in [5] is how and under which circumstances the partial evaluation technique can be used to evaluate generic data selecting XPath queries. The need for the extension is evident since most XPath queries in practice are data-selecting queries, which find a set of XML elements rather than return a single truth value. This highlights the demand for an efficient algorithm to evaluate generic XML queries in distributed settings. Of course, such an extension should not come at a high price. It would be desirable if, while supporting generic XPath queries, we could retain properties comparable to those of the ParBoX algorithm.

It is, however, nontrivial to extend the technique to data-selecting queries. Consider a data selecting XPath query \( Q' = \\text{/broker/*/stock/code/text} = \text{"GOOG"}/\text{name} \). A direct extension of the technique of [5] would send \( Q' \) to every site and partially evaluate \( Q' \). However, what should be returned by each site as a partial answer, namely, residual function? Note that the final answer of \( Q' \) is a set of elements, and thus using the technique of [5], each site may send a set of elements that are probably in the answer of \( Q' \), and let the coordinator do the checking. This leads to excessive data shipping and in the worst case may end up shipping the entire tree to the coordinator. It is challenging to identify precisely what elements are in the final answer at each site before they are shipped to the coordinator. Furthermore, for data-selecting XPath queries even centralized evaluation algorithms require two traversals of the tree [20], instead of a single pass as for Boolean XPath queries.

To this end we develop algorithms and optimizations for evaluating generic data selecting XPath queries in the same fragmented and distributed setting of [5]. Based on partial evaluation, our algorithms retain the properties of [5]. Our contributions can be summarized as follows:

- We introduce the first partial evaluation algorithm for generic XPath queries in the distributed setting. We support a fragment of XPath with the downward axes, which covers most XPath queries used in practice [15] and includes the Boolean XPath of [5] as a special case. Our first evaluation algorithm guarantees that each site is visited at most three times and our optimized algorithm at most twice, irrespectively of the number of fragments stored there (recall that sophisticated centralized algorithms would require two passes of XML trees to evaluate data-selecting queries). Moreover, the algorithm has the same worst-case computation complexity as ParBoX and optimal communication cost, in spite of its support of more complex data-selecting queries. In particular, each site ships to the coordinator only elements that are certainly in the answer of a query, such that the union of these partial answers is the answer of the query.
- Our second contribution is an optimization technique on the performance of our basic evaluation algorithm which reduces the communication and computation costs.
- Our third contribution is a detailed experimental study, using our implemented partial-evaluation XPath query engine. Our experiments demonstrate the full potential of partially evaluating XPath queries on distributed stores.

We believe that our proposed techniques are promising for efficiently evaluating generic XPath queries in distributed systems with performance guarantees. We expect that the evaluation algorithms also shed light on evaluating XPath queries on large-scale datasets in centralized systems. Indeed, when the whole tree does not fit in main memory, through fragmentation we are able to load each time from secondary storage a different fragment of the tree into main memory. Our partial evaluation techniques help reduce at least the cost of swapping the fragments.

**Organization.** Section 2 discusses XML tree fragmentation and reviews the class of XPath queries considered in this paper. Section 3 presents a basic evaluation algorithm for XPath queries based on partial evaluation, referred to as PaX3, which directly employs ParBoX to evaluate XPath qualifiers and requires at most three visits at each site. We start with this basic algorithm for its conceptual clarity and to simplify the discussion. In Section 4 we then present an improved version of PaX3, which visits each site at most twice. An optimization technique is given in Section 5. An experimental study is provided in Section 6, followed by related work in Section 7 and conclusions in Section 8.

2. Preliminaries

We next discuss the fragmentation of XML documents and present the class of XPath queries studied in this paper.
We consider settings in which an XML tree $T$ is decomposed into a set $\mathcal{F}_T$ of disjoint trees, or fragments. Each fragment $F_i \in \mathcal{F}_T$ can be stored in a different site. We do not impose any constraints on the fragmentation: we allow for an arbitrary “nesting” of fragments. Fragments can appear at any level of the tree, and different fragments may have different sizes (in terms of number of nodes). Furthermore, we do not impose any constraints on how the fragments are distributed: this is determined by the system. Hence our fragmentation setting is the most generic possible. As an example, the tree in Fig. 1 consists of five fragments, $\mathcal{F}_T = \{F_0, F_1, F_2, F_3, F_4\}$, each fragment represented by a dotted polygon. Observe that a fragmentation of a tree $T$ induces another tree, called a fragment tree, which represents the relationship between the different fragments of $T$. We are going to use $\mathcal{F}_T$ to denote both the set of fragments of a tree $T$ and the corresponding induced fragment tree representation. It will be clear from the context which of the two notions we refer to. The fragment tree $\mathcal{F}_T$ for tree $T$ of Fig. 1 is shown to the right of Fig. 2. We call root fragment, the fragment at the root of the fragment tree which also contains the root of tree $T$. In Fig. 2, this is fragment $F_0$. Furthermore, given two fragments $F_j$ and $F_k$, we say that $F_k$ is a sub-fragment of $F_j$ if $F_k$ is a child of $F_j$ in the fragment tree. If $F_k$ is a sub-fragment of $F_j$ then there exists a node $v \in F_j$ such that the root node $w$ of $F_k$ is a child of $v$ in the original tree $T$. In Fig. 1, fragment $F_2$ is a sub-fragment of $F_1$ which, in turn, is a sub-fragment of fragment $F_0$. In the original tree $T$, node broker of fragment $F_1$ is the parent of root node market of fragment $F_2$.

Each fragment is possibly stored in a different site, as shown to the right of Fig 2. For example, fragment $F_3$ is stored in site $S_3$ while both fragments $F_2$ and $F_4$ are stored in site $S_2$. Given the distribution of fragments, we need to maintain the relationship between a fragment and its subfragments so as to preserve the structure of the original tree $T$. To this end, given a fragment $F_j$ and its sub-fragment $F_k$, we add a virtual node in $F_j$, which we label as $F_k$, in place of the missing tree fragment $F_k$. Under normal circumstances, while traversing fragment $F_j$, we know that if we reach the virtual node $F_k$, we need to “pass control” to the site holding fragment $F_k$ in order to continue the traversal of the tree. For example, in Fig. 3(a) fragment $F_0$ has virtual nodes representing fragments $F_1, F_2$ and $F_4$. While traversing fragment $F_0$ stored in site $S_0$, when we reach virtual node $F_4$, we know that the control for the traversal of the tree must pass to site $S_2$ holding fragment $F_4$, shown in Fig. 3(b). We refer to a fragment that has no sub-fragments as a leaf fragment. In Fig. 3(b), fragment $F_4$ is a leaf fragment and therefore it has no virtual nodes.

### 2.2 XPath

We consider a class of XPath queries, denoted by $\mathcal{X}$, that is defined as follows:

$$Q := \varepsilon \mid A \mid * \mid Q//Q \mid Q/Q \mid Q[q],$$

$$q := Q \mid q/text() = \str \mid q/val() \text{ op } num \mid \neg q \mid q \wedge q \mid q \vee q$$

where $Q$ is a path expression defined in terms of the empty path $\varepsilon$ (self), label $A$ (tag), wildcard $*$, the descendant-or-self-axis ‘//’, child ‘/’, and qualifier $q$. In the qualifier $q$, $\str$ is a string constant, $\text{op}$ stands for one of the arithmetic comparison operators $=, \neq, <, \leq, >, \geq$, $\text{num}$ is a number, and $\neg, \wedge, \vee$ are the Boolean negation, conjunction and disjunction operators, respectively.

For example, to get the name of a broker through which GOOG stocks are purchased, but no YHOO stocks, we write a query $Q_1$: $(//broker//stock/code/text() = \text{“GOOG”} \wedge \neg ((//stock/code/text() = \text{“YHOO”})/name)$.

At a context node $v$ in an XML tree $T$, the evaluation of a query $Q$ at $v$, denoted by $\text{val}(Q, v)$, yields the set of nodes of $T$ reachable via $Q$ from $v$. On a centralized XML tree $T$, i.e., when $T$ is not decomposed and distributed, $\text{val}(Q, r)$ can be computed in $O(|T|/|Q|)$ time [11], where $r$ is the root of $T$.

The class $\mathcal{X}$ subsumes twig queries [4] and Boolean XPath studied in [5]. Note that although our query language supports only the self, child and descendant XPath axes, this is usually sufficient since the majority of XPath queries use the downward axes [15].

Similar to [5], we normalize each query $Q$ in $\mathcal{X}$ and convert it into a normal form (although our normal form is slightly different here). Specifically, we convert each query $Q$ in $\mathcal{X}$ to a normal form $\beta_1/\ldots/\beta_n$, where $\beta_i$ is one of $A, *$, $//$ or $\varepsilon[q]$. Function $\text{normalize}(Q)$ inductively normalizes, in linear-time, a query $Q$ as follows:

1. $\text{normalize}(\varepsilon) = \varepsilon$; similarly for ‘$\varepsilon’’, ‘$//$’ and $A$;
2. $\text{normalize}(Q_1/Q_2) = \text{normalize}(Q_1)/\text{normalize}(Q_2)$;
3. $\text{normalize}(Q[q]/\str) = \text{normalize}(Q)/\text{normalize}(\str)$;
4. $\text{normalize}(Q[q/\text{val()} = \text{\text{num}}]) = \text{normalize}(Q)/\text{normalize}(\text{num})$;
5. $\text{normalize}(Q[q_1 \wedge q_2]) = \text{normalize}(q_1) \wedge \text{normalize}(q_2)$;
6. $\text{normalize}(Q[q_1/\ldots/q_n]) = \varepsilon[\text{normalize}(q_1) \wedge \ldots \wedge \text{normalize}(q_n)]$;

where the last rule is to combine a sequence of $\varepsilon$’s into one. In the sequel, we consider $\mathcal{X}$ queries in the normal form.

Striking out all the qualifiers in $\beta_1/\ldots/\beta_n$, we get a “path” $\eta_1/\ldots/\eta_n$, where $\eta_i$ is either $A$, $\varepsilon$, $*$, or $//$.

For reasons that will become clear in the next section, we decouple the evaluation of qualifiers for each query $Q$, from the evaluation of its selection path. We use a vector-based representation of our queries. More specifically, we use a vector $\text{SVect}(Q)$ to store the prefixes of the selection path $\eta_1/\ldots/\eta_n$, such that $\text{SVect}(Q)[i]$ indicates the query $\eta_1/\ldots/\eta_i$. Obviously, vector $\text{SVect}(Q)$ is linear in the size of $Q$. We use another Boolean vector $\text{QVect}(Q)$ to store the list of all sub-queries of the qualifiers of $Q$. We sort $\text{QVect}(Q)$ in a topological order such that for any sub-queries $q_1, q_2$, if $q_1$ is a sub-query of $q_2$ then $q_1$ precedes $q_2$ in $\text{QVect}(Q)$.

Again, vector $\text{QVect}(Q)$ is linear in the size of $Q$.

**Example 2.1:** Consider query $Q = \text{client}/\text{country}/text() = \text{“US”}///\text{broker}/\text{market}/\text{name}/text() = \text{“NASDAQ”}/\text{name}$ which returns the name of brokers of US clients that trade in the NASDAQ market. Then:
We decouple the selection path /client/broker/ of the query from the qualifiers *[/country/[@text()="US"]] and *[/market/[@text()="NASDAQ"]]. Then, vectors $SVec(Q)$ and $QVect(Q)$ are as follows:

\[
SVec(Q) = [q_1, q_2, q_3] \quad \text{where} \quad q_1 = \text{client}, \quad q_2 = q_1/\text{broker}, \quad q_3 = q_2/\text{name}
\]

\[
QVec(Q) = [q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9] \quad \text{where} \quad q_1 = \text{country}, q_2 = [\text{[text()="US"]}], q_3 = q_1/\text{q}_2, q_4 = */\text{/q}_3, q_5 = \text{name}, q_6 = [\text{[text()="NASDAQ"]}], q_7 = q_5/\text{q}_6, q_8 = \text{market/q}_7, q_9 = */\text{/q}_9
\]

where the first four entries in $QVec(Q)$ are for the first qualifier, while the remaining five are for the second. □

3. Query Evaluation

Consider an $A$ query $Q$ submitted to a site $S_Q$ and the query is to be evaluated over a fragmented and distributed XML tree $T$. A naive evaluation of $Q$ requires collecting to site $S_Q$ all the fragments of tree $T$, identified by the fragment tree $T_F$. Once collected, tree $T$ is reconstructed by its fragments and a centralized algorithm is used to evaluate $Q$ over $T$. We refer to this approach as NaiveCentralized. The authors of [5] have already shown experimentally that this approach is inefficient, even for the case of simple Boolean queries. Although there are a number of efficient XPath evaluation engines for centralized XML stores, in research (e.g., the algorithm of [11]) and beyond (e.g., Saxon [24], Xalan [29]), the efficiency of these evaluators becomes apparent only after site $S_Q$ gets all the data. This incurs large overhead in network traffic, since fragments are transmitted over the network each time a query is executed. Furthermore, the NaiveCentralized algorithm assumes that site $S_Q$ has main memory big enough to fit $T$. Worse still, as remarked earlier, security or privacy reasons may prevent entire fragments to be shipped from one site to another.

To cope with this we propose the Parallel XPath (PaX3) evaluation algorithm, based on partial evaluation. The PaX3 Algorithm guarantees the following:

1. Each site is visited only at most three times, irrespectively of the number of fragments stored in it.
2. Query processing is performed in parallel, on all the participating sites.
3. The total computation on all sites is comparable to what is needed by the best-known centralized algorithm.
4. The total network traffic, in any practical setting, is determined by the size of the query and the size of the query answer rather than the XML tree.

Thus algorithm PaX3 maintains all the desirable properties of the ParBoX algorithm, while it allows evaluating generic data-selecting XPath queries, instead of just Boolean ones.

We start by presenting the PaX3 algorithm to focus on the main idea of the partial evaluation technique. In the next section we will present a refined version of PaX3, called PaX2, which guarantees that each site is visited at most twice while retaining the other properties of PaX3.

The PaX3 algorithm is initiated at site $S_Q$ where the query $Q$ is issued. Without loss of generality, we assume $S_Q$ to be the site storing the root fragment of the tree $T$. As shown in Fig. 4, the algorithm consists of three stages where each stage corresponds to a single visit of a site holding tree fragments. In turn, each visit makes a single pass of each tree fragment and therefore, overall, algorithm PaX3 makes three passes over the XML tree $T$. Specifically:

Stage 1: The objective of this stage is to (partially) evaluate the qualifiers of query $Q$. For each node of each fragment, we partially evaluate qualifiers by using our own extension of the ParBoX algorithm [5]. Intuitively, at the end of this stage for some nodes we know the actual value of each qualifier, while for other nodes the value for some qualifiers is a Boolean formula whose value is yet to be determined. The value of each qualifier, i.e., Boolean formula, will be fully known for all nodes by the beginning of the next stage.

Stage 2: The objective of this stage is to (partially) evaluate the selection part of query $Q$. Intuitively, this means that at the end of this stage, for each node of each fragment, we know one of two things: (a) whether or not the node is part of the answer of query $Q$; or (b) that the node is a candidate to be part of the answer. Again, candidacy depends on the value of a Boolean formula.

Stage 3: For this latter set of candidate nodes, we will need one additional pass during this stage to determine which candidate answer nodes are true answer nodes. At the same time, at this stage, all nodes belonging to the answer of $Q$ are transmitted to site $S_Q$.

Note that the ParBoX algorithm of [5] corresponds only to the first stage of PaX3. The tricky part, namely, finding candidate answer nodes and identifying true answer nodes, is done in Stages 2 and 3.

In what follows, we describe each stage in more detail.

3.1 Qualifier Evaluation

During Stage 1 of Algorithm PaX3, the evaluation of qualifiers is performed through our extension of the ParBoX algorithm [5], which we present here briefly since the extended ParBoX is only one part of our evaluation algorithm and it is not a main contribution of our work. Our extensions include a more expressive language in the qualifiers, which includes arithmetic comparisons, and the ability to evaluate multiple top-level qualifiers in a query while in contrast, ParBoX only needs to compute a single top-level qualifier. Obviously, if a query has no qualifiers, the whole stage can be skipped and the evaluation proceeds with Stage 2.

Consider a simple query $Q$ over $T$. For simplicity, assume that $Q$ has a single qualifier. An efficient evaluation of $QVec(Q)$ for all the nodes $v$ of $T$ requires a single bottom-up traversal of $T$. During the traversal, at a node $v$ we compute the values of all the sub-queries in $QVec(Q)$ and store them in a vector $QV_v$ associated with $v$. For this computation, we often need to consult the (already computed) values of the $QVec(Q)$ sub-queries at the children and descendants of $v$. Only two additional vectors are required to store these values, namely, vectors $QCV_v$ and $QDV_v$, respectively. Intuitively, for each sub-query $q$ in $QVec(Q)$, $QCV_v(q)$ is true if and only if there exists some child $u$ of $v$ such that $QV_u(q)$ is true, and similarly, $QDV_v(q)$ is true if and only if either $QV_u(q)$ is true or there exists some descendant $w$ of $v$ such that $QV_w(q)$ is true. Now, if $Q$ has a single qualifier, then the value of the qualifier at $v$ is determined by the value of the last entry in $QV_v$. If $Q$ has more than one qualifiers, as is the case in Example 2.1, then the truth value of each qualifier is determined by the entry in $QV_v$ corresponding to
Procedure PaX3
Input: An XPath query Q and a fragmented tree T
Output: The answer (set of nodes) ans of Q over T
/* Stage 1 */
1. for each site Si in FT do
2. exec(Si.evalQual(Qvect(Q))) in parallel;
3. for each fragment Fj stored in Si do
4. annotate FT with (QVFj, QCVFj, QDVFj);
5. evaIFT(FT);
/* Stage 2 */
6. for each fragment Fj stored in Si do
7. for each sub-fragment Fk of Fj do
8. send (QVFj, QCVFj, QDVFj) to Si;
9. for each site Si in FT do
10. exec(Si.evalSelQ, SVect(Q)) in parallel;
11. for each fragment Fj stored in Si do
12. for each sub-fragment Fk of Fj do
13. annotate FT with QVfk;
14. evaIFT(FT);
/* Stage 3 */
15. for each subfragment Fk of a fragment Fj do
16. send SVvk to site Si storing Fj;
17. for each site Si in FT do exec(Si.collectAns);
18. receive ans from each site Si.

(a) PaX3 algorithm executed at participating site SQ

Procedure evaSelQ
Input: A node SVQ of (sub-)queries
Output: Set returnSet = {SVQk} where Fk is a subfragment of Fj on Si
1. for each fragment Fj assigned to Si do
2. cans := Ø;
3. tempSet := topDown(root(Fj), SV(Q));
4. returnSet := returnSet ∪ tempSet;
5. send returnSet to site SQ;

Procedure topDown
Input: A node v and a vector SV(Q) of (sub-)queries
Output: Vector SVv* of formulas for node v
1. initStack();
2. for each query qi in SV(Q) from left to right do
3. case qi of
4. t. SV.v(qi) := term(v, t);
5. qj/t. SV.v(qj) := evalFM(stack(SV(qj)), term(v, t), λ);
6. qj/. SV.v(qj) := evalFM(stack(SV(qj)), SV(qj), v);
7. SV.v := evalFM(SV(v), assocQual(QVv), λ);
8. pushStack(v, SVv);
9. if SVv.(SV(Q)) = true then ans := ans v;
10. else if SVv.(SV(Q)) is a formula then
11. case ans := ans v.
12. if v is a virtual node then returnSet := returnSet ∪ SVv;
13. for each child w of v do
14. SVw := topDown(w, SV(Q));
15. popStack();

Procedure term
Input: A node v and a temporal symbol t in the normal of q
Output: The truth value of evaluating t on v
1. case t of
2. ε. return true;
3. * or return true;
4. label() = t. return compareString(label(), t);
5. text() = str. return compareString(text(), str);

Procedure collectAns
Input: Vector SVvk of a fragment Fk of site Si
Output: The set of nodes ans of q over Fk
1. for each pair (v, SVv.(SV(Q))) in cans do
2. if unify(SVv.(SV(Q))) = true then ans := ans v;
3. return ans.

(b) PaX3 algorithm executed at participating site

Figure 4: The PaX3 Algorithm

Recall from Example 2.1 that qa = */e[qa]. That is, at a node v the value of the first qualifier (entry qa) depends on the value of entry qa at the child nodes of v. Indeed, notice that the first qualifier (value of qa) is true in QVclient since client has a child node whose corresponding text is "US" (entry qa in QVcountry). We also consider the value of the second qualifier (value of qa) in QVclient although, in practice, this qualifier is associated with broker nodes in our query. From Example 2.1 we know that qa = */e[qa], that is, at a node v the value of the second qualifier (entry qa) depends on the value of entry qa at the child nodes of v. In the specific client node, the value of this qualifier depends on variable x4, that is, it depends on whether or not the virtual node F1 represents such a fragment that its root node is market with a name child that has a value of "NASDAQ". Note that for virtual node F1, we introduce fresh variables since we do not know the value for any of the entries in the vector. Note that there are dependencies between some of the introduced variables. For example, the value of the first qualifier in F1, that is variable x4, is equal to cx3 since x4 is true in node F1 as long as node F1 has a child node whose label is country and whose value is "US" (variable cx3). □

In the algorithm (shown in Fig. 4), the bottom-up partial evaluation of qualifiers is initiated in site SQ which makes a remote procedure call (Fig. 4(a), line 2) to all the sites holding at least one tree fragment. The actual evaluation, outlined in the previous paragraphs, is performed by Procedure evalQual (not shown due to space constraints) which returns the tuple of vectors (QVF, QCVF, QDVF) corresponding to the root node of fragment F1. Each site returns its triplet(s) to site SQ, one triplet per fragment. Then in site SQ, Procedure evaFT is executed (not shown due to space constraints). Remember that we introduce variables during the partial evaluation of fragments, for each virtual node of a fragment. During the computation of qualifiers, these variables are often composed and result in complex
Boolean formulas that appear as values in some vector entry [5]. The objective of Procedure \texttt{evalFT} is to use information from all the fragments and, by unifying variables, to compute the Boolean values for these vector entries. In more detail, Procedure \texttt{evalFT} considers the fragment tree which is now annotated with vector triplets. The procedure requires a single bottom-up traversal of the fragment tree to unify all variables and compute the Boolean formula truth values. Note that vectors of leaf fragments in the fragment tree contain no variables (since they do not have any virtual nodes), as shown in our example for fragments \( F_2, F_3 \) and \( F_4 \). During the bottom-up traversal of \( \mathcal{F} \), Procedure \texttt{evalFT} uses the Boolean values of the leaf fragments to unify the variables of the vectors that belong to the parent fragments in \( \mathcal{F} \). The procedure continues in this fashion until it reaches the root of \( \mathcal{F} \).

**Example 3.2:** After Procedure \texttt{evalQual} concludes in fragments \( F_1 \) and \( F_2 \), the following are some of the vectors that are computed for nodes of the two fragments.

Vectors computed in fragment \( F_1 \):

- \( QV_{\text{broker}} = <0, 0, 0, y_5, 0, 0, 0, 0, y_6> \)
- \( QV_{F_2} = <y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9> \)

Vectors computed in fragment \( F_2 \):

- \( QV_{\text{market}} = <0, 0, 0, 0, 0, 1, 0> \)
- \( QV_{\text{name}} = <0, 0, 0, 0, 0, 1, 0> \)

Vectors \( QV_{\text{broker}} \) and \( QV_{\text{market}} \) are computed for the roots of the two fragments, respectively. Vector \( QV_{F_2} \) corresponds to virtual node \( F_2 \) of fragment \( F_1 \) where a distinct variable is introduced for each unknown vector value. Vector \( QV_{\text{name}} \) is computed at the name node of fragment \( F_2 \). Note that \( q_1 \) is true in \( QV_{\text{market}} \), i.e., the \( \text{market} \) node has a \( \text{name} \) child node with value “NASDAQ”. The computation of \( q_1 \) relies, in turn, on the precomputed (by the bottom-up evaluation) value of entry \( q_1 \) from \( QV_{\text{name}} \).

As described earlier, vectors \( QV_{\text{broker}} \) and \( QV_{\text{market}} \) are sent to site \( S_Q \) and are used as input to Procedure \texttt{evalFT}. The procedure unifies the variables in the received vectors by matching virtual nodes to root fragment nodes. That is, it determines that the vector which introduced the \( y_i \) variables in fragment \( F_1 \) (resp. vector \( QV_{F_2} \) for virtual node \( F_2 \)) is essentially vector \( QV_{\text{market}} \) corresponding to the root node \( \text{market} \) of fragment \( F_2 \). By matching the two vectors, the procedure unifies variable \( y_i \) to true (entry \( q_1 \) of \( QV_{\text{market}} \)). This, in turn, implies that entry \( q_1 \) in \( QV_{\text{broker}} \) is also true and therefore the second qualifier of the initial query is true for the corresponding \( \text{broker} \) node.

**3.2 Selection Path Evaluation**

Stage 2 of Algorithm PaX3 is initiated again at site \( S_Q \) by having site \( S_Q \) notifying each site \( S_i \) holding a fragment \( F_j \) about the result of Procedure \texttt{evalFT} (Fig 4(a), line 6), i.e., the truth values of vector triplets \( (QV_{F_2}, QCV_{F_2}, QDV_{F_2}) \) for each sub-fragment \( F_2 \) of \( F_1 \) (Fig 4(a), lines 6-8). The received vector triplets will be used by site \( S_i \) to determine the values of qualifiers for all the nodes in \( F_j \). As a next step, site \( S_Q \) initiates the partial evaluation of the query selection path by making a remote procedure call (Fig 4(a), lines 9-10) to all the sites holding at least one tree fragment.

We now examine in more detail the partial evaluation of a selection path. Consider a query \( Q \) over \( T \) and, in particular, consider the selection path vector \( SVect(Q) \) of \( Q \). An efficient evaluation of \( SVect(Q) \) over a tree \( T \) requires a single top-down (depth-first) traversal of \( T \) (Fig 4(b)).

For each node \( v \) encountered during the traversal and for each sub-query \( q_i \) in \( SVect(Q) \), we decide whether or not \( v \) can be reached from the root of the entire tree by following \( q_i \). We store the results for all sub-queries of \( SVect(Q) \) in a Boolean vector \( SV_v \) associated with node \( v \) (Procedure \texttt{topDown}, lines 2-6). This computation often requires to consult the (already computed) values of the \( SVect(Q) \) sub-queries at the ancestors of \( v \). To this end, we use a stack to store the values of \( SV_v \) for every node \( u \) that is an ancestor of \( v \). At the same time, we make sure that each time the vector at the top of the stack summarizes the information for all vectors in the stack. More specifically, when \( v \) is being processed, \( SV_v \) is the top of the stack, where \( p \) is the parent of \( v \). With this we compute \( SV_v(q_i) \) as follows. If \( q_i \) is a basic term \( A, SV_v(q_i) \) is set true if the label of \( v \) is \( A \) (line 4 of Procedure \texttt{topDown} and Procedure term of Fig 4(b)). If \( q_i \) is \( q_i/t \) for some query \( q_j \) and a basic term \( t \), then \( SV_v(q_i) \) is true if both \( SV_v(q_j) \) and term \( (v, t) \) are true. Here function \texttt{evalFM} (not shown) is used to compute the Boolean formula of \( SV_v(q_i) \land \text{term}(v, t) \). If \( q_i \) is \( q_i/f \), then \( SV_v(q_i) \) is true if either \( SV_v(q_i) \) or \( SV_v(q_i) \) is true (note that \( SV_v(q_i) \) is already computed since it precedes \( q_i \) in \( SVect(Q) \)). We do this without imposing any additional overhead, in terms of space, to our vectors which are still linear in the size of the query \( Q \). At the same time, we manage to save computation time since we do not have to go through the whole stack for each node under consideration. Also note that unlike the evaluation of qualifiers which require three vectors per node, here we often maintain a single vector per node, an improvement of our algorithm compared to ParaBox [5].

At end of the computation at node \( v \), we consult the last entry in \( SV_v \), denoted by \( SV_v(\langle SVect(Q)\rangle) \). If the value of this entry is true then node \( v \) is part of the answer for query \( Q \) and is put in the set \( \text{ans} \), otherwise it is not.

There are two more things to consider during the evaluation of selection paths. First, we need to consider qualifiers. Recall that qualifiers and selection paths are evaluated independently by our algorithm. Also, note that the truth values of all qualifiers, in all the nodes, are known after the end of Stage 1. We maintain the relationship between the selection \( SVect(Q) \) and qualifier part \( QVect(Q) \) of query \( Q \) through Procedure \texttt{assocQual} (not shown). The procedure returns for each entry of \( SVect(Q) \) the qualifier entry in \( QVect(Q) \) for which the truth value we need to check. To determine the truth value the qualifier associated to \( q_i \) at a node \( v \) and entry \( q_i \) of \( SVect(Q) \) (Procedure \texttt{topDown}, line 7), we only need to instantiate the variables in \( QV_i \) with the corresponding truth values in \( QV_{F_i} \) for sub-fragments \( F_i \), which were received from site \( S_Q \) at the beginning of the second stage.

**Example 3.3:** For simplicity, ignore for a moment the fragmentation of the tree in Fig 1 and assume that we evaluate \( SVect(Q) \) from Example 2.1 over the three clients of the tree. Then, the following are the \( SVect(Q) \) vectors computed for some of the nodes:

<table>
<thead>
<tr>
<th>Vectors</th>
<th>leftmost client</th>
<th>middle client</th>
<th>rightmost client</th>
</tr>
</thead>
<tbody>
<tr>
<td>( SV_{client} )</td>
<td>(&lt;1, 0, 0&gt;)</td>
<td>(&lt;1, 0, 0&gt;)</td>
<td>(&lt;0, 0, 0&gt;)</td>
</tr>
<tr>
<td>( SV_{broker} )</td>
<td>(&lt;0, 1, 0&gt;)</td>
<td>(&lt;0, 1, 0&gt;)</td>
<td>(&lt;0, 0, 0&gt;)</td>
</tr>
<tr>
<td>( SV_{name} )</td>
<td>(&lt;0, 0, 1&gt;)</td>
<td>(&lt;0, 0, 1&gt;)</td>
<td>(&lt;0, 0, 0&gt;)</td>
</tr>
</tbody>
</table>

For the two leftmost clients all vectors are identical since for both clients the qualifiers on \( client \) and \( country \) evaluate to true (from Stage 1). As a result, the name nodes of \( broker \) are answers to the query, verified by the fact that
SV_{name}(SVect(Q)) is true. For the rightmost client, although there exists a client/broker/name path, the vector entries are all false since all qualifiers evaluate to false.

The second thing we need to consider is fragmentation. At the beginning of the top-down traversal, given the root node \( r \) of a fragment \( F_k \), we do not know the SVect(Q) vector which summarizes the ancestors of \( r \) (located in some other fragment). Similar to the evaluation of qualifiers, we address this issue by introducing Boolean variables, one for each value of the unknown SVect(Q) vector. We initialize the stack used in the traversal to include the vector with the variables (Procedure topDown, line 1). An immediate effect from the introduction of variables is that for some nodes, say node \( v \), the last entry in vector \( SV_v \) might be a Boolean formula. Since we are not sure about whether or not node \( v \) is an answer to \( Q \) we add \( v \) to the set of candidate answers of \( Q \) (Procedure topDown, lines 10-11).

The third stage of algorithm PaX3 will determine which of the candidate answers is an actual answer to \( Q \).

Similar to Stage 1, Stage 2 concludes by having each site \( S_i \) returning to site \( S_Q \) a set of SVect(Q) vectors, namely, returnSet, one vector for each sub-fragment (virtual node) \( F_k \) of a fragment \( F_j \) of site \( S_i \). Note that returnSet consists of at most \( k \) Boolean vectors, where \( k \) is the number of virtual nodes in the fragment. Neither ans nor cans is sent to \( S_Q \). At site \( S_Q \) Procedure evalFT unifies the variables in the received vectors, through a single top-down traversal of \( F_T \).

**Example 3.4:** After Procedure topDown concludes in fragment \( F_1 \), the following SVect(Q) vectors are computed for the nodes in \( F_1 \).

\[
\begin{align*}
SV_{init} &= < z_1, z_2, z_3 > \\
SV_{broker} &= < 0, 0, 0 > \\
SV_{name} &= < 0, 0, z_1 >
\end{align*}
\]

Vector \( SV_{init} \) is inserted into the stack in Procedure topDown (line 1). This is because we are not sure during the parallel processing of fragments what path precedes node broker. In this particular case, we are interested in whether the parent node of broker, which is stored in fragment \( F_0 \), is client (variable \( z_1 \)). Even if we know that the parent node of broker must be client, we are not certain whether there are any qualifiers in the parent node or whether any such qualifiers evaluate to true or false. One of the advantages of partial evaluation is that query processing proceeds, even in the presence of uncertainty, and information about qualifiers and selection paths is kept local to each fragment rather than being sent to the coordinator site \( S_Q \). This results in, as we will prove in Section 3.4, minimum network traffic while computation costs remain optimal.

The uncertainty of what precedes node broker is propagated in both SVect(Q) vectors \( SV_{broker} \) and \( SV_{name} \) through Boolean variable \( z_1 \). Note that node name is a candidate answer due to the last entry in \( SV_{name} \). After Procedure topDown concludes in all fragments, Procedure evalFT uses vector \( SV_{client}= <1, 0, 0> \) from fragment \( F_0 \) to unify vector \( SV_{init}= <1, 0, 0> \) from fragment \( F_1 \). Variable \( z_1 \) is unified to true and thus node name is an answer to \( Q \).

### 3.3 Retrieving query answers

The last stage of Algorithm PaX3 is initiated at site \( S_Q \) by having site \( S_Q \) notifying each site \( S_i \) holding a fragment \( F_k \) about the result of Procedure evalFT (Fig 4(a), line 15), i.e., the truth values of vector \( SV_{F_k} \). Note that although vector \( SV_{F_k} \) was sent to \( S_Q \) from the site \( S_i \) holding the parent fragment \( F_j \) of \( F_k \), the vector is sent to \( S_i \) in which \( F_j \) is stored, instead of \( S_i \). The received vectors are used by each site to determine which candidate answers in cans are real answers of \( Q \) (Fig. 4(b), Procedure collectAns). Referring to our last example, after site \( S_Q \) determines that variable \( z_1 \) unifies to true, it sends this information to fragment \( F_1 \). Fragment \( F_1 \) determines, in turn, that node name is an answer to query \( Q \) and thus it sends this node back to site \( S_Q \).

### 3.4 Analysis

To compare with [5], for the analysis of the PaX3 algorithm, we consider the communication cost of the algorithm as well as its total and parallel computation costs. The total computation cost is the sum of the computation performed at all the sites that participate in the evaluation. The parallel computation cost is the time needed for evaluating the query at different sites in parallel. Since a large part of the evaluation is performed in parallel, the parallel computation cost measures the perceived execution time of the algorithm and therefore it more accurately describes its performance.

**Communication cost.** As shown in [5], the communication cost for the first stage is \( O(|Q| |F_T|) \), that is, communication is independent of the initial tree \( T \) and it only depends on the size of the query \( Q \). It is easy to see that the second stage of PaX3 also has communication cost \( O(|Q| |F_T|) \). Finally, the last stage has communication cost \( O(|Q| |F_T|) + |ans| \). Therefore, the total communication cost of PaX3 is \( O(|Q| |F_T|) + |ans| \).

Note that when we execute a query \( Q \) in a distributed environment with at most \( |F_T| \) sites, it is obvious that we are willing to pay at least the cost of transmitting our query over the various sites (cost \( O(|Q| |F_T|) \)). In addition, it is obvious that one cannot avoid the cost of retrieving the actual answers to our query (cost \( O(|ans|) \)). In this sense, a communication cost \( O(|Q| |F_T|) + |ans| \) is optimal.

**Total computation cost.** At each stage, each fragment \( F_j \) is traversed only once. During the traversal, at each node \( v \) of \( F_j \) at most \( |F_T| \) operations are performed (one operation per vector entry). Therefore, at each stage, the total computation for each fragment is \( O(|Q| |F_T|) \). At the end, in each stage and overall, the total computation for all fragments is \( O(|Q| |T|) \). Note that this coincides with the cost of executing a query \( Q \) over a tree \( T \) in a central site [11]. Therefore, the distribution of computation does not incur much extra costs in terms of total computation.

**Parallel computation cost.** As more than one fragments can be assigned to a site, we use \( |F_S| \) to denote the cumulative size of the fragments in site \( S_i \). As computation is performed in parallel at all sites, the parallel computation cost at each stage is determined by the site holding the largest cumulative fragment. That is, the parallel computation cost for the first two stages and overall is \( O(|Q| \max |F_S|) \).

**Correctness.** One can verify, by induction on the structure of \( X \) queries \( Q \), that algorithm PaX3 computes the correct answer \( Q(T) \) on any XML tree \( T \) no matter how \( T \) is fragmented and distributed.

**Summary.** With the exception of the necessary \( O(|ans|) \) cost incurred by transmitting query answers and the necessary at most two visits per site, the costs of PaX3 coincide with those of ParBoX. In short, we have proposed an algorithm to partially evaluate a larger, and more useful,
fragment of queries than those considered by \textit{ParBoX} yet we provided comparable performance guarantees. Moreover, we guarantee minimum tree data transmission since the only tree data transmitted by \textit{PaX3} are the actual query answers.

4. Improved Algorithm

We next present Algorithm \textit{PaX2}, which needs two stages and at most two visits of each site, one less than \textit{PaX3}.

The main idea behind algorithm \textit{PaX2} is to \textit{combine} the first two stages of algorithm \textit{PaX3}, i.e., evaluation of qualifiers and that of selection paths, into a single stage. As shown in Fig. 5, the algorithm starts with letting querying identifiers and that of selection paths, into a single stage. As an example, \textit{evalXPath} combines computation over fragment \(F\) has already been computed in Stage 1 by \textit{PaX3}. (shown in Fig. 3(a)) results in the \(SV\) vector \(SV_{\text{client}} = < q_{z1}, 0, 0 >\). Here variable \(q_{z1}\) indicates that although the node label is indeed client, the qualifier for the node is yet to be determined. Contrast this with Example 3.3 that, at the same node, Algorithm \textit{PaX3} results in vector \(< 1, 0, 0 >\) since the value of qualifiers has already been computed in Stage 1 by \textit{PaX3}.

For a more complex example, consider the pre-order computation over fragment \(F_1\). The computation results in \(SV_{\text{broker}} = < 0, z_1 \wedge q_{z2}, 0 >\). Here variable \(z_1\) is due to the initialization of the vector stack (see Example 3.4), while variable \(q_{z2}\) is due to the qualifier for the node which is still undetermined. In terms of node \textit{name}, the computation results in \(SV_{\text{name}} = < 0, 0, z_1 \wedge q_{z2} >\). The post-order computation at a node \(v\) starts once every node in the sub-tree rooted at \(v\) is visited (always within a fragment). At that point, the qualifiers have been computed for all the nodes in the sub-tree, through a procedure that is similar in spirit with Stage 1 of \textit{PaX3}. Given the qualifier values at node \(v\), we can unify some of the variables that have been introduced during the pre-order computation.

Example 4.2: Continuing with our last example, after we traverse the sub-tree rooted at the leftmost node \textit{client} of fragment \(F_0\), the qualifiers for the sub-tree rooted at \textit{client} have been computed and are shown in Example 3.1. Given these qualifiers, and since the \textit{client} node is associated only with the first qualifier (entry \(q_0\) of \(QV_{\text{client}}\)), we can unify variable \(q_{z1}\) to true. This yields vector \(SV_{\text{client}} = < 1, 0, 0 >\), the same vector that Algorithm \textit{PaX3} computes but only after two passes.

For fragment \(F_1\), the qualifiers for the sub-tree rooted at \textit{broker} are shown in Example 3.2. Since node \textit{broker} is associated only with the second qualifier (entry \(q_0\) of \(QV_{\text{broker}}\)), we can unify variable \(q_{z2}\) to \(y_8\). Then, \(SV_{\text{broker}}\) is now \(< 0, z_1 \wedge y_8, 0 >\) while \(SV_{\text{name}}\) is now \(< 0, 0, z_1 \wedge y_8 >\). The values for both variables \(z_1\) and \(y_8\) will be determined in the next stage of \textit{PaX2}.

Stage 1 concludes by having each site \(S_i\) returning to site \(S_0\) a set of \(SV\) and \(QV\) vectors, one \(SV\) vector for each virtual node \(F_i\) of a fragment \(F_i\) of site \(S_i\), and one \(QV\) vector for each fragment \(F_i\) of site \(S_i\). Then at \(S_0\) Procedure \textit{evalFT} unifies the variables in the received vectors. Stage 2 of \textit{PaX2} is similar to Stage 3 of \textit{PaX3}. The unified vectors are sent from \(S_0\) to the appropriate sites, and the sites sent to \(S_0\) the query answers.

Example 4.3: Continuing with our last example, site \(S_1\) holding fragment \(F_1\) receives the following vectors from \(S_0\):

\[
SV_{\text{init}} = < 1, 0, 0 >,\quad QV_{\text{F2}} = < 0, 0, 0, 0, 0, 0, 1, 0 >
\]

With these vectors, site \(S_1\) unifies variable \(z_1\) to true (entry \(q_1\) in \(SV_{\text{init}}\)) and variable \(y_8\) to true (entry \(q_9\) in \(QV_{\text{F2}}\)). Then the vector for node \textit{broker} becomes \(< 0, 1, 0 >\) and the vector for node \textit{name} becomes \(< 0, 0, 1 >\). Therefore, node \textit{name} is an answer node for query \(Q\).

Analysis. While the worst-case complexity of algorithms \textit{PaX3} is the same as its \textit{PaX2} counterpart, algorithm \textit{PaX2} requires one less visit. As will be seen in Section 6, our experimental results demonstrate that \textit{PaX2} outperforms \textit{PaX3}.

5. Optimizing Query Evaluation

We now present an optimization that identifies fragments which do not contain any nodes that are in the query answer. The optimization is used by both \textit{PaX3} and \textit{PaX2} to rule out the identified fragments from any further processing.

To do this, we require that each edge \((F_i, F_k)\) of the fragment tree \(F_T\) of \(T\) is annotated with a simple XPath expression describing the path in \(T\) connecting the root of fragment \(F_i\) with the root of fragment \(F_k\). As an example, Fig. 6 shows the XPath-annotated fragment tree from our motivating example. The \((F_0, F_1)\) edge is annotated with \textit{client/broker/market} since the root of fragment \(F_0\) is reachable from the \textit{client} node through this expression. Note that the additional XPath-annotation requirement imposes negligible space overhead for the fragment tree \(F_T\).

To see how XPath-annotation can help during query evaluation, consider a query \(Q\) and a top-down evaluation of
Figure 6: XPath-annotated fragment tree

the selection path of \( Q \). Both algorithms PaX3 (in its Stage 2) and PaX2 (in its Stage 1) perform an evaluation of this type. We propose to first use a top-down evaluation of the selection path of \( Q \) over the XPath-annotated fragment tree. Intuitively, performing such an evaluation results in a set of nodes which correspond to fragments. Each returned fragment potentially contains actual tree nodes that are in the answer of \( Q \). Our objective is to evaluate PaX3 or PaX2 only on these fragments and skip fragments that we know contain no nodes relevant to the answering of \( Q \).

Example 5.1: Consider a simple query \( \text{client/name} \) over tree \( T \) which returns the names of all clients. Evaluating this query over the fragment tree of Fig. 6 finds fragments \( F_0 \) and \( F_4 \). Fragment \( F_0 \) is considered since the procedure cannot determine with certainty whether the fragment contains or not any paths satisfying the query. Fragment \( F_4 \) is considered since the procedure knows that a client subtree is included in \( F_4 \), although it is not certain whether a name node is also included. On the other hand, the procedure determines with certainty that fragments \( F_1, F_2 \) and \( F_3 \) should not be considered. Fragment \( F_1 \) is ruled out since the path \( \text{client/broker} \) between fragments \( F_0 \) and \( F_1 \) does not satisfy the query. Similarly, the path \( \text{client/broker/market} \) from \( F_0 \) to \( F_2 \) and \( F_3 \) does not satisfy the query and therefore fragments \( F_2 \) and \( F_3 \) are both ruled out.

XPath-annotations are used before the beginning of Stage 2 in PaX3 and before Stage 1 in PaX2 to identify fragments that are relevant to a query. Apart from ruling out irrelevant fragments, XPath-annotations can also be used to reduce the number of passes of our algorithms. In more detail, if the input query \( Q \) has no qualifiers then we can use XPath-annotations to skip the step of both algorithms PaX3 and PaX2. Intuitively, through the XPath-annotations we can guarantee that any candidate answers identified by Stage 2 and 1, respectively, of the algorithms are real answers to the query and can be sent back to site \( S_Q \). Recall from Section 3.2 that without XPath-annotations, for each fragment, we need to initialize the stack in Procedure topDown with variables, since we have no information about the ancestor nodes of each fragment root. XPath-annotations encapsulate precisely the information about the ancestors of a fragment root and they can be used to initialize the stack in Procedure topDown with concrete Boolean values, instead of variables. Thus every answer to query \( Q \) can be identified with certainty. As will be shown in Section 6, XPath-annotations are effective in reducing query evaluation time.

Figure 7: Sample queries

<table>
<thead>
<tr>
<th>Query</th>
<th>XPath Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>/sites/site/people/person</td>
</tr>
<tr>
<td>Q2</td>
<td>/sites/site/open_auctions/annotation</td>
</tr>
<tr>
<td>Q3</td>
<td>/sites/site/people/person/profile/age &gt; 20 ∧ /address/country=“US”/creditcard</td>
</tr>
<tr>
<td>Q4</td>
<td>/sites/people/person/profile/age &gt; 20 ∧ /address/country=“US”/creditcard</td>
</tr>
</tbody>
</table>

Figure 8: (Annotated) fragment trees used


6. Experimental Study

We next present our experimental results. For our experiments we used ten Linux machines (fragment sites), distributed over a local LAN. Each machine has a 3GHz CPU and 1GB of memory. Our datasets consist of trees whose root node is called “sites” and each child node of the root node is a whole XMark [25] “site”. We generated multiple XMark “sites” and in each experiment we assigned (fragments of) XMark “sites” to different machines. In terms of queries, Fig. 7 shows a sample of the executed queries over our fragmented tree. The choice of presented queries will become clear shortly. In all experiments, reported times are averaged over multiple runs of each experiment. For each reported time, computation time dominates over communication time, that is, the time it takes to send query answers to the query site is negligible compared to the time to compute these answers by running PaX3 or PaX2. For consistency, algorithm PaX3 is always plotted using solid lines, while algorithm PaX2 is plotted with dotted lines. During the evaluation of an algorithm, if no XPath-annotations (NA) are used then the line is plotted using a box symbol, while if XPath-annotations (XA) are used the line is plotted using the black diamond symbol.

Experiment 1: The objective of this series of experiments is twofold. First, we want to illustrate the benefits of fragmentation. Second, we want to verify the effectiveness and scalability (in number of fragments) of PaX3 and PaX2 both with and without XPath-annotations in the fragment tree. We consider a simple fragment tree like FT1, shown to the left of Fig. 8. Our conclusions carry over to more complex fragment trees (with the same number of fragments) since in both PaX3 and PaX2, irrespective of the structure of the fragment tree and the presence of XPath annotations, a site holding a fragment at any level of the tree communicates directly with site \( S_Q \). Each fragment in FT1 corresponds to an XMark “site” and is assigned to a different machine. Throughout this experiment, the cumulative size of all fragments in FT1 is constant and equal to approx. 100MB. In more detail, in the first iteration of the experiment we consider a single fragment \( F_0 \) of size 100MB, then iteration two considers two fragments \( F_0 \) and \( F_1 \) of 50MB each and, in general, in iteration \( j \) we consider \( j \) fragments each of size (100/j)MB. For this experiment, we focus on two queries, one without qualifiers (Q1) and one with qualifiers (Q4).

In Fig. 9(a), we show the evaluation times of query Q1 at each iteration of the experiment for algorithm PaX3. The top line in the graph shows the evaluation times of Q1 in the absence of XPath-annotations, while the bottom line uses XPath-annotations. In general, note that regardless of the presence of XPath-annotations, fragmentation of trees is beneficial since as fragmentation increases query evaluation time decreases, due to parallelism. The figure also val-
Figure 9: Evaluation vs. Fragmentation

At each iteration, we increase the size of each fragment, while maintaining constant the relative ratio of sizes between different fragments. The increase is such that the cumulative size of the data is augmented by 20MB, per iteration. At the last iteration, the tree is approximately 280MB. We consider four queries in this experiment, such that (a) two do not have qualifiers (Q1 and Q2), while the other two do (Q3 and Q4): (b) two are without a ‘/’ in the selection part of the query (Q1 and Q3) while the other two do have a ‘/’ (Q2 and Q4). Therefore, the queries cover all four possible combinations and are representative of a large class of common queries.

Figure 10(a) clearly shows that algorithm PaX3 scales linearly for query Q1, as the data size increases (with or without XPath annotations). The running times for PaX2 are almost identical with the two lines from PaX3, and therefore are not shown in the figure. As explained earlier, this is because both algorithms execute two passes over each fragment, due to lack of qualifiers. The figure also illustrates that evaluation times are more than halved from using XPath-annotations during query evaluation. Specifically, due to the annotations in FT2, the evaluation only considers the data in fragments F0, F1, F2 and F3. Figure 10(b) shows that the situation is similar, even in the presence of a ‘/’. Here in spite of the ‘/’ in the query, due to the fragmentation in FT2, only fragments F0, F1, F2, F3, F6 and F8 are considered during the evaluation.

Let us look now at Fig. 10(c). Again, PaX3 scales linearly and its evaluation times are almost identical regardless of whether or not XPath-annotations are in place. The reason for this is that here PaX3 must execute Phase 1 over all fragments, since Q3 has qualifiers. Our experiments show that the cost of evaluating qualifiers (Phase 1) is dominant in PaX3 and therefore any gains made from XPath-annotations are minor, for this query, compared to the total execution time. Observe that this is not the case for queries like Q1 and Q2 where no qualifiers are present. There, the gains are significant, compared to the total execution time.

The second line in Fig. 10(c) shows that PaX2 scales also linearly and is faster than PaX3, illustrating once more the benefits of combining the two passes into one. Note that the use of XPath-annotations (third line in the figure), the evaluation time of PaX2 is improved even further. In contrast to PaX3, which computes qualifiers in all the fragments, PaX2 is more sophisticated in that it uses XPath-annotations to decide on which fragments it executes the combined pass.

The last query considered, query Q4, has a ‘/’ in its selection path and given the fragmentation of FT2, we must evaluate the query (and its qualifiers) over all the fragments of FT2. Here, XPath-annotations do not help in ruling out any fragments. Therefore, as the figure shows, the only gains in evaluation time are from combining the two passes of PaX3 into one pass of PaX2.

<table>
<thead>
<tr>
<th>Fragments</th>
<th>F0, F1, F2, F3</th>
<th>F4, F5, F6, F8</th>
<th>F7</th>
<th>F9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>5MB, 12MB, 28MB</td>
<td>5MB, 28MB</td>
<td>8MB</td>
<td>8MB</td>
</tr>
</tbody>
</table>

Experiment 2: The objective of this series of experiments is to study the scalability of our algorithms in terms of query evaluation times, as we increase the size of data. Here we consider a more natural fragment tree for our data, shown to the right in Fig. 8. The tree contains four XMark “sites” that are fragmented in different ways. Fragments F0 (which includes the root of the whole tree) and F3 contain two whole XMark “sites”, while the other two XMark “sites” are in fragments F1 and F2 and are further fragmented as shown in the figure. Unlike the previous experiment, not all fragments have the same size. The table below shows the approximate sizes of the various fragments in the first experiment iteration. Each fragment is assigned to a different machine and the cumulative size of the data is 100MB.
considering XPath-annotations in the fragment tree while evaluating query Q1, the parallel computation cost was almost halved. Figure 11(a) shows that, in addition, the total computation cost was reduced by two-thirds. This is because the machines holding fragments that were termed irrelevant to the query did not perform any computation. Therefore, by using XPath-annotations not only the query evaluated faster, but also it required less processing power to do so. Similarly, for Q2 XPath-annotations saved two-thirds, in terms of parallel computation, and almost three-quarters, in terms of total computation.

Figures 11(c) and 11(d) illustrate that, in the absence of XPath-annotations, both algorithms are evaluated over each fragment of the tree. However, in the presence of XPath-annotations (last line in Fig. 10(c) and 11(c)), the savings in total computation are again even more significant from those in parallel computation.

**Concluding remarks:** We have shown that distributing tree fragments over various sites proves an effective strategy with significant reductions in evaluation times. Obviously, always using the (optimized) algorithm PaX2 along with XPath-annotations in the fragment tree is sufficient to consistently give the best results in terms of query evaluation time. We should temper the claim about the effectiveness of XPath-annotations with the following observation. In the presence of a ‘//’ in the selection path of a query, XPath-annotations might not help much, as shown earlier for queries like Q4. However, it is not the case that the presence of ‘//’ makes XPath-annotations useless. As shown above for query Q2, if ‘//’ appears after a prefix of the selection path of the query that matches a path in the XPath-annotations, then a considerable number of fragments might be ruled out, thus improving query evaluation times.

7. Related Work

Closest to this work is the ParBoX algorithm proposed in [5], which possesses performance guarantees on the total network traffic, computation and communication steps (once). While this work is an extension of [5], the new technical development presented here is substantial and nontrivial. As remarked in Section 1, ParBoX is restricted to Boolean XPath queries, which return a single truth value of constant size, and thus the partial answers in their setting are easy to characterize. Indeed, ParBoX is a special case of Stage 1 of the PaX3 algorithm (Section 3.1). In contrast, this work deals with data-selecting XPath queries, which return a set of XML elements with variable sizes depending on the selectivity of the queries. A direct application of the partial evaluation technique or a direct extension of the ParBoX algorithm will lead to shipping a superset of the answer of a query from each participating site, and thus to excessive network traffic that in the worst case is as large as the entire tree rather than the answer of the query. The focus of this work is to figure out what appropriate partial answers should be used, identify precisely elements in the answer of the query, and ship only those in the final answer to the coordinator site. Despite that this works tackles a far more challenging problem, it achieves the same performance guarantee as its Boolean-query counterpart.

Close to this work are also [27, 2], both with nice performance bounds on total network traffic, computation and communication steps. While [27] studies distributed (data-selecting) query evaluation on semi-structured data, [2] provides algorithms for evaluating (aggregate and Boolean) queries on hierarchical distributed catalogs [26]. This work deals with a different problem, namely, (data-selecting) XPath evaluation on XML data. It also differs from [27, 2] in technical approaches. The algorithms of [27, 2] employ query decomposition and focus on query plan generation, which rewrite an input query into sub-queries appropriate for individual sites (using, e.g., the accessibility information of the distributed data); in contrast, this work avoids this overhead by sending the whole query to each relevant site; in addition, the algorithms in this paper characterize partial
answers as Boolean expressions rather than concrete data as found in [27, 2]. Moreover, an algorithm was given in [19] for evaluating XPath queries on tree fragments distributed in disk pages, employing a nice notion of partial path instances. It differs from this work in the following: (a) it is based on query decomposition and scheduling of the execution of sub-queries, instead of evaluating the same query at all sites and taking expressions as partial answers; (b) it considers a simpler set of XPath queries without qualifiers.

There has been a large body of work on distributed query processing (see, e.g., [21] for a nice survey). The key issue is minimizing communication cost [21]. Based on partial evaluation, this work minimizes both data movement and communication steps by shipping residual functions (Boolean formula) rather than data. Existing techniques for distributed query processing can benefit our algorithms, notably hybrid shipping, two-phase optimization [21], replication [1], parallel query evaluation [8, 14] and mutant query plans [22]. Partial evaluation can also be combined with recent techniques developed for P2P systems (e.g., [7, 13, 17, 9]) and be applied to P2P query processing.

A number of algorithms have been developed for evaluating XPath queries in centralized systems (e.g., [11, 20]). As remarked in Section 1, these algorithms may not work well in a distributed setting, and in addition, the partial evaluation technique may improve the performance of processing XML queries on large XML documents stored in secondary storage in a centralized system. In particular, since partial evaluation of XML queries conducts XML tree traversal, the most time consuming part, in parallel, it suggests potential optimizations for XML query processing in native XML stores by exploring parallelism. On the other hand, XPath optimizations (e.g., [6, 23]) are complementary to this work. Partial evaluation has been proven useful in a variety of areas including compiler generation, code optimization and dataflow evaluation (see [18]). Its relevance to query evaluation has surfaced from time to time, most notably in the Disco system [28] and in query rewriting with views and deductive databases [10, 12].

8. Conclusions

We have developed algorithms and optimizations for evaluating generic XPath queries on arbitrarily fragmented and distributed XML trees. We have shown both analytically and experimentally that our techniques are scalable and efficient for handling complex XPath queries on large datasets. We remark that it is far more challenging to partially evaluate data-selecting queries than Boolean queries, and our techniques are among the first for distributed processing of data-selecting XPath queries with performance guarantees.

The first topic for future work is naturally the application of partial evaluation to processing XML updates and more expressive XML queries in distributed systems. The second topic, as remarked earlier, is to use the technique to evaluate XML queries on large XML documents in native XML stores. A third topic is to integrate partial evaluation with other optimization techniques for distributed query processing.

Acknowledgment. Wenfei Fan is supported in part by EPSRC GR/S63285/01, GR/T27433/01, and BBSRC BB/D006473/1. The work was done when Gao Cong was at the University of Edinburgh.

9. References