Structural Characterizations of the Semantics of XPath as Navigation Tool on a Document

Marc Gyssens
Hasselt University & Transnational University of Limburg
marc.gyssens@uhasselt.be

Jan Paredaens
University of Antwerp
jan.paredaens@ua.ac.be

Dirk Van Gucht
George H.L. Fletcher
Indiana University, Bloomington
vgucht@cs.indiana.edu
gefletch@cs.indiana.edu

ABSTRACT

Given a document D in the form of an unordered labeled tree, we study the expressibility on D of various fragments of XPath, the core navigational language on XML documents. We give characterizations, in terms of the structure of D, for when a binary relation on its nodes is definable by an XPath expression in these fragments. Since each pair of nodes in such a relation represents a unique path in D, our results therefore capture the sets of paths in D definable in XPath. We refer to this perspective on the semantics of XPath as the "global view." In contrast with this global view, there is also a "local view" where one is interested in the nodes to which one can navigate starting from a particular node in the document. In this view, we characterize when a set of nodes in D can be defined as the result of applying an XPath expression to a given node of D. All these definability results, both in the global and the local view, are obtained by using a robust two-step methodology, which consists of first characterizing when two nodes cannot be distinguished by an expression in the respective fragments of XPath, and then bootstrapping these characterizations to the desired results.

Categories and Subject Descriptors

H.2.3 [**Database Management**]: Languages—*query languages*

General Terms

Languages, Theory

Keywords

XPath, expressibility, definability

1. INTRODUCTION

XPath is a simple language for navigation in XML documents which is at the heart of standard XML transformation languages and other XML technologies [4].

PODS'06, June 26–28, 2006, Chicago, Illinois, USA.

XPath can be viewed as a query language in which an expression associates to every document a binary relation on its nodes representing all navigation paths in the document defined by that expression [3, 11, 18]. From that query-level perspective, several natural semantic issues have been investigated in recent years for various fragments of XPath. These include expressibility, closure properties, and complexity of evaluation [3, 12, 18], as well as decision problems such as satisfiability, containment, and equivalence [2, 19].

Alternatively, we can view XPath as a navigational tool on a particular given document, and study expressiveness issues from this document-level perspective. (A similar duality exists in the relational database model, where Bancilhon [1] and Paredaens [21] considered and characterized expressiveness at the instance level, which, subsequently, Chandra and Harel [7] contrasted with expressiveness at the query level.)

In this setting, our goal is to characterize, for various natural fragments of XPath, when a binary relation on the nodes of a given document (i.e., a set of navigation paths) is definable by an expression in the fragment.

To achieve this goal, we develop a robust two-step methodology. The first step consists of characterizing when two nodes in a document cannot be distinguished by an expression in the fragment under consideration. It turns out for those fragments we consider that this notion of expression equivalence of nodes is equivalent to an appropriate generalization of bisimilarity. The second step of our methodology then consists of bootstrapping this result to a characterization for when a binary relation on the nodes of a given document is definable by an expression in the fragment (in the sense of the previous paragraph).

We refer to this perspective on the semantics of XPath at the document level as the "global view." In contrast with this global view, there is also a "local view" which we consider. In this view, one is only interested in the nodes to which one can navigate starting from a particular given node in the document under consideration. From this perspective, a set of nodes of that document can be seen as the end points of a set of paths starting at the given node. For each of the XPath fragments considered, we characterize when such a set represents the set of *all* paths starting at the given node defined by some expression in the fragment. These characterizations are derived from the corresponding characterizations in the "global view," and turn out to be particularly elegant in the important special case where the starting node is the root.

In this paper, we study four XPath fragments. The most expressive among them is the *XPath-algebra* which permits the self, parent, and child operators, predicates, compositions, and the boolean operators union, intersection, and difference. (Since we work at the document level, i.e., the document is given, there is no need to include the ancestor and descendant operators as primitives.) We also consider the *core XPath-algebra*, which is the XPath-algebra without intersection and difference at the expression level. The core XPath-algebra is the adaptation to our setting of Core XPath of Gottlob et al. [11]. Of both of these algebras, we consider the fragments without the parent operator, called the *downward XPath-algebra* and *downward core XPath-algebra*, respectively.

The robustness of the characterizations provided in this paper is further strengthened by their feasibility. As discussed in Section 8, the global and local definability problems for each of the XPath fragments are decidable in polynomial time. This feasibility hints towards efficient partitioning and reduction techniques on both the set of nodes and the set of paths in a document. Such techniques may be fruitfully applied towards document compression [6], access control [9], and designing indexes for query processing [10, 14, 20, 22].

The remainder of this paper is organized as follows. In Section 2, we formally define the four XPath fragments as well as expression equivalence of nodes, and introduce some terminology. In Section 3, we propose our two-step methodology by applying it to both downward fragments of XPath, because these allow the simplest exposition. In particular, it will turn out that both fragments are equivalent, and that, in these cases, expression equivalence is the same as bisimilarity. In Section 4, we present the generalizations of bisimilarity required to deal with the XPath-algebra and the core XPath-algebra, which are studied in Sections 5 and 6, respectively. The structural characterizations of the semantics of the four XPath fragments in Sections 3, 5 and 6 pertain to the "global view" only. In Section 7, we derive the corresponding characterizations for the "local view." In Section 8, finally, we discuss some ramifications of our results as well as directions for future research.

Because of space considerations, several proofs are either omitted or only sketched. The proofs of Section 4, many of which require a case analysis, have been moved to an Appendix.

2. NOTATION AND TERMINOLOGY

In this paper, *documents* are finite *unordered* node-labeled trees. More formally, a document D is a 4-tuple $(V, \mathit{Ed}, r, \lambda)$, with V the finite set of nodes, $\mathit{Ed} \subseteq V \times V$ the set of edges, $r \in V$ the root and $\lambda : V \to \mathcal{L}$ the node-labeling function into an infinite enumerable set $\mathcal{L}$ of labels.

We next define the fragments of XPath [4] considered in this paper. As observed in the Introduction, we can prune the set of operators considerably, since we are only concerned with (1) *expressibility* on (2) a *single* document.

Definition 1. The *XPath-algebra* consists of the primitives ε, $\hat{\ell}$ ($\ell \in \mathcal{L}$), $\emptyset$, $\downarrow$, and $\uparrow$, together with the operators. E_1/E_2, $E_1[E_2]$, $E_1 \cup E_2$, $E_1 \cap E_2$, and $E_1 - E_2$.

Given a document $D = (V, \mathit{Ed}, r, \lambda)$, the *semantics*, $E(D)$, of an XPath-algebra expression E is a binary relation over V, defined as follows:

- $\varepsilon(D) = \{(n, n) \mid n \in V\}$; $\hat{\ell}(D) = \{(n, n) \mid n \in V \text{ and } \lambda(n) = \ell\}$; $\emptyset(D) = \emptyset$;
- $\downarrow(D) = \mathit{Ed}$; $\uparrow(D) = \mathit{Ed}^{-1}$;
- $E_1/E_2(D) = \pi_{1,4}\sigma_{2=3}(E_1(D) \times E_2(D))$; $E_1[E_2](D) = \pi_{1,2}\sigma_{2=3}(E_1(D) \times E_2(D))$;
- $E_1 \star E_2(D) = E_1(D) \star E_2(D)$, where "$\star$" stands for "$\cup$", "$\cap$", or "$-$".

Actually, we can show (proof omitted) that the predicate operator "$E_1[E_2]$" is superfluous in the XPath-algebra, but we leave it in because it cannot be omitted in the XPath fragments we define next:

- The *downward XPath-algebra* is the XPath-algebra without "$\uparrow$".
- The *core XPath-algebra* has the same primitives as the XPath-algebra, together with the operators E_1/E_2, $E_1[E_2]$ with E_2 a boolean combination[1] of core XPath-algebra expressions, and $E_1 \cup E_2$.
- The *downward core XPath-algebra* is the core XPath-algebra without "$\uparrow$".

Definition 1 reflects the "global" perspective of XPath as working on an entire document, rather than the "local" perspective of XPath as working on a particular node, reflected in Definition 2.

Definition 2. Let E be an XPath-algebra expression, and let $D = (V, \mathit{Ed}, r, \lambda)$ be a document. For $m \in V$, $E(D)(m) := \{n \in V \mid (m, n) \in E(D)\}$.

As the first step in our two-step methodology, we are interested in which nodes in a document we can or cannot distinguish by XPath. Therefore, we define the following equivalence relation:

Definition 3. Let $D = (V, \mathit{Ed}, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. Then m_1 and m_2 are *expression-equivalent* (denoted $m_1 \equiv_e m_2$) if, for each XPath-algebra expression E, $E(D)(m_1) = \emptyset$ if and only if $E(D)(m_2) = \emptyset$.

Similarly, we can also define *downward expression equivalence* (denoted as $m_1 \equiv_{e\downarrow} m_2$), *core expression equivalence* (denoted $m_1 \equiv_{e^-} m_2$), and *downward core expression equivalence* (denoted $m_1 \equiv_{e^-\downarrow} m_2$), each corresponding to one of the XPath-algebra fragments introduced above.

Next, we introduce the notion of *signature* of a pair of a nodes in a document.

Definition 4. Let $D = (V, \mathit{Ed}, r, \lambda)$ be a document, and let $m, n \in V$. The *signature* $\mathrm{sig}(m, n)$ is an XPath-algebra expression defined as follows:

1. If n is a descendant (ancestor) of m, then $\mathrm{sig}(m, n) :=\downarrow^k$ ($\mathrm{sig}(m, n) :=\uparrow^k$), with k the length of the path between m and n.[2]
2. Otherwise, let $\mathrm{top}(m, n)$ be the least common ancestor of m and n. Then

$$\mathrm{sig}(m, n) := \mathrm{sig}(m, \mathrm{top}(m, n))/\mathrm{sig}(\mathrm{top}(m, n), n).$$

The sequence $m = p_1, \ldots, p_k = n$ of all the intermediate nodes encountered upon computing $\mathrm{sig}(m, n)(D)(m)$ is called the *path* from m to n.

Note that, for $m_1, m_2, n_1, n_2 \in V$, $(m_2, n_2) \in \mathrm{sig}(m_1, n_1)(D)$ in general does *not* imply that $\mathrm{sig}(m_1, n_1) = \mathrm{sig}(m_2, n_2)$ unless n_1 is a descendant of m_1, or vice-versa. For example, in the document D in Figure 1, *top left*, $(m_1, m_1) \in \mathrm{sig}(m_1, m_3)$, while $\mathrm{sig}(m_1, m_1) = \varepsilon$ and $\mathrm{sig}(m_1, m_3) =\uparrow^2 / \downarrow^2$.

We therefore define the following comparison between the signatures of pairs of nodes:

[1] Obtained using union, intersection, and complementation with respect to $V \times V$.

[2] The exponent notation denotes repeated composition ("/"). If $m = n$, then $\mathrm{sig}(m, n) := \varepsilon$.

Definition 5. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2 \in V$. We say that $\text{sig}(m_1, n_1) \geq \text{sig}(m_2, n_2)$ if $(m_2, n_2) \in \text{sig}(m_1, n_1)(D)$.

We conclude this section with the following observation:

PROPOSITION 1. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2 \in V$. There exists an XPath-algebra expression Sig(m_1, n_1) such that $(m_2, n_2) \in$ Sig$(m_1, n_1)(D)$ if and only if sig$(m_1, n_1) =$ sig(m_2, n_2).*

PROOF. If n_1 is a descendant of m_1, or vice-versa, choosing $\text{Sig}(m_1, n_1) := \text{sig}(m_1, n_1)$ clearly satisfies all requirements. Otherwise, $\text{Sig}(m, n) := \text{sig}(m, n) - \text{sig}(\text{parent}(m), \text{parent}(n))$ satisfies all requirements. ☐

3. CHARACTERIZING THE SEMANTICS OF THE DOWNWARD AND THE DOWNWARD CORE XPATH-ALGEBRAS

In this section, we are concerned with the downward XPath-algebra and the downward core XPath-algebra, since their semantics have the simplest characterizations. In subsequent sections, we generalize our results to the full XPath-algebra and the core XPath-algebra.

Our first goal is to characterize both downward expression equivalence and downward core expression equivalence in terms of the structure of the document under consideration. Thereto, we define another equivalence relation on the nodes of a document, this time purely in terms of the structure of that document.

Definition 6. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. Then m_1 and m_2 are *downward* 1*-equivalent* (denoted $m_1 \equiv^1_\downarrow m_2$) if

1. $\lambda(m_1) = \lambda(m_2)$; and
2. for each child n_1 of m_1, there exists a child n_2 of m_2 such that $n_1 \equiv^1_\downarrow n_2$, and vice versa.

In the literature, downward 1-equivalence is usually referred to as *bisimilarity* [5]. For the sake of generalization in Section 4, we use a different terminology in this paper.

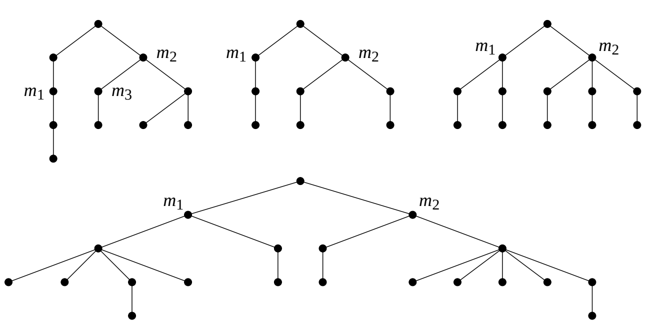

Figure 1: Example documents. All nodes are assumed to have the same label.

Example 1. Consider the document in Figure 1, *top left*. By Definition 6 the nodes m_1 and m_2 are downward 1-equivalent, whereas the nodes m_1 and m_3 are *not* downward 1-equivalent.

We generalize downward 1-equivalence to *pairs* of nodes.

Definition 7. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2 \in V$ such that n_1 is descendant of m_1 and n_2 is a descendant of m_2. Then, (m_1, n_1) and (m_2, n_2) are *downward* 1*-equivalent* (denoted $(m_1, n_1) \equiv^1_\downarrow (m_2, n_2)$) if

1. $\text{sig}(m_1, n_1) = \text{sig}(m_2, n_2)$; and
2. for each pair of nodes p_1 and p_2 with
   (a) p_1 on the path from m_1 to n_1;
   (b) p_2 on the path from m_2 to n_2; and
   (c) $\text{sig}(m_1, p_1) = \text{sig}(m_2, p_2)$[3],

   we have that $p_1 \equiv^1_\downarrow p_2$.

By repeatedly applying Definition 6, the following connection between between downward 1-equivalence of nodes and pairs of nodes can be established:

LEMMA 1. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1 \in V$ such that n_1 is a descendant of m_1 and $m_1 \equiv^1_\downarrow m_2$. Then there exists a descendant n_2 of m_2 such that $(m_1, n_1) \equiv^1_\downarrow (m_2, n_2)$.*

Using Lemma 1, the following key lemma can now be proved by structural induction.

LEMMA 2. *Let E be a downward XPath-algebra expression, let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2 \in V$ such that $(m_1, n_1) \equiv^1_\downarrow (m_2, n_2)$. If $(m_1, n_1) \in E(D)$, then $(m_2, n_2) \in E(D)$.*

Combining Lemmas 1 and 2 immediately yields

COROLLARY 1. *Let E be a downward XPath-algebra expression, let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1 \in V$ such that $m_1 \equiv^1_\downarrow m_2$ and $(m_1, n_1) \in E(D)$. Then there exists $n_2 \in V$ such that $(m_2, n_2) \in E(D)$.*

We can now present a characterization of downward (core) expression equivalence.

THEOREM 1. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. Then, $m_1 \equiv_{e\downarrow} m_2$ if and only if $m_1 \equiv_{e-\downarrow} m_2$ if and only if $m_1 \equiv^1_\downarrow m_2$.*

PROOF. Since $m_1 \equiv_{e\downarrow} m_2$ implies $m_1 \equiv_{e-\downarrow} m_2$, it remains to prove that (1) $m_1 \equiv^1_\downarrow m_2$ implies $m_1 \equiv_{e\downarrow} m_2$ and (2) $m_1 \equiv_{e-\downarrow} m_2$ implies $m_1 \equiv^1_\downarrow m_2$.

For (1), let $m_1 \equiv^1_\downarrow m_2$, and let E be a downward XPath-algebra expression such that $E(D)(m_1) \neq \emptyset$. Hence, there exists $n_1 \in V$ such that $(m_1, n_1) \in E(D)$. By Corollary 1, there exists $n_2 \in V$ such that $(m_2, n_2) \in E(D)$, whence $E(D)(m_2) \neq \emptyset$. By symmetry, the same holds vice-versa.

For (2), let $m_1 \equiv_{e-\downarrow} m_2$. By induction on the height of m_1, we show that $m_1 \equiv^1_\downarrow m_2$.

If m_1 is a leaf, then m_2 is a leaf, for, otherwise, $\downarrow(D)(m_1) = \emptyset$ and $\downarrow(D)(m_2) \neq \emptyset$, a contradiction. In addition, we also have that $\lambda(m_1) = \lambda(m_2)$, for, otherwise, $\widehat{\lambda(m_1)}(D)(m_1) \neq \emptyset$ and $\widehat{\lambda(m_1)}(D)(m_2) = \emptyset$, a contradiction. By Definition 6, $m_1 \equiv^1_\downarrow m_2$.

If m_1 is not a leaf, m_2 is not a leaf either, and $\lambda(m_1) = \lambda(m_2)$, by the same arguments as in the base case. Now, let n_1^1 be a child of m_1, and let $n_2^1, \ldots, n_2^\ell$ be all children of m_2. Suppose that,

[3] Or, equivalently, $\text{sig}(p_1, n_1) = \text{sig}(p_2, n_2)$.

for all i, $1 \leq i \leq \ell$, $n_1^1 \not\equiv_{e-\downarrow} n_2^i$. Hence, there exists a downward core XPath-algebra expression E_i such that $E_i(D)(n_1^1) \neq \emptyset$ and $E_i(D)(n_2^i) = \emptyset$.[4] Let $F := \varepsilon[\varepsilon[E_1] \cap \ldots \cap \varepsilon[E_\ell]]$. Then $\downarrow /F(D)(m_1) \neq \emptyset$ and $\downarrow /F(D)(m_2) = \emptyset$, a contradiction. Hence, there exists a child n_2^j of m_2, $1 \leq j \leq \ell$, such that $n_1^1 \equiv_{e-\downarrow} n_2^j$. By the induction hypothesis, $n_1^1 \equiv_\downarrow^1 n_2^j$. Of course, the same holds vice-versa. □

As a consequence of Theorem 1, downward (core) expression equivalence is decidable.

We next turn to the second step of our two-step methodology by bootstrapping Theorem 1 to characterize those binary relations over the nodes of a document that can be defined as the evaluation of a downward (core) XPath-algebra expression.[5] For that purpose, we need the following lemma.

LEMMA 3. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. There exists a downward core XPath-algebra expression E_{m_1} such that $E_{m_1}(D)(m_2) \neq \emptyset$ if and only if $m_1 \equiv_\downarrow^1 m_2$.*

PROOF. Let $p_2 \in V$ be a node such that $m_1 \not\equiv_\downarrow^1 p_2$. By Theorem 1, $m_1 \not\equiv_{e-\downarrow} p_2$. Hence, there exists a downward core XPath-algebra expression F_{m_1,p_2} such that $F_{m_1,p_2}(D)(m_1) \neq \emptyset$ and $F_{m_1,p_2}(D)(p_2) = \emptyset$. It is now easily seen that

$$E_{m_1} := \varepsilon\Big[\bigcap_{p_2 \in V \text{ and } m_1 \not\equiv_\downarrow^1 p_2} \varepsilon[F_{m_1 p_2}] \Big].$$

is the required downward core XPath-algebra expression. □

We now prove the main theorem of this section.

THEOREM 2. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $J \subseteq V \times V$. The following statements are equivalent:*

1. *There exists a core downward XPath-algebra expression E such that $E(D) = J$.*
2. *There exists a downward XPath-algebra expression E such that $E(D) = J$.*
3. (a) *for all $m, n \in V$, $(m, n) \in J$ implies n is a descendant of m; and*
   (b) *for all $m_1, n_1, m_2, n_2 \in V$ with n_1 a descendant of m_1, n_2 a descendant of m_2, and $(m_1, n_1) \equiv_\downarrow^1 (m_2, n_2)$, $(m_1, n_1) \in J$ implies $(m_2, n_2) \in J$.*

PROOF. Clearly (1) ⇒ (2). The implication (2) ⇒ (3) has been shown in Lemma 2. It remains to show that (3) ⇒ (1). Thereto, consider the downward core XPath-algebra expression

$$E := \bigcup_{(m_1,n_1) \in J} \ \bigcap_{\substack{p_1 \text{ on the path} \\ \text{from } m_1 \text{ to } n_1}} \text{sig}(m_1, p_1)/\varepsilon[E_{p_1}]/\text{sig}(p_1, n_1),$$

with E_{p_1} as in Lemma 3. It is now easily seen that condition (3) above implies that $E(D) = J$. □

We immediately conclude that the downward core XPath-algebra and the downward XPath-algebra are equally expressive as navigation tools on a given document.[6]

[4]Alternatively, if E_i' is an expression such that $E_i'(D)(n_1^1) = \emptyset$ and $E_i'(D)(n_2^i) \neq \emptyset$, then put $E_i := \varepsilon[\varepsilon - \varepsilon[E_i']]$.

[5]In Section 7, we consider this second step for the local view.

[6]Using an involved argument (omitted), we can actually show that both fragments are equivalent as query languages.

4. DOWNWARD *K*-EQUIVALENCE AND *K*-EQUIVALENCE

We now generalize downward 1-equivalence to downward k-equivalence, for arbitrary $k \geq 1$. The values of k that will interest us most are 1, 2, and 3.

Definition 8. Let $k \geq 1$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. Then m_1 and m_2 are *downward k-equivalent* (denoted $m_1 \equiv_\downarrow^k m_2$) if

1. $\lambda(m_1) = \lambda(m_2)$;
2. for each child n_1 of m_1, there exists a child n_2 of m_2 such that $n_1 \equiv_\downarrow^k n_2$, and vice versa; and
3. for each child n_1 of m_1 and each child n_2 of m_2 such that $n_1 \equiv_\downarrow^k n_2$, $\min(|\bar{n}_1|, k) = \min(|\bar{n}_2|, k)$, where, for $i = 1, 2$, $\bar{n}_i = \{p \mid (m_i, p) \in Ed \text{ and } p \equiv_\downarrow^k n_i\}$.[7]

Clearly, Definition 8 reduces to Definition 6 for $k = 1$. It can be shown (proof omitted) that downward k-equivalence is the coarsest equivalence relation satisfying conditions (1), (2), and (3) above.

In order to deal with the presence of the "↑" operator in both the XPath-algebra and the core XPath-algebra, we need a more restrictive kind of "k-equivalence" than downward k-equivalence.

Definition 9. Let $k \geq 1$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. Then m_1 and m_2 are *k-equivalent* (denoted $m_1 \equiv^k m_2$) if

1. $m_1 \equiv_\downarrow^k m_2$;
2. m_1 is the root if and only if m_2 is the root;
3. if m_1 and m_2 are not the root, and p_1 and p_2 are the parents of m_1 and m_2, respectively, then $p_1 \equiv^k p_2$.

In other words, m_1 and m_2 are k-equivalent if they are at the same depth in the document, and each pair of same-generation ancestors of m_1 and m_2 is downward k-equivalent. As a consequence, we see that same-generation ancestors of k-equivalent nodes are k-equivalent themselves.

Example 2. In Figure 1, *top left*, m_1 and m_2 are downward 1-equivalent, but *not* 1-equivalent. In Figure 1, *top center*, m_1 and m_2 are 1-equivalent, but *not* 2-equivalent. In Figure 1, *top right*, m_1 and m_2 are 2-equivalent, but *not* 3-equivalent. Finally, in Figure 1, *bottom*, m_1 and m_2 are 3-equivalent, but *not* 4-equivalent.

Definition 10. Let $k \geq 1$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2 \in V$. Then (m_1, n_1) and (m_2, n_2) are *k-equivalent* (denoted $(m_1, n_1) \equiv^k (m_2, n_2)$) if

1. $\text{sig}(m_1, n_1) = \text{sig}(m_2, n_2)$; and
2. for each pair of nodes p_1 and p_2 with
   (a) p_1 on the path from m_1 to n_1;
   (b) p_2 on the path from m_2 to n_2; and
   (c) $\text{sig}(m_1, p_1) = \text{sig}(m_2, p_2)$,

   we have that $p_1 \equiv^k p_2$.

Similarly, (m_1, n_1) and (m_2, n_2) are *k-related* (denoted $(m_1, n_1) \Rrightarrow^k (m_2, n_2)$) if

[7]For a set A, $|A|$ denotes the cardinality of A.

1. sig$(m_1, n_1) \geq$ sig(m_2, n_2); and
2. for each pair of nodes p_1 and p_2 with
   (a) p_1 on the path from m_1 to n_1;
   (b) p_2 either on the path from m_2 to n_2 or an ancestor of top(m_2, n_2); and
   (c) sig$(m_1, p_1) \geq$ sig(m_2, p_2),

   we have that $p_1 \equiv^k p_2$.

Notice that k-equivalence and k-relatedness coincide if n_1 is a descendant of m_1, or vice-versa. In general, downward k-relatedness is *not* symmetric.

The following technical lemmas are very practical. The second is the generalization of Lemma 1.

LEMMA 4. *Let $k \geq 1$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2 \in V$. Then $(m_1, n_1) \equiv^k (m_2, n_2)$ (respectively $(m_1, n_1) \Rrightarrow^k (m_2, n_2)$) if and only if $m_1 \equiv^k m_2$, $n_1 \equiv^k n_2$, and sig(m_1, n_1) = sig(m_2, n_2) (respectively sig$(m_1, n_1) \geq$ sig(m_2, n_2)).*

PROOF. For each pair of nodes p_1 and p_2 for which $p_1 \equiv^k p_2$ must hold according to Definition 10, p_1 is either an ancestor of m_1 or an ancestor of n_1 and p_2 a same-generation ancestor of m_2 or of n_2. As observed earlier, same-generation ancestors of k-equivalent nodes are k-equivalent. □

LEMMA 5. *Let $k \geq 1$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1 \in V$ such that n_1 is a descendant of m_1 and $m_1 \equiv^k_\downarrow m_2$. Then, there exists a descendant n_2 of m_2 such that $(m_1, n_1) \equiv^k_\downarrow (m_2, n_2)$.*

The following properties play a crucial role in proving the analogues of Lemma 2 and Corollary 1, used in characterizing the semantics of the downward (core) XPath-algebra, for characterizing the semantics of the XPath-algebra (Lemma 6 and Corollary 2) and the core XPath-algebra (Lemma 11 and Corollary 3). Their proofs are in the Appendix.

PROPOSITION 2. *Let $k \geq 1$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1 \in V$ with $m_1 \equiv^k m_2$. Then, there exists $n_2 \in V$ such that $(m_1, n_1) \Rrightarrow^k (m_2, n_2)$.*

PROPOSITION 3. *Let $k \geq 1$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2, p_1 \in V$ such that $(m_1, n_1) \Rrightarrow^k (m_2, n_2)$. Then, there exists $p_2 \in V$ such that $(m_1, p_1) \Rrightarrow^k (m_2, p_2)$ and $(p_1, n_1) \Rrightarrow^k (p_2, n_2)$.*

PROPOSITION 4. *Let $k \geq 2$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1 \in V$ with $m_1 \equiv^k m_2$. Then, there exists $n_2 \in V$ such that $(m_1, n_1) \equiv^k (m_2, n_2)$.*

PROPOSITION 5. *Let $k \geq 3$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2, p_1 \in V$ such that $(m_1, n_1) \equiv^k (m_2, n_2)$. Then, there exists $p_2 \in V$ such that $(m_1, p_1) \equiv^k (m_2, p_2)$ and $(p_1, n_1) \equiv^k (p_2, n_2)$.*

5. CHARACTERIZING THE SEMANTICS OF THE XPATH-ALGEBRA

Lemma 6, below, is the analogue of Lemma 2 for the full XPath-algebra.

LEMMA 6. *Let E be an XPath-algebra expression, let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2 \in V$ such that $(m_1, n_1) \equiv^3 (m_2, n_2)$. If $(m_1, n_1) \in E(D)$, then also $(m_2, n_2) \in E(D)$.*

PROOF. The proof goes by induction on the structure of E. The induction step for the composition E_1/E_2 relies on Proposition 5; the induction step for the predicate operator $E_1[E_2]$ relies on Proposition 4; and the induction step for the difference operator $E_1 - E_2$ relies on the symmetry of 3-equivalence on pairs of nodes. The rest of the proof is straightforward. □

Combining Proposition 4 and Lemma 6 immediately yields

COROLLARY 2. *Let E be an XPath-algebra expression, let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1 \in V$ such that $m_1 \equiv^3 m_2$ and $(m_1, n_1) \in E(D)$. Then there exists $n_2 \in V$ such that $(m_2, n_2) \in E(D)$.*

Using the same argument used for statement (1) in the proof of Theorem 1, we obtain

LEMMA 7. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. If $m_1 \equiv^3 m_2$, then $m_1 \equiv_e m_2$.*

The reverse implication, however, requires more work. We first show that expression equivalence implies downward 3-equivalence, and then bootstrap this result to show that, actually, expression equivalence implies 3-equivalence.

LEMMA 8. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. If $m_1 \equiv_e m_2$, then $m_1 \equiv^3_\downarrow m_2$.*

PROOF. Since downward 3-equivalence is the coarsest equivalence relation satisfying conditions (1), (2), and (3) of Definition 8, it suffices to prove that expression equivalence satisfies these conditions.

For conditions (1) and (2), this requires the same arguments as used for statement (2) in the proof of Theorem 1. We therefore restrict ourselves to condition (3). Thus, let $n_1^1, \ldots, n_1^k$ be all children of m_1 and $n_2^1, \ldots, n_2^\ell$ be all children of m_2, and assume that $n_1^1 \equiv_e n_2^1$. We have to show that $\min(|\tilde{n}_1^1|, 3) = \min(|\tilde{n}_2^1|, 3)$, where, for $i = 1, 2$, $\tilde{n}_i^1 = \{p \mid (m_i^1, p) \in Ed$ and $p \equiv_e n_i^1\}$. To do so, we have to show that the following situations cannot occur:

1. $|\tilde{n}_1^1| = 1$ and $|\tilde{n}_2^1| > 1$, or vice-versa; and
2. $|\tilde{n}_1^1| = 2$ and $|\tilde{n}_2^1| > 2$, or vice-versa.

By symmetry, it suffices to consider the former situation in each of these cases.

1. $|\tilde{n}_1^1| = 1$ *and* $|\tilde{n}_2^1| > 1$. Hence, $\tilde{n}_1^1 = \{n_1^1\}$ and, without loss of generality, we may assume that $\tilde{n}_2^1 \supseteq \{n_2^1, n_2^2\}$. Since, for all $i = 2, \ldots, k$, $n_1^1 \not\equiv_e n_1^i$, there exists an XPath-algebra expression E_i such that $E_i(D)(n_1^1) \neq \emptyset$ and $E_i(D)(n_1^i) = \emptyset$. By definition of expression equivalence, we also have, for $j = 1, 2$, that $E_i(D)(n_2^j) \neq \emptyset$.

   Let $F := \varepsilon[E_2] \cap \ldots \cap \varepsilon[E_k]$, and let $G := F/\uparrow/\downarrow/F$. One can easily verify that $\varepsilon[G - \varepsilon](D)(n_1^1) = \emptyset$, while $\varepsilon[G - \varepsilon](D)(n_2^1) \neq \emptyset$, a contradiction.[8] So, this case cannot occur.

[8] Of course, one could also have use the expression $G - \varepsilon$ instead of $\varepsilon[G - \varepsilon]$. However, our choice allows reuse of this part of the proof in a subsequent proof.

2. $|\tilde{n}_1^1| = 2$ *and* $|\tilde{n}_2^1| > 2$. Without loss of generality, we may assume that $\tilde{n}_1^1 = \{n_1^1, n_1^2\}$ and $\tilde{n}_2^1 \supseteq \{n_2^1, n_2^2, n_2^3\}$. Since, for all $i = 3, \ldots, k$, $n_1^1 \not\equiv_e n_1^i$, there exists an XPath-algebra expression E_i such that $E_i(D)(n_1^1) \neq \emptyset$ and $E_i(D)(n_1^i) = \emptyset$. By definition of expression equivalence, we also have, for $j = 1, 2, 3$, that $E_i(D)(n_2^j) \neq \emptyset$.

   Now, let $F := \varepsilon[E_3] \cap \ldots \cap \varepsilon[E_k]$, let $G := F/\uparrow/\downarrow/F$, and let $H := \varepsilon[G - \varepsilon]$. One can easily verify that $((H/H) - \varepsilon)(D)(n_1^1) = \emptyset$, while $((H/H) - \varepsilon)(D)(n_2^1) \neq \emptyset$, a contradiction. So, this case cannot occur either.

We may thus conclude that expression equivalence also satisfies condition (3) of Definition 8. ☐

LEMMA 9. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. If $m_1 \equiv_e m_2$, then $m_1 \equiv^3 m_2$.*

PROOF. By induction on the depth of m_1 in the document.

If m_1 is the root, then m_2 is also the root, for, otherwise, $\uparrow(D)(m_1) = \emptyset$ and $\uparrow(D)(m_2) \neq \emptyset$. Equal nodes are of course 3-equivalent.

If m_1 is not the root, then m_2 cannot be the root either, for, otherwise, we could derive a contradiction as in the base case. Thus, condition (2) of Definition 9 is met. To prove that condition (3) is met, let p_1 be the parent of m_1 and p_2 the parent of m_2. If $p_1 \not\equiv_e p_2$, there exists an XPath-algebra expression E such that $E(D)(p_1) \neq \emptyset$ and $E(D)(p_2) = \emptyset$. Obviously, then $\uparrow/E(D)(m_1) \neq \emptyset$ and $\uparrow/E(D)(m_2) = \emptyset$, a contradiction. Thus, $p_1 \equiv_e p_2$. By the induction hypothesis, $p_1 \equiv^3 p_2$. Finally, Lemma 8 yields condition (1). We may thus conclude that $m_1 \equiv^3 m_2$. ☐

Lemmas 7 and 9 are both directions of a characterization of expression equivalence:

THEOREM 3. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. Then, $m_1 \equiv_e m_2$ if and only if $m_1 \equiv^3 m_2$.*

As a consequence of Theorem 3, expression equivalence is decidable.

We next turn to characterizing those binary relations over the nodes of a document that can be defined as the evaluation of an XPath-algebra expression. For that purpose, we need the following lemma, which is the analogue for the full XPath-algebra of Lemma 3 for the downward (core) XPath-algebra. The proof is completely analogous.

LEMMA 10. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. There exists an XPath-algebra expression E_{m_1} such that $E_{m_1}(D)(m_2) \neq \emptyset$ if and only if $m_1 \equiv^3 m_2$.*

We now prove the main theorem of this section.

THEOREM 4. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $J \subseteq V \times V$. There exists an XPath-algebra expression E such that $E(D) = J$ if and only if, for all $m_1, m_2, n_1, n_2 \in V$ such that $(m_1, n_1) \equiv^3 (m_2, n_2)$, $(m_1, n_1) \in J$ implies $(m_2, n_2) \in J$.*

PROOF. The "only if" follows immediately from Lemma 6. Therefore, we focus on the "if". Thereto, consider the XPath-algebra expression

$$E := \bigcup_{(m_1,n_1)\in J} \varepsilon[E_{m_1}]/\mathrm{Sig}(m_1, n_1)/\varepsilon[E_{n_1}],$$

with E_{m_1} and E_{n_1} as in Lemma 10 and $\mathrm{Sig}(m_1, n_1)$ as in Proposition 1. It is now easily seen that the condition above imposed on J implies that $E(D) = J$. ☐

6. CHARACTERIZING THE SEMANTICS OF THE CORE XPATH-ALGEBRA

Lemma 11, below, is the analogue of Lemma 6 for the core XPath-algebra.

LEMMA 11. *Let E be a core XPath-algebra expression, let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2 \in V$ such that $(m_1, n_1) \Rrightarrow^2 (m_2, n_2)$. If $(m_1, n_1) \in E(D)$, then $(m_2, n_2) \in E(D)$.*

PROOF. The proof goes by induction on the structure of E. The proof of the base case is straightforward. The induction step for the composition E_1/E_2 relies on Proposition 3. The induction step for the union operator $E_1 \cup E_2$ is straightforward. We discuss the induction step for the predicate operator $E_1[E_2]$, with E_1 a core XPath-algebra expression and E_2 a boolean combination of core XPath-algebra expressions, in more detail.

Since E_2 can be normalized in disjunctive normal form, and since set union can be pushed out from the predicate to the expression level, we may assume that E_2 is of the form $F_1 \cap \ldots \cap F_k \cap \overline{G}_1 \cap \ldots \cap \overline{G}_\ell$. If $(m_1, n_1) \in E(D)$, there exists $p_1 \in V$ such that $(m_1, n_1) \in E_1(D)$, $(n_1, p_1) \in F_1(D)$, ..., $(n_1, p_1) \in F_k(D)$, $(n_1, p_1) \notin G_1(D)$ ..., $(n_1, p_1) \notin G_\ell(D)$. By the induction hypothesis, $(m_2, n_2) \in E_1(D)$. By Proposition 4, there exists $p_2 \in V$ such that $(n_1, p_1) \equiv^2 (n_2, p_2)$. In particular, $(n_1, p_1) \Rrightarrow^2 (n_2, p_2)$, whence, by the induction hypothesis, $(n_2, p_2) \in F_1(D)$, ..., $(n_2, p_2) \in F_k(D)$. Since $(n_1, p_1) \equiv^2 (n_2, p_2)$, we also have $(n_2, p_2) \Rrightarrow^2 (n_1, p_1)$.[9] If there were i, $1 \leq i \leq \ell$, such that $(n_2, p_2) \in G_i(D)$, then, by the induction hypothesis, $(n_1, p_1) \in G_i(D)$, a contradiction. We conclude that $(n_2, p_2) \notin G_1(D)$, ..., $(n_2, p_2) \notin G_\ell(D)$, whence $(m_2, n_2) \in E(D)$. ☐

Notice that the absence of difference at the expression level is crucial for this proof to work, as an induction step for the difference operator would fail because of the asymmetry of "$\Rrightarrow^2$".

Combining Proposition 2 and Lemma 11 immediately yields

COROLLARY 3. *Let E be a core XPath-algebra expression, let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1 \in V$ such that $m_1 \equiv^2 m_2$ and $(m_1, n_1) \in E(D)$. Then there exists $n_2 \in V$ such that $(m_2, n_2) \in E(D)$.*

Using the same argument used for statement (1) in the proof of Theorem 1, we obtain

LEMMA 12. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. If $m_1 \equiv^2 m_2$, then $m_1 \equiv_{e^-} m_2$.*

To prove the reverse direction, we proceed in the same way as for the XPath-algebra.

LEMMA 13. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. If $m_1 \equiv_{e^-} m_2$, then $m_1 \equiv^2_\downarrow m_2$.*

PROOF. The proof is completely analogous to that of Lemma 8, except that, in order to prove that core expression equivalence satisfies condition (3) of Definition 9, we must only show that the case "$|\tilde{n}_1^1| = 1$ and $|\tilde{n}_2^1| > 1$" cannot occur. Since the expression exhibited for this case is actually a core XPath-algebra expression, the argument used there can be reused here. ☐

Notice that the expression exhibited in the proof of Lemma 8 to show that the case "$|\tilde{n}_1^1| = 2$ and $|\tilde{n}_2^1| > 2$" can*not* be transformed into a core XPath-algebra expression.

Lemma 13 can be bootstrapped to Lemma 14, in the same way as Lemma 8 to Lemma 9:

[9] Remember that, while "$\equiv^2$" is symmetric, "$\Rrightarrow^2$" in general is *not*!

LEMMA 14. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. If $m_1 \equiv_{e^-} m_2$, then $m_1 \equiv^2 m_2$.*

Lemmas 12 and 14 are both directions of a characterization of core expression equivalence:

THEOREM 5. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. Then, $m_1 \equiv_{e^-} m_2$ if and only if $m_1 \equiv^2 m_2$.*

As a consequence of Theorem 5, core expression equivalence is decidable.

We next turn to characterizing those binary relations over the nodes of a document that can be defined as the evaluation of a core XPath-algebra expression. For that purpose, we need the following lemma, which is the analogue for the core XPath-algebra of Lemma 3 for the downward (core) XPath-algebra. The proof is completely analogous.

LEMMA 15. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2 \in V$. There exists a core XPath-algebra expression E_{m_1} such that $E_{m_1}(D)(m_2) \neq \emptyset$ if and only if $m_1 \equiv^2 m_2$.*

We now prove the main theorem of this section.

THEOREM 6. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $J \subseteq V \times V$. There exists a core XPath-algebra expression E such that $E(D) = J$ if and only if, for all $m_1, m_2, n_1, n_2 \in V$ such that $(m_1, n_1) \Rrightarrow^2 (m_2, n_2)$, $(m_1, n_1) \in J$ implies $(m_2, n_2) \in J$.*

PROOF. The proof is completely analogous to the proof of Theorem 4, except that, in the expression E exhibited, "Sig(m_1, n_1)"—which is *not* a core XPath-algebra expression—is replaced by "sig(m_1, n_1)". □

7. THE LOCAL PERSPECTIVE

Theorems 2, 4, and 6 settle the definability of XPath from a global perspective. Starting from these results, we can now also settle the definability of XPath from a local perspective.

COROLLARY 4. *Let $D = (V, Ed, r, \lambda)$ be a document, let $m \in V$, and let $N \subseteq V$.*

1. *There exists a downward (core) XPath-algebra expression E such that $E(D)(m) = N$ if and only if, for $n_1, n_2 \in V$ with $(m, n_1) \equiv^1_\downarrow (m, n_2)$, $n_1 \in N$ implies $n_2 \in N$.*
2. *There exists an XPath-algebra expression E where $E(D)(m) = N$ if and only if, for $n_1, n_2 \in V$ with $n_1 \equiv^3 n_2$ and sig$(m, n_1) =$ sig(m, n_2), $n_1 \in N$ implies $n_2 \in N$.*
3. *There exists a core XPath-algebra expression E such that $E(D)(m) = N$ if and only if, for $n_1, n_2 \in V$ with $n_1 \equiv^2 n_2$ and sig$(m, n_1) \geq$ sig(m, n_2), $n_1 \in N$ implies $n_2 \in N$.*

For the important special case where the node m is the root, the statements of Corollary 4 can be simplified.

COROLLARY 5. *Let $D = (V, Ed, r, \lambda)$ be a document, and let $N \subseteq V$.*

1. *There exists a downward (core) XPath-algebra expression E such that $E(D)(r) = N$ if and only if, for $n_1, n_2 \in V$ with $n_1 \equiv^1 n_2$, $n_1 \in N$ implies $n_2 \in N$.*
2. *There exists an XPath-algebra expression E where $E(D)(r) = N$ if and only if, for $n_1, n_2 \in V$ with $n_1 \equiv^3 n_2$, $n_1 \in N$ implies $n_2 \in N$.*
3. *There exists a core XPath-algebra expression E such that $E(D)(r) = N$ if and only if, for $n_1, n_2 \in V$ with $n_1 \equiv^2 n_2$, $n_1 \in N$ implies $n_2 \in N$.*

8. DISCUSSION

In this paper, we characterized the expressive power of four natural fragments of XPath at the document level. Of course, it is possible to consider other fragments or extensions of the XPath-algebra and its data model. Analyzing these using our two-step methodology in order to further improve our understanding of XPath is one possible research direction which we are currently pursuing.

Another future research direction is refining the links between XPath and finite-variable first-order logics [16]. Recently, such links have been established at the level of query semantics. For example, Marx [17, 18] has shown that Core XPath [11] is equivalent to $\mathrm{FO}^2_{\mathtt{tree}}$—first-order logic using at most two variables over *ordered* node-labeled trees—interpreted in the signature `child`, `descendant`, and `following_sibling`. Our results establish new links to finite-variable first-order logics at the document level. For example, we can show that, on a given document, the XPath-algebra and FO^3—first-order logic with at most three variables—are equivalent in expressive power. Indeed, we can show that, at the document level, the XPath-algebra is equivalent with Tarski's relation algebra [23] over trees. Tarski and Givant [24] established the link between Tarski's algebra and FO^3. Theorem 3 can then be used to give a new characterization, other than via pebble-games [8, 15], of when two nodes in an unordered tree are indistinguishable in FO^3. In this light, connections between other fragments of the XPath-algebra and finite-variable logics must be examined.

The connection between the XPath-algebra and FO^3 also has ramifications with regard to complexity issues. Indeed, using a result of Grohe [13] which establishes that expression equivalence for FO^3 is decidable in polynomial time, it follows readily from Theorem 4 and Corollary 4 that the global and local definability problems for the XPath-algebra are decidable in polynomial time. By other arguments, based on the syntactic characterizations in this paper, one can also establish that the global and local definability problems for the other fragments of the XPath-algebra are decidable in polynomial time. As mentioned in the Introduction, this feasibility suggests efficient partitioning and reduction techniques on the set of nodes and the set of paths in a document. Such techniques may be successfully leveraged for various aspects of XML document processing such as indexing, access control, and document compression. This is another research direction which we are currently pursuing.

9. ACKNOWLEDGMENTS

We thank Floris Geerts, Jan Hidders, Changqing Lin, Frank Neven, Jan Van den Bussche, and Yuqing Wu for useful discussions. We especially thank Maarten Marx for clarifications about the links between various fragments of XPath and finite-variable logics at the query level.

10. REFERENCES

[1] F. Bancilhon. On the Completeness of Query Languages for Relational Data Bases. In *MFCS*, pages 112–123, Zakopane, Poland, 1978. Springer LNCS 64.

[2] M. Benedikt, W. Fan, and F. Geerts. XPath Satisfiability in the Presence of DTDs. In *ACM PODS*, pages 25–36, Baltimore, MD, USA, 2005.

[3] M. Benedikt, W. Fan, and G. M. Kuper. Structural Properties of XPath Fragments. In *ICDT*, pages 79–95, Siena, Italy, 2003. Springer LNCS 2572.

[4] A. Berglund, S. Boag, D. Chamberlin, M. F. Fernández, M. Kay, J. Robie, and J. Siméon. XML Path Language (XPath) Version 2.0. Technical report, W3C, 2005.

[5] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, Cambridge, UK, 2001.

[6] P. Buneman, M. Grohe, and C. Koch. Path Queries on Compressed XML. In *VLDB*, pages 141–152, Berlin, Germany, 2003.

[7] A. K. Chandra and D. Harel. Computable Queries for Relational Data Bases. *J. Comp. Sys. Sci.*, 21(2):156–178, 1980.

[8] H.-D. Ebbinghaus and J. Flum. *Finite Model Theory*. Springer Verlag, Berlin, 1995.

[9] I. Fundulaki and M. Marx. Specifying Access Control Policies for XML Documents with XPath. In *ACM SACMAT*, pages 61–69, New York, NY, USA, 2004.

[10] R. Goldman and J. Widom. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. In *VLDB*, pages 436–445, Athens, Greece, 1997.

[11] G. Gottlob and C. Koch. Monadic Queries over Tree-Structured Data. In *IEEE LICS*, pages 189–202, Copenhagen, Denmark, 2002.

[12] G. Gottlob, C. Koch, and R. Pichler. Efficient Algorithms for Processing XPath Queries. *ACM Trans. Database Syst.*, 30(2):444–491, 2005.

[13] M. Grohe. Equivalence in Finite-Variable Logics is Complete for Polynomial Time. *Combinatorica*, 19(4):507–532, 1999.

[14] R. Kaushik, P. Shenoy, P. Bohannon, and E. Gudes. Exploiting Local Similarity for Indexing Paths in Graph-Structured Data. In *IEEE ICDE*, pages 129–140, San Jose, CA, USA, 2002.

[15] Ł. Krzeszczakowski. Pebble Games on Trees. In *EACSL CSL*, pages 359–371, Vienna, Austria, 2003. Springer LNCS 2803.

[16] L. Libkin. Logics for Unranked Trees: An Overview. In *EATCS ICALP*, pages 35–50, Lisbon, Portugal, 2005. Springer LNCS 3580.

[17] M. Marx. Conditional XPath, the First Order Complete XPath Dialect. In *ACM PODS*, pages 13–22, Paris, France, 2004.

[18] M. Marx and M. de Rijke. Semantic Characterizations of Navigational XPath. *SIGMOD Record*, 34(2):41–46, 2005.

[19] G. Miklau and D. Suciu. Containment and Equivalence for a Fragment of XPath. *J. ACM*, 51(1):2–45, 2004.

[20] T. Milo and D. Suciu. Index Structures for Path Expressions. In *ICDT*, pages 277–295, Jerusalem, Israel, 1999.

[21] J. Paredaens. On the Expressive Power of the Relational Algebra. *Inf. Process. Lett.*, 7(2):107–111, 1978.

[22] P. Ramanan. Covering Indexes for XML Queries: Bisimulation - Simulation = Negation. In *VLDB*, pages 165–176, Berlin, Germany, 2003.

[23] A. Tarski. On the Calculus of Relations. *J. Symb. Log.*, 6(3):73–89, 1941.

[24] A. Tarski and S. Givant. *A Formalization of Set Theory Without Variables*. American Mathematical Society, Providence, RI, USA, 1987.

APPENDIX

In this Appendix, we provide more details regarding the proofs of Propositions 2–5. For the convenience of the reader, the statements of the results are repeated.

PROPOSITION 2 1. *Let* $k \geq 1$. *Let* $D = (V, Ed, r, \lambda)$ *be a document, and let* $m_1, m_2, n_1 \in V$ *with* $m_1 \equiv^k m_2$. *Then, there exists* $n_2 \in V$ *such that* $(m_1, n_1) \Rrightarrow^k (m_2, n_2)$.

PROOF. The case that n_1 is a descendant of m_1 follows from Lemma 5.

If n_1 is *not* a descendant of m_1, then let $t_1 := \text{top}(m_1, n_1)$. Since $m_1 \equiv^k m_2$, Definition 9 implies that there exists an ancestor t_2 of m_2 such that $t_1 \equiv^k t_2$ and $\text{sig}(m_1, t_1) = \text{sig}(m_2, t_2)$ (1). By the previous case, there exists a descendant n_2 of t_2 such that $n_1 \equiv^k n_2$ and $\text{sig}(t_1, n_1) = \text{sig}(t_2, n_2)$ (2). From (1) and (2), we deduce $\text{sig}(m_1, n_1) \geq \text{sig}(m_2, n_2)$. Lemma 4 now yields the desired result. □

Before proceeding to the proof of Proposition 3, we like to point attention to Figure 2.

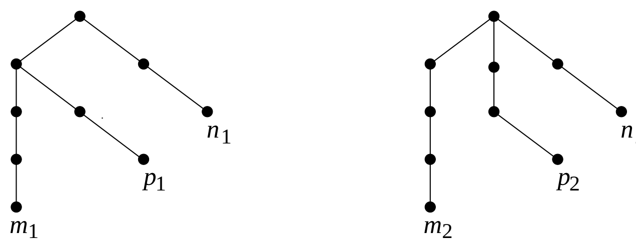

Figure 2: Example document. All nodes are assumed to have the same label

In this document, $\text{sig}(m_1, n_1) = \text{sig}(m_2, n_2)$ and $\text{sig}(n_1, p_1) = \text{sig}(n_2, p_2)$, but $\text{sig}(m_1, p_1) \not\geq \text{sig}(m_2, p_2)$. This example shows that, in the proof of Proposition 3—as well as in the proof of Proposition 5 to follow later—care must be taken in choosing p_2. Therefore, both proofs proceed via an exhaustive case analysis. Once the correct choice for p_2 is made, however, the remainder of the proof for that case is technical but straightforward and, therefore, omitted.

PROPOSITION 3 1. *Let* $k \geq 1$. *Let* $D = (V, Ed, r, \lambda)$ *be a document, and let* $m_1, m_2, n_1, n_2, p_1 \in V$ *such that* $(m_1, n_1) \Rrightarrow^k (m_2, n_2)$. *Then, there exists* $p_2 \in V$ *such that* $(m_1, p_1) \Rrightarrow^k (m_2, p_2)$ *and* $(p_1, n_1) \Rrightarrow^k (p_2, n_2)$.

PROOF. We distinguish three principal cases:

1. *top*(m_1, p_1) *is a strict ancestor of top*(m_1, n_1).
   Then, $\text{top}(p_1, n_1) = \text{top}(m_1, p_1)$. Let p_2 be any node satisfying $(m_1, p_1) \Rrightarrow^k (m_2, p_2)$. (Proposition 2). It can now be shown that $(p_1, n_1) \Rrightarrow^k (p_2, n_2)$. Figure 3 illustrates the constructions in this case for one possible configuration of m_2, n_2, and p_2.
2. *top*(m_1, p_1) *is a strict descendant of top*(m_1, n_1).
   Then, $\text{top}(p_1, n_1) = \text{top}(m_1, n_1)$. Now, let p_2 be any node satisfying $(m_1, p_1) \Rrightarrow^k (m_2, p_2)$ (Proposition 2). Again, it can now be shown that $(p_1, n_1) \Rrightarrow^k (p_2, n_2)$. Figure 4 illustrates the constructions in this case for one possible configuration of m_2, n_2, and p_2.
3. $\text{top}(m_1, p_1) = \text{top}(m_1, n_1)$. We distinguish two subcases:
   (a) *top*(p_1, n_1) *is a strict descendant of top*(m_1, n_1). This situation is shown in Figure 5. This case is the same as the second principal case, with the roles of m_1 and n_1 (and hence the roles of m_2 and n_2) interchanged. Since the statement of the lemma is symmetric in this respect, we may consider this case solved.

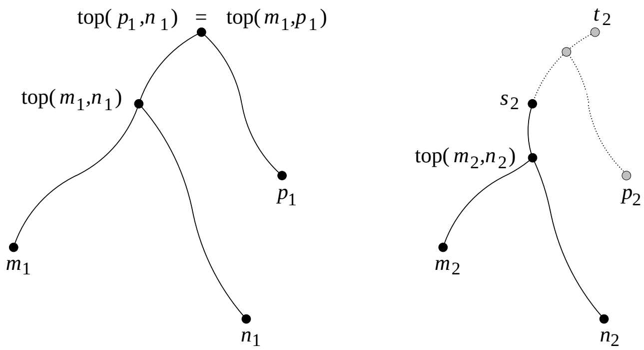

Figure 3: Illustration of the constructions in the first principal case of the proof of Proposition 3. The nodes s_2 and t_2 are the ancestors of m_2 satisfying top$(m_1, n_1) \equiv^k s_2$ and top$(m_1, p_1) \equiv^k t_2$.

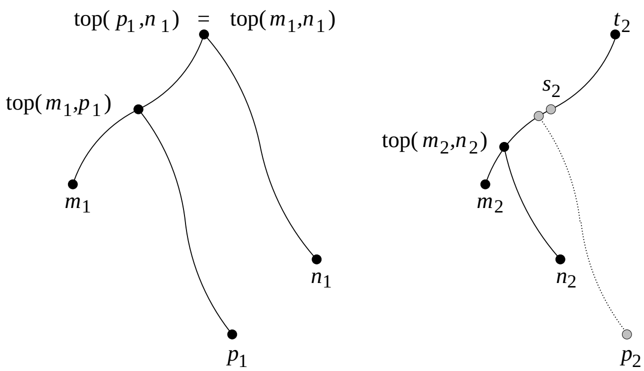

Figure 4: Illustration of the constructions in the second principal case of the proof of Proposition 3. The nodes s_2 and t_2 are the ancestors of m_2 satisfying top$(m_1, n_1) \equiv^k t_2$ and top$(m_1, p_1) \equiv^k s_2$.

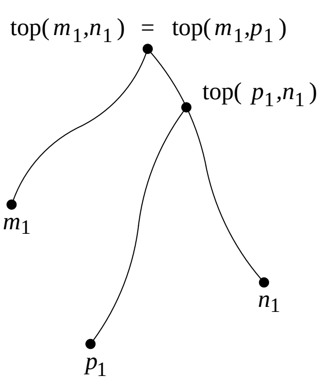

Figure 5: The first subcase of the third principal case of the proof of Proposition 3.

(b) top(p_1, n_1) = top(m_1, p_1) = top(m_1, n_1). This situation is shown in Figure 6. Since, in this subcase, we have, in particular, that top(m_1, n_1) = top(p_1, n_1) and top(m_1, p_1) is on the path from m_1 to top(m_1, n_1), this subcase can be dealt with in the same way as the second principal case.

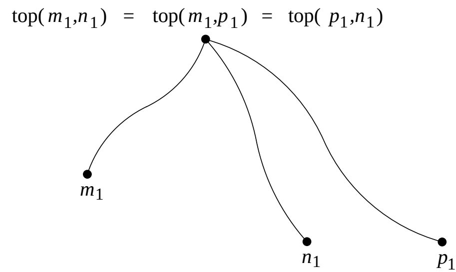

Figure 6: The second subcase of the third principal case of the proof of Proposition 3.

□

PROPOSITION 4 1. *Let $k \geq 2$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1 \in V$ with $m_1 \equiv^k m_2$. Then, there exists $n_2 \in V$ such that $(m_1, n_1) \equiv^k (m_2, n_2)$.*

PROOF. By Lemma 4, it suffices to show that there exists $n_2 \in V$ such that $n_1 \equiv^k n_2$ and sig(m_1, n_1) = sig(m_2, n_2).

The case where n_1 is a descendant of m_1 is covered by Proposition 2, since k-equivalence and k-relatedness coincide in this case.

If n_1 is *not* a descendant of m_1, then consider $t_1 :=$ top(m_1, n_1). Since $m_1 \equiv^k m_2$, Definition 9 implies that there exists an ancestor t_2 of m_2 such that $t_1 \equiv^k t_2$ and sig(m_1, t_1) = sig(m_2, t_2).

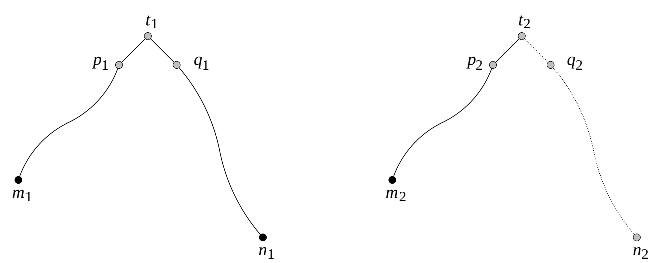

Figure 7: Illustration of the constructions in the proof of Proposition 4.

If $n_1 = t_1$, then, clearly, $n_2 := t_2$ satisfies all requirements. Otherwise, let p_1 be the child of t_1 on the path to m_1, and let q_1 be the child of t_1 on the path to n_1. Clearly, $p_1 \neq q_1$. Also, let p_2 be the child of t_2 on the path to m_2. Clearly, $p_1 \equiv^k p_2$. In particular, $p_1 \equiv^k_\downarrow p_2$. We now distinguish two cases:

1. $p_1 \not\equiv^k_\downarrow q_1$. By Definition 8, there exists a child q_2 of t_2 such that $q_1 \equiv^k_\downarrow q_2$ (whence $q_1 \equiv^k q_2$). Since $p_1 \not\equiv^k_\downarrow q_1, p_2 \not\equiv^k_\downarrow q_2$. In particular, $p_2 \neq q_2$.
2. $p_1 \equiv^k_\downarrow q_1$. By Definition 8, and because of $k \geq 2$, there exists a child q_2 of t_2 such that $p_2 \neq q_2$ and $q_1 \equiv^k_\downarrow q_2$ (whence $q_1 \equiv^k q_2$).

In both cases, $p_2 \neq q_2$ and $q_1 \equiv^k q_2$. Since n_1 is a descendant of q_1, there exists a descendant n_2 of q_2 such that $n_1 \equiv^k n_2$ and sig(q_1, n_1) = sig(q_2, n_2). Since $p_2 \neq q_2$, it follows that sig(m_1, n_1) = sig(m_2, n_2). □

PROPOSITION 5 1. *Let $k \geq 3$. Let $D = (V, Ed, r, \lambda)$ be a document, and let $m_1, m_2, n_1, n_2, p_1 \in V$ such that $(m_1, n_1) \equiv^k (m_2, n_2)$. Then, there exists $p_2 \in V$ such that $(m_1, p_1) \equiv^k (m_2, p_2)$ and $(p_1, n_1) \equiv^k (p_2, n_2)$.*

PROOF. We distinguish three principal cases:

1. *top(m_1, p_1) is a strict ancestor of top(m_1, n_1).*
   Then, top$(p_1, n_1) =$ top(p_2, n_2). Let p_2 be any node satisfying $(m_1, p_1) \equiv^k (m_2, p_2)$. (Such a node exists, by Proposition 4.) It is now readily seen that $(p_1, n_1) \equiv^k (p_2, n_2)$.

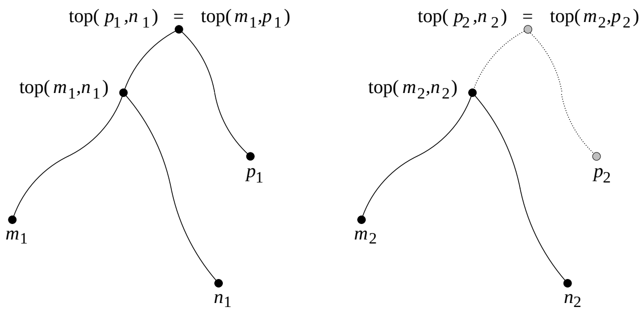

Figure 8: Illustration of the constructions in the first principal case of the proof of Proposition 5.

2. *top(m_1, p_1) is a strict descendant of top(m_1, n_1).*
   Let p_2 be any node satisfying $(m_1, p_1) \equiv^k (m_2, p_2)$. (Such a node exists, by Proposition 4.) It is now readily seen that $(p_1, n_1) \equiv^k (p_2, n_2)$.

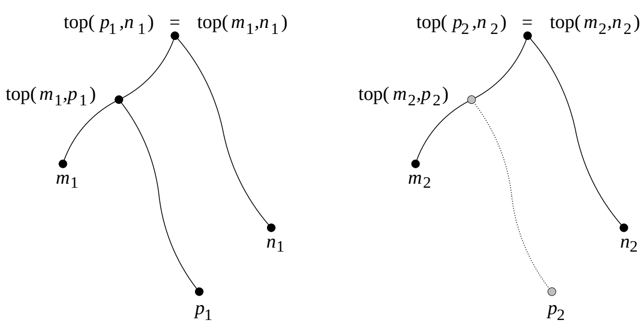

Figure 9: Illustration of the constructions in the second principal case of the proof of Proposition 5.

3. top$(m_1, p_1) =$ top(m_1, n_1). We distinguish two subcases:
   (a) *top(p_1, n_1) is a strict descendant of top(m_1, n_1).*
   Let p_2 be any node satisfying $(p_1, n_1) \equiv^k (p_2, n_2)$. (Such a node exists, by Proposition 4.) It is now readily seen that $(m_1, p_1) \equiv^k (m_2, p_2)$.

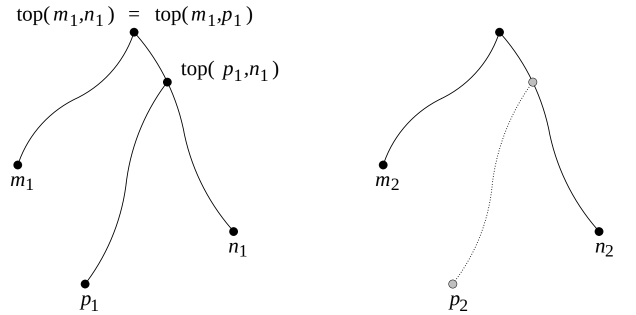

Figure 10: Illustration of the constructions in the first subcase of the third principal case of the proof of Proposition 5.

   (b) top$(p_1, n_1) =$ top$(m_1, p_1) =$ top(m_1, n_1).
   If p_1 equals this top node, then let $p_2 := top(m_2, n_2)$. If m_1 equals this top node, then let p_2 be any node satisfying $(p_1, n_1) \equiv^k (p_2, n_2)$. Finally, if n_1 equals this top node, then let p_2 be any node satisfying $(m_1, p_1) \equiv^k (m_2, p_2)$. (Such nodes exist, by Proposition 4.) It is readily seen that, in all these border cases, p_2 satisfies all requirements.
   If none of these bordercases occur, we are in the situation shown in Figure 11.

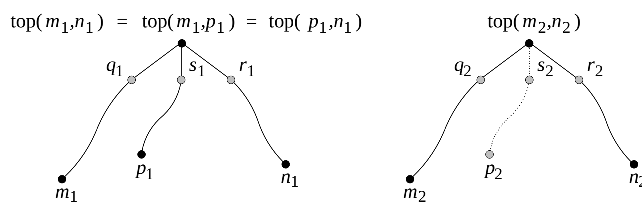

Figure 11: Illustration of the constructions in the second subcase of the third principal case of the proof of Proposition 5.

   Let q_1, r_1, and s_1 be the children of top(m_1, n_1) on the paths to m_1, n_1, and p_1, respectively, and let q_2 and r_2 be the children of top(m_2, n_2) on the paths to m_2 and n_2, respectively. Clearly, top$(m_1, n_1) \equiv^k$ top(m_2, n_2), whence, in particular, top$(m_1, n_1) \equiv^k_{\downarrow}$ top(m_2, n_2). By Definition 8, and since $k \geq 3$, it can be seen in an analogous way as in the proof of Proposition 4 that there exists a child s_2 of top(m_2, n_2) such that $s_1 \equiv^k_{\downarrow} s_2$ (whence $s_1 \equiv^k s_2$), $s_2 \neq q_2$, and $s_2 \neq r_2$.
   Finally, let $p_2 \in V$ be any descendant of s_2 satisfying $(s_1, p_1) \equiv^k (s_2, p_2)$. (Since k-equivalence and k-relatedness coincide for ancestor-descendant pairs, such a node exists, by Proposition 2.) Obviously, sig(m_1, p_1) = sig(m_2, p_2) and sig(p_1, n_1) = sig(p_2, n_2), whence $(m_1, p_1) \equiv^k (m_2, p_2)$ and $(p_1, n_1) \equiv^k (p_2, n_2)$.

□