Disjoint Pattern Matching and Implication in Strings

Leonid Libkin* Cristina Sirangelo†

Abstract

We deal with the problem of deciding whether a given set of string patterns implies the presence of a fixed pattern. While checking whether a set of patterns occurs in a string is solvable in polynomial time, this implication problem is well-known to be intractable. Here we consider a version of the problem when patterns in the set are required to be disjoint. We show that for such a version of the problem the situation is reversed: checking whether a set of patterns occurs in a string is NP-complete, but the implication problem is solvable in polynomial time.

1 Introduction and the main result

The problem we consider in this note was motivated by answering queries in incompletely specified XML documents. Suppose that $\mathcal{L}$ is a set of letters, or labels, assumed to be countably infinite, and that _ is a special symbol (wildcard) not in $\mathcal{L}$. By $\mathcal{L}_{\_}$ we denote $\mathcal{L} \cup \{\_\}$. A *pattern* is a finite string over $\mathcal{L}_{\_}$. If a string s over $\mathcal{L}$ matches a pattern π, we write $s \models \pi$. More precisely, if $s = a_0 \dots a_{n-1}$ and $\pi = b_0 \dots b_{m-1}$, then $s \models \pi$ if there is a number $j \leq n - m$ so that for each b_i that is not a wildcard, we have $b_i = a_{i+j}$ (i.e., a wildcard can match every symbol). In this case we also say s matches π from j to $j + m - 1$. If s matches every pattern in a set Π of patterns, we write $s \models \Pi$.

We say that a set Π of patterns *implies* a pattern π, written as $\Pi \vdash \pi$ if $s \models \pi$ whenever $s \models \Pi$, for every finite string s over $\mathcal{L}$. Now, for each pattern π, consider the *implication problem* $\textsc{Impl}(\pi)$: its input is a set Π of patterns, and the question is whether $\Pi \vdash \pi$.

This problem is known to be coNP-complete, even for very simple patterns π. To see hardness, consider the following well-known NP-complete problem [4]: given strings $s_1, \dots, s_l$ and a number K in unary, is there a string s of length K so that each s_i is a substring of s? For the reduction, assume that \$ is in $\mathcal{L}$ but not used in any of the s_i's, and take each s_i as a pattern, as well as the string \$_ . . . _\$ (consisting of exactly K wildcards between a leading and a trailing \$-symbol). Then this set of patterns implies _\$_ iff every string that contains all the s_i's has length greater than K (as it has symbols both left and right of a \$); this shows coNP-hardness of $\textsc{Impl}(\_\$\_)$.

Notice, however, that in this reduction the matches for the s_i's may well overlap. We are concerned with the implication when pattern matches are required to be *disjoint*. More precisely, we say that s *disjointly matches* $\{\pi_0, \dots, \pi_{n-1}\}$ if there are disjoint intervals $[i_0, j_0]$, ..., $[i_{n-1}, j_{n-1}]$ so that s matches each π_l from i_l to j_l. We denote this by $s \models_{\mathrm{d}} \{\pi_0, \dots, \pi_{n-1}\}$. We then say that $\{\pi_0, \dots, \pi_{n-1}\}$ *disjointly implies* π, written as $\{\pi_0, \dots, \pi_{n-1}\} \vdash_{\mathrm{d}} \pi$, if, for every s, we have $s \models \pi$ whenever $s \models_{\mathrm{d}} \{\pi_0, \dots, \pi_{n-1}\}$.

The *disjoint implication problem* $\textsc{Impl}_{\mathrm{d}}(\pi)$ then takes a set of patterns Π as an input and checks whether $\Pi \vdash_{\mathrm{d}} \pi$.

The motivation for this notion comes from the study of querying incompletely described XML documents [1, 2, 3]. For instance, one may have an incomplete description of children of a node, and a

*School of Informatics, University of Edinburgh, libkin@inf.ed.ac.uk

†LSV, ENS-Cachan, cristina.sirangelo@lsv.ens-cachan.fr

query given by a pattern over a set of possible labels of nodes. The standard database approach is to look for *certain answers*, i.e., query evaluation should return true if every XML document matching the incomplete description matches the query pattern as well. This is precisely the pattern implication problem. Very often, XML documents are accessed using the DOM interface, in which every node has its own unique id. Then matches of different subpatterns cannot overlap, as this would lead to id clashes. Hence, query answering in incomplete XML documents with node ids corresponds to the problem of disjoint pattern implication. It is also common to consider the setting where a query is fixed, while data varies (the notion of data complexity).

For the usual matching notion, checking whether $s \models \Pi$ is polynomial, but $\textsc{Impl}(\pi)$ could be coNP-hard, as we have just seen. For the disjoint versions of these problems, the situation is exactly the opposite: checking whether $s \models_{\mathrm{d}} \Pi$ is intractable, while $\textsc{Impl}_{\mathrm{d}}(\pi)$ is solvable in polynomial time for every π.

Theorem 1
- *The problem of checking, for a string s and a set of patterns Π, whether $s \models_{\mathrm{d}} \Pi$ is* NP*-complete. It remains* NP*-complete even if the patterns do not use the wildcard.*
- *For each pattern π, the problem $\textsc{Impl}_{\mathrm{d}}(\pi)$ is solvable in polynomial time.*

2 Proof of Theorem 1

We start with the first item. NP membership is immediate – just guess a disjoint matching. For hardness, we use a reduction from the following version of binpacking: given a set $N = \{u_0, \ldots, u_{n-1}\} \subset \mathbb{N}$, as well as $\ell \in \mathbb{N}$ (number of bins) and $m \in \mathbb{N}$ (capacity of bins), is there a partition of N into ℓ sets so that the sum of numbers in each set is at most m? The problem is NP-hard even if all the numbers are in unary. Now assume that 0 and 1 are elements of $\mathcal{L}$; define π_i as 0^{u_i} for each $i \leq n-1$, and define $s = (0^m 1)^\ell$. Then $s \models_{\mathrm{d}} \{\pi_0, \ldots, \pi_{n-1}\}$ iff there exists a solution to the binpacking problem.

We now move to the second item. We start with some additional notations and auxiliary results.

Given a string s and a position $i < |s|$, we shall refer to i as the *prefix* of that position in s and to $|s| - i - 1$ as the *suffix* of that position. If $s \models \pi$ from i to j, we will also say that s matches π with prefix i and suffix $|s| - j - 1$.

We apply the notion of pattern matching, as is, to pairs of patterns: given two patterns $\pi_1 = a_0 \ldots a_{n-1}$ and $\pi_2 = b_0 \ldots b_{m-1}$ in $\mathcal{L}_\_^*$, we write $\pi_1 \models \pi_2$ iff there is a number $j \leq n - m$ so that for each b_i that is not a wildcard, we have $b_i = a_{i+j}$.

We also extend the notion of disjoint pattern matching to ordered sets of patterns. Given a string $s \in \mathcal{L}^*$, a set of patterns $\Pi = \{\pi_0, \ldots \pi_{n-1}\}$ in $\mathcal{L}_\_^*$, a subset $O \subseteq \Pi$, and a total order $<_O$ over O, we say that s disjointly matches Π under $<_O$, and write $s \models_{\mathrm{d}} (\Pi, <_O)$, if there are disjoint intervals $[i_0, j_0]$, …, $[i_{n-1}, j_{n-1}]$ so that:

- s matches each π_l from i_l to j_l;
- for all $\pi_l, \pi_r \in O$, if $\pi_l <_O \pi_r$ then $j_l < i_r$.

This naturally leads to the notion $(\Pi, <_O) \vdash_{\mathrm{d}} \pi$.

Assume that π is a pattern in $\mathcal{L}_\_^*$. Let $\Pi = \{\pi_0, \ldots, \pi_{n-1}\}$ be a set of patterns in $\mathcal{L}_\_^*$. Without loss of generality, let

$$\pi = (\_)^{k_{pre}} \pi_c (\_)^{k_{suf}}$$

where $k_{pre}, k_{suf} \geq 0$ and π_c is either the empty string or

$$\pi_c = \ell_0 \ell_2 \cdots \ell_{p-1}$$

with $\ell_0, \ell_{p-1} \in \mathcal{L}$ and $p \geq 1$.

If π_c is empty, $\textsc{Impl}_{\mathrm{d}}(\pi)$ is straightforward: $\Pi \vdash_{\mathrm{d}} (\_)^k$ iff $\sum_{\pi_i \in \Pi} |\pi_i| \geq k$. In fact if $\sum_{\pi_i \in \Pi} |\pi_i| \geq k$, given a string s such that $s \models_{\mathrm{d}} \Pi$ we know $|s| \geq \sum_{\pi_i \in \Pi} |\pi_i| \geq k$. Then $s \models (\_)^k$. On the other hand, for each $\pi_i \in \Pi$, by definition of the matching relation, there exists always a string s_i such that $s_i \models \pi_i$ and $|s_i| = |\pi_i|$. Then the string $s = s_0 s_1 \cdots s_{n-1}$ disjointly matches Π and has size $|s| = \sum_{\pi_i \in \Pi} |\pi_i|$. Thus if $\sum_{\pi_i \in \Pi} |\pi_i| < k$, clearly $s \not\models (\_)^k$, and therefore $\Pi \nvdash_{\mathrm{d}} (\_)^k$.

Hence, in the rest of the proof π_c is assumed to be non-empty. Then, given a string s, clearly $s \models \pi$ iff $s \models \pi_c$ with prefix at least k_{pre} and suffix at least k_{suf}.

We will also need the following:

Claim 1 *If $O = \{\rho_0, \ldots, \rho_{m-1}\}$ is an ordered set of patterns (with ordering $<_O$), then there exists a string s over $\mathcal{L}$ such that $s \models_{\mathrm{d}} (O, <_O)$ and, if $s \models \pi$, there exists an index $i \in [0, \ldots, m-1]$ so that the following hold:*

- *$\rho_i \models \pi_c$;*
- *if $i = 0$ then $\rho_i \models \pi_c$ with prefix at least k_{pre};*
- *if $i = m-1$ then $\rho_i \models \pi_c$ with suffix at least k_{suf}.*

Proof Assume w.l.o.g. that in the ordered set O, we have $\rho_i <_O \rho_j$ whenever $i < j$. We construct the string s as follows. Let l be an arbitrary label of $\mathcal{L}$ not occurring in π_c. For each ρ_i, if $\rho_i = a_0 \cdots a_k$, construct the string $s_i = b_0 \cdots b_k$, where $b_i = a_i$ if $a_i \in \mathcal{L}$, and $b_i = l$ if a_i is a wildcard. Fix arbitrarily an integer $w \geq |\pi_c|$ and define $s = s_0 l^w s_1 l^w \ldots l^w s_{m-1}$ (which gives $s = s_0$ in the case $m = 1$). Clearly $s \models_{\mathrm{d}} (O, <_O)$.

Now assume that $s \models \pi$. This implies that $s \models \pi_c$ from some position i_c to some position j_c with prefix at least k_{pre} and suffix at least k_{suf}. Since $\pi_c = \ell_0 \ell_1 \cdots \ell_{p-1}$ with $\ell_0, \ell_{p-1} \in \mathcal{L}$, the following holds:

- positions i_c and j_c occur within some of the substrings s_i's; in fact $s(i_c) = \ell_0$ and $s(j_c) = \ell_{p-1}$ and this can occur only in the s_i's as all other positions have label l not occurring in π_c.
- positions i_c and j_c occur in the same substring s_i of s, since $w \geq |\pi_c|$.

Fix i, with $0 \leq i \leq m-1$, and assume that i_c and j_c occur in s_i; then $s_i \models \pi_c$ starting from some position j. As a consequence, for all ℓ_k which is not a wildcard we have $s_i(j+k) = \ell_k$. Then, by construction, also $\rho_i(j+k) = \ell_k$, which shows that $\rho_i \models \pi_c$ from position j to position $j + |\pi_c| - 1$. In particular, if $i = 0$, then $j = i_c \geq k_{pre}$; thus $\rho_i \models \pi_c$ with prefix at least k_{pre}. Similarly, if $i = m-1$, then the suffix of position $j + |\pi_c| - 1$ in s_i (as well as in ρ_i) equals the suffix of position j_c in s, which is at least k_{suf}; therefore $\rho_i \models \pi_c$ with suffix at least k_{suf}.

This implies the claim. □

We now describe the procedure to verify whether $\Pi \vdash_{\mathrm{d}} \pi$. Let B be the set of all patterns $\pi_k \in \Pi$ such that $\pi_k \models \pi_c$ (and $|B|$ its cardinality). Let H_B be the set of all possible matchings of π_c in patterns of B, i.e., H_B is the set of all pairs $(\pi_k, [i, j])$ such that $\pi_k \in B$ and $\pi_k \models \pi_c$ from i to j. These sets can be computed in time polynomial in the size of Π and π; indeed for each pattern $\pi_k \in \Pi$, and for each position in π_k, we simply check whether $\pi_k \models \pi_c$ starting from that position.

We now show that if $B = \emptyset$, then $\Pi \nvdash_{\mathrm{d}} \pi$. Let $<_\Pi$ be an arbitrary total order on Π. We know that there exists a string $s \models_{\mathrm{d}} (\Pi, <_\Pi)$ satisfying conditions of Claim 1. Clearly $s \models_{\mathrm{d}} \Pi$, but on the other hand $s \not\models \pi$, otherwise by Claim 1, we would have $\pi_i \models \pi_c$ for some $\pi_i \in \Pi$. Then π_i would belong to B, which would contradict the fact that B is empty. Hence $\Pi \nvdash_{\mathrm{d}} \pi$ for empty B.

Now assume $B \neq \emptyset$. Observe that for each string s, we have $s \models_{\mathrm{d}} \Pi$ iff $s \models_{\mathrm{d}} (\Pi, <_B)$ for some total order $<_B$ on B.

For each total order $<_B$ on B and for each matching $\mu = (\pi_k, [i, j]) \in H_B$, define integers $pre_{<_B}(\mu)$ and $suf_{<_B}(\mu)$ as the prefix and suffix of the interval $[i, j]$ in μ, ordered with $<_B$. More precisely:

$$pre_{<_B}(\mu) = i + \sum_{\pi_l \in B,\ \pi_l <_B \pi_k} |\pi_l|$$

$$suf_{<_B}(\mu) = |\pi_k| - j - 1 + \sum_{\pi_l \in B,\ \pi_k <_B \pi_l} |\pi_l|$$

Since H_B can be computed in polynomial time in the size of Π and π, for each given total order $<_B$, the integers $pre_{<_B}(\mu)$ and $suf_{<_B}(\mu)$ for all $\mu \in H_B$ can be computed in polynomial time as well.

Claim 2 *For each string s such that $s \models_{\mathrm{d}} (\Pi, <_B)$ and pattern $\pi_k \in B$, there exists a partition (S, S') of $\Pi \setminus B$ and integers $P \geq \sum_{\pi_i \in S} |\pi_i|$ and $X \geq \sum_{\pi_i \in S'} |\pi_i|$ such that the following holds:*

if $\mu \in H_B$ is a matching in π_k, then $s \models \pi_c$ with prefix $pre_{<_B}(\mu)+P$ and suffix $suf_{<_B}(\mu)+X$;

Proof Since $s \models_{\mathrm{d}} (\Pi, <_B)$, there exist disjoint intervals $[i_k, j_k]$ in s which match patterns π_k of Π, for $k = 0, \ldots, n-1$. In particular, among these intervals of s, the ones matching patterns of B follow the order induced by $<_B$. Therefore since $\pi_k \in B$, intervals $[i_r, j_r]$ such that $\pi_r <_B \pi_k$ precede $[i_k, j_k]$. Similarly intervals $[i_r, j_r]$ such that $\pi_k <_B \pi_r$ follow $[i_k, j_k]$. Moreover there exists a partition (S, S') of $\Pi \setminus B$ such that intervals $[i_r, j_r]$ with $\pi_r \in S$ precede $[i_k, j_k]$ and intervals $[i_r, j_r]$ with $\pi_r \in S'$ follow $[i_k, j_k]$. Since all those intervals are disjoint, and since the cardinality of $[i_r, j_r]$ is equal to $|\pi_r|$, we have $i_k \geq \sum_{\pi_r <_B \pi_k} |\pi_r| + \sum_{\pi_r \in S} |\pi_r|$ and $|s| - j_k - 1 \geq \sum_{\pi_k <_B \pi_r} |\pi_r| + \sum_{\pi_r \in S'} |\pi_r|$. In particular let

$$i_k = \sum_{\pi_r <_B \pi_k} |\pi_r| + P \text{ with } P \geq \sum_{\pi_r \in S} |\pi_r|$$

and

$$|s| - j_k - 1 = \sum_{\pi_k <_B \pi_r} |\pi_r| + X \text{ with } X \geq \sum_{\pi_r \in S'} |\pi_r|$$

Let $\mu = (\pi_k, [i_c, j_c])$ be a matching in π_k, i.e., π_k matches π_c from i_c to j_c. By transitivity of the matching relation, s matches π_c from $i_k + i_c$ to $i_k + j_c$. Since $\pi_k \in B$, the prefix of $i_k + i_c$ is

$$i_k + i_c = \sum_{\pi_r <_B \pi_k} |\pi_r| + P + i_c = pre_{<_B}(\mu) + P$$

and the suffix of $i_k + j_c$ is

$$|s| - 1 - (i_k + j_c) = |s| - 1 - j_k + j_k - (i_k + j_c) = \sum_{\pi_k <_B \pi_r} |\pi_r| + X + j_k - i_k - j_c = suf_{<_B}(\mu) + X$$

Then s matches π_c with prefix $pre_{<_B}(\mu) + P$ and suffix $suf_{<_B}(\mu) + X$, as claimed. □

We now distinguish two cases. If $|B| \geq k_{pre} + k_{suf} + 1$, then $\Pi \vdash_{\mathrm{d}} \pi$, as shown below.

Claim 3 *If $|B| \geq k_{pre} + k_{suf} + 1$, then $\Pi \vdash_{\mathrm{d}} \pi$.*

Proof Given a string s such that $s \models_{\mathrm{d}} \Pi$ then, as observed above, $s \models_{\mathrm{d}} (\Pi, <_B)$ for some total order $<_B$ on B. W.l.o.g. let B, totally ordered with $<_B$, be the sequence $\pi_0 \dots \pi_{|B|-1}$ of patterns of Π and let $r = k_{pre}$. We now take the pattern π_r which must belong to B and have, according to $<_B$, exactly k_{pre} preceding patterns in B and at least k_{suf} following patterns in B. Then we take an arbitrary matching $\mu = (\pi_r, [i, j]) \in H_B$ and have $pre_{<_B}(\mu) \geq |\pi_0| + \dots + |\pi_{r-1}| \geq k_{pre}$ and $suf_{<_B}(\mu) \geq |\pi_{r+1}| + \dots + |\pi_{|B|-1}| \geq k_{suf}$. It follows from Claim 2 that $s \models \pi_c$ with prefix at least k_{pre} and suffix at least k_{suf}. This shows that $s \models \pi$ and concludes the proof of the claim. □

Now assume $|B| < k_{pre} + k_{suf} + 1$. In this case $|B|$ is bounded by a constant that only depends on the number of leading and trailing wildcards in π (since π is fixed); this allows us to consider all possible total orders $<_B$ on B. That is, we can check whether $\Pi \vdash_{\mathrm{d}} \pi$ by checking $(\Pi, <_B) \vdash_{\mathrm{d}} \pi$ for each total order $<_B$ on B. Now we show how to check whether $(\Pi, <_B) \vdash_{\mathrm{d}} \pi$.

Assume again w.l.o.g. that B ordered with $<_B$ is given by the sequence $\pi_0, \dots \pi_{|B|-1}$ of patterns of Π. We first check in time polynomial in the size of Π and π whether there exists a matching $\mu \in H_B$ such that $pre_{<_B}(\mu) \geq k_{pre}$ and $suf_{<_B}(\mu) \geq k_{suf}$. The claim below says that this allows to conclude that $(\Pi, <_B) \vdash_{\mathrm{d}} \pi$.

Claim 4 *If there exists a matching $\mu \in H_B$ such that $pre_{<_B}(\mu) \geq k_{pre}$ and $suf_{<_B}(\mu) \geq k_{suf}$, then $(\Pi, <_B) \vdash_{\mathrm{d}} \pi$*

Proof Given a string s such that $s \models_{\mathrm{d}} (\Pi, <_B)$, by Claim 2, $s \models \pi_c$ with prefix at least $pre_{<_B}(\mu) \geq k_{pre}$ and suffix at least $suf_{<_B}(\mu) \geq k_{suf}$. Therefore $s \models \pi$. □

If there does not exist a matching satisfying conditions of Claim 4, there are two possibilities covered by the following three claims.

Claim 5 *If for all $\pi_i \in B$, either all matchings $\mu \in H_B$ into π_i have $pre_{<_B}(\mu) < k_{pre}$ or all matchings $\mu \in H_B$ into π_i have $suf_{<_B}(\mu) < k_{suf}$, then $(\Pi, <_B) \nvdash_{\mathrm{d}} \pi$.*

Proof We show that there must exist a partition of the sequence $(\pi_0, \dots, \pi_{|B|-1})$ into two (possibly empty) subsequences $\sigma_1 = (\pi_0, \dots, \pi_j)$ and $\sigma_2 = (\pi_{j+1}, \dots, \pi_{|B|-1})$ such that

- for each π_i in σ_1, all matchings $\mu \in H_B$ into π_i have $pre_{<_B}(\mu) < k_{pre}$ and
- for each π_i in σ_2, all matchings $\mu \in H_B$ into π_i have $suf_{<_B}(\mu) < k_{suf}$.

Let j be the maximum index in $[0, \dots, |B| - 1]$ such that all the matchings $\mu \in H_B$ into π_j have $pre_{<_B}(\mu) < k_{pre}$, if it exists. If j exists, let $\sigma_1 = (\pi_0, \dots, \pi_j)$ and $\sigma_2 = (\pi_{j+1}, \dots, \pi_{|B|-1})$, otherwise let σ_1 be empty and $\sigma_2 = (\pi_0, \dots, \pi_{|B|-1})$. Clearly, for all π_i in σ_1 and for all matchings $\mu \in H_B$ into π_i, we also have $pre_{<_B}(\mu) < k_{pre}$. Moreover, by definition of σ_1 and σ_2 and by the hypothesis of the claim, for all π_i in σ_2, all matchings $\mu \in H_B$ into π_i have $suf_{<_B}(\mu) < k_{suf}$. This proves the above mentioned properties of σ_1 and σ_2.

Now let π_{pre} (resp. π_{suf}) be the pattern obtained as the concatenation of patterns of σ_1 (resp., σ_2), in the same order as they appear in σ_1 (resp., σ_2). We take the totally ordered set of patterns $O = \{\pi_{pre}, \pi_{|B|}, \dots, \pi_{n-1}, \pi_{suf}\}$ and let $<_O$ be the corresponding total order. We know that there exists a string s satisfying conditions of Claim 1 with O. Notice that $s \models_{\mathrm{d}} (\Pi, <_B)$; we now show that $s \not\models \pi$, thus showing that $(\Pi, <_B) \nvdash_{\mathrm{d}} \pi$.

Assume to the contrary that $s \models \pi$. Then there exists an index $i \in [0, \dots, |O| - 1]$ where the conditions of Claim 1 are satisfied. This index i must be equal to either 0 or $|O| - 1$, otherwise there would exists a pattern in $\{\pi_{|B|}, \dots, \pi_{n-1}\}$ matching π_c; then this pattern would belong to B, contradicting the hypothesis that $B = \{\pi_0, \dots, \pi_{|B|-1}\}$. Assume first $i = 0$.

In this case, by Claim 1, $\pi_{pre} \models \pi_c$ with prefix $p \geq k_{pre}$. On the other hand p must be at most equal to the maximum value of $pre_{<_B}(\mu)$ for all matchings $\mu \in H_B$ in patterns of σ_1. By definition of σ_1, this maximum value is strictly less than k_{pre}, then $p < k_{pre}$. This is a contradiction.

With a symmetric argument we reach a contradiction also in the case that $i = |O| - 1$. This proves that $(\Pi, <_B) \nvdash_{\mathrm{d}} \pi$, as claimed. □

If conditions of Claim 5 are not satisfied, there must exist $\pi_r \in B$ and two matchings $\mu_1, \mu_2 \in H_B$ in π_r such that $suf_{<_B}(\mu_1) \geq k_{suf}$ and $pre_{<_B}(\mu_2) \geq k_{pre}$. On the other hand we must have $pre_{<_B}(\mu_1) < k_{pre}$ and $suf_{<_B}(\mu_2) < k_{suf}$, otherwise either μ_1 or μ_2 would satisfy conditions of Claim 4. (Notice that such a π_r, as well as μ_1 and μ_2 can be found by simply scanning H_B.) In this situation there are again two cases depending on whether $\sum_{\pi \in \Pi \setminus B} |\pi_i| \geq k_{pre} + k_{suf}$ holds.

Claim 6 *If $\sum_{\pi \in \Pi \setminus B} |\pi_i| \geq k_{pre} + k_{suf}$, then $(\Pi, <_B) \vdash_{\mathrm{d}} \pi$.*

Proof Let $s \models_{\mathrm{d}} (\Pi, <_B)$, then s with the pattern π_r satisfies properties of Claim 2. In particular, since both μ_1 and μ_2 are matchings in π_r, Claim 2 implies that there exist integers P and X such that $P + X \geq \sum_{\pi_i \in \Pi \setminus B} |\pi_i| \geq k_{pre} + k_{suf}$ and both the following conditions hold:

- $s \models \pi_c$ with prefix $p_1 = pre_{<_B}(\mu_1) + P$ and suffix $x_1 = suf_{<_B}(\mu_1) + X \geq k_{suf}$;
- $s \models \pi_c$ with prefix $p_2 = pre_{<_B}(\mu_2) + P \geq k_{pre}$ and suffix $x_2 = suf_{<_B}(\mu_2) + X$.

Now there are two cases. If $P \geq k_{pre}$, then $p_1 \geq k_{pre}$, therefore $s \models \pi_c$ with prefix $p_1 \geq k_{pre}$ and suffix $x_1 \geq k_{suf}$. If conversely $P < k_{pre}$, then $X > k_{suf}$, hence $x_2 > k_{suf}$; therefore $s \models \pi_c$ with prefix $p_2 \geq k_{pre}$ and suffix $x_2 \geq k_{suf}$. In both cases $s \models \pi$. This shows that $(\Pi, <_B) \vdash_{\mathrm{d}} \pi$. □

If $\sum_{\pi_i \in \Pi \setminus B} |\pi_i| < k_{pre} + k_{suf}$, this quantity is a constant depending only on the number of leading and trailing wildcards in the pattern π. Then we consider all possible partitions of $\Pi \setminus B$ into two sets S and S'; there is a constant number of them. Checking all these partitions, we can verify whether $(\Pi, <_B) \vdash_{\mathrm{d}} \pi$ as follows.

Claim 7 *If for each partition (S, S') of $\Pi \setminus B$ there exists a matching $\mu \in H_B$ into π_r such that $pre_{<_B}(\mu) + \sum_{\pi_i \in S} |\pi_i| \geq k_{pre}$ and $suf_{<_B}(\mu) + \sum_{\pi_i \in S'} |\pi_i| \geq k_{suf}$, then $(\Pi, <_B) \vdash_{\mathrm{d}} \pi$. Otherwise $(\Pi, <_B) \nvdash_{\mathrm{d}} \pi$.*

Proof Assume $s \models_{\mathrm{d}} (\Pi, <_B)$. Then we know that for s and π_r there exists a partition (S, S') of $\Pi \setminus B$ and integers P and X as stated in Claim 2. By the hypothesis, we also know that there exists a matching $\mu \in H_B$ into π_r such that $pre_{<_B}(\mu) + \sum_{\pi_i \in S} |\pi_i| \geq k_{pre}$ and $suf_{<_B}(\mu) + \sum_{\pi_i \in S'} |\pi_i| \geq k_{suf}$. Then by Claim 2, $s \models \pi_c$ with prefix $pre_{<_B}(\mu) + P \geq pre_{<_B}(\mu) + \sum_{\pi_i \in S} |\pi_i| \geq k_{pre}$ and suffix $suf_{<_B}(\mu) + X \geq suf_{<_B}(\mu) + \sum_{\pi_i \in S'} |\pi_i| \geq k_{suf}$. This shows that $s \models \pi$ and therefore $(\Pi, <_B) \vdash_{\mathrm{d}} \pi$.

Conversely, assume there exists a partition (S, S') of $\Pi \setminus B$ such that all matchings $\mu \in H_B$ into π_r have either $pre_{<_B}(\mu) + \sum_{\pi_i \in S} |\pi_i| < k_{pre}$ or $suf_{<_B}(\mu) + \sum_{\pi_i \in S'} |\pi_i| < k_{suf}$. Notice that in particular μ_1 must have $pre_{<_B}(\mu_1) + \sum_{\pi_i \in S} |\pi_i| < k_{pre}$, because it satisfies $suf_{<_B}(\mu_1) \geq k_{suf}$. Similarly we must have $suf_{<_B}(\mu_2) + \sum_{\pi_i \in S'} |\pi_i| < k_{suf}$.

Let π_S and $\pi_{S'}$ be the concatenations of patterns of S and of S', respectively, in an arbitrary order. Let $\pi_{SS'}$ be $\pi_S \pi_0 \cdots \pi_{|B|-1} \pi_{S'}$. We know there exists a string s satisfying the conditions of Claim 1 with $O = \{\pi_{SS'}\}$; notice that $s \models_{\mathrm{d}} (\Pi, <_B)$. We now show that $s \not\models \pi$, thus showing that $(\Pi, <_B) \nvdash_{\mathrm{d}} \pi$.

Assume to the contrary that $s \models \pi$. Then, by Claim 1, $\pi_{SS'} \models \pi_c$ with prefix $p \geq k_{pre}$ and suffix $x \geq k_{suf}$. Moreover notice that the existence of matchings μ_1 and μ_2 in π_r implies that $\pi_{SS'} \models \pi_c$ with prefix $|\pi_S| + pre_{<_B}(\mu_i)$ and suffix $|\pi_{S'}| + suf_{<_B}(\mu_i)$, for both $i = 1$ and $i = 2$. Now there are three cases.

- If $p \leq |\pi_S| + pre_{<_B}(\mu_1)$ then $p < k_{pre}$, which is a contradiction.
- If $x \leq |\pi_{S'}| + suf_{<_B}(\mu_2)$ then $x < k_{suf}$, which is a contradiction.
- Otherwise the matching into $\pi_{SS'}$ with prefix p and suffix x corresponds to some matching into π_r. More precisely there exists a matching $\mu \in H_B$ into π_r such that $p = |\pi_S| + pre_{<_B}(\mu)$ and $x = |\pi_{S'}| + suf_{<_B}(\mu)$. Then either $p < k_{pre}$ or $x < k_{suf}$. This is also a contradiction.

This proves that $s \not\models \pi$ and therefore $(\Pi, <_B) \nvdash_{\mathrm{d}} \pi$, thus concluding the proof of the claim. □

Since one of the cases considered in Claims 3, 4, 5, 6 and 7 has to occur, the above results define a procedure for checking whether $\Pi \vdash_{\mathrm{d}} \pi$.

The cost of this procedure is dominated by the cost of the following computation:

- compute sets B and H_B;
- check $|B| \geq k_{pre} + k_{suf} + 1$;
- in the case $|B| < k_{pre} + k_{suf} + 1$, for each total order $<_B$ on B
  - compute $pre_{<_B}(\mu)$ and $suf_{<_B}(\mu)$ for all $\mu \in H_B$;
  - check conditions of Claims 4, 5, 6 and 7.

We have already observed that sets B and H_B, as well as $pre_{<_B}(\mu)$ and $suf_{<_B}(\mu)$, for all $\mu \in H_B$, can be computed in polynomial time in the sizes of Π and π. Moreover, conditions of Claims 4, 5, 6 can be checked in time polynomial in the sizes of Π and π. This is also true for Claim 7 if k_{pre} and k_{suf} are fixed (which is our case, as π is fixed). This implies that the cost of the above computation is $O(p(|\Pi|+|\pi|))$, for some polynomial p depending only on the number of leading and trailing wildcards in π, and concludes the proof of Theorem 1. □

Remark 1 The proof shows that the problem of checking whether $\Pi \vdash_{\mathrm{d}} \pi$, with π being a part of the input, is in PTIME if the number of leading and trailing wildcards in π is fixed. That is, for every fixed k, ℓ, the problem of checking whether $\Pi \vdash_{\mathrm{d}} (\_)^k a \pi b (\_)^\ell$, with $a, b \in \mathcal{L}$ and $\pi \in \mathcal{L}_{\_}^*$, is solvable in polynomial time when both Π and π are inputs.

Acknowledgment Supported by EPSRC grants E005039 and F028288 and by the FET-Open FOX project, grant agreement 233599. We thank Pablo Barceló and Gonzalo Navarro for their comments.

References

[1] S. Abiteboul, L. Segoufin, V. Vianu. Representing and querying XML with incomplete information. *ACM TODS*, 31 (2006), 208–254.

[2] P. Barceló, L. Libkin, A. Poggi, C. Sirangelo. XML with incomplete information: models, properties, and query answering. In *PODS'09*, pages 237–246.

[3] H. Björklund, W. Martens, T. Schwentick. Conjunctive query containment over trees. *DBPL'07*, pages 66–80. Full version to appear in *JCSS*.

[4] D. Maier. The complexity of some problems on subsequences and supersequences. *J. ACM* 25(2): 322–336 (1978).

[5] Document Object Model (DOM). W3C Recommendation, April 2004. http://www.w3.org/TR/DOM-Level-3-Core.