Extended Dualization: Application to Maximal Pattern Mining

Lhouari Nourine, Jean-Marc Petit

Extended Dualization: Application to Maximal Pattern Mining

Lhouari Nourine*and Jean-Marc Petit†

Abstract

The dualization in arbitrary posets is a well-studied problem in combinatorial enumeration and is a crucial step in many applications in logics, databases, artificial intelligence and pattern mining.
The objective of this paper is to study *reductions* of the dualization problem on arbitrary posets to the dualization problem on boolean lattices, for which output quasi-polynomial time algorithms exist. Quasi-polynomial time algorithms are algorithms which run in $n^{o(logn)}$ where n is the size of the input and output. We introduce *convex embedding* and *poset reflection* as key notions to characterize such reductions. As a consequence, we identify posets, which are not boolean lattices, for which the dualization problem remains in quasi-polynomial time and propose a classification of posets with respect to dualization.
From these results, we study how they can be applied to maximal pattern mining problems. We deduce a new classification of pattern mining problems and we point out how known problems involving sequences and conjunctive queries patterns, fit into this classification. Finally, we explain how to adapt the seminal Dualize & Advance algorithm to deal with such patterns.
As far as we know, this is the first contribution to explicit non-trivial reductions for studying the hardness of maximal pattern mining problems and to extend the Dualize & Advance algorithm for complex patterns.

Keywords : Hypergraph Dualization, Enumeration algorithms, Patterns mining

*Clermont-Université, Université Blaise Pascal, LIMOS, CNRS, France (nourine@isima.fr)
†Université de Lyon, CNRS, INSA-Lyon, LIRIS, France(jean-marc.petit@insa-lyon.fr)

1 Introduction

The dualization in arbitrary finite[1] partially ordered sets (poset for short) is well-studied in combinatorial enumeration such as minimal transversals of a hypergraph, the blocker of a clutter, minimal dominating sets and maximal cliques of a graph. The dualization problem is the following: Given a compact representation of a poset P and an antichain $\mathcal{B}^+$ of P, find another antichain $\mathcal{B}^-$ of P such that the union of the ideal induced by $\mathcal{B}^+$ and the filter induced by $\mathcal{B}^-$ is exactly P.[2] The dualization problem has been popularized largely through the work on artificial intelligence and pattern mining [9, 12, 21, 10, 18, 19], where P is a boolean lattice and $\mathcal{B}^+$ is given by a monotocally decreasing predicate. This link has been done through the well known DUALIZE & ADVANCE algorithm [21, 15, 3, 4, 23, 22]. Many authors have investigated the existence of an output-polynomial time algorithm for listing without duplications the antichain $\mathcal{B}^-$. An output-polynomial algorithm is an algorithm whose running time is bounded by a polynomial depending on the sum of the sizes of the input and output. The existence of an output-polynomial algorithm for the enumeration of minimal transversals of a hypergraph is a widely open question and is closely related to many data mining problems [11].

In this paper, we are interested in characterizing posets for which the dualization is equivalent to the enumeration of minimal transversals of a hypergraph. The strategy is based on *reductions* of the dualization problem on arbitrary posets to the dualization problem on boolean lattices. On posets, the dualization problem can be stated as follows:

DualizeOnPoset
Input: A representation of a poset $(P, \leq)$, B^+ an antichain of P.
Output: B^- such that (B^+, B^-) are dual sets.

On boolean lattices, it is stated as follows:

DualizeOnSet
Input: A finite set E, B^+ an antichain of $\mathcal{P}(E)$ (the powerset of E).
Output: B^- such that (B^+, B^-) are dual sets in $\mathcal{P}(E)$.

The complexity of **DualizeOnSet** is known to be quasi-polynomial time while the complexity of **DualizeOnPoset** is still open in most posets (for example, lattice)

[1] It also works for infinite partially ordered sets that are well ordered, i.e. all antichains are finite.

[2] $\mathcal{B}^+$ and $\mathcal{B}^-$ are dual sets, also known as blocker and anti-blocker or positive and negative borders.

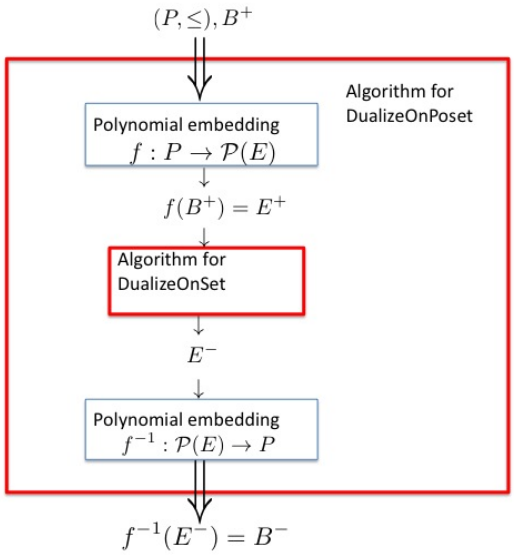

Figure 1: Reduction from **DualizeOnPoset** to **DualizeOnSet**

[11]. In this setting, we are interested in studying the *reduction* from **DualizeOnPoset** to **DualizeOnSet**, i.e. under which conditions **DualizeOnSet** is polynomially equivalent to **DualizeOnPoset**. Notice that reductions for the hardness of enumeration problems are not well established as for decision problems. In this paper, we consider only polynomial time reductions, inspired from classic polynomial reductions of decision problems (cf. Figure 1).

Contribution on dualization We introduce *convex embedding* and *poset reflection* as key notions to characterize such reductions. As a consequence, we identify posets, which are not boolean lattices, for which the dualization problem remains quasi-polynomial time (cf. Figure 1) and propose a classification of posets with respect to dualization.

From these results, we study how they can be applied to maximal (or more specific) pattern mining problems. Mining interesting patterns in databases has been extensively studied in the data mining community over the last twenty years, from association rules and frequent itemset mining to frequent graph mining or functional dependency inference to mention a few. For studying their complexity, the underlying dualization problem has been identified as the main bottleneck in [21, 15, 23].

Roughly speaking, for a partial order $(P, \leq)$ (representing patterns) and some monotonically decreasing predicate Q over P, the dualization consists in identifying all maximal elements of P verifying Q from all minimal elements of P not verifying Q, and vice versa.

The seminal work of Mannila and Toivonen [21] proposes a general framework, especially they classify pattern mining problems that are (isomorphically) equivalent to frequent itemset mining (FIM). Nevertheless, the isomorphism requirement is too restrictive for many "complex" patterns such as sequences, episodes or graphs to mention a few. The ambition of this paper is to take into account such complex patterns and to propose a new framework for studying their complexity. From a practical point of view, the idea is to be able to re-use as much as possible the myriad of techniques and algorithms developed for FIM to such complex patterns.

Contribution on maximal pattern mining From the contributions on dualization, we deduce a new classification of pattern mining problems. We point out how known problems involving sequences and conjunctive queries patterns, fit into this classification. Finally, we explain how to adapt the seminal Dualize & Advance [15] algorithm to deal with such complex patterns.

As far as we know, this is the first contribution to explicit non-trivial reductions for studying the hardness of maximal pattern mining problems and to extend the Dualize & Advance algorithm for complex patterns.

This paper is a consolidation and an extension of two papers published in the proceedings of the conference ECAI 2012 [23] and the workshop ISIP 2014 [24].

2 Preliminaries

We briefly recall definitions on partial orders, embeddings, borders and pattern mining problems [6, 23].

A partial order is a binary relation $\leq$ over a set P which is reflexive, antisymmetric, and transitive. Let x, y be elements of P, if $x \leq y$ or $y \leq x$, then x and y are comparable, otherwise they are incomparable. A partial order under which every pair of elements is comparable is called a *chain*. A subset of a poset in which no two distinct elements are comparable is called an *antichain*. We say that y *covers* x if whenever $x \leq z \leq y$ then $z = x$ or $y = z$; we denote by $\prec$ the covering relation. For $S \subseteq P$, $\downarrow S$ (resp. $\uparrow S$) is the downward (resp. upward) closed set of S under

the relation $\leq$ (i.e. $\downarrow S$ is an ideal and $\uparrow S$ a filter of $(P, \leq)$)[3]. In case of ambiguity, $\downarrow S$ (resp. $\uparrow S$) will be denoted by $\downarrow^{\leq} S$ (resp. $\uparrow^{\leq} S$). A subset $S \subseteq P$ is *convex* in P if for all $x, y, z \in P$, $x, y \in S$ and $x \leq z \leq y$ imply $z \in S$. We denote by $Max_{\leq}(S)$ (resp. $Min_{\leq}(S)$) the maximal (resp. minimal) elements of S with respect to $\leq$. When $\leq$ is clear from context, $(P, \leq)$ (resp. $Max_{\leq}(S)$ and $Min_{\leq}(S)$) will be denoted by P (resp. $Max(S)$, $Min(S)$).

Let $(P, \leq_P)$ and $(Q, \leq_Q)$ be posets and $f : P \to Q$ a mapping. f is an *embedding* if for all $x, y \in P$, $x \leq_P y$ iff $f(x) \leq_Q f(y)$. The mapping f is an *isomorphism* if f is a bijective embedding. In this case P and Q are said to be *isomorphic*. f is a *convex embedding* if f is an *embedding* and $f(P)$ is convex in $(Q, \leq_Q)$. Since f is injective, there exists another mapping $g : f(P) \to P$ such that $g \circ f = Id$, the identity function. A *reflection*[4] of a poset $(P, \leq)$ is a poset $(P, \leq')$ on the same ground set P such that for all $x, y \in P$, $x \leq' y \Rightarrow x \leq y$, i.e. a reflection preserves incomparibilities only.

Two antichains $(\mathcal{B}^+, \mathcal{B}^-)$ of P are said to be *dual* if $\downarrow \mathcal{B}^+ \cup \uparrow \mathcal{B}^- = P$ and $\downarrow \mathcal{B}^+ \cap \uparrow \mathcal{B}^- = \emptyset$. The relationship between these dual sets is known as the dualization, i.e. given $\mathcal{B}^+$, compute $\mathcal{B}^-$ (or inversely). In the sequel, $(\mathcal{B}^+, \mathcal{B}^-)$ will be referred to as a *border*.

Let $f : P \to Q$ be a mapping and $(\mathcal{B}^+, \mathcal{B}^-)$ a border in P. The border $(\mathcal{E}^+, \mathcal{E}^-)$ in Q is an *extension* of $(\mathcal{B}^+, \mathcal{B}^-)$ with respect to f, if $f(\mathcal{B}^+) \subseteq \mathcal{E}^+$ and $f(\mathcal{B}^-) \subseteq \mathcal{E}^-$. The extension $(\mathcal{E}^+, \mathcal{E}^-)$ is said to be a *polynomial extension* of $(\mathcal{B}^+, \mathcal{B}^-)$ if $|\mathcal{E}^+| + |\mathcal{E}^-|$ is polynomial in $|\mathcal{B}^+| + |\mathcal{B}^-|$.

The intuition of the reduction of enumeration problems used in this paper is based on finding a mapping between posets such that borders are polynomially preserved, i.e. every border has a polynomial extension.

2.1 Pattern mining problem

We recall the framework of Mannila and Toivonen [21]: Given a database $\mathbf{d}$, a description $\mathcal{L}$ of the language of patterns $\mathcal{L}^*$ derivable from $\mathcal{L}$, and a predicate Q for evaluating whether a pattern $\varphi \in \mathcal{L}^*$ is "interesting" in $\mathbf{d}$. We assume that the size of $\mathcal{L}^*$ is exponential in the size of the description of $\mathcal{L}$, otherwise the size of $\mathcal{L}^*$ is polynomial and the mining problem is obviously polynomial by brute force enumeration. For instance, the description of the language for frequent itemset

[3]Here, an ideal or filter does not have to be upward directed

[4]A reflection of a poset P is a new poset obtained from P by deleting a subset of its comparabilities

mining is given by the set of items, the set of itemsets being exponential.

The discovery task is to find the theory of $\mathbf{d}$ with respect to $\mathcal{L}$ and Q, i.e. the set $Th(\mathcal{L}, \mathbf{d}, Q) = \{\varphi \in \mathcal{L}^* | Q(\mathbf{d}, \varphi) \text{ holds}\}$. We assume that the set of patterns $\mathcal{L}^*$ is structured with a partial order $\preceq$ and the predicate Q is monotonically decreasing wrt $\preceq$, i.e. for all $\theta, \varphi \in \mathcal{L}^*, \varphi \preceq \theta,\ Q(\mathbf{d}, \theta) \Rightarrow Q(\mathbf{d}, \varphi)$. The antichain $\mathcal{B}^+ \subseteq Th(\mathcal{L}, \mathbf{d}, Q)$ such that $\downarrow \mathcal{B}^+ = Th(\mathcal{L}, \mathbf{d}, Q)$ is known as the positive border of $Th(\mathcal{L}, \mathbf{d}, Q)$. The antichain $\mathcal{B}^-$ such that $\uparrow \mathcal{B}^- = (\mathcal{L}^* \setminus Th(\mathcal{L}, \mathbf{d}, Q))$ is known as the negative border of $Th(\mathcal{L}, \mathbf{d}, Q)$. $(\mathcal{B}^+, \mathcal{B}^-)$ is a border of $(\mathcal{L}^*, \preceq)$.

2.1.1 Problem statement

In the data mining context, the problem statement is quite different since the antichain $\mathcal{B}^+$ is known implicitly (given by a monotically decreasing predicate) and it can be defined as follows:

> **Pattern mining problem (PMP)**
> *Given a pattern mining problem $(\mathcal{L}, \preceq, \mathbf{d}, Q)$, enumerate the positive border of $Th(\mathcal{L}, \mathbf{d}, Q)$.*

We now make explicit some usual assumptions, valid for almost every pattern mining problem:

- The predicate Q is computable in polynomial time in the size of $\mathbf{d}$,
- Given two patterns $\theta, \phi \in \mathcal{L}^*$, checking $\theta \preceq \phi$ (or $\phi \preceq \theta$) is computable in polynomial time.

In this setting, the **PMP** problem can be simplified by ignoring both the database $\mathbf{d}$ and the predicate Q. In the sequel, we shall use $(\mathcal{L}^*, \preceq)$ instead of $(\mathcal{L}, \preceq, \mathbf{d}, Q)$.

An enumeration algorithm is said to be incremental quasi-polynomial time, when it enumerates one by one the positive border such that the time it takes to list another one is quasi-polynomial in the size of the input and the previous output, i.e. $n^{o(logn)}$ where n is the sum of the sizes of the input and the output. It is worth noting that whenever $(\mathcal{L}^*, \preceq)$ is isomorphic to $(\mathcal{P}(E), \subseteq)$ for some set E, **PMP** can be solved in incremental quasi-polynomial time [15] (known as DUALIZE & ADVANCE algorithm) in the size of the positive and negative borders.

In the next section, we shall consider dualization on arbitrary posets; we will come back to maximal pattern mining problems in Section 4.

3 Classification of posets with respect to dualization

In this section, we describe two properties of posets that lead us to have polynomial time reductions to **DualizeOnSet**. First we show that a convex embedding from a poset $(\mathcal{L}^*, \preceq)$ to $\mathcal{P}(E)$ for some set E is sufficient to re-use algorithms of **DualizeOnSet**. Second, we show that the convex embedding is not a necessary condition and introduce the reflection of a poset $(\mathcal{L}^*, \preceq)$ to obtain a new poset $(\mathcal{L}^*, \preceq')$ in which there is a convex embedding. Indeed, a reflection of a poset $(\mathcal{L}^*, \preceq)$ corresponds to an embedding which preserves incomparabilities only, i.e. some comparabilities could be lost. As a consequence, extra-elements may appear to the dualization. Intuitively, whenever the number of extra-elements is bounded by a polynomial, the dualization can be reduced to **DualizeOnSet**. To do so, we ask the following questions:

Given a poset of patterns $(\mathcal{L}^*, \preceq)$*,*

- *Does there exist a convex embedding of* $(\mathcal{L}^*, \preceq)$ *into* $(\mathcal{P}(E), \subseteq)$ *for some finite set* E*? If not,*
- *Does there exist a reflection* $(\mathcal{L}^*, \preceq')$ *of* $(\mathcal{L}^*, \preceq)$ *such that there exists a convex embedding of* $(\mathcal{L}^*, \preceq')$ *into* $(\mathcal{P}(E), \subseteq)$ *for some finite set* E*?*

These two questions and their associated computational costs allow to come up with new classes of posets with respect to dualization. Figure 2 gives an illustration of the proposed framework. In the sequel, we first give some definitions of pattern structures based on convex embedding. Then, we propose poset reflection as a key notion to reach the convex embedding constraint.

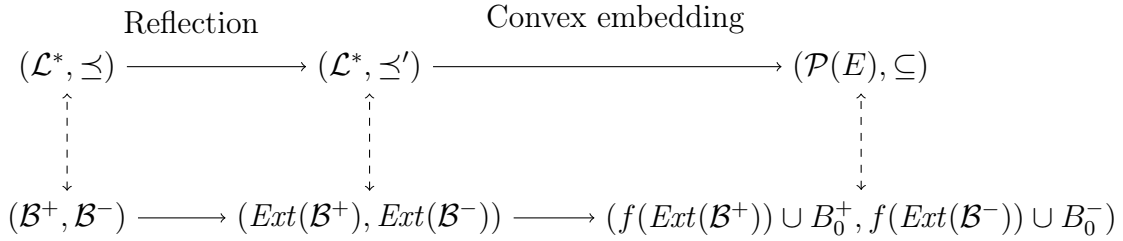

Figure 2: Overview of the main transformations

3.1 Convex embedding

Wild [26] studied some sufficient and necessary conditions for posets to have a cover preserving embedding into boolean lattices, but none of them can be re-used in this

paper since the input poset is given implicitly in our case.

First, let us recall that any poset has an embedding into a boolean lattice.

Proposition 1 *[6] For any poset $(\mathcal{L}^*, \preceq)$, there exists an embedding from $(\mathcal{L}^*, \preceq)$ to $(\mathcal{P}(E), \subseteq)$, for some finite set E.*

It follows that any poset has a set representation but obviously the dualization on $(\mathcal{L}^*, \preceq)$ may be more difficult than the dualization on $(\mathcal{P}(E), \subseteq)$ [23]. We define the set of posets representable as sets ($\mathcal{RAS}$ for short) as follows:

Definition 1 $(\mathcal{L}^*, \preceq) \in \mathcal{RAS}$ *iff* $(\mathcal{L}^*, \preceq)$ *and* $(\mathcal{P}(E), \subseteq)$ *are isomorphic, for some finite set E.*

This class of posets gathers together many patterns such as frequent itemsets (FIM) [2], functional dependencies (FD) [20], inclusion dependencies (IND) [8]. This class is known as *representation as sets* class of pattern mining problems defined in [21] (also known as *strong duality* [25]).

Nevertheless, requirements to be in $\mathcal{RAS}$ are restrictive, since the poset must be isomorphic to a boolean lattice, and therefore its size has to be equal to 2^n where $n = |E|$. One may remark that $(\mathcal{L}^*, \preceq)$ is trivially convex in $(\mathcal{P}(E), \subseteq)$, suggesting to relax the bijective constraint of $\mathcal{RAS}$ without losing the convexity on the set representation. Hence, we extend $\mathcal{RAS}$ to a new class, called $\mathcal{XRAS}$, for *conve*$\mathcal{X}$ $\mathcal{RAS}$.

Definition 2 $(\mathcal{L}^*, \preceq) \in \mathcal{XRAS}$ *iff there exists a* convex embedding *from* $(\mathcal{L}^*, \preceq)$ *to* $(\mathcal{P}(E), \subseteq)$*, for some finite set E.*

The idea is still to require an isomorphism but just between the poset of patterns and some subset of $\mathcal{P}(E)$, instead of the entire set $\mathcal{P}(E)$ (see Figure 3). Moreover, the subset of $\mathcal{P}(E)$ has to be convex, which was true but implicit in the definition of $\mathcal{RAS}$. Note also that f is injective since f is an embedding. The following proposition points out a simple yet important characterization of $\mathcal{XRAS}$ problems. An illustration is given in Figure 3.

Proposition 2 $(\mathcal{L}^*, \preceq) \in \mathcal{XRAS}$ *iff there exist two antichains B_0^+ and B_0^- of $\mathcal{P}(E)$ such that $\downarrow B_0^+ \cup \uparrow B_0^- = \emptyset$ and $(\mathcal{L}^*, \preceq)$ is isomorphic to $(\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-), \subseteq)$.*

Proof: Let f be a convex embedding from $(\mathcal{L}^*, \preceq)$ to $(\mathcal{P}(E), \subseteq)$ and $\mathcal{F} = f(\mathcal{L}^*)$. Let $B_0^+ = Max(\downarrow \mathcal{F} \setminus \mathcal{F})$ and $B_0^- = Min(\mathcal{P}(E) \setminus \downarrow \mathcal{F})$.

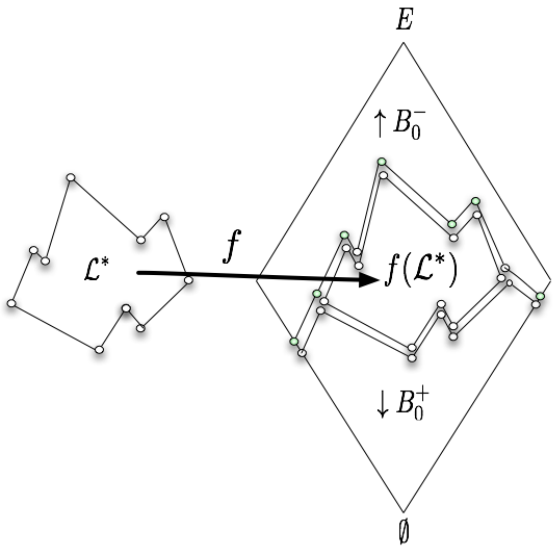

Figure 3: The class $\mathcal{XRAS}$

We show that $\downarrow B_0^+$, $\uparrow B_0^-$ and $\mathcal{F}$ is a partition of $\mathcal{P}(E)$.
Clearly $\downarrow B_0^+ \cup \uparrow B_0^- \cup \mathcal{F} = \mathcal{P}(E)$, $\downarrow B_0^+ \cap \uparrow B_0^- = \emptyset$ and $\mathcal{F} \cap \uparrow B_0^- = \emptyset$. Now we show that $\mathcal{F} \cap \downarrow B_0^+ = \emptyset$. Suppose that exists $F \in \mathcal{F} \cap \downarrow B_0^+$. The fact that $B_0^+ \cap \mathcal{F} = \emptyset$ implies that exist $X \in B_0^+$ and $F' \in \mathcal{F}$ with $F \subset X \subset F'$. This contradicts the fact that $\mathcal{F}$ is a convex set in $\mathcal{P}(E)$ since $X \notin \mathcal{F}$.

The other direction holds since $\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-)$ is a convex set of $\mathcal{P}(E)$. □

The proof of Proposition 2 gives a particular pair of antichains (B_0^+, B_0^-), but many such pairs exist to achieve the convex embedding. We also note that the size of the sets B_0^+ and B_0^- could be exponential in the size of the description $\mathcal{L}$.

The next definition introduces computationally *efficient* version of problems of $\mathcal{XRAS}$, called $\mathcal{EXRAS}$.

Definition 3 $(\mathcal{L}^*, \preceq) \in \mathcal{EXRAS}$ *if* $(\mathcal{L}^*, \preceq) \in \mathcal{XRAS}$ *and there is a pair* (B_0^+, B_0^-) *of antichains such that* $|B_0^+ \cup B_0^-|$ *is polynomial in the size of the description* $\mathcal{L}$.

The following proposition points out that a polynomial extension of any border of $(\mathcal{L}^*, \preceq)$ exists if $(\mathcal{L}^*, \preceq) \in \mathcal{EXRAS}$.

Proposition 3 *Let $(\mathcal{L}^*, \preceq) \in \mathcal{EXRAS}$ with $f : \mathcal{L}^* \to \mathcal{P}(E)$ a convex embedding, for some set E, and (B_0^+, B_0^-) two antichains such that $(\mathcal{L}^*, \preceq)$ is isomorphic to $(\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-), \subseteq)$ and $|B_0^+ \cup B_0^-|$ is polynomial in the size of $\mathcal{L}$.*
Then for every border $(\mathcal{B}^+, \mathcal{B}^-)$ of $(\mathcal{L}^*, \preceq)$, $(Max(B_0^+ \cup f(\mathcal{B}^+)), Min(B_0^- \cup f(\mathcal{B}^-)))$ is a polynomial extension of $(\mathcal{B}^+, \mathcal{B}^-)$ in $(\mathcal{P}(E), \subseteq)$.

Proof: Let $(\mathcal{B}^+, \mathcal{B}^-)$ be a border of $(\mathcal{L}^*, \preceq)$. By definition of $\mathcal{EXRAS}$, we have $(\mathcal{B}^+, \mathcal{B}^-)$ as a border of $(\mathcal{L}^*, \preceq)$ iff $(f(\mathcal{B}^+), f(\mathcal{B}^-))$ as a border of $(\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-), \subseteq)$ since $(\mathcal{L}^*, \preceq)$ is isomorphic to $(\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-), \subseteq)$.

We show that $(Max(B_0^+ \cup f(\mathcal{B}^+)), Min(B_0^- \cup f(\mathcal{B}^-)))$ is a border of $\mathcal{P}(E)$. First, we show that $\downarrow Max(B_0^+ \cup f(\mathcal{B}^+) \cap \uparrow Min(B_0^- \cup f(\mathcal{B}^-)) = \emptyset$, or equivalentely $(\downarrow B_0^+ \cup \downarrow f(\mathcal{B}^+)) \cap (\uparrow B_0^- \cup \uparrow f(\mathcal{B}^-)) = \emptyset$. By Proposition 2 we have $\downarrow B_0^+ \cap \uparrow \mathcal{B}_0^- = \emptyset$. Moreover, $\downarrow f(\mathcal{B}^+) \cap \uparrow f(\mathcal{B}^-) = \emptyset$ since f is a bijective embedding of $(\mathcal{L}^*, \preceq)$ into $(\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-), \subseteq)$ and $(\mathcal{B}^+, \mathcal{B}^-)$ a border of $(\mathcal{L}^*, \preceq)$. Finally, we have $\downarrow f(\mathcal{B}^+) \cap \uparrow \mathcal{B}^- = \uparrow f(\mathcal{B}^-) \cap \downarrow \mathcal{B}^+ = \emptyset$ since $\downarrow f(\mathcal{B}^+) \cup \uparrow f(\mathcal{B}^-) \subseteq (\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-)$.

Now we show that $\downarrow Max(B_0^+ \cup f(\mathcal{B}^+) \cap \uparrow Min(B_0^- \cup f(\mathcal{B}^-)) = \mathcal{P}(E)$. Since $(f(\mathcal{B}^+), f(\mathcal{B}^-))$ as a border of $(\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-), \subseteq)$, we have $\downarrow f(\mathcal{B}^+) \cup \uparrow f(\mathcal{B}^-) = \mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-)$ and it union with $\downarrow B_0^+ \cup \uparrow B_0^-)$ gives $\mathcal{P}(E)$.

Thus $(Max(B_0^+ \cup f(\mathcal{B}^+)), Min(B_0^- \cup f(\mathcal{B}^-)))$ is a border in $\mathcal{P}(E)$ which is a polynomial extension of $(\mathcal{B}^+, \mathcal{B}^-)$ since $|B_0^+ \cup B_0^-|$ is polynomial in the size of the description $\mathcal{L}$.

□

3.2 Polynomial reflection of posets

We now consider posets that are not in $\mathcal{XRAS}$. Our idea is to transform the initial poset to a new poset over the same ground set, in order to get a convex embedding. As a consequence, two natural questions arise:

(1) For a given poset $(\mathcal{L}^*, \preceq)$, does there exist a "polynomial reflection" $(\mathcal{L}^*, \preceq')$ such that $(\mathcal{L}^*, \preceq')$ belongs to $\mathcal{EXRAS}$?

(2) How to quantify the "loss of comparabilities" induced by a reflection?

In the sequel, we study poset reflection to give answers to the previous questions. Since the reflection of a poset induces the lost of some comparabilities in the original poset, we have to recover them efficiently.

Before that, we consider different examples of posets over sequences.

3.2.1 Examples with different posets of sequences

Let us consider sequences with or without wildcard (see e.g. [3]).

Let Σ be a totally ordered alphabet which contains the wildcard $\star$ as a minimal letter. A sequence is an element of Σ^* and a rigid sequence is of the form $P = P[1]\dots P[m] \in \Sigma^*$ such that $P[1] \neq \star$ and $P[m] \neq \star$. Let Σ_R^* be the set of rigid sequences and Σ^* the set of sequences. We denote by Σ^n (resp. Σ_R^n) the set of all (resp. rigid) sequences in Σ^* (resp. Σ_R^n) of size at most n. We define different partial orders over Σ_R^* and Σ^*:

- $(\Sigma^*, \leq_s)$: Let $P[1..m], Q[1..n] \in \Sigma^*$ with $m \leq n$. P is subsequence of Q, denoted $P \leq_s Q$, if P can be obtained from Q by deleting $n - m$ letters. Formally, $P \leq_s Q$ if there exist integers $i_1 < \dots < i_m$ in $[1..n]$ such that $P[j] = Q[i_j]$ for all $j \in [1..m]$.

  P is a prefix of Q, denoted by $P \leq_p Q$, if $P[1..m] = Q[1..m]$.

- $(\Sigma_R^*, \sqsubseteq)$: Let $P[1..m], Q[1..n] \in \Sigma_R^*$: P *occurs* in Q at position $j \in [1..n]$, denoted by $P \sqsubseteq_j Q$, if for every $i \in [1..m]$, either $P[i] = Q[j + i - 1]$ or $P[i] = \star$. We say that P *occurs as prefix* in Q if $P \sqsubseteq_1 Q$. We write $P \sqsubseteq Q$ if there exists $j \in [1..m]$ such that $P \sqsubseteq_j Q$.

It is worth noticing, that $(\Sigma^*, \leq_s), (\Sigma^*, \leq_p), (\Sigma_R^*, \sqsubseteq)$ and $(\Sigma_R^*, \sqsubseteq_1)$ are partial orders. Moreover $(\Sigma_R^*, \sqsubseteq_1)$ (resp. $(\Sigma^*, \leq_p)$) is a reflection of $(\Sigma_R^*, \sqsubseteq)$ (resp. $(\Sigma^*, \leq_s)$), since $P \sqsubseteq_1 Q$ (resp. $P \leq_p Q$) implies $P \sqsubseteq Q$ (resp. $P \leq_s Q$).

The following example shows that the posets $(\Sigma^*, \leq_s)$ and $(\Sigma_R^*, \sqsubseteq)$ cannot have a convex embedding in the boolean lattice $\mathcal{P}(E)$ for some set E. Indeed, consider the set $A = \{a, b, ab, ba\} \subseteq \Sigma^* \cap \Sigma_R^*$. Then for any embedding f of $(\Sigma^*, \leq_s)$ into $(\mathcal{P}(E), \subseteq)$, there exists a set $X \subseteq E$ such that $f(a) \subseteq X \subseteq f(ab)$, $f(b) \subseteq X \subseteq f(ba)$ and X is not an image of f. Otherwise $(\mathcal{P}(E), \subseteq)$ will not be closed under intersection, since $f(ab) \cap f(ba)$ must contain $f(a)$ and $f(b)$ and there is no word $w \in \Sigma^*$ such that $a \leq_s w \leq_s ab$ and $f(w) = f(ab) \cap f(ba)$. Thus the set of all images of f is not a convex set in $(\mathcal{P}(E), \subseteq)$. The same reasoning applies for $(\Sigma_R^*, \sqsubseteq)$.

Now consider the posets $(\Sigma_R^*, \sqsubseteq_1)$ and $(\Sigma^*, \leq_p)$ which are refections of $(\Sigma_R^*, \sqsubseteq)$ and $(\Sigma^*, \leq_s)$ respectively. In the following, we first show that $(\Sigma_R^*, \sqsubseteq_1)$ has a convex embedding and that no convex embedding exists for $(\Sigma^*, \leq_p)$ into $(\mathcal{P}(E), \subseteq)$.

(1) Let $f : \Sigma_R^* \to \mathcal{P}(E)$ be an embedding, for some finite set E [23, 3]: f associates to each letter of a sequence a pair (index, letter). For instance, let ab and ba be two patterns. Then $f(ab) = \{(1, a), (2, b)\}$ and $f(ba) = \{(1, b), (2, a)\}$. It is easy to verify

that $f(\Sigma_R^*)$ is convex in $(\mathcal{P}(E), \subseteq)$ [23]. Let us again consider $A = \{a, b, ab, ba\}$: we only have $a \sqsubseteq_1 ab, b \sqsubseteq_1 ba$, (two comparabilities are lost) allowing to reach the convexity constraint.

(2) Consider the set $A = \{a, aa, aaa\}$. A is convex in $(\Sigma^*, \leq_p)$ but its image by any embedding cannot be convex in $(\mathcal{P}(E), \subseteq)$ since $\mathcal{P}(E)$ cannot contain a convex set which is a chain of length 3.

3.2.2 Reaching convexity by poset reflection

The two previous examples point out that we have to transform the poset (through reflection) to get a chance to obtain some convex embedding. Note that for any poset there exists a reflection which has a convex embedding in a boolean lattice. But the difficulty is how to retrieve lost comparabilities?

Given a poset reflection, we define the lost successors and lost predecessors induced by a poset reflection for each element of the poset.

Definition 4 *Let $(P, \leq')$ be a reflection of a poset $(P, \leq)$ and $x \in P$. The lost predecessors of x in the reflection of $(P, \leq)$ to $(P, \leq')$, denoted by $LostPred(x)$, are defined by:*
$LostPred(x) = Max_{\leq'}\{y \in P | y \leq x, y \not<' x\}$. Similarly, the lost successors are defined by: $LostSucc(x) = Min_{\leq'}\{y \in P | x \leq y, x \not<' y\}$.

By extension, we note $LostPred(X) = \bigcup_{x \in X} LostPred(x)$ (resp. $LostSucc(X)$) for $X \subseteq P$.

Example 1 *Let us consider the previous reflection $(\Sigma_R^n, \sqsubseteq_1)$ of $(\Sigma_R^n, \sqsubseteq)$. Let $S \in \Sigma_R^n$. We have $LostPred(S) = Max_{\sqsubseteq_1}\{S[i..|S|] \mid 1 \leq i \leq |S|, S[i] \neq \star\}$ and $LostSucc(S) = Min_{\sqsubseteq_1}\{\underbrace{x \star \ldots \star}_{i} S \mid 0 \leq i \leq n - |S|, x \in \Sigma \setminus \{\star\}\}$. For instance with $n = 5$ and $\Sigma = \{\star, a, b\}$, $LostPred(a \star ba) = Max_{\sqsubseteq_1}\{a \star ba, ba, a\}$ and $LostSucc(ba) = Min_{\sqsubseteq_1}\{ba, aba, bba, a \star ba, b \star ba, a \star \star ba, b \star \star ba\}$.*

As shown in the following lemma, we can recover the initial poset from any reduced poset with *LostPred* and *LostSucc*.

Lemma 1 *Let $x \in P$ and $(P, \leq')$ a reflection of $(P, \leq)$. Then:*

1. $\downarrow^{\leq} x = \downarrow^{\leq'} LostPred(x)$ *and*
2. $\uparrow^{\leq} x = \uparrow^{\leq'} LostSucc(x)$.

Proof: (1) Let $y \in{\downarrow}^{\leq} x$. We have either $y \leq' x$ or $y \not<' x$. If $y \leq' x$, then $y \in{\downarrow}^{\leq'} \mathit{LostPred}(x)$ since $x \in \mathit{LostPred}(x)$. If $y \not<' x$, then there exists $z \in P$ such that $y \leq' z, z \in \mathit{LostPred}(x)$, i.e. $y \in{\downarrow}^{\leq'} \mathit{LostPred}(x)$.

Now let $y \in{\downarrow}^{\leq'} \mathit{LostPred}(x)$. Then exists $z \in \mathit{LostPred}(x)$ such that $y \leq' z$. Since f is a reflection we have $z \leq x$ and thus $y \in{\downarrow}^{\leq} x$.

The same reasoning applies for (2). □

Some remarks have to be made: First, for any poset, there always exists a reflection that has a convex embedding into a boolean lattice. It suffices to take a reflection which is an antichain, i.e. that deletes all comparabilities. In this case, the number of lost comparabilities can be exponential in the size of the description of the poset. Second, we would like to be able to recover lost comparabilities in polynomial time. This is formalized with the notion of *poly-reflection* as follows.

Definition 5 *$(\mathcal{L}^*, \preceq')$ is a* poly-reflection *of $(\mathcal{L}^*, \preceq)$ if $(\mathcal{L}^*, \preceq')$ is a reflection of $(\mathcal{L}^*, \preceq)$ and for all $x \in \mathcal{L}^*$, $\mathit{LostPred}(x)$ and $\mathit{LostSucc}(x)$ are computable in polynomial time in the size of the description $\mathcal{L}$.*

Example 2 *Continuing the previous example, for all $S \in \Sigma_R^n$, $|\mathit{LostPred}(S)|$ is polynomial in n and for all $s \sqsubseteq S$, there exists $s' \in \mathit{LostPred}(S)$ such that $s \sqsubseteq_1 s'$. Therefore, $(\Sigma_R^n, \sqsubseteq_1)$ is a poly-reflection of $(\Sigma_R^n, \sqsubseteq)$.*

Now we show the relationship between borders in a poset and its reflection. For a given border on the initial poset, we define its extension in the reduced poset to take into account lost comparabilities.

Definition 6 *Let $(\mathcal{L}^*, \preceq')$ be a poly-reflection of $(\mathcal{L}^*, \preceq)$ and $(\mathcal{B}^+, \mathcal{B}^-)$ a border of $(\mathcal{L}^*, \preceq)$. The extension of $(\mathcal{B}^+, \mathcal{B}^-)$ in $(\mathcal{L}^*, \preceq')$, denoted by $(\mathit{Ext}(\mathcal{B}^+), \mathit{Ext}(\mathcal{B}^-))$, is defined by:*

$\mathit{Ext}(\mathcal{B}^+) = \mathit{Max}_{\leq'}\{\mathit{LostPred}(x) \mid x \in \mathcal{B}^+\}$

$\mathit{Ext}(\mathcal{B}^-) = \mathit{Min}_{\leq'}\{\mathit{LostSucc}(x) \mid x \in \mathcal{B}^-\}$.

The "preservation" of borders can now be formally stated.

Proposition 4 *Let $(\mathcal{L}^*, \preceq')$ be a poly-reflection of $(\mathcal{L}^*, \preceq)$ and $(\mathcal{B}^+, \mathcal{B}^-)$ a border of $(\mathcal{L}^*, \preceq)$. Then $(\mathit{Ext}(\mathcal{B}^+), \mathit{Ext}(\mathcal{B}^-))$ is a polynomial extension of $(\mathcal{B}^+, \mathcal{B}^-)$.*

Proof: We have to show:

1. $(\mathit{Ext}(\mathcal{B}^+), \mathit{Ext}(\mathcal{B}^-))$ is a border of $(\mathcal{L}^*, \preceq')$ with $\mathcal{B}^+ \subseteq \mathit{Ext}(\mathcal{B}^+)$ and $\mathcal{B}^- \subseteq \mathit{Ext}(\mathcal{B}^-)$,

2. $|Ext(\mathcal{B}^+)| + |Ext(\mathcal{B}^-)|$ is polynomial in $|\mathcal{B}^+| + |\mathcal{B}^-|$.

(1) Any reflection preserves all incomparabilities and $x \in LostPred(x)$ (resp $x \in LostSucc(x)$) for all $x \in \mathcal{L}^*$. Since $\mathcal{B}^+$ and $\mathcal{B}^-$ are antichains in $(\mathcal{L}^*, \preceq)$, we have $\mathcal{B}^+ \subseteq Ext(\mathcal{B}^+)$ and $\mathcal{B}^- \subseteq Ext(\mathcal{B}^-)$. By definition, $Ext(\mathcal{B}^+)$ and $Ext(\mathcal{B}^-)$ are antichains in $(\mathcal{L}^*, \preceq')$.

We show that $(Ext(\mathcal{B}^+), Ext(\mathcal{B}^-))$ is a border of $(\mathcal{L}^*, \preceq')$. Since $(\mathcal{B}^+, \mathcal{B}^-)$ is a border of $(\mathcal{L}^*, \preceq)$, we have $\downarrow \mathcal{B}^+ \cup \uparrow \mathcal{B}^- = \mathcal{L}^*$ and $\downarrow \mathcal{B}^+ \cap \uparrow \mathcal{B}^- = \emptyset$. By definition of the polynomial extension, for any $x \in \mathcal{B}^+$ (resp. $x \in \mathcal{B}^-$) we have $LostPred(x) \subseteq \downarrow Ext(\mathcal{B}^+)$ (resp. $LostSucc(x) \subseteq \uparrow Ext(\mathcal{B}^-)$), and, by Lemma 1, $\downarrow^{\preceq} x = \downarrow^{\preceq'} LostPred(x)$ (resp. $\uparrow^{\preceq} x = \uparrow^{\preceq'} LostSucc(x)$). Since $\mathcal{B}^+ \subseteq Ext(\mathcal{B}^+)$ and $\mathcal{B}^- \subseteq Ext(\mathcal{B}^-)$, we have $\downarrow Ext(\mathcal{B}^+) \cup \uparrow Ext(\mathcal{B}^+) = \mathcal{L}^*$.

Now suppose that there is an element $x \in \downarrow Ext(\mathcal{B}^+) \cap \uparrow Ext(\mathcal{B}^-)$. Then there exist $y \in \mathcal{B}^+$ and $z \in \mathcal{B}^-$ $x \in \downarrow^{\preceq'} LostPred(y) \cap \uparrow^{\preceq'} LostSucc(z)$. By Lemma 1, we deduce that $x \in (\downarrow^{\preceq} y) \cap (\uparrow^{\preceq} z)$ which is a contradiction with $(\mathcal{B}^+, \mathcal{B}^-)$ a border of $(\mathcal{L}^*, \preceq)$.
(2) $|Ext(\mathcal{B}^+)| + |Ext(\mathcal{B}^-)|$ is polynomial in $|\mathcal{B}^+| + |\mathcal{B}^-|$ since $(\mathcal{L}^*, \preceq')$ is a poly-reflection of $(\mathcal{L}^*, \preceq)$ since computing $LostPred(x)$ and $LostSucc(x)$ can be done in polynomial time in the size of the description of $\mathcal{L}$. □

The notion of poly-reflection allows to define the last class of posets, called $\mathcal{EWRAS}$, meaning *Efficient weak representation as sets*. $\mathcal{EWRAS}$ is the more general class ensuring the existence of quasi-polynomial time algorithms. It combines both poly-reflection of posets and $\mathcal{EXRAS}$.

Definition 7 $(\mathcal{L}^*, \preceq) \in \mathcal{EWRAS}$ *iff there exists a poly-reflection* $(\mathcal{L}^*, \preceq')$ *of* $(\mathcal{L}^*, \preceq)$ *such that* $(\mathcal{L}^*, \preceq') \in \mathcal{EXRAS}$.

Then, this definition means that if some comparabilities can be forgotten – up to a polynomial cost to recover them – to get a new poset satisfying the condition of $\mathcal{EXRAS}$, then the dualization problem on the initial poset can be reduced to **DualizeOnSet**.

Example 3 *Continuing previous examples, we have* $(\Sigma_R^*, \sqsubseteq_1)$ *is a poly-reflection of* $(\Sigma_R^*, \sqsubseteq)$ *and* $(\Sigma_R^*, \sqsubseteq_1)$ *belongs to* $\mathcal{EXRAS}$. *Then,* $(\Sigma_R^*, \sqsubseteq)$ *belongs to* $\mathcal{EWRAS}$.

From Propositions 3 and 4 we deduce the main result concerning the $\mathcal{EWRAS}$ class.

Theorem 1 *Let $(\mathcal{L}^*, \preceq) \in \mathcal{EWRAS}$. Assume that $(\mathcal{L}^*, \preceq')$ is a poly-reflection of $(\mathcal{L}^*, \preceq)$ such that $(\mathcal{L}^*, \preceq')$ belongs to $\mathcal{EXRAS}$, i.e. $(\mathcal{L}^*, \preceq')$ isomorphic to $\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-)$ for some antichains $B_0^+, B_0^- \subseteq \mathcal{P}(E)$. Then, for every border $(\mathcal{B}^+, \mathcal{B}^-)$ of $(\mathcal{L}^*, \preceq)$ the pair $(Max(B_0^+ \cup f(Ext(\mathcal{B}^+))), Min(B_0^- \cup f(Ext(\mathcal{B}^-))))$ is a border in $(\mathcal{P}(E), \subseteq)$ and it is a polynomial extension of $(\mathcal{B}^+, \mathcal{B}^-)$*

The following corollary gives the relationship between all the classes introduced so far.

Corollary 1 *Let $(\mathcal{L}^*, \preceq')$ be a poly-reflection of $(\mathcal{L}^*, \preceq)$ such that $(\mathcal{L}^*, \preceq')$ belongs to $\mathcal{XRAS}$, i.e. there exists two antichains B_0^+ and B_0^- of $\mathcal{P}(E)$ such that $(\mathcal{L}^*, \preceq')$ isomorphic to $\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-)$. We have:*

1. *$(\mathcal{L}^*, \preceq) \in \mathcal{RAS}$ if $(\mathcal{L}^*, \preceq) = (\mathcal{L}^*, \preceq')$ and $B_0^+ = B_0^- = \emptyset$.*
2. *$(\mathcal{L}^*, \preceq) \in \mathcal{XRAS}$ if $(\mathcal{L}^*, \preceq) = (\mathcal{L}^*, \preceq')$.*
3. *$(\mathcal{L}^*, \preceq) \in \mathcal{EXRAS}$ if $(\mathcal{L}^*, \preceq) = (\mathcal{L}^*, \preceq')$ and the size of $B_0^+ \cup B_0^-$ is polynomial in the size of the description $\mathcal{L}$.*
4. *$(\mathcal{L}^*, \preceq) \in \mathcal{EWRAS}$ if the size of $B_0^+ \cup B_0^-$ is polynomial in the size of the description $\mathcal{L}$.*

To sum up, Figure 4 gives inclusion between classes of posets introduced in this paper.

In the rest of this section, we point out a surprising result: the dualization problem on sequences is equivalent to the dualization problem on sets.

3.3 DualizeOnSeq is equivalent to DualizeOnSet

Recall that we consider rigid sequences only. The dualization problem can be stated as follows:
DualizeOnSeq
Input: Let $(\Sigma_R^n, \sqsubseteq)$ the poset of rigid sequences, where Σ a totally ordered alphabet with a minimal element $\star$, n a positive integer and B^+ a positive border of Σ_R^n.
output: B^- such that (B^+, B^-) is a border of Σ_R^n.

We have shown in the previous section that **DualizeOnSeq** is at most as hard as **DualizeOnSet** (see also [23]). In the sequel, we point out that **DualizeOnSet**

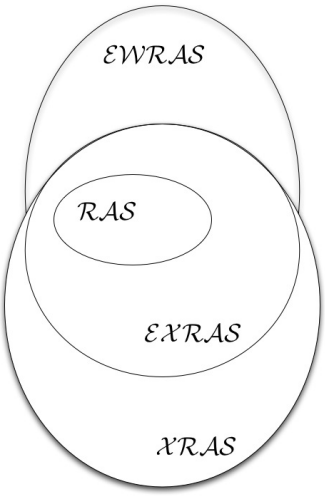

Figure 4: Posets classification with respect to the dualization problem

is at most as hard as **DualizeOnSeq**, and therefore the two problems are polynomially equivalent. Indeed, we show that **DualizeOnSet** is a particular case of **DualizeOnSeq**.

Let $\Sigma = \{1, 2, ..., n, \star\}$ be an ordered alphabet (i.e. $\star < 1 < 2... < n$) and $S \in \Sigma^n$. The sequence S is said to be an *ordered sequence* if for every $i, j \in [1..n]$ such that $i < j$, $S[i] \neq \star$ and $S[j] \neq \star$ we have $S[j] - S[i] = j - i$. We denote $\Sigma_O^n \subseteq \Sigma_R^n$ the set of all ordered sequences of size at most n. For example, the sequence $2 \star \star 5$ is an ordered sequence but $2 \star 5$ is not.

We define the set of forbidden sequences $B_0^- = \{i *^k j \mid i, j \in \Sigma, i \geq j, k \in [0..n-2]\} \cup \{i *^k j \mid i, j \in \Sigma, i < j, k \in [0..n-2], k \neq j - i - 1\}$. For example for $\Sigma = \{1, 2, 3, *\}$ and $n = 3$, we have $B_0^- = \{11, 1*1, 22, 2*2, 33, 3*3, 21, 2*1, 31, 3*1, 32, 3*2\} \cup \{13, 1*2, 2*3\}$.

The following lemma characterizes ordered sequences.

Lemma 2 *Let $\Sigma = \{1, 2, ..., n, \star\}$ be an ordered alphabet and $S \in (\Sigma_R^n, \sqsubseteq)$. Then $S \in \Sigma_O^n$ iff for any $S' \in B_0^-$, $S' \not\sqsubseteq S$. Furthermore, $\Sigma_R^n \setminus \uparrow B_0^- = \Sigma_O^n$.*

Proof:

Let $S = x_1 x_2 ... x_m \in \Sigma_R^n$. Suppose that $S \in \Sigma_O^n$. Then for all $1 \leq i < j \leq m$ such that $x_j \neq \star$, $x_i \neq \star$ and $x_j - x_i = j - i$. Since $j > i$ then $x_j > x_i$. Moreover

$x_i \underbrace{\star \ldots \star}_{k} x_j$ with $k = j - i - 1$ cannot belong to B_0^- (by construction of B_0^-) and thus for any $S' \in B_0^-$, $S' \not\sqsubseteq S$

Now suppose that $S \notin \Sigma_O^n$. Then there exists $1 \leq i < j \leq m$ such that $x_j \neq \star$, $x_i \neq \star$ and $x_j - x_i \neq j - i$. Let $k = j - i - 1 \in [0, n-2]$. We distinguish two cases:

- $x_i \geq x_j$: then the sequence $x_i \underbrace{\star \ldots \star}_{k} x_j \in B_0^-$ with $x_i \underbrace{\star \ldots \star}_{k} x_j \sqsubseteq S$.
- $x_i < x_j$: the sequence $x_i \underbrace{\star \ldots \star}_{k} x_j \in B_0^-$ since $k \neq x_j - x_i - 1$, and, $x_i \underbrace{\star \ldots \star}_{k} x_j \sqsubseteq S$.

Finally we conclude that $\Sigma_R^n \setminus \uparrow B_0^- = \Sigma_O^n$. $\square$

Let $E = \{1, ..., n\}$ be a set. We define the mapping $f : \mathcal{P}(E) \to \Sigma_R^n$ such that for any $X \in \mathcal{P}(E)$, $f(X) = S$ with $S[i] = i$ if $i \in X$ and $S[i] = \star$ otherwise. Without loss of generality, we delete the symbols $\star$ that are prefix or suffix of $f(X)$. Note that $f(X)[i] = \star$ means that $i \notin X$. For example $f(\{2, 5\}) = 2 \star \star 5$ and $f(\{\})$ is the empty sequence.

Proposition 5 *Let $E = \{1, ..., n\}$ be a set. Then the mapping f is a convex embedding of $\mathcal{P}(E)$ into Σ_R^n. Moreover $f(\mathcal{P}(E)) = \Sigma_O^n$.*

Proof:

Let P, Q be two sequences that are images of two sets $A \subset B \subseteq E$, i.e. $f(A) = P$ and $f(B) = Q$. Clearly $f(A) \sqsubset f(B)$.

Now suppose there is a sequence S such that $P \sqsubset S \sqsubset Q$.

For every $i, j \in [1..n], i \neq j$, we have either $S[i] \neq S[j]$ or $S[i] = S[j] = \star$, by definition of the embedding f. Then the set $C = \{x \in E \mid S[i] = x, i \in [1..n]\}$ is clearly defined. Moreover $f(C) = S$ and $A \subset C \subset B$, since for any $x \in \Sigma$, $x \not\sqsubset \star$, but $\star \sqsubset x$.

We have $f(\mathcal{P}(E)) = \Sigma_O^n$ by construction. $\square$

Now we are able to show that the dualization on rigid sequences with wildcard is equivalent to the dualization on set, i.e. enumerating minimal transversals of a given hypergraph.

Theorem 2 ***DualizeOnSeq*** *and* ***DualizeOnSet*** *are polynomially equivalent.*

Proof:

First we show that **DualizeOnSeq** is polynomially reducible to **DualizeOnSet** [23]. Consider an instance of **DualizeOnSeq**: Let $(\Sigma_R^n, \sqsubseteq)$ be the poset of rigid sequences and $\mathcal{B}^+$ a positive border of Σ_R^n. We construct an instance of **DualizeOnSet**, i.e. a set E, $\mathcal{E}^+$ an antichain of $\mathcal{P}(E)$ as follows:

Consider first the reflection $(\Sigma_R^n, \sqsubseteq_1)$ of $(\Sigma_R^n, \sqsubseteq)$. By Lemma 1, we have $\downarrow^{\sqsubseteq} x = \downarrow^{\sqsubseteq_1} \mathit{LostPred}(x)$ and $\uparrow^{\sqsubseteq} x = \uparrow^{\sqsubseteq_1} \mathit{LostSucc}(x)$ with $\mathit{LostPred}(S) = \mathit{Max}_{\sqsubseteq_1}\{S[i..|S|] \mid 1 \leq i \leq |S|, S[i] \neq \star\}$ and $\mathit{LostSucc}(S) = \mathit{Min}_{\sqsubseteq_1}\{\underbrace{x \star \ldots \star}_{i} S \mid 0 \leq i \leq n - |S|, x \in \Sigma\}$. Clearly $(\Sigma_R^n, \sqsubseteq_1)$ is a poly-reflection of $(\Sigma_R^n, \sqsubseteq)$ since $|\mathit{LostPred}(S)| \in O(n)$ and $|\mathit{LostSucc}(x)| \in O(n.m)$, i.e. their sizes are polynomial in the size of the description Σ_R^n and n.

Now consider the embedding $f : (\Sigma_R^n, \sqsubseteq_1) \to \mathcal{P}(E)$ that associates for every $P[1..m] \in \Sigma_R^n$ a set of pairs of $E = (\{1, ..., m\} \times (\Sigma \setminus \{\star\}))$ as follows:

$$f(P) = \{(i, P[i]) \mid i \in [1..m], P[i] \neq \star\}$$

Then f is a convex embedding with $B_0^+ = \{\{(i, x) \mid x \in \Sigma, i \in [2..n]\}\}$ and $B_0^- = \{\{(1, x), (1, y) \mid x, y \in \Sigma, x \neq y\}\} \cup \{\{(1, x), (i, y), (i, z)\} \mid x, y, z \in \Sigma, y \neq z, i \in [2..n]\}$. Hence $|B_0^+ \cup B_0^-|$ is polynomial in m and n since $|B_0^+| \in O(n.m)$ and $|B_0^-| \in O(n.m^3)$, where $m = |\Sigma_R^n|$ and $n = |S|$.

From Theorem 1, we conclude that **DualizeOnSeq** is polynomially reducible to **DualizeOnSet**, i.e. $(\mathcal{B}^+, \mathcal{B}^-)$ are duals in $(\Sigma_R^n, \sqsubseteq)$ iff $(B_0^+ \cup f(\mathit{Ext}(\mathcal{B}^+)), B_0^- \cup f(\mathit{Ext}(\mathcal{B}^-)))$ are duals in $\mathcal{P}(E)$.

Conversely, Proposition 5 shows the existence of a convex embedding from $\mathcal{P}(E)$ into Σ_R^n. Moreover we have shown that $B_0^+ = \emptyset$ and according to Lemma 2 the size of B_0^- is bounded by $O(n^3)$. Thus **DualizeOnSet** is polynomially reducible to **DualizeOnSeq**. □

4 Applications to Maximal Pattern Mining Problems

We are now interested in the **PMP** problems given in Section 2.1.1, i.e. the mining of maximal interesting patterns. In the simplest case, it is well-known that maximal interesting patterns(i.e. the positive border) can be computed incrementally by using the Dualize & Advance algorithm [15] as a subroutine. In this section, we study how to modify the Dualize & Advance algorithm in more complex cases.

Table 1: A new classification of pattern mining problems

DualizeOnPoset	**PMP** problems
$\mathcal{RAS}$	$\mathcal{RAS}^{PM}$
$\mathcal{XRAS}$	$\mathcal{XRAS}^{PM}$
$\mathcal{EXRAS}$	$\mathcal{EXRAS}^{PM}$
$\mathcal{EWRAS}$	$\mathcal{EWRAS}^{PM}$

From the classification of posets given in previous sections (see Figure 4), it follows immediately a similar classification for **PMP** problems, as depicted in Table 1.

Note that most of **PMP** problems are known to be NP-Hard [5, 15]. Nevertheless, some of them can be solved in incremental quasi-polynomial time in the size of the positive and negative borders (see [15] for $\mathcal{RAS}^{PM}$ problems).

In the sequel, we recall basic results of the simplest case, i.e. pattern mining problems said to be *representable as sets* [21]. Then we show how to apply the materials defined in previous sections to the problem of frequent simple conjunctive queries in databases [14] and the problem of frequent rigid sequences.

4.1 Known results

We recall $\mathcal{RAS}$ problems defined in Section 3.1, and we give the relationship between the positive and negative borders through the notion of minimal transversal of hypergraphs.

Let $\mathcal{H} \subseteq \mathcal{P}(E)$ be a hypergraph on a set E. We denote by $Max(\mathcal{H})$ (resp. $Min(\mathcal{H})$) the set of maximal (resp. minimal) hyperedges of $\mathcal{H}$ with respect to set inclusion. $\mathcal{H}$ is said to be *simple* if $\mathcal{H} = Min(\mathcal{H}) = Max(\mathcal{H})$. A minimal transversal of $\mathcal{H}$ is a set of elements $X \subseteq E$ such that (1) X has a non-empty intersection with every hyperedge of $\mathcal{H}$ and (2) X is minimal w.r.t. this property. We denote by $\mathbf{MinTr}(\mathcal{H})$ the set of minimal transversals of $\mathcal{H}$ and $\overline{\mathcal{H}} = \{E \setminus e \mid e \in \mathcal{H}\}$ the set of complements in E of the edges of $\mathcal{H}$.

Let $\mathcal{S} \subseteq \mathcal{P}(E)$. We define $\mathcal{B}^+(\mathcal{S}) = Max(\mathcal{S})$ and $\mathcal{B}^-(\mathcal{S}) = Min(\mathcal{P}(E) \setminus \downarrow \mathcal{S})$. Note that for a given set $\mathcal{S}$ of $\mathcal{P}(E)$, the sets $\mathcal{B}^+(\mathcal{S})$ and $\mathcal{B}^-(\mathcal{S})$ are duals in $\mathcal{P}(E)$.

The relationship between the positive and negative borders is given as follows [21]:

Theorem 3 *[21] Let $(\mathcal{L}^*, \boldsymbol{d}, Q) \in \mathcal{RAS}^{\mathcal{PM}}$, $S \subseteq \mathcal{L}^*$ and $f : \mathcal{L}^* \to \mathcal{P}(E)$ a bijective mapping. Then $\mathcal{B}^-(\downarrow S) = f^{-1}(\mathbf{MinTr}(\overline{f(\mathcal{B}^+(\downarrow S))}))$*

We recall the complexity of the DUALIZE & ADVANCE algorithm proposed in [16] for problems representable as set (aka $\mathcal{RAS}^{PM}$). Let us denote by $width(\mathcal{L}^*, \preceq)$ the maximal number of immediate successors on $(\mathcal{L}^*, \preceq)$, and $t(n) = n^{o(\log n)}$.

Corollary 2 *[15] Let $(\mathcal{L}^*, \preceq, \boldsymbol{d}, Q) \in \mathcal{RAS}^{\mathcal{PM}}$ and Th its theory.*
$\mathcal{B}^+(Th)$ can be computed in time polynomial in $|\mathcal{B}^+(Th)|$ and $t(|\mathcal{B}^+(Th)|+|\mathcal{B}^-(Th)|)$, while using at most $|\mathcal{B}^-(Th)|+ width(\mathcal{L}^*, \preceq)\times|\mathcal{B}^+(Th)|$ queries.

4.2 Application to Frequent simple conjunctive queries

The frequent queries problem has been studied for instance in [7, 13, 14, 17]. We consider the simple problem statement defined in [14][5] and we assume a set semantic for the relational model. Let $\mathbf{R} = \{R_1, \ldots, R_n\}$ be a database schema, $\mathcal{D}$ the domain of $\mathbf{R}$ and $sch(\mathbf{R}) = \{R_i.A | R_i \in \mathbf{R}, A \in R_i\}$. A (simple) conjunctive queries Q over $\mathbf{R}$ is of the form $\pi_X(\sigma_F(R_1 \times \cdots \times R_n))$ where $X \subseteq sch(\mathbf{R})$, F a conjunction of equalities of the form $R_i.A = R_j.B$ or $R_i.A = c$ with $i \neq j$, $R_i.A, R_j.B \in sch(\mathbf{R})$ and $c \in \mathcal{D}$ and $(R_1 \times \cdots \times R_n)$ a cartesian product between n relation schemas (with no repetition). When clear from context, $\pi_X(\sigma_F(R_1 \times \cdots \times R_n))$ will be noted $\pi_X(\sigma_F)$. The reader may refer to [1] for more details on database notations.

Let $\mathcal{Q}_r$ be the set of all possible simple conjunctive queries over $\mathbf{R}$. Note that any query which is syntactically correct belongs to $\mathcal{Q}_r$. For a given database $\mathbf{d}$ over $\mathbf{R}$, we note $Adom(\mathbf{d}) \subseteq \mathcal{D}$ the active domain of $\mathbf{d}$ and $Q(\mathbf{d})$ the result of the evaluation of Q against $\mathbf{d}$.

Let Q_1, Q_2 be two simple conjunctive queries over $\mathbf{R}$. Q_1 is *contained* in Q_2, denoted $Q_1 \subseteq Q_2$, if for every database $\mathbf{d}$ over $\mathbf{R}$, $Q_1(\mathbf{d}) \subseteq Q_2(\mathbf{d})$. Q_1 is *diagonally contained* in Q_2, denoted $Q_1 \subseteq^{\Delta} Q_2$, if Q_1 is contained in a projection of Q_2, i.e. for instance $Q_1 \subseteq \pi_X(Q_2)$

The frequency of $Q = \pi_X(\sigma_F(R_1 \times \cdots \times R_n))$ in $\mathbf{d}$, denoted by $freq(Q, \mathbf{d})$, is defined by: $freq(Q, \mathbf{d}) = |\pi_X(\sigma_F(R_1 \times \cdots \times R_n))(\mathbf{d})|$. A query Q is *frequent* in $\mathbf{d}$ with respect to a threshold ϵ if $freq(Q, \mathbf{d}) \geq \epsilon$. $freq(Q, d)$ is monotonically decreasing with respect to $\subseteq^{\Delta}$ [13].

The problem statement can now be given:

FQ problem: Given a database $\mathbf{d}$ over $\mathbf{R}$ and a threshold ϵ, enumerate the positive border of ϵ-frequent queries of $(\mathcal{Q}_r, \subseteq^{\Delta})$ in $\mathbf{d}$.

[5]Query containment being NP-Complete, we restrict ourselves to a simple fragment, the so-called "Simple Conjunctive Queries" where every relation occurs only once. Thus, the number of Simple Conjunctive Queries is finite and deciding containment is possible in linear time (syntactic check of the where clause).

In the following, we show that $(\mathcal{Q}_r, \subseteq^{\Delta})$ is isomorphic to a boolean lattice, i.e. **FQ** belongs to $\mathcal{RAS}$.

To do so, we denote by $\mathcal{F}$ the finite set of all possible selection formulas over $\mathbf{R}$ and $Adom(\mathbf{d})$, i.e. $\mathcal{F} = \{\{A, B\} \mid A \in sch(\mathbf{R}), B \in sch(\mathbf{R}) \cup Adom(\mathbf{d})\}$. First, notice that a bijection trivially exists between $\mathcal{Q}_r$ and $\mathcal{P}(\mathcal{F} \cup \mathbf{R})$. So it remains to show that $\mathcal{Q}_r$ ordered under $\subseteq^{\Delta}$ is a boolean lattice.

Proposition 6 *Let $Q_1 = \pi_{X_1}(\sigma_{F_1})$ and $Q_2 = \pi_{X_2}(\sigma_{F_2})$ be two queries of $\mathcal{Q}_r$. Then $Q_1 \subseteq^{\Delta} Q_2$ iff $X_1 \subseteq X_2$ and $F_2 \subseteq F_1$. Equivalently, $Q_1 \subseteq^{\Delta} Q_2$ iff $X_1 \cup (\mathcal{F} \setminus F_1) \subseteq X_2 \cup (\mathcal{F} \setminus F_2)$.*

Proof: ($\Rightarrow$) Suppose $Q_1 \subseteq^{\Delta} Q_2$. Then there exists $X \subseteq \mathbf{R}$ such that $Q_1 \subseteq \pi_X(Q_2)$. Clearly $X \subseteq X_2$ and $X = X_1$, which implies that $X_1 \subseteq X_2$. Moreover, for every database d over $\mathbf{R}$, $\pi_{X_1}(\sigma_{F_1}) \subseteq \pi_{X_1}(\sigma_{F_2})$ and it follows $\sigma_{F_1} \subseteq \sigma_{F_2}$ and then $F_2 \subseteq F_1$.

($\Leftarrow$) Suppose $X_1 \subseteq X_2$ and $F_2 \subseteq F_1$. Let $\mathbf{d}$ be a database over $\mathbf{R}$ and $t \in Q_1(\mathbf{d})$ a tuple over X_1. Then t satisfies F_1 and also F_2 since $F_2 \subseteq F_1$. Since $X_1 \subseteq X_2$, we have $t \in \pi_{X_1}(Q_2)(\mathbf{d})$ and therefore $Q_1 \subseteq^{\Delta} Q_2$. □

From the Proposition 6, the mapping $f : \mathcal{Q}_r \to \mathcal{P}(\mathbf{R} \cup \mathcal{F})$ with $f(\pi_X(\sigma_F)) = X \cup (\mathcal{F} \setminus F)$ is a bijective embedding and **FQ** belongs to $\mathcal{RAS}$.

Example 4 *Let $\mathbf{R} = \{R_1, R_2\}$ be a database schema, d a database over $\mathbf{R}$, $Adom(d) = \{1, 2\}$ and $R = R_1 \times R_2$ such that $sch(R) = \{A, B, C\}$. The set of possible conditions $\mathcal{F}$ is $\{\{A, B\}, \{A, C\}, \{B, C\}, \{A, 1\}, \{B, 1\}, \{C, 1\}, \{A, 2\}, \{B, 2\}, \{C, 2\}\}$.*
Let us consider two queries:
$Q_1 = \pi_{\{B\}}\sigma_{(A=B \wedge C=1)}(R)$ and $Q_2 = \pi_{\{A,B\}}\sigma_{(C=1)}(R)$.
Then, $f(Q_1) = \{B, \{A, C\}, \{B, C\}, \{A, 1\}, \{B, 1\}, \{A, 2\}, \{B, 2\}, \{C, 2\}\}$
and $f(Q_2) = \{A, B, \{A, B\}, \{A, C\}, \{B, C\}, \{A, 1\}, \{B, 1\}, \{A, 2\}, \{B, 2\}, \{C, 2\}\}$. Remark that $Q_1 \subseteq^{\Delta} Q_2$ and $f(Q_1) \subseteq f(Q_2)$.

Clearly, **FQ** belongs to $\mathcal{RAS}^{PM}$ and from Corollary 2, the following result is straightforward. Let Th be the theory of $(\mathcal{Q}_r, \subseteq^{\Delta}, d, freq(Q, d))$.

Corollary 3 *$\mathcal{B}^+(Th)$ can be computed in time $t(|\mathcal{B}^+(Th)| + |\mathcal{B}^-(Th)|)$, where $t(k) = k^{o(\log k)}$, while using at most $(|\mathcal{B}^-(Th)|) + (n^2 + pn).(|\mathcal{B}^+(Th)|)$ queries, where $n = |\mathbf{R}|$ and $p = |Adom(d)|$.*

It means that the complexity of mining frequent simple conjunctive queries in databases is incremental quasi-polynomial time in the size of the two borders, i.e. classic algorithms like Dualize & Advance can be re-used up to a polynomial

transformation. Nevertheless, this result is rather artificial since whenever we consider *consistent* simple conjunctive queries only, the problem is no longer in $\mathcal{RAS}^{PM}$.

Indeed, the set of queries $\mathcal{Q}_r$ allows queries which are not consistent, i.e. there is no database such that their evaluation returns a value different from zero. For instance, $\sigma(B = 1 \wedge B = 2)$ or $\sigma(A = B \wedge A = 1 \wedge B = 2)$ are not consistent but belong to $\mathcal{Q}_r$.

In the following, let us consider the set of all consistent queries, denoted by $\mathcal{Q}_C$, defined by $\mathcal{Q}_C = \{Q \in \mathcal{Q}_r |$ there exists a database $\mathbf{d}$ such that $freq(Q, \mathbf{d}) \neq 0\}$. A new problem statement, slightly modified from **FQ**, can be given as follows:

CFQ problem: Given a database $\mathbf{d}$ over $\mathbf{R}$ and a threshold ϵ. Enumerate the positive border of ϵ-frequent queries of $(\mathcal{Q}_C, \subseteq^{\Delta})$ in $\mathbf{d}$.

Consider the mapping $f' : \mathcal{Q}_C \to \mathcal{P}(\mathbf{R} \cup \mathcal{F})$ with $f'(Q) = f(Q)$. f' is no longer bijective and the problem does not belong to $\mathcal{RAS}^{PM}$. Now we show that **CFQ** belongs to $\mathcal{XRAS}^{PM}$, but does not belong to $\mathcal{EXRAS}^{PM}$.

Lemma 3 *Let $Q_1 = \pi_{X_1}(\sigma_{F_1})$ and $Q_2 = \pi_{X_2}(\sigma_{F_2})$ be two queries of $\mathcal{Q}_r$ such that $Q_1 \subseteq^{\Delta} Q_2$. Then Q_1 is consistent implies Q_2 is consistent.*

Proof: $Q_1 \subseteq^{\Delta} Q_2$ implies $F_2 \subseteq F_1$ or equivalently $\sigma(F_1) \subseteq \sigma(F_2)$. Q_1 consistent means there exists a database $\mathbf{d}$ such that $Q_1(\mathbf{d}) \neq \emptyset$. Let $t \in Q_1(\mathbf{d})$. Then $t \in \pi_{X_1}(\pi_{X_2}(\sigma_{F_2}))(\mathbf{d})$, i.e. Q_2 is consistent. □

We now point out how the restriction to consistent queries changes the classification of the initial problem.

Proposition 7 **CFQ** $\in \mathcal{XRAS}^{PM}$ *but* **CFQ** $\notin \mathcal{EXRAS}^{PM}$.

Proof: From lemma 3, we deduce that $(Qc, \subseteq^{\Delta})$ has a convex embedding, and therefore **CFQ** belongs to $\mathcal{XRAS}^{PM}$. From Proposition 2, there exists two antichains B_0^+ and B_0^- such that $(Qc, \subseteq^{\Delta})$ isomorphic to $\mathcal{P}(\mathbf{R} \cup \mathcal{F}) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-)$. Clearly, the set B_0^- is empty whereas the set B_0^+ contains all maximal non-consistent queries. Without loss of generality, we assume a total order $<$ on attribute set. A maximal non-consistent query has the following form:
$\pi_{\mathbf{R}}(\sigma_{A_1=A_2 \wedge \cdots \wedge A_{n-1}=A_n \wedge A_1=v \wedge A_n=v'})$ where $v \neq v'$ and $A_1, A_2, \ldots A_{n-1}, A_n$ is a chain with $A_i < A_{i+1}$ and A_i, A_{i+1} pairwise distinct for all $i \in [1..n-1]$.

Thus, the size of B_0^+ is exponential in the size of $\mathbf{R} \cup Adom(\mathbf{d})$ since the number of chains is exponential in the number of attributes. We conclude that **CFQ** does not belong to $\mathcal{EXRAS}^{PM}$. □

4.2.1 Application to Frequent Rigid Sequences with Wildcard

Continuing examples given in section 3.2.1, we consider the poset $(\Sigma_R^n, \sqsubseteq)$ of rigid sequences with wildcard. Let us fix a finite string, called an *input sequence*, $S = S[1] \dots S[n] \in \Sigma_R^n$ of length $n \geq 0$. Consider $P \in \Sigma_R^n$. The location list for P in S is the set $L_S(P) = \{p \in [1..n] \mid P \sqsubseteq_p S\}$. The *frequence* of P in S is defined by: $freq(P, S) = |L_S(P)|$, i.e. the number of times that P occurs in S. A motif P is said to be *k-frequent* in S if $freq(P, S) \geq k$. A k-frequent motif P in S is said to be *maximal* if for every motif Q such that $P \sqsubseteq Q$, Q is not k-frequent in S.

The corresponding frequent sequence mining problem is the following:

SEQ problem: Given a sequence $S \in \Sigma_R^n$ and a threshold ϵ. Enumerate the maximal ϵ-frequent rigid sequences of S.

From the proof of Theorem 2 and [23], we deduce that **SEQ** $\in \mathcal{EWRAS}^{PM}$ and then, we have the following complexity result. Let Th be the theory of $(\Sigma_R^n, \sqsubseteq, S, freq(P, S))$.

Corollary 4 **SEQ** *can be computed in time* $t(nm+n.|\mathcal{B}^+(Th)|+nm^3+nm.|\mathcal{B}^-(Th)|)$, *where* $t(k) = k^{o(\log k)}$, *while using at most* $|\mathcal{B}^-(Th)| + nm.|\mathcal{B}^+(Th)|$ *queries, where* $m = |\Sigma_R^n|$ *and* $|S| = n$.

5 Algorithms for Dualization

In this section, we assume that the reader is familiar with the Dualize & Advance algorithm proposed for $\mathcal{RAS}^{PM}$ in [15]. Dualize & Advance has the best known complexity for this class of problems, i.e. incremental quasi-polynomial time in the sum of the sizes of the positive and negative borders. Dualize & Advance takes as input a set of items E, a database $\mathbf{d}$ and a Interestingness predicate $Q : \mathcal{P}(E) \times \mathbf{d} \to \{0, 1\}$, referred to as Dualize & Advance$(E, \mathbf{d}, Q)$. A set X is said to be interesting if $Q(X, \mathbf{d})$ holds.

Suppose now we are given a pattern mining problem $P = (\mathcal{L}^*, \preceq, \mathbf{d}, Q)$. We distinguish the three efficient classes of problems identified in the previous sections, leading to three algorithms derived from Dualize & Advance. They are given hereafter.

5.1 $P \in \mathcal{RAS}^{PM}$

By definition of $\mathcal{RAS}$, there is an isomorphism $f : (\mathcal{L}^*, \preceq) \to (\mathcal{P}(E), \subseteq)$. Given $A \subseteq E$, let us consider the predicate $\mathcal{P}red_1$ defined as follows:

$$\mathcal{P}red_1(A, \mathbf{d}) = Q(f^{-1}(A), \mathbf{d})$$

Algorithm 1: MINING FOR $\mathcal{RAS}^{PM}$

Input: $(\mathcal{L}^*, \preceq, \mathbf{d}, Q) \in \mathcal{RAS}^{\mathcal{PM}}$, E
Output: The positive border $\mathcal{B}^+$
$\mathcal{C}$ = DUALIZE & ADVANCE$(E, \mathbf{d}, \mathcal{P}red_1)$
Return $f^{-1}(\mathcal{C})$

Algorithm 1 is the DUALIZE&ADVANCE on the powerset of E with the corresponding predicate $\mathcal{P}red_1$.

5.2 $P \in \mathcal{EXRAS}^{PM}$

By definition of $\mathcal{EXRAS}$, there are two antichains B_0^+ and B_0^- of $\mathcal{P}(E)$ such that $(\mathcal{L}^*, \preceq)$ is isomorphic to $\mathcal{P}(E) \setminus (\downarrow B_0^+ \cup \uparrow B_0^-)$. Let $f : \mathcal{L}^* \to \mathcal{P}(E)$ be the associated convex embedding. Given $A \subseteq E$, let us consider the predicate $\mathcal{P}red_2$ defined as follows:

$$\mathcal{P}red_2(A, \mathbf{d}) = \begin{cases} 1 & \text{if} \quad A \in \downarrow B_0^+ \\ 0 & \text{if} \quad A \in \uparrow B_0^- \\ Q(f^{-1}(A), \mathbf{d}) & \text{otherwise} \end{cases}$$

Again, Algorithm 2 is also similar to DUALIZE&ADVANCE, with two key differences: First, the predicate $\mathcal{P}red_2$ has to verify whether or not a given pattern is covered by B_0^+ or B_0^- before accessing to the data; Second, at the end, elements of B_0^+ have to be removed from the output.

5.3 $P \in \mathcal{EWRAS}^{PM}$

This case corresponds to the more sophisticated extension of DUALIZE&ADVANCE, since we have to deal with the poly-reflection of the initial poset.

Let $(\mathcal{L}^*, \preceq, \mathbf{d}, Q) \in \mathcal{EWRAS}$. Then, we consider the following notations:

- $\mathcal{B}^+$ (resp. $\mathcal{B}^-$) is the positive (resp. negative) border of $Th(\mathcal{L}^*, \preceq, \mathbf{d}, Q)$

Algorithm 2: MINING FOR $\mathcal{EXRAS}^{PM}$

Input: $(\mathcal{L}^*, \preceq, \mathbf{d}, Q) \in \mathcal{EXRAS}^{\mathcal{PM}}$, E, B_0^+
Output: The positive border $\mathcal{B}^+$
$\mathcal{C}$ = DUALIZE & ADVANCE$(E, \mathbf{d}, \mathcal{P}red_2)$
Return $f^{-1}(\mathcal{C} \setminus B_0^+)$

- E is a finite set, $f : \mathcal{L}^* \to \mathcal{P}(E)$ a map and B_0^+, B_0^- two antichains of $\mathcal{P}(E)$ such that $(Max(B_0^+ \cup f(Ext(\mathcal{B}^+))), Min(B_0^- \cup f(Ext(\mathcal{B}^-))))$ is a border of $(\mathcal{P}(E), \subseteq)$
- For $\theta \in \mathcal{L}^*$, $LostPred(\theta)$ is an algorithm which computes the set of predecessors of θ in $\mathcal{L}^*$ that are lost in the associated reflection, from which $Ext(\mathcal{B}^+)$ can be computed.

Conveniently, we also define $f^{-1}(\mathcal{E}) = \cup_{X \in \mathcal{E}} f^{-1}(X)$, for $\mathcal{E} \subseteq f(\mathcal{L}^*)$.

Algorithm 3: DUALIZE & ADVANCE revisited

Input: $(\mathcal{L}^*, \preceq, \mathbf{d}, Q) \in \mathcal{EWRAS}$, $(E, f, LostPred, B_0^+, B_0^-)$
Output: The positive border $\mathcal{B}^+$ of $Th(\mathcal{L}, \mathbf{d}, Q)$.

```
1  C_0 = B_0^+; i = 0; Continue = true;
   while Continue do
2      D_i = {complements of sets in C_i}
3      Continue = false
4      for each minimal transversal X of D_i do
5          if X ∉ B_0^- then
6              if Q(d, f^{-1}(X)) holds then
7                  θ = AMSS(f^{-1}(X));
8                  C_{i+1} = Max_⊆(C_i ∪ f(LostPred(θ)));
9                  i = i + 1; Continue = true;
                   quit for loop
10     if Not(Continue) then
11         return Max_⪯(f^{-1}(C_i \ B_0^+)) and exit;
```

The main changes in Algorithm 3 with respect to the original version of DUALIZE & ADVANCE [15] are:

Algorithm 4: AMSS revisited

Input: ϕ, a counter-example
Output: $\theta \in \mathcal{B}^+(\mathit{Th}(\mathcal{L}^*, \preceq, \mathbf{d}, Q))$ such that $\phi \preceq \theta$
$\theta = \phi$;
while *there exists* α *such that* $\theta \prec \alpha$ *and* $Q(\boldsymbol{d}, \alpha)$ *holds* **do**
  $\theta = \alpha$
return θ

- C_0 is initialized with B_0^+ instead of $\emptyset$ (line 1),
- a minimal transversal of $\overline{D_i}$ may belong to B_0^- and will not be evaluated against the database (line 5),
- finding a maximal superset θ of a counter example $f^{-1}(X)$ (line 7): Algorithm 4 performs a search in $\mathcal{L}^*$ instead of $\mathcal{P}(E)$,
- the lost comparabilities of a maximal pattern have to be added to the border of $\mathcal{P}(E)$ thanks to *LostPred*() (line 8),
- the final result has to delete all patterns in $f^{-1}(C_i \setminus B_0^+)$ that are not maximal in $(\mathcal{L}^*, \preceq)$ (line 11), those extra elements were added to the current solution by *LostPred*().

Consider the two particular cases:

- If P belongs to $\mathcal{RAS}$ then $\mathit{LostPred}(\theta) = \{\theta\}$, $B_0^+ = B_0^- = \emptyset$ and f is bijective since $f(\mathcal{L}^*) = \mathcal{P}(E)$. Algorithm 3 is the classic DUALIZE & ADVANCE algorithm of [15].
- If P belongs to $\mathcal{EXRAS}$ then $\mathit{LostPred}(\theta) = \{\theta\}$ and ($B_0^+ \neq \emptyset$ or $B_0^- \neq \emptyset$). Algorithm 3 starts with B_0^+ (line 1) and it avoids to check (in line 5) the predicate for sets in B_0^-. Note also that elements of B_0^+ are first enumerated and then removed at the end in Algorithm 2 whereas Algorithm 3 avoids to enumerate them with the initialization (line 1).

The proofs are extensions of the proofs given in [15].

Lemma 4 *At any step i of Algorithm 3, $\downarrow (B_0^+ \cup f(\mathit{Ext}(\mathcal{B}^+))) \not\subseteq \downarrow C_i$ iff at least one minimal transversal of $\overline{D_i}$ is interesting.*

Proof: We have $\downarrow (B_0^+) \subseteq \downarrow C_i$ since $C_0 = B_0^+$ (see Line 1 in Algorithm 3) and $\downarrow C_i \subseteq \downarrow C_{i+1}$. Moreover, only elements in $\downarrow f(Ext(\mathcal{B}^+))$ corresponding to $\downarrow \mathcal{B}^+$ are interesting.

Suppose $\downarrow f(Ext(\mathcal{B}^+)) \not\subseteq \downarrow C_i$. Then exists a set $X \in (\downarrow f(Ext(\mathcal{B}^+)) \setminus \downarrow C_i)$. Since $\preceq$ is monotone, there is a minimal interesting set not in C_i, that is, there is an interesting set Y in the negative border of C_i, i.e. Y is minimal transversal of $\overline{D_i}$.

Conversely, suppose $f(Ext(\mathcal{B}^+)) \subseteq C_i$. From Theorem 1, we have any minimal transversal of $\overline{D_i}$ is not interesting. □

Lemma 5 *At any step of Algorithm 3, the size of minimal transversals computed at line 4 is bounded by* $|B_0^-| + |Ext(\mathcal{B}^-)|$.

Proof: Each enumerated set X (line 4) either matches an element of B_0^-, or an element of $Ext(\mathcal{B}^-)$ or is interesting, i.e. $Q(\mathbf{d}, f^{-1}(X))$ holds. □

Theorem 4 *Let* $(\mathcal{L}^*, \preceq, \boldsymbol{d}, Q) \in \mathcal{EWRAS}^{\mathcal{PM}}$. *Then Algorithm 3 computes* $\mathcal{B}^+$ *in time polynomial in* $|\mathcal{B}^+|$ *and* $t(|Ext(\mathcal{B}^+)| + |Ext(\mathcal{B}^-)| + |B_0^+| + |B_0^-|)$, *while using at most* $(|Ext(\mathcal{B}^-)|) + width(\mathcal{L}^*, \preceq) \times high(\mathcal{L}^*) \times |\mathcal{B}^+|$ *queries, where* $high(\mathcal{L}^*)$ *is the maximal size of patterns in* $\mathcal{L}^*$.

Proof: According to Lemma 4, Algoritm 3 computes $f(Ext(\mathcal{B}^+))$. In line 11, Algorithm 3 returns the border $\mathcal{B}^+$ which corresponds to maximal patterns in C_i with respect to $(\mathcal{L}^*, \preceq)$, i.e. it deletes the added predecessors sets and the set B_0^+. Therefore the algorithm is correct.

By Lemma 5, in each iteration the algorithm calls the minimal transversals subroutine that enumerates the negative border of C_i.

Each minimal transversal X is either a set in $f(Ext(\mathcal{B}^-)) \cup B_0^-$ or is interesting. Thus if we keep all sets in $f(Ext(\mathcal{B}^-)) \cup B_0^-$ that we have already found, then for each minimal transversal X we check if it is already found. If yes we ignore it, otherwise we check if is interesting. If yes we compute a maximal interesting set θ using Algorithm 4. The extension of X to θ needs a search in the poset $(\mathcal{L}^*, \preceq)$ of depth $high(\mathcal{L})$ the maximal size of a pattern, and for each level in the search we check at most $width(\mathcal{L}^*, \preceq)$ the maximum number of immediate successors of a pattern in $(\mathcal{L}^*, \preceq)$. Thus the total number of queries is bounded by $|Ext(\mathcal{B}^-)| + width(\mathcal{L}^*, \preceq) \times high(\mathcal{L}^*) \times |\mathcal{B}^+|$ since the predicate is not checked for elements in $LostPred()$.

Finally, since the number of calls of the minimal transversals subroutine is bounded by $|\mathcal{B}^+|$, the result follows. □

6 Conclusion

In this paper, we have defined the reduction of the dualization problem on arbitrary posets to the dualization problem on boolean lattices. We have proposed two key properties, namely convex embedding and poset reflections, to identify "good" reductions. Classes of posets have been identified to define the hardness of the dualization problem. Moreover, a new and somewhat surprising result have been given: dualization on set is equivalent to dualization on rigid sequences. We have also applied these results to maximal pattern mining problems. We have studied the complexity of two problems: mining of frequent conjunctive queries and mining of frequent rigid sequences. Finally, from the classification on posets, we have deduced a new classification on maximal pattern mining problems and we have shown how the best enumeration algorithm for maximal itemset mining, Dualize & Advance, can be adapted to those new classes.

Many perspectives remain to be addressed. Complex patterns (e.g. trees, graphs) have to be studied with respect to these new classes. From a theoretical point of view, an interesting open question is to identify the smallest number of comparabilities to be broken on a poset to ensure the existence of a convex embedding. It could be also interesting to study dualization of product of chains, known to have a quasi-polynomial time complexity [11]. Another interesting point would be to study the duality gap [25] with respect to the polynomial extension introduced in this paper.

Acknowledgment: This work has been funded by the french national research agency (ANR DAG project, 2009-2013) and CNRS (Mastodons PETASKY project, 2012-2015).

References

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. *Fondements des bases de données*. Addison Wesley, 2000.

[2] Rakesh Agrawal, Tomaz Imielinski, and Arun Swami. Mining associations between sets of items in massive databases. In *ACM SIGMOD 1993, Washington D.C.*, pages 207–216, 1993.

[3] Hiroki Arimura and Takeaki Uno. Polynomial-delay and polynomial-space algorithms for mining closed sequences, graphs, and pictures in accessible set systems. In *SDM*, pages 1087–1098, 2009.

[4] Mario Boley, Tamás Horváth, Axel Poigné, and Stefan Wrobel. Listing closed sets of strongly accessible set systems with applications to data mining. *Theor. Comput. Sci.*, 411(3):691–700, 2010.

[5] E. Boros, V. Gurvich, L. Khachiyan, and K. Makino. On maximal frequent and minimal infrequent sets in binary matrices. *Annals of Mathematics and Artificial Intelligence*, 39(3):211–221, November 2003.

[6] Brian A. Davey and Hilary A. Priestley. *Introduction to Lattices and Order*. Cambridge Press, 1990.

[7] Luc Dehaspe and Hannu Toivonen. Discovery of frequent datalog patterns. *Data Min. Knowl. Discov.*, 3(1):7–36, 1999.

[8] Fabien De Marchi and Jean-Marc Petit. Zigzag: a new algorithm for mining large inclusion dependencies in databases. In *ICDM'03, USA*, pages 27–34, November 2003.

[9] Thomas Eiter and Georg Gottlob. Identifying the minimal transversals of a hypergraph and related problems. *SIAM J. Comput.*, 24(6):1278–1304, 1995.

[10] Thomas Eiter, Georg Gottlob, and Kazuhisa Makino. New results on monotone dualization and generating hypergraph transversals. *SIAM J. Comput.*, 32:514–537, February 2003.

[11] Khaled M. Elbassioni. Algorithms for dualization over products of partially ordered sets. *SIAM J. Discrete Math.*, 23(1):487–510, 2009.

[12] Michael L. Fredman and Leonid Khachiyan. On the complexity of dualization of monotone disjunctive normal forms. *J. Algorithms*, 21(3):618–628, 1996.

[13] Bart Goethals and Jan Van den Bussche. Relational association rules: Getting warmer. In *Pattern Detection and Discovery*, pages 125–139, 2002.

[14] Bart Goethals, Wim Le Page, and Heikki Mannila. Mining association rules of simple conjunctive queries. In *SDM*, pages 96–107, 2008.

[15] Dimitrios Gunopulos, Roni Khardon, Heikki Mannila, Sanjeev Saluja, Hannu Toivonen, and Ram Sewak Sharm. Discovering all most specific sentences. *ACM Trans. Database Syst.*, 28(2):140–174, 2003.

[16] Dimitrios Gunopulos, Roni Khardon, Heikki Mannila, and Hannu Toivonen. Data mining, hypergraph transversals, and machine learning. In *Proceedings of the* 16^{th} *ACM Symposium on Principles of Database Systems, Tucson, Arizona*, pages 209–216, 1997.

[17] Tao-Yuan Jen, Dominique Laurent, and Nicolas Spyratos. Mining all frequent projection-selection queries from a relational table. In *EDBT*, pages 368–379, 2008.

[18] Mamadou Moustapha Kanté, Vincent Limouzy, Arnaud Mary, and Lhouari Nourine. On the enumeration of minimal dominating sets and related notions. *SIAM J. Discrete Math.*, 28(4):1916–1929, 2014.

[19] Mamadou Moustapha Kanté and Lhouari Nourine. Minimal dominating set enumeration. In *Encyclopedia of Algorithms*. 2015.

[20] Heikki Mannila and Kari-Jouko Räihä. Algorithms for inferring functional dependencies from relations. *Data and Knowledge Engineering*, 12(1):83–99, February 1994.

[21] Heikki Mannila and Hannu Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258, 1997.

[22] Benjamin Negrevergne, Alexandre Termier, Marie-Christine Rousset, and Jean-François Méhaut. Para miner: a generic pattern mining algorithm for multi-core architectures. *Data Mining and Knowledge Discovery*, 2013.

[23] Lhouari Nourine and Jean-Marc Petit. Extending Set-Based Dualization: Application to Pattern Mining. In IOS Press, editor, *ECAI 2012*, August 2012.

[24] Lhouari Nourine and Jean-Marc Petit. Dualization on partially ordered sets: Preliminary Results. In *Post-Proceedings of the ISIP 2014 Workshop*, 2014.

[25] Lhouari Nourine and Jean-Marc Petit. Beyond hypergraph dualization. In *Encyclopedia of Algorithms*. 2015.

[26] Marcel Wild. Cover-preserving order embeddings into boolean lattices. *Order*, 9:209–232, 1992.