On BP-complete query languages on \( \mathcal{K} \)-relations

Floris Geerts\(^1\) and Antonella Poggi\(^2\)

\(^1\) University of Edinburgh, UK
\(^2\) SAPIENZA, Università di Roma, Italy

Abstract. The relational model has recently been extended to so-called \( \mathcal{K} \)-relations in which tuples are assigned a unique value in a semiring \( \mathcal{K} \). A query language, denoted by \( \mathcal{RA}^+_{\mathcal{K}} \), similar to the classical positive relational algebra, allows for the querying of \( \mathcal{K} \)-relations. We study the completeness of \( \mathcal{RA}^+_{\mathcal{K}} \) in the sense of Bancilhon and Paredaens (BP) and show that \( \mathcal{RA}^+_{\mathcal{K}} \) is, in general, not BP-complete. Faced with this incompleteness, we identify two necessary and sufficient operators that need to be added to \( \mathcal{RA}^+_{\mathcal{K}} \) to make it BP-complete: difference (\(-\)) and duplicate elimination (\(\delta\)). We investigate conditions on semirings under which these constructs can be added in a natural way, and investigate basic properties of our query languages.

1 Introduction

Several forms of annotated relations have appeared in various contexts in the database literature. The querying of such relations involves the generalization of the relational algebra (\(\mathcal{RA}\)) to perform corresponding operations on the annotations. Recently, a general data model, called \(\mathcal{K}\)-relations, has been proposed for annotated relations in which tuples in a relation are assigned a unique value coming from a semiring \(\mathcal{K}\) [6]. By varying the semiring \(\mathcal{K}\), \(\mathcal{K}\)-relations can model the standard relational model with both set [1] and bag semantics [8], incomplete databases (Boolean \(c\)-tables to be more precise) [7] and probabilistic databases [5, 11]. The generality of semirings also allows for the definition of new data models which are of particular interest for the study of provenance of data. Moreover, the operations that queries in the relational algebra perform on annotations naturally translate into algebraic operations in these semirings. This lead to the definition of the positive relational algebra on \(\mathcal{K}\)-relations, or \(\mathcal{RA}^+_{\mathcal{K}}\) for short [6].

In this paper, we consider completeness of \(\mathcal{RA}^+_{\mathcal{K}}\) on \(\mathcal{K}\)-relations. Recall that Bancilhon [3] and Paredaens [10] independently proposed a language-independent characterization for completeness of the relational algebra, known as BP-completeness. This characterizes the set of relations that can be computed by applying any generic query in the relational algebra to a particular database instance, based on classical symmetry properties of databases. Specifically, BP-completeness can be stated as follows: a relation \(R_2\) is the result of a generic relational query applied to a database \(R_1\) if and only if (i) the active domain of \(R_2\) is included in the domain of \(R_1\); and (ii) every automorphism of \(R_1\) is also an automorphism of \(R_2\).

The main contributions of our work are summarized in the following.

– First, we naturally extend the notion of BP-completeness to \(\mathcal{K}\)-relations. Then, in contrast to the standard relational case, we show that \(\mathcal{RA}^+_{\mathcal{K}}\) is not BP-complete for arbitrary semirings \(\mathcal{K}\).
Second, we expand $\mathcal{RA}_K^+$ with the difference operator, resulting in the algebra $\mathcal{RA}_K$. This, however, is a non-trivial task as it requires the definition of a monus operator on the underlying semirings $K$. Specifically, since not every semiring can be equipped with a monus operator, in order for $\mathcal{RA}_K$ to make sense, we are forced to restrict the class of semirings under consideration. We show that most semirings encountered in practice fall into this class.

Third, while the addition of difference to $\mathcal{RA}_K^+$ is desirable by its own, we show it is not sufficient to obtain BP-completeness; even more expressive power is needed. Hence, motivated by the study of bag query languages [8] (bags being a special case of $K$-relations as we will see in the next section), we introduce the concept of duplicate elimination on $K$-relations. In particular, we consider finitely generated semirings $K$ and for each of the generators $k$ of $K$, we define a corresponding duplicate elimination operator $\delta_k$.

Fourth, we demonstrate that $\mathcal{RA}_K^*$, i.e., the extension of $\mathcal{RA}_K$ with duplicate elimination operators, is BP-complete.

Finally, we show that the semantics of the evaluation of queries in $\mathcal{RA}_K$ is less transparent than in the positive ($\mathcal{RA}_K^+$) case. In particular, we provide some preliminary insights in the intricacies involved in the evaluation of queries in $\mathcal{RA}_K$ on $K$-relations.

The paper is organized as follows. After recalling in Section 2 the basic notions about the $K$-relations data model and the positive query language $\mathcal{RA}_K^+$ that were introduced in [6], we extend in Section 3 BP-completeness to $K$-relations, and show that $\mathcal{RA}_K^+$ is not BP-complete. In Section 4, we define the language $\mathcal{RA}_K$ obtained by adding a difference operator to $\mathcal{RA}_K^+$. Then, in Section 5, the language $\mathcal{RA}_K$ is further extended with duplicate elimination and shown to be BP-complete. In Section 6 some crucial properties of $\mathcal{RA}_K$ are established. Section 7 concludes the paper.

2 Preliminaries

In this section we recall the definition of $K$-relations and query language $\mathcal{RA}_K^+$ that were introduced in [6].

2.1 $K$-relations

A semiring $K= (\mathbb{K}, \oplus, \otimes, 0, 1)$ is an algebraic structure consisting of a set $\mathbb{K}$ equipped with two binary operations $\oplus$ and $\otimes$ such that $(\mathbb{K}, \oplus, 0)$ is a commutative monoid with identity element $0$; $(\mathbb{K}, \otimes, 1)$ is a monoid with identity element $1$; the operation $\otimes$ distributes over $\oplus$; and finally $0$ is an annihilating element. A semiring $K$ is called commutative if $(\mathbb{K}, \otimes, 1)$ is a commutative monoid. In the following, we only consider commutative semirings.

To formally introduce semirings into the relational data model, we now recall the definition of $K$-relations (see [6] for more details). Let $\mathbb{D}$ be a domain of data values and $U$ be a finite set of attributes. We define a $U$-tuple $t$ to be mapping from $U \rightarrow \mathbb{D}$. The set of $U$-tuples is denoted by $U$-Tup. A relation over $U$ is a subset of $U$-Tup. Let now $\mathbb{K}= (\mathbb{K}, \oplus, \otimes, 0, 1)$ be a semiring. A $K$-relation over $U$ is a function $R: U$-Tup $\rightarrow \mathbb{K}$. The
Fig. 1. Examples of $K$-relations.

support of a $K$-relation $R$, denoted by $\text{supp}(R)$, is defined as $\text{supp}(R) = \{ t | R(t) \neq 0 \}$. The active domain of a $K$-relation $R$, denoted by $\text{adom}(R)$, is defined as the set of data values occurring in $\text{supp}(R)$.

Example 1. In Fig. 1, we illustrate $K$-relations for five different semirings. Specifically, $R_1$ is a $K_3$-relation, where $K_3=\langle \mathbb{B}, \lor, \land, \text{false}, \text{true} \rangle$ is the Boolean semiring. $K_3$-relations correspond to the standard set semantics of relational tables. $R_2$ is a $K_{N\mathbb{N}}$-relation, for the semiring of natural number $K_{N\mathbb{N}}=\langle \mathbb{N}, +, \times, 0, 1 \rangle$. $K_{N\mathbb{N}}$-relations correspond to the bag semantics of relational tables. $R_3$ is a $K_{c\text{-table}}$ for the semiring $K_{c\text{-table}}=(\text{PosBool}(X), \lor, \land, \text{false}, \text{true})$, where $\text{PosBool}(X)$ is the set of all positive Boolean expressions over a finite set $X$ in which each two equivalent expressions are identified. $K_{c\text{-table}}$-relations correspond to the $c$-tables of Lipski and Imieliński [7]. $R_4$ is a $K_{\text{prob}}$-relation for the probabilistic semiring $K_{\text{prob}}=(\mathcal{P}(\Omega), \cup, \cap, \emptyset, \Omega)$, where $\Omega$ is a finite set of events. $K_{\text{prob}}$-relations correspond to the event tables [5, 11]. Finally, $R_5$ is a $K_{\text{prov}}$-relation for the provenance semiring $K_{\text{prov}}=(\mathcal{P}(\mathcal{P}(X)), \cup, \cap, \emptyset, \emptyset)$, where $X$ is a set source tuple id’s. $K_{\text{prov}}$-relations consist of tuples that are tagged with polynomials that “describe” the creation of the tuples in terms of source id’s in $X$. Other examples of semirings include: the lineage semiring $K_{\text{lin}}=(\mathcal{P}(X), \cup, \cap, \emptyset, \emptyset)$ where $X$ is again a finite set representing source tuple id’s. $K_{\text{lin}}$-relations consist of tuples that are “tagged” with the id’s of source tuples that contributed to the creation of that tuple; and finally the witness semiring $K_{\text{witness}}=(\mathcal{P}(\mathcal{P}(X)), \cup, \cap, \emptyset, \emptyset)$, where $X$ is again a set of source tuple id’s and $X \cup Y = \{ s \cup t | s \in X, t \in Y \}$.$^1$ $K_{\text{witness}}$-relations consist of tuples that are tagged with combinations of tuple id’s of the contributing source tuples. $
$
2.2 The query language $\mathcal{RA}_K^+$

We now recall the definition of the positive relational algebra on $K$-relations, denoted by $\mathcal{RA}_K^+$. Let $K=\langle \mathbb{K}, \oplus, \otimes, 0, 1 \rangle$ be a commutative semiring. Then $\mathcal{RA}_K^+$ includes the following operators:

**union** If $R_1, R_2 : U\dashv\text{Tup} \rightarrow \mathbb{K}$ then $R_1 \cup R_2 : U\dashv\text{Tup} \rightarrow \mathbb{K}$ is defined by

$$(R_1 \cup R_2)(t) = R_1(t) \oplus R_2(t).$$

**projection** If $R : U\dashv\text{Tup} \rightarrow \mathbb{K}$ and $V \subseteq U$ then $\pi_V(R) : V\dashv\text{Tup} \rightarrow \mathbb{K}$ is defined by

$$\pi_V(R)(t) = \bigoplus_{t = t' \text{ on } V \text{ and } R(t') \neq 0} R(t').$$

$^1$ This semiring was reported to us by James Cheney.
selection If $R:U\text{-Tup} \rightarrow \mathbb{K}$ and the selection predicate $P$ maps each $U$-tuple to either 0 or 1 depending on the equality or inequality of pairs of attributes, then $\sigma_P(R): U\text{-Tup} \rightarrow \mathbb{K}$ is defined by

$$(\sigma_P(R))(t) = R(t) \otimes P(t).$$

natural join If $R_i: U_i\text{-Tup} \rightarrow \mathbb{K}$, for $i=1,2$, then $R_1 \bowtie R_2$ is the $\mathbb{K}$-relation over $U_1 \cup U_2$ defined by

$$(R_1 \bowtie R_2)(t) = R_1(t_1) \otimes R_2(t_2),$$

where $t_1 = t$ on $U_1$ and $t_2 = t$ on $U_2$.

renaming If $R:U\text{-Tup} \rightarrow \mathbb{K}$ and $\beta: U \rightarrow U'$ is a bijection then $\rho_\beta(R)$ is the $\mathbb{K}$-relation over $U'$ defined by

$$(\rho_\beta R)(t) = R(t \circ \beta).$$

It is shown in [6] that the semantics of $\mathcal{RA}_K^+$ coincides with the positive relational algebra for semirings encountered in the database literature, i.e., for the semirings $\mathbb{K}_B$ [1], $\mathbb{K}_N$ [8], $\mathbb{K}_c$-tables [7] and $\mathbb{K}_{prob}$ [5, 11].

We recall the following two important properties of $\mathcal{RA}_K^+$ [6]:

Homomorphism property Let $\mathbb{K} = (\mathbb{K}, \oplus_{\mathbb{K}}, \otimes_{\mathbb{K}}, 0_{\mathbb{K}}, 1_{\mathbb{K}})$ and $\mathbb{K}' = (\mathbb{K}', \oplus_{\mathbb{K}'}, \otimes_{\mathbb{K}'}, 0_{\mathbb{K}'}, 1_{\mathbb{K}'})$ be two (commutative) semirings and let $h: \mathbb{K} \rightarrow \mathbb{K}'$ be a mapping. Then the transformation induced by $h$ from $\mathbb{K}$-relations to $\mathbb{K}'$-relations satisfies the property that $Q(h(R)) = h(Q(R))$ for any $Q \in \mathcal{RA}_K^+$ iff $h$ is a semiring homomorphism. That is, we have that $h(0_{\mathbb{K}}) = 0_{\mathbb{K}'}$, $h(1_{\mathbb{K}}) = 1_{\mathbb{K}'}$, and for any $x, y \in \mathbb{K}$, $h(x \oplus_{\mathbb{K}} y) = h(x) \oplus_{\mathbb{K}'} h(y)$ and $h(x \otimes_{\mathbb{K}} y) = h(x) \otimes_{\mathbb{K}'} h(y)$.

Factorization property Let $\mathbb{K}$ be a semiring, $R$ a $\mathbb{K}$-relation and $Q \in \mathcal{RA}_K^+$. Let $X$ be the set of tuple ids of the tuples in $R$. Let $ar{R}$ be the abstractly tagged version of $R$, obtained by letting $\bar{R}(t) = x$ iff $x$ is the tuple id of $t$ in $R$ and 0 otherwise. Let $v: X \rightarrow \mathbb{K}$ be the valuation that maps $x$ to $R(t)$ where $x$ is the tuple id of $t$. It is easily verified that there is unique homomorphism $\text{Eval}_v: \mathbb{N}[X] \rightarrow \mathbb{K}$ such that for one-variable monomials we have that $\text{Eval}_v(x) = v(x)$.

We then have that for all tuples $t$

$$Q(R)(t) = \text{Eval}_v \circ (Q(\bar{R}))(t).$$

In other words, the semantics of queries in $\mathcal{RA}_K^+$ over arbitrary semirings factors through its semantics in the provenance semiring.

3 BP-Completeness for query languages on $\mathbb{K}$-relations

In this section we initiate our study of BP-completeness in the setting of $\mathbb{K}$-relations. After defining the BP-completeness for query languages on $\mathbb{K}$-relations, we show that $\mathcal{RA}_K^+$ is not BP-complete. By looking at the bag case, we provide some insights on how to extend $\mathcal{RA}_K^+$ in order to obtain a BP-complete query language for $\mathbb{K}$-relations.

As stated in the introduction, BP-completeness establishes a language-independent criterion for completeness of relational algebra, based on symmetry properties of databases. The symmetries of classical databases are exactly captured by their automorphisms. Specifically, two values are “indistinguishable” in a database instance if the
In order to formulate BP-completeness in the setting of $K$-relations, we have to reconsider the notion of automorphisms. This is illustrated by the following example.

Example 2. Consider the $K_0$-relations given in Figure 2. When viewed as ordinary relations, i.e., by just looking at their support, it is clear that $R_3$ is not definable from $R_1$. Indeed, the mapping $\{a\mapsto b, b\mapsto a\}$ is an automorphism of $R_1$ but not of $R_3$ (any relation $R'_1$ definable from $R_1$ that contains the tuple $(b,b)$ must also contain the tuple $(a,a)$). Similarly, one can show that $R_3$ is not definable from $R_2$. Still regarded as ordinary relations, $R_1$ and $R_5$ have the same set of automorphisms and are therefore mutually definable in the classical sense. Nevertheless, when regarded as $K$-relations, the values $a$ and $b$ are distinguishable in $R_5$ since the tuples $(a,a),(b,b)$ have distinct multiplicities (i.e., $R_5((a,a)) \neq R_5((b,b))$). In contrast, in $R_5$ these two tuples are indistinguishable. This shows that the classical notion of automorphism needs to be extended in order to capture the symmetries of $K$-relations. Specifically, an automorphism of a $K$-relation should be an automorphism of its support (i.e., an automorphism in the classical sense) and it should preserve the $K$-values of the tuples. With this new notion of automorphism in mind, $R_5$ is indeed not definable from $R_1$. Similarly, $R_4$ is not definable from $R_2$ whereas $R_4$ is definable from $R_1$ and $R_5$ is definable from $R_2$. \[\square\]

We are now ready to formally extend BP-completeness and related definitions to $K$-relations.

**Definition 1.** Let $S$ be a classical relation. The set of automorphisms of $S$, denoted by $\text{Aut}(S)$, consists of all permutations $h$ of $\text{adom}(S)$ such that $h(i) \in S$ iff $i \in S$. Let $R$ be a $K$-relation. The set of automorphisms of $R$, denoted by $\text{Aut}_K(R)$, is defined as

$$\text{Aut}_K(R) = \{ h | h \in \text{Aut}(\text{supp}(R)) \text{ and } R(h(t)) = R(t), \forall t \in \text{supp}(R) \}.$$  

**Definition 2.** Let $R$ be a $K$-relation. The set of $K$-relations that are definable from $R$, denoted by $\text{DEF}(R)$, is defined as:

$$\text{DEF}(R) = \{ S | \text{adom}(S) \subseteq \text{adom}(R) \text{ and } \text{Aut}_K(R) \subseteq \text{Aut}_K(S) \}.$$  

Now, the expressiveness of a query language can be described in terms of the information that can be deduced from a $K$-relation using queries. Thus, by extending the notion of [10], we have:

**Definition 3.** Let $Q$ be a query language, and $R$ a $K$-relation. The basic information of $R$ with respect to $Q$ is the set of $K$-relations:

$$\text{BI}_Q(R) = \{ S | Q(R) = S \text{ for some generic query } Q \in Q \}.$$  

}\[\text{Table 2. Examples of definable and non-definable } K\text{-relations.}\]  

\[\begin{array}{|c|c|}
\hline
R_1 & \begin{pmatrix} A & B \\ a & a \\ b & b \end{pmatrix} \\
\hline
R_2 & \begin{pmatrix} A & B \\ a & a \\ b & b \end{pmatrix} \\
\hline
R_3 & \begin{pmatrix} A & B \\ a & a \\ b & b \end{pmatrix} \\
\hline
R_4 & \begin{pmatrix} A & B \\ a & a \\ b & b \end{pmatrix} \\
\hline
R_5 & \begin{pmatrix} A & B \\ a & a \\ b & b \end{pmatrix} \\
\hline
\end{array}\]
Finally, the BP-completeness for $\mathcal{K}$-relations links the notions of basic information and definable relations together:

**Definition 4.** A query language $Q$ is $BP$-complete if $\text{Bl}_Q(R) = \text{DEF}(R)$.

It is worth noting that the above definitions coincide with the standard definitions of automorphisms, definability and BP-completeness in the relational setting under the sets semantics, when considering $\mathcal{K} = \mathcal{K}_{\mathbb{B}}$.

We are now interested in establishing BP-completeness for $\mathcal{R}_\mathcal{K}^+$. A straightforward induction on the structure of queries in $\mathcal{R}_\mathcal{K}^+$ shows that the inclusion of $\text{Bl}_{\mathcal{R}_\mathcal{K}^+}(R) \subseteq \text{DEF}(R)$ holds for any semiring $\mathcal{K}$ and $\mathcal{K}$-relation $R$:

**Lemma 1.** For any semiring $\mathcal{K}$, any $Q \in \mathcal{R}_\mathcal{K}^+$ and any $\mathcal{K}$-relation $R$, we have that (i) $\text{adom}(Q(R)) \subseteq \text{adom}(R)$ and (ii) $\text{Aut}_\mathcal{K}(R) \subseteq \text{Aut}_\mathcal{K}(Q(R))$.

The other direction, i.e., whether $\text{DEF}(R) \subseteq \text{Bl}_{\mathcal{R}_\mathcal{K}^+}(R)$ for any semiring $\mathcal{K}$ and $\mathcal{K}$-relation $R$, does not hold. Indeed, the following example provides a counterexample for the semiring $\mathcal{K} = (\mathbb{N}, +, \times, \div, 0, 1)$.

**Example 3.** Consider again the $\mathcal{K}$-relations given in Fig. 2. As already noticed, $R_4$ is definable from $R_1$. However, it is easily verified, by induction on the structure of queries, that for every query $Q \in \mathcal{R}_\mathcal{K}^+$, $Q(R_1)$ is either: (i) empty, or (ii) the empty tuple, or (iii) such that it contains only tuples having *even* multiplicity. In other words, $R_4$ cannot be the result of a query in $\mathcal{R}_\mathcal{K}^+$, i.e. $R_4 \notin \text{Bl}_{\mathcal{R}_\mathcal{K}^+}(R_1)$.

From the previous example it follows:

**Proposition 1.** There exists a semiring $\mathcal{K}$ such that $\mathcal{R}_\mathcal{K}^+$ is not BP-complete on $\mathcal{K}$-relations.

In fact, the counterexample of Example 3 shows that more expressive power is needed to make $\mathcal{R}_\mathcal{K}^+$ BP-complete for the semiring $\mathcal{K} = (\mathbb{N}, +, \times, \div, 0, 1)$. To gain some intuition of what extra power is needed, let us consider the bag semiring (i.e., $\mathcal{K} = \mathcal{K}_{\mathbb{B}}$). Motivated by the literature on bag query languages [8], a natural choice is to add *duplicate elimination* to $\mathcal{R}_\mathcal{K}^+$. The duplicate elimination operator $\delta$, when applied on a $\mathcal{K}$-relation $R$, is defined as $\delta(R)(t) = 1$ for all $t \in \text{supp}(R)$. It is easy to see that considering the language $\mathcal{R}_\mathcal{K}^{+;\delta}$ obtained by adding $\delta$ to $\mathcal{R}_\mathcal{K}^+$, would resolve the specific counterexample of Example 3. Indeed, $R_4 = \delta(R_1)$ and therefore $R_4 \in \text{Bl}_{\mathcal{R}_\mathcal{K}^{+;\delta}}(R_1)$.

Still considering the bag case, it turns out that the combination $\mathcal{R}_\mathcal{K}^{+;\delta}$ of the positive algebra with duplicate elimination is neither BP-complete. This is illustrated by the following example.

**Example 4.** Consider again the $\mathcal{K}$-relations of Figure 2. It was already noticed that $R_5$ is definable from $R_2$. However, it is easily verified that for any $Q \in \mathcal{R}_\mathcal{K}^{+;\delta}$, the tuples in $Q(R_2)$ satisfy the property that for any two tuples $t_1$ and $t_2$ in $Q(R_2)$, $t_1$ occurs with less or equal multiplicity than $t_2$ if and only if $t_1$ contains a less or equal number of $b$’s than $t_2$. Hence, $R_5 \notin \text{Bl}_{\mathcal{R}_\mathcal{K}^{+;\delta}}(R_2)$.

As anticipated, from the previous example it follows that:
**Proposition 2.** There exists a semiring \( K \) such that \( \mathcal{R}A_K^+ \delta \) is not BP-complete on \( K \)-relations.

Finally, still focusing on the bag case, one may wonder whether the addition of the difference operator to \( \mathcal{R}A_K^+ \delta \) would help. The following example shows that adding it does look like a step towards the “right” direction. Let \( \mathcal{R}A_K^* \) be the relational algebra with difference and duplicate elimination.

**Example 5.** It is easy to check that the following holds:

\[
R_5 = Q(R_2) = (((\delta(R_2) \cup \delta(R_2)) - R_2) \cup \delta(R_2)).
\]

In other words, in this case, we have that \( R_5 \in BI_{\mathcal{R}A_K^*}(R_1) \).

Actually, we will see in Section 5 that the fact that \( \mathcal{R}A_K^* \) resolves both counterexamples 3 and 4 is not a coincidence.

### 4 Adding difference to \( \mathcal{R}A_K^+ \)

Motivated by the observation at the end of the previous section, we now investigate under which conditions on semirings one can naturally introduce the difference operator on \( K \)-relations. More specifically, we first identify a large class of semirings that can be equipped with a monus operator \( \ominus \). The addition of the monus operator on semirings \( K \) will then allow to extend \( \mathcal{R}A_K^+ \) with a difference operator, resulting in the relational algebra \( \mathcal{R}A_K^\ominus \) over \( K \)-relations.

#### 4.1 Semirings with monus

We follow here the standard approach for introducing \( \ominus \) into additive commutative monoids [2]. As we will see shortly, when introducing \( \ominus \) one has to pose some restriction on the class of semirings. More specifically, we first assume that \( K \) is naturally ordered. That is, the quasi-order \( x \prec y \) on \( K \) defined as \( x \prec y \) iff there exists a \( z \in K \) such that \( x \oplus z = y \), must define a partial order on \( K \). All semirings described in Section 2 are naturally ordered. We additionally require the following property (†): for each pair of elements \( x, y \in K \), the set \( \{ z \in K \mid x \prec y \oplus z \} \) has a smallest element. Note that the fact that \( \prec \) defines a partial order guarantees that \( \{ z \in K \mid x \prec y \oplus z \} \) has a unique smallest element, provided that it exists.

**Definition 5.** Let \( K \) be a naturally ordered semiring that satisfies property (†). For any \( x, y \in K \), we define \( x \ominus y \) to be the smallest element \( z \) such that \( x \prec y \oplus z \).

From here on, we call a commutative semiring \( K \) which can be equipped with a monus operator an \( m \)-semiring. We next revisit the semirings described in Section 2 and see whether they can be extended to \( m \)-semirings.

**Example 6.** It is easily verified that for \( K_B \), \( 1 \ominus 0 = 1 \), \( 1 \ominus 1 = 0 \), \( 0 \ominus 1 = 0 \) and \( 0 \ominus 0 = 0 \). In other words, \( K_B \) can be turned into an \( m \)-semiring and moreover, \( K_{\overline{B}} \) becomes a Boolean algebra. For the lineage semiring \( K_{lin} \), \( s \ominus t \) for \( s, t \in \mathcal{P}(X) \) coincides with
the set difference $s\setminus t$. Similarly, for the witness semiring $K_{\text{witness}}$. Here, $s\odot t$ for $s,t\in P(\mathcal{P}(X))$ coincides with the set difference $s\setminus t$. Finally, the monus operator coincides with set difference in the case of $K_{\text{prob}}$ as well. Moreover, the addition of $\oplus$ turns $K_{\text{lin}}$, $K_{\text{witness}}$ and $K_{\text{prob}}$ into Boolean algebras.

For $K_\mathbb{N}$, $x\odot y$ clearly coincides with the truncated minus $x \div y = \max\{0, x - y\}$ on $\mathbb{N}$. We next consider $K_{\text{prob}}$ for $\mathbb{N}[X]$ with $X = \{x_1, \ldots, x_n\}$. Let $\alpha$ denote some element in $\mathbb{N}^n$. If $\alpha = (a_1, \ldots, a_n)$, then $x^\alpha$ stands for the monomial $x_1^{a_1}x_2^{a_2}\cdots x_n^{a_n}$. Given two polynomials $f = \sum_{\alpha \in I} f_\alpha x^\alpha$ and $g = \sum_{\alpha \in J} g_\alpha x^\alpha$ it is easily verified that $f \oplus g = \sum_{\alpha \in I} (f_\alpha - g_\alpha) x^\alpha$.

The fact that $\oplus$ can be defined for $K_{\mathbb{B}}$, $K_{\text{lin}}$, $K_{\text{witness}}$ and $K_{\text{prob}}$, and turns these semirings into Boolean algebras is not a coincidence. Indeed, it can be shown that any additive commutative monoid $K = (\mathbb{K}, \oplus, 0)$ that can be equipped with $\odot$ and satisfies the following two conditions:

\begin{enumerate}
  \item[(B1)] The addition is idempotent, i.e., for any $x \in \mathbb{K}$, $x + x = x$; and
  \item[(B2)] For any $x, y \in \mathbb{K}$, $x \odot (x \odot y) = y \odot (y \odot x)$,
\end{enumerate}

is in fact either a Boolean algebra or a prime ideal of a Boolean algebra [2]. Observe that both conditions are desirable to hold for any definition of $\odot$ on $K_{\mathbb{B}}$, $K_{\text{lin}}$, $K_{\text{witness}}$ and $K_{\text{prob}}$. Moreover, these two conditions should also hold for any sensible definition of monus on $K_{\text{c-table}}$. The above result, however, shows that this is impossible as this semiring is neither a Boolean algebra nor a prime ideal thereof (it is just a distributive lattices but not every element has a unique complement). The following example shows how to modify $K_{\text{c-table}}$ such that it can be equipped with a monus.

**Example 7.** The semiring $K_{\text{c-table}} = (\text{PosBool}(X), \lor, \land, \text{false}, \text{true})$ was originally defined for positive queries only and therefore only positive Boolean expressions over $X$ where allowed [6]. The original definition of $c$-tables, however, does allow for arbitrary Boolean expressions [7]. Hence, we redefine the semiring $K_{\text{c-table}}$ as $(\text{Bool}(X), \lor, \land, \text{false}, \text{true})$, where $\text{Bool}(X)$ is the set of Boolean expressions over $X$ in which any two equivalent expressions are identified. This semiring allows for the definition of $\odot$. Indeed, for any two expressions $\phi_1, \phi_2$ in $\text{Bool}(X)$, we have that $\phi_1 \odot \phi_2$ is a Boolean expression that is equivalent to $\phi_1 \land \neg \phi_2$. Observe that the addition of $\oplus$ to $K_{\text{c-table}}$ again turns it into a Boolean algebra.

\[\Box\]

4.2 **The difference operator**

We are now ready to extend $\mathcal{R}\mathcal{A}^+_K$ with the difference operator. Let $K$ be an $m$-semiring. Then, we obtain $\mathcal{R}\mathcal{A}^+_K$ by extending $\mathcal{R}\mathcal{A}^+_K$ with the operator $\text{difference}$

\[
\text{difference } \text{If } R_1, R_2 : U\text{-Tup} \to K \text{ then } R_1 \ominus R_2 : U\text{-Tup} \to K \text{ is defined by } (R_1 - R_2)(t) = R_1(t) \odot R_2(t). 
\]

From Examples 6 and 7 it immediately follows that $\mathcal{R}\mathcal{A}^+_K$ coincides with the (full) relational algebra on sets for $K_{\mathbb{B}}$, and the bag algebra with the monus operator for $K_{\text{lin}}$ [8]. Furthermore, in the case of $K_{\text{c-table}}$ it coincides with the semantics of the relational algebra on $c$-tables [7] and for $K_{\text{prob}}$ it coincides with the semantics of the relational algebra provided on event tables [5, 11].
5 Extending $\mathcal{RA}_K$ to a BP-complete language

Coming back to the observation made at the end of Section 3, we next investigate how to further equip the language $\mathcal{RA}_K$ with a duplicate elimination operator, and obtain the language $\mathcal{RA}_K^\ast$. Then we show the main result of the paper, i.e. we prove that $\mathcal{RA}_K^\ast$ is BP-complete.

5.1 Adding duplicate elimination to $\mathcal{RA}_K$

To introduce a duplicate elimination operator to $\mathcal{RA}_K$, we need to further restrict our attention to $m$-semirings $K = (K, \oplus, \otimes, \ominus, 0, 1)$ that are finitely generated, i.e. every element in $K$ can be written as a finite sequence of sums, monus and products of a finite set of elements $k_1, \ldots, k_m$, called generators. We denote the set of generators of $K$ by $\text{Gen}(K)$ and assume it is minimal.

Example 8. The $m$-semirings considered so far are all finitely generated. Indeed, it is easily verified that $\text{Gen}(\mathbb{B}) = \{\text{true}\}$; $\text{Gen}(\mathbb{N}) = \{1\}$; $\text{Gen}(\text{Bool}(X)) = X$; $\text{Gen}(\mathcal{P}(\Omega)) = \Omega$; $\text{Gen}(\mathcal{P}(X)) = X$; $\text{Gen}(\mathbb{N}[X]) = \{1\} \cup X$; and finally $\text{Gen}(\mathcal{P}(\mathcal{P}(X))) = \{\mathcal{P}(X)\}$. An example of an $m$-semiring that is not finitely generated is $(\mathbb{R}, +, \times, -, 0, 1)$. Such semirings can be seen as generated by infinitely many (in this case uncountably many) generators, but we omit such semirings from our discussion.

We now formally define the notion of duplicate elimination. Given a finitely generated $m$-semiring $K = (K, \oplus, \otimes, \ominus, 0, 1)$ with generators $\text{Gen}(K) = \{k_1, \ldots, k_m\}$, we define the following set duplicate elimination operators:

**duplicate elimination** If $R: U\to \mathbb{K}$ and $k_i$ is a generator of $K$ then $\delta_{k_i}(R)$ is the $K$-relation over $U$ defined by

$$(\delta_{k_i}(R))(t) = k_i$$

We denote by $\mathcal{RA}_K^\ast$ the query language obtained by extending $\mathcal{RA}_K$ with the duplicate elimination operators for the $m$-semiring $K$ under consideration. Observe that when $K$ is the Boolean $m$-semiring, i.e., when considering the standard relational algebra with the set semantics, $\delta_{k_i}(R) = R$ is just the identity mapping. Hence, duplicate elimination does not add expressive power in this case.

5.2 BP-Completeness of $\mathcal{RA}_K^\ast$

We are now finally ready to state the main result of our work: The addition of difference and duplicate elimination to $\mathcal{RA}_K^\ast$ turned it into a BP-complete query language on $K$-relations.

**Theorem 1.** $\mathcal{RA}_K^\ast$ is BP-complete on $K$-relations for arbitrary finitely generated $m$-semirings $K$. 
Proof. (Sketch) The proof is constructive by nature. That is, given two $\mathcal{K}$ relations $R_1$ and $R_2$, the proof of Theorem 1 provides the construction of a $\mathcal{R}\mathcal{A}_\mathcal{K}$-query $Q$ such that $Q(R_1) = R_2$ provided that $\text{adom}(R_2) \subseteq \text{adom}(R_1)$ and $\text{Aut}_\mathcal{K}(R_1) \subseteq \text{Aut}_\mathcal{K}(R_2)$. It is interesting to observe that in case of $\mathcal{K} = \mathbb{N}_2$, i.e., when considering the standard relational algebra with the set semantics, the construction of $Q$ reduces to the construction given by Paredaens [10]. More specifically, neither difference nor duplicate elimination are needed in this case to obtain BP-completeness, in accordance with [10].

6 Homomorphism and factorization properties of $\mathcal{R}\mathcal{A}_\mathcal{K}$

We next investigate the semantics of the evaluation of queries in the presence of the difference operator. Specifically, we revisit the homomorphism and factorization properties (see Section 2) in $\mathcal{R}\mathcal{A}_\mathcal{K}$.

Let $\mathcal{K} = (\mathbb{K}, \oplus, \ominus, \otimes, 0, 1)$ and $\mathcal{K}' = (\mathbb{K}', \oplus', \ominus', \otimes', 0', 1')$ be two m-semirings. A mapping $h : \mathbb{K} \rightarrow \mathbb{K}'$ is an m-homomorphism if it is a semiring homomorphism and moreover it also preserves $\ominus$, i.e, for any two elements $x, y \in \mathbb{K}$ we have that $h(x) \ominus h(y) = h(x \ominus y)$.

Proposition 3. Let $\mathcal{K}$ and $\mathcal{K}'$ be two m-semirings. Let $h : \mathbb{K} \rightarrow \mathbb{K}'$ be a mapping. Then the transformation given by $h$ from $\mathcal{K}$-relations to $\mathcal{K}'$-relations commutes with any $\mathcal{R}\mathcal{A}_\mathcal{K}$-query $Q$, i.e., $Q(h(R)) = h(Q(R))$ if and only if $h$ is an m-homomorphism. 

In contrast to $\mathcal{R}\mathcal{A}_\mathcal{K}$, the following example shows that $\mathcal{K}'_{\text{prov}} = (\mathbb{N}[X], +, \times, \ominus, 0, 1)$ does not capture the semantics of queries in $\mathcal{R}\mathcal{A}_\mathcal{K}$ over arbitrary m-semirings.

Example 9. Let $\mathcal{R}$ be the $\mathbb{N}_2$-relation consisting of the tuple $(a)$ and such that $\mathcal{R}(a) = 2$, i.e., its multiplicity equals two. Consider the query $Q = (\mathcal{R} \ominus \mathcal{R}) - \mathcal{R}$. It is easily verified that $Q(R)$ consists of a single tuple $(a)$ and $Q(R)(a) = 2$. In contrast, the evaluation of $Q$ on the abstractly tagged relation $\bar{R}$, consisting of the tuple $(a)$ tagged with the variable $x$, results in the tuple $(a)$ tagged with the polynomial $x^2$. Since the valuation $v$ maps $x$ to 2, it is not the case that $Q(\bar{R}) = \text{Eval}_v(Q(\bar{R}))$. Indeed, $\text{Eval}_v(Q(\bar{R})(a)) = v(x)^2 = 2^2 = 4$ while $Q(\bar{R})(a) = 2$. The same counterexample works when we regard $\bar{R}$ as a $\mathcal{K}_2$-relation. Indeed, in this case $Q(D)(a) = \text{false}$, i.e., $(a)$ is not in the query result, while $\text{Eval}_v(Q(D)(a)) = \text{true} \land \text{true} = \text{true}$.

In the remainder of this section, we establish factorization results for two large classes of m-semirings that cover all m-semiring examples seen so far. We distinguish between the following two cases: (i) $\mathcal{K}$ is an m-semiring that is Boolean algebra or a prime ideal of a Boolean algebra; or (ii) $\mathcal{K}$ is an m-semiring that is cancellative, i.e., it satisfies the condition that for all $x, y \in \mathbb{K}$, $(x \oplus y) \ominus x = y$.

The first class of m-semirings includes all m-semirings considered so far, except for $\mathcal{K}_N$ and $\mathcal{K}_{\text{prov}}$. It is easily verified, however, that both $\mathcal{K}_N$ and $\mathcal{K}_{\text{prov}}$ fall into the second class of m-semirings. We next treat these two classes separately.

Let $\mathcal{K}$ be an m-semiring that is a Boolean algebra or a prime ideal of a Boolean algebra. Let $X = \{x_1, \ldots, x_k\}$ and $v : X \rightarrow \mathbb{K}$ be a valuation of $X$ in $\mathbb{K}$. Let $\text{FrBool}(X)$

\footnote{We slightly abuse notation here since $Q$ is regarded as both a $\mathcal{R}\mathcal{A}_\mathcal{K}$ and a $\mathcal{R}\mathcal{A}_\mathcal{K}'$-query.}
be the free Boolean algebra over $X$. Recall that FrBool$(X)$ can be realized by the set of Boolean functions on the variables in $X$ [4]. Moreover, it is known that FrBool$(X)$ has the universality property, i.e., any mapping $v : X \to \mathbb{K}$ can be uniquely extended to a homomorphism

$$\text{Eval}_v : \text{FrBool}(X) \to \mathbb{K},$$

that coincides with $v$ on $X$. It is easily verified that a homomorphism between Boolean algebras is an $m$-homomorphism when seen as $m$-semirings. From the universality of FrBool$(X)$ and Proposition 3 we get

**Proposition 4.** If $K$ is a Boolean algebra or a prime ideal of a Boolean algebra, then for any query $Q \in \mathcal{RA}_K$ and any $K$-relation $R$ with tuple id set $X$, $Q(\bar{R}) = \text{Eval}_v \circ Q(\bar{R})$, where $\bar{R}$ denotes the FrBool$(X)$-relation obtained by tagging each tuple in $R$ with its own tuple id.

**Example 10.** If we revisit Example 9 for the case that $K = \mathbb{K}_2$, then $Q(\bar{R})(a) = (x \land y) \land \neg x$ since we are now looking at $\bar{R}$ as a FrBool$(X)$-relation. It is clear that $\text{Eval}_v \circ Q(\bar{R})(a) = (\text{true} \land \text{true}) \land \text{false} = \text{false}$. In other words, $(a)$ is removed from the result, as desired. \hfill $\square$

Next, let $K$ be an $m$-semiring that is cancellative. The general theory of such $m$-semirings tells that they can always be seen as the positive cone of some lattice-ordered rings [9]. Here, for an ordered ring $(\mathcal{R}, \prec)$ the positive cone is defined as $\{ x \in \mathcal{R} | 0 \prec x \}$. In particular, we have that $\mathbb{N} = \mathbb{Z}^+$ and $\mathbb{N}[X] = (\mathbb{Z}[X])^+$. Now, given a cancellative $m$-semiring $K$, let $\mathcal{R}$ be the ring such that $K = \mathcal{R}^+$. It is known that the ring of polynomials $\mathcal{R}[X]$ has the universality property that for any other ring $\mathcal{S}$, such that there exists a (ring) homomorphism $h$ from $\mathcal{R}$ to $\mathcal{S}$, any valuation $v : X \to \mathcal{S}$ can be uniquely extended to a homomorphism

$$\text{Eval}_v : \mathcal{R}[X] \to \mathcal{S},$$

that coincides with $v$ on $X$ and with $h$ on $\mathcal{R}$. Due to the fact that we work over rings, it is easily verified that ring homomorphisms are $m$-homomorphisms. Indeed, let $h$ be a ring homomorphism. Then, $h(x \ominus y) = h(x \ominus (-y)) = h(x) \oplus h(-y) = h(x) \oplus (-h(y)) = h(x) \ominus h(y)$. Here, $-y$ denotes the unique additive inverse element of $y$ which exists since we work over rings.

The above universality property together with Proposition 3 gives us the following factorization property for $K_{\mathbb{N}}$ and $K_{\text{prov}}$.

**Theorem 2.** For $K = K_{\mathbb{N}}$ or $K = K_{\text{prov}}$, we have that for any query $Q \in \mathcal{RA}_K$ and any $K$-relation $R$ with tuple id set $X$, $Q(\bar{R})(t) = \text{max}(0, \text{Eval}_v \circ Q(\bar{R})(t))$, where $\bar{R}$ denotes the $\mathbb{Z}[X]$-relation obtained by tagging each tuple in $R$ with its own tuple id.

The crux here is that we do the computation in $\mathbb{Z}[X]$ but in case that we get negative numbers, we truncate them.

**Example 11.** If we revisit Example 9 for the case that $K = K_{\mathbb{N}}$, then $Q(\bar{R})(a) = x^2 - x$ since we are now looking at $\bar{R}$ as a $\mathbb{Z}[X]$-relation. It is clear that $\text{Eval}_v \circ Q(\bar{R})(a) = 2^2 - 2 = 2$ now gives the desired result. \hfill $\square$
7 Conclusion

Our main goal was to investigate a language-independent criterion for completeness of \( \mathcal{RA}_K^+ \) on \( K \)-relations, in the same spirit of BP-completeness of relational algebra. To this aim, we naturally extended BP-completeness to \( K \)-relations, and showed that for some semirings \( K \), however, \( \mathcal{RA}_K^+ \) is not BP-complete. We therefore extended \( \mathcal{RA}_K^+ \) with a difference operator, resulting in the query language \( \mathcal{RA}_K^* \), and additional duplicate elimination operators \( \delta \), resulting in the query language \( \mathcal{RA}_K^* \). Our main result is that \( \mathcal{RA}_K^* \) is BP-complete on \( K \)-relations for a general class of semirings \( K \). More specifically, \( \mathcal{RA}_K^* \) is BP-complete for semirings that can be extended with a monus operator and that are finitely generated. This class of semirings covers most of the semirings considered in the database literature so far. We showed that neither the difference nor duplicate elimination can be omitted while still retaining BP-completeness.

While in this paper we showed that both the homomorphism and factorization properties hold for \( \mathcal{RA}_K^* \), in future work, we plan to study whether these hold for \( \mathcal{RA}_K^* \) as well. Also, we aim at finding an exact characterization of when two \( K \)-relations are related by means of a query in \( \mathcal{RA}_K^* \). This will require to add some further restriction to BP-completeness as defined in the present work. Finally, we will investigate what is the exact expressive power of our query languages on \( K \)-relations.

Acknowledgments

The authors would like to thank James Cheney and Leonid Libkin for helpful discussions.

References