Containment of Conjunctive Queries on Annotated Relations

Todd J. Green
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
tjgreen@cis.upenn.edu

ABSTRACT

We study containment and equivalence of (unions of) conjunctive queries on relations annotated with elements of a commutative semiring. Such relations and the semantics of positive relational queries on them were introduced in a recent paper as a generalization of set semantics, bag semantics, incomplete databases, and databases annotated with various kinds of provenance information. We obtain positive decidability results and complexity characterizations for databases with lineage, why-provenance, and provenance polynomial annotations, for both conjunctive queries and unions of conjunctive queries. At least one of these results is surprising given that provenance polynomial annotations seem "more expressive" than bag semantics and under the latter, containment of unions of conjunctive queries is known to be undecidable. The decision procedures rely on interesting variations on the notion of containment mappings. We also show that for any positive semiring (a very large class) and conjunctive queries without self-joins, equivalence is the same as isomorphism.

1. INTRODUCTION

K-relations, which are relations whose tuples are annotated with elements from a commutative semiring K, were introduced in a recent paper [19] as a generalization of sets, bags, the Boolean c-tables used in incomplete databases [22, 20], probabilistic databases [15, 34], databases with lineage [13] or why-provenance [4] information, and other kinds of annotated relations. The semantics of positive relational algebra queries extends to K-relations via definitions in terms of the abstract "+" and "$\cdot$" operations of K. For $K = \mathbb{B}$, the Boolean semiring, this specializes to the usual set semantics, while for $K = \mathbb{N}$, the semiring of natural numbers, it is bag semantics.

The introduction of annotations on relations presents new challenges in query reformulation and optimization, however, as queries that are semantically *equivalent* when posed over ordinary relations may become *inequivalent* when posed over K-relations. Indeed, this phenomenon was already observed for the case of bag semantics [8, 23], where, e.g., adding a "redundant" self-join to a query actually changes the query's meaning. The need to compare query equivalence for different kinds of provenance annotations was also emphasized from early on in [4, 5] and reiterated in [3]. A central theme of this paper is to compare different provenance-annotated semantics among themselves and with the standard set and bag semantics. The comparison is done w.r.t. containment[1] and equivalence of conjunctive queries (CQs) and unions of conjunctive queries (UCQs), leading to four different hierarchies among these semantics. Whether the steps in these hierarchies are strict or not is always informative and sometimes surprising.

We consider in this paper five different kinds of provenance information that can be captured using semiring annotations. These range from the very simple data warehousing *lineage* of [13], in which a tuple in the output is annotated with a set of tuple ids of all "contributing" source tuples, to the *why-provenance* of [4], in which output tuples are annotated with a *set of sets* of contributing source tuples, to the *provenance polynomials* $\mathbb{N}[X]$ of [19], in which the annotations are polynomial expressions over the source tuple ids which fully "document" how an output tuple is produced in the result of a query. Provenance polynomials are as "general" as any other commutative semiring, hence this is the most informative form of provenance annotations. $\mathbb{N}[X]$-relations are not just of theoretical interest, but also have practical applications as the foundation of trust policies and incremental maintenance algorithms in systems for collaborative data sharing [18]. We also consider a new form of provenance, the *Boolean provenance polynomials* $\mathbb{B}[X]$, as well as the form of lineage used in the Trio project [30], which we show can also be captured using a semiring. These two forms of provenance are intermediate between why-provenance and $\mathbb{N}[X]$.

To illustrate, for the source database R and query Q

A B C	
$a\ b\ c$	p
$d\ b\ e$	r
$f\ g\ e$	s

$$Q(R) \overset{\text{def}}{=} \pi_{\text{AC}}\big(\pi_{\text{AB}}R \bowtie \pi_{\text{BC}}R \ \cup \ \pi_{\text{AC}}R \bowtie \pi_{\text{BC}}R \big)$$

(where p, r, and s are the tuple ids) the data warehousing lineage of (d, e) in the output is $\{r, s\}$, the why-provenance

[1]We define inclusion of K-relations by the *natural order* present in the semirings of interest to us (see Section 4).

ICDT 2009, March 23–25, 2009, Saint Petersburg, Russia.
Copyright 2009 ACM 978-1-60558-423-2/09/0003 ...$5.00

of (d, e) is $\{\{r\}, \{r, s\}\}$, the $\mathbb{B}[X]$-provenance is $r^2 + rs$, the Trio-style lineage is $r + rs$, and the $\mathbb{N}[X]$-provenance is $2r^2 + rs$. Thus, lineage tells us *which* source tuples were involved in producing a given output tuple; why-provenance tells us which *sets* of source tuples were involved in producing the output tuple; $\mathbb{B}[X]$ tells us which *bags* of source tuples were involved; Trio-style lineage tells us how many times a given *set* of source tuples was involved; and $\mathbb{N}[X]$-provenance tells us exactly *how* the tuple was produced from the source tuples. Note also that by "plugging in" numeric values for the variables (e.g., $p = 2, r = 3, s = 1$) and evaluating the $\mathbb{N}[X]$-provenance of an output tuple, we obtain the multiplicity of the tuple under bag semantics (e.g., 15 for (d, e)).

Another central theme of this paper is to establish the complexity of containment and equivalence of CQs/UCQs for various semirings. For the semirings $\mathbb{B}$ and $\mathbb{N}$ this corresponds to set, respectively bag semantics and the questions were studied in the past [6, 29, 8] as was the case of bag-set semantics [28, 11] (In section 7.5 we discuss the relationship between the latter and our results.) Results for an entire class of semirings (the distributive lattices) have already been established in [16, 19]. This paper focuses primarily on the provenance semirings.

A priori, it is not clear that containment and equivalence for queries on relations with provenance annotations should even be decidable, as bag containment is known to be undecidable for UCQs [23], and $\mathbb{N}[X]$ seems related to bags. Nevertheless, we are able to show that containment is decidable for all the forms of provenance annotations we consider, for both CQs and UCQs (with the exception of containment of UCQs with Trio-style lineage, which we leave open). We also establish interesting connections with the same problems for bag semantics. In particular our contributions are:

- We show that the various forms of provenance annotations we consider are related by *surjective semiring homomorphisms*, which yields easy bounds on their relative behavior with respect to query containment.
- We show that for UCQs, $\mathbb{N}[X]$- containment implies K-containment for *any* semiring K, and for any *positive* K (a very large class that includes all the semirings we consider in this paper, see Section 4), K-containment implies containment under the usual set semantics.
- For the case of CQs without self-joins, we show that for any positive K, K-equivalence is the same as *isomorphism*, and thus its complexity is complete for the class GI of problems polynomial time reducible to *graph isomorphism*.[2]
- We show that containment of CQs and UCQs is decidable for lineage, why-provenance, $\mathbb{B}[X]$, and $\mathbb{N}[X]$ annotations. The decision procedures involve interesting variations on the concept of *containment mappings*, or (in the case of $\mathbb{N}[X]$-containment of UCQs) establishing a *small counterexample property* (see Section 7.4). We also identify the complexity in each case as NP-complete (with the exception of $\mathbb{N}[X]$-containment of UCQs, where we give a PSPACE upper bound).
- We show that for why-provenance, $\mathbb{B}[X]$, and $\mathbb{N}[X]$, equivalence of CQs implies isomorphism, and the complexity is therefore somewhat lower than for containment (GI-complete). $\mathbb{N}[X]$-equivalence of UCQs is also shown to be the same as isomorphism and GI-complete. Lineage-equivalence of CQs and why-prov. and $\mathbb{B}[X]$-equivalence of UCQs are shown to remain NP-complete.
- We show that for CQs, why-prov. containment implies bag-containment, and bag-containment implies lineage-containment. We also show that for UCQs $\mathbb{N}[X]$-equivalence is the same as bag equivalence hence providing a proof that the latter is the same as isomorphism and therefore GI-complete.

[2]Graph isomorphism is known to be in NP, but is not known or believed to be either NP-complete or in PTIME, see [25].

Figure 1 summarizes the complexity results mentioned above (for completeness we include previously known results in the shaded boxes). Figure 2 summarizes the logical relationships for containment/equivalence among the various semirings we consider.

The rest of this paper is organized as follows. We define K-relations and the semantics of queries on them in Section 2. We define the various semirings for provenance in Section 3; we also establish there the existence of semiring homomorphisms relating the various models. We define containment of queries on K-relations in terms of the *natural order* in Section 4 and discuss the connections with semiring homomorphisms. We review the background concepts of containment mappings and canonical databases in Section 5. We derive the bounds on containment based on surjective semiring homomorphisms in Section 6. We present the main results on containment and equivalence in Section 7. We discuss related work in Section 8. Finally, we conclude with some ideas for future work in Section 9.

2. QUERIES ON K-RELATIONS

Fix a countable domain $\mathbb{D}$ of constants. Let $(K, +, \cdot, 0, 1)$ be a commutative semiring, i.e., $(K, +, 0)$ and $(K, \cdot, 1)$ are commutative monoids, $\cdot$ is distributive over $+$ and $\forall a$, $0 \cdot a = a \cdot 0 = 0$. An n-ary *K-relation* is a function $R : \mathbb{D}^n \to K$ such that its *support* defined by $\mathsf{supp}(R) \overset{\text{def}}{=} \{t : R(t) \neq 0\}$ is finite. A *K-instance* is a mapping from predicate symbols to K-relations.

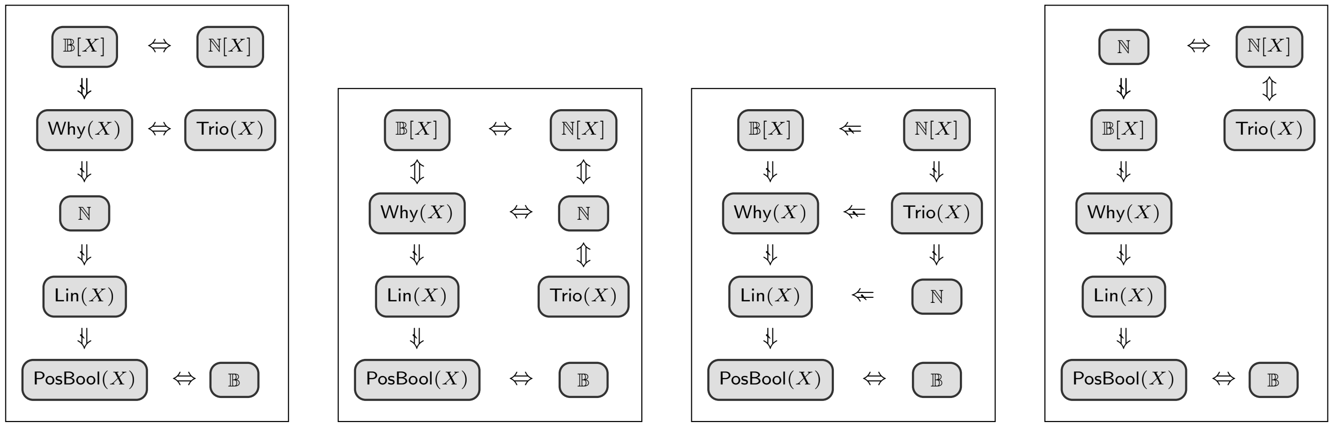
We use Datalog-style syntax for conjunctive queries and unions of conjunctive queries. A *conjunctive query* (CQ) is an expression of the form

$$Q(\bar{u}) \quad \text{:-} \quad R_1(\bar{u}_1), \ldots, R_n(\bar{u}_n)$$

where $Q(\bar{u})$ is the *head* of the query, denoted $\mathsf{head}(Q)$, the *multiset* (bag) of *atoms* $R_1(\bar{u}_1), \ldots, R_n(\bar{u}_n)$ is the *body* of the query, denoted $\mathsf{body}(Q)$, $\bar{u}$ is the tuple of *distinguished variables* and constants, $\bar{u}_1, \ldots, \bar{u}_n$ are tuples of variables and constants whose arities are consistent with their associated predicate symbols, and each variable appearing in the head also appears somewhere in the body. We denote the set of variables appearing in Q by $\mathsf{vars}(Q)$ and the set of constants

		$\mathbb{B}$	$\mathsf{PosBool}(X)$	$\mathsf{Lin}(X)$	$\mathsf{Why}(X)$	$\mathsf{Trio}(X)$	$\mathbb{B}[X]$	$\mathbb{N}[X]$	$\mathbb{N}$
CQs	cont	NP	NP	NP	NP	NP	NP	NP	? (Π_2^p-hard)
	equiv	NP	NP	NP	GI	GI	GI	GI	GI
UCQs	cont	NP	NP	NP	NP	?	NP	in PSPACE	undec
	equiv	NP	NP	NP	NP	GI	NP	GI	GI

Figure 1: Complexity of containment and equivalence. Non-shaded boxes indicate contributions of this paper. NP is short for NP-complete. GI is short for GI-complete (i.e., complete for the class of problems polynomial time reducible to graph isomorphism).



(a) CQ containment (b) CQ equivalence (c) UCQ containment (d) UCQ equivalence

Figure 2: Logical implications of containment and equivalence. $K_1 \Rightarrow K_2$ indicates that K_1-containment (equivalence) implies K_2-containment (equivalence). A ticked arrow "$\not\Rightarrow$" indicates that the implication is strict.

by $\mathsf{consts}(Q)$. When $\bar{u}$ is empty we say that Q is a *Boolean conjunctive query*; for these we will sometimes drop the parentheses in the head and write $Q \quad \text{:-} \quad R_1(\bar{u}_1), \ldots, R_n(\bar{u}_n)$. We say that a CQ has a *self-join* if some predicate symbol appears more than once in the body of a CQ.

A *union of conjunctive queries* (UCQ) is a bag $\bar{Q} = (Q_1, \ldots, Q_n)$ of CQs. The arities of the heads of the CQs in a UCQ must all agree.

The semantics of CQs on K-relations is based on the notion of *valuations*. A valuation is a function $\nu : \mathsf{vars}(Q) \to \mathbb{D}$ extended to be the identity on constants. Valuations operate component-wise on tuples in the expected way. Let Q be a CQ

$$Q(\bar{u}) \quad \text{:-} \quad R_1(\bar{u}_1), \ldots, R_n(\bar{u}_n)$$

and let I be a K-instance of the same schema. The *result* of evaluating Q on I is the K-relation

$$Q(I) \stackrel{\text{def}}{=} \lambda t. \sum_{\substack{\nu \text{ s.t.} \\ \nu(\bar{u})=t}} \mathsf{prod}_\nu^Q(I) \tag{1}$$

where $\mathsf{prod}_\nu^Q(I) \stackrel{\text{def}}{=} \prod_{i=1}^n R_i(\nu(\bar{u}_i))$ and the sums and products are in K. A valuation ν which maps $\bar{u}$ to t such that $\mathsf{prod}_\nu^Q(I) \neq 0$ is called a *derivation* of t, and we say that it *justifies* the associated product. The meaning of (1) is unchanged if we assume the sum ranges only over derivations of t.

We extend the semantics to UCQs as follows. If $\bar{Q} = (Q_1, \ldots, Q_n)$ is a UCQ, then the result of evaluating $\bar{Q}$ on a K-instance I is the K-relation

$$\bar{Q}(I) \stackrel{\text{def}}{=} \lambda t. \sum_{i=1}^n Q_i(I)(t)$$

For the commutative semiring $(\mathbb{B}, \vee, \wedge, \mathsf{false}, \mathsf{true})$ this specializes to the set semantics for UCQs. For $(\mathbb{N}, +, \cdot, 0, 1)$ it is bag semantics. For $(\mathsf{PosBool}(X), \vee, \wedge, \mathsf{false}, \mathsf{true})$ (see Section 3) it is the positive Boolean c-tables used in incomplete databases [22].

A subtlety in the preceding definitions is that we allow the same atom to appear multiple times in the body of a CQ (and similarly, we allow the same CQ to appear multiple times in a UCQ). With set semantics the distinction is immaterial, but for other K, where idempotence of multiplication and addition may not hold, the distinction does matter. The classic example is adding a "redundant" self-join to a query in the case of $K = \mathbb{N}$.

In contrast to repetitions, the order of atoms in the body of a CQ (and order of CQs in a UCQ) is not important, since we are considering only K-relations where K is *commutative* (cf. Proposition 3.4 in [19]). Thus the body of a CQ can be viewed a *bag* of atoms. When comparing the bodies of CQs, we will use the notation $\mathsf{body}(P) \leq_{\mathbb{N}} \mathsf{body}(Q)$ to mean

$$\begin{tikzcd}
& \mathbb{N}[X] \arrow[dl, no head] \arrow[dr, no head] & \\
\mathbb{B}[X] \arrow[dr, no head] & & \mathsf{Trio}(X) \arrow[dl, no head] \\
& \mathsf{Why}(X) \arrow[dl, no head] \arrow[dr, no head] & \\
\mathsf{Lin}(X) \arrow[dr, no head] & & \mathsf{PosBool}(X) \arrow[dl, no head] \\
& \mathbb{B} &
\end{tikzcd}$$

Figure 3: Provenance hierarchy. A path downward from K_1 to K_2 indicates that there exists a surjective semiring homomorphism $h : K_1 \to K_2$.

bag containment of the query bodies. We will also identify queries which are the same up to reordering of atoms in the body, i.e., $P = Q$ means $\mathsf{head}(P) = \mathsf{head}(Q)$, $\mathsf{body}(P) \leq_{\mathbb{N}} \mathsf{body}(Q)$, and $\mathsf{body}(Q) \leq_{\mathbb{N}} \mathsf{body}(P)$.

We use the notation $P \cong Q$ ($\bar{P} \cong \bar{Q}$) to denote that P and Q ($\bar{P}$ and $\bar{Q}$) are *isomorphic*, i.e., syntactically identical up to renaming of variables and reordering of terms (and, for UCQs, reordering of CQs).

3. SEMIRINGS FOR PROVENANCE

In this section we define several kinds of provenance annotations that can be captured in the semiring framework. We will also observe that the various models are related by *surjective semiring homomorphisms* (see Appendix for definition), as summarized in Figure 3. In Section 6, we will use the existence of surjective semiring homomorphisms to establish some basic relationships among the provenance models with respect to query containment.

We fix a countable set X of *variables*, which can be thought of as *tuple identifiers*, and parametrize all of the provenance models by this set X.

The most informative form of provenance annotations in the framework of K-relations is the semiring of *provenance polynomials* [19]:

Definition 3.1 (Provenance Polynomials). *The* provenance polynomials semiring *for X is the semiring of polynomials with variables from X and coefficients from $\mathbb{N}$, with the operations defined as usual:* $(\mathbb{N}[X], +, \cdot, 0, 1)$.

The provenance polynomials are the "most informative" among semiring annotation by dint of their *universality*: any function $\nu : X \to K$ (call it a "valuation") can be extended uniquely to a semiring homomorphism $\mathsf{Eval}_\nu : \mathbb{N}[X] \to K$. Intuitively, Eval_ν operates by assigning the value $\nu(x)$ to each variable x in a polynomial expression, then evaluating the resulting expression in K. Combined with the commutation with homomorphisms property (cf. Proposition 6.1), this allows the computations for any commutative semiring K to *factor* through the computations for the provenance polynomials (see [19]).

To illustrate, consider the $\mathbb{N}[X]$-relation R in Figure 4(a) and consider the UCQ $\bar{Q}$ defined by

$$\begin{aligned}
\bar{Q}(x, z) &\ \text{:-}\ R(x, y, u), R(v, y, z) \\
\bar{Q}(x, z) &\ \text{:-}\ R(x, u, z), R(v, y, z)
\end{aligned}$$

Figure 4(b) shows the result of $\bar{Q}$ applied to R.

The second provenance model we consider is obtained from the provenance polynomials by replacing natural number coefficients with Boolean coefficients:

Definition 3.2 (Boolean Provenance Polynomials). *The* Boolean provenance polynomials semiring *for X is the semiring of polynomials over variables X with Boolean coefficients:* $(\mathbb{B}[X], +, \cdot, 0, 1)$.

Considering the same UCQ $\bar{Q}$ as before, Figure 4(c) shows the result of applying $\bar{Q}$ to R, where R is interpreted as a $\mathbb{B}[X]$-relation. Note that the annotations in Figure 4(c) can be obtained from those in Figure 4(b) by simply dropping the numeric coefficients. In fact, one can check that the operation $f : \mathbb{N}[X] \to \mathbb{B}[X]$ which "drops coefficients" (i.e., by replacing non-zero coefficients with true) is a *surjective semiring homomorphism*.

The third provenance model we consider, $\mathsf{Trio}(X)$, is inspired by the form of lineage used in the Trio project [30]. Like $\mathbb{B}[X]$, this semiring can be viewed as being obtained from $\mathbb{N}[X]$, but instead of "dropping coefficients," this time we "drop exponents." We formalize this using the notion of *quotient semirings* (see Appendix for definition). Let $f : \mathbb{N}[X] \to \mathbb{N}[X]$ be the mapping that "drops exponents," e.g., f maps $2x^2y + 3xy + 2z^3 + 1$ to $5xy + 2z + 1$. Denote by $\approx_f$ the equivalence relation on $\mathbb{N}[X]$ defined by $a \approx_f b \overset{\text{def}}{\Longleftrightarrow} f(a) = f(b)$. One can check that $\approx_f$ is a *congruence relation* (see Appendix for definition). This justifies the following:

Definition 3.3 (Trio Semiring). *The* Trio semiring *for X is the* quotient semiring *of* $\mathbb{N}[X]$ *by* $\approx_f$*, denoted* $\mathsf{Trio}(X)$.

As an example, considering again the same UCQ $\bar{Q}$, Figure 4(d) shows the result of applying $\bar{Q}$ to R, where R is interpreted as a $\mathsf{Trio}(X)$-relation, and an annotation A is understood to represent its equivalence class $A/\approx_f$ in $\approx_f$. Note that the mapping $h : \mathbb{N}[X] \to \mathsf{Trio}(X)$ defined by $h(A) \mapsto A/\approx_f$ is a surjective semiring homomorphism.

The fourth provenance model we consider is the *why-provenance* of [4]. The why-provenance of a tuple is the *set of sets* of "contributing" source tuples, which is called the *proof witness basis* in [4]. This can be captured using a semiring [3] (called the *proof why-provenance* semiring in [3]):

Definition 3.4 (Why-Provenance). *The* why-provenance semiring *for X is* $(\mathsf{Why}(X), \cup, \Cup, \emptyset, \{\emptyset\})$ *where* $\mathsf{Why}(X) \overset{def}{=}$

a	b	c	p
d	b	e	r
f	g	e	s

(a) Source R

a	c	$2p^2$
a	e	pr
d	c	pr
d	e	$2r^2 + rs$
f	e	$2s^2 + rs$

(b) $\mathbb{N}[X]$

a	c	p^2
a	e	pr
d	c	pr
d	e	$r^2 + rs$
f	e	$s^2 + rs$

(c) $\mathbb{B}[X]$

a	c	$2p$
a	e	pr
d	c	pr
d	e	$2r + rs$
f	e	$2s + rs$

(d) $\mathsf{Trio}(X)$

a	c	$\{\{p\}\}$
a	e	$\{\{p,r\}\}$
d	c	$\{\{p,r\}\}$
d	e	$\{\{r\},\{r,s\}\}$
f	e	$\{\{s\},\{r,s\}\}$

(e) $\mathsf{Why}(X)$

a	c	p
a	e	$p \wedge r$
d	c	$p \wedge r$
d	e	r
f	e	s

(f) $\mathsf{PosBool}(X)$

a	c	$\{p\}$
a	e	$\{p,r\}$
d	c	$\{p,r\}$
d	e	$\{r,s\}$
f	e	$\{r,s\}$

(g) $\mathsf{Lin}(X)$

Figure 4: Provenance Annotations

$\mathcal{P}(\mathcal{P}(X))$ *and* $\Cup$ *denotes* pairwise union*:* $A \Cup B \overset{def}{=} \{a \cup b : a \in A, b \in B\}$

Considering again the same query $\bar{Q}$, we can interpret the source relation in Figure 4(a) as a why-provenance relation by doubly-nesting the variables (e.g., p becomes $\{\{p\}\}$). Figure 4(e) shows the query output and the resulting why-provenance annotations. Note that these annotations can be obtained from the $\mathbb{B}[X]$-annotations by dropping exponents (and writing the result as a set of sets rather than sum of monomials). One can check that the corresponding operation $g : \mathbb{B}[X] \to \mathsf{Why}(X)$ which "drops exponents" is in fact a surjective semiring homomorphism. Note also that the annotations can be obtained from the $\mathsf{Trio}(X)$-annotations by dropping coefficients, and it is easy to verify that the corresponding operation $h : \mathsf{Trio}(X) \to \mathsf{Why}(X)$ which does this is also a surjective semiring homomorphism.

An interesting variation on the why-provenance semiring is obtained by requiring that the witness basis for an output tuple be *minimal*. Here the domain is $\mathsf{irr}(\mathcal{P}(X))$ the set of *irredundant* subsets of $\mathcal{P}(X)$, i.e., W is in $\mathsf{irr}(\mathcal{P}(X))$ if for any A, B in W neither is a subset of the other. We can associate with any $W \subseteq \mathcal{P}(X)$ a unique irredundant subset $\mathsf{irr}(W)$ by repeatedly looking for elements A, B such that $A \subseteq B$ and deleting B from W. Then we define a semiring $(\mathsf{irr}(\mathcal{P}(X)), +, \cdot, 0, 1)$ as follows:

$$\begin{array}{rclcrcl} I + J & \overset{\text{def}}{=} & \mathsf{irr}(I \cup J) & \quad & I \cdot J & \overset{\text{def}}{=} & \mathsf{irr}(I \Cup J) \\ 0 & \overset{\text{def}}{=} & \emptyset & & 1 & \overset{\text{def}}{=} & \{\emptyset\} \end{array}$$

This is the semiring in which we compute the *minimal witness basis* [4]. It is a well-known semiring: the construction above is the construction for the free distributive lattice generated by the set X. Moreover, it is isomorphic to the semiring of *positive Boolean expressions* $(\mathsf{PosBool}(X), \vee, \wedge, \mathsf{false}, \mathsf{true})$ used in *incomplete databases* [22].[3] The domain of this semiring is the set of all Boolean expressions over variables X which are *positive*, i.e., they involve only disjunction, conjunction, and constants for true and false.[4]

[3]This characterization of minimal witness basis and its relationship to $\mathsf{PosBool}(X)$ are due to Val Tannen.

[4]Also, we identify those expressions that are equivalent modulo the axioms of Boolean algebra.

Containment of UCQs for $\mathsf{PosBool}(X)$ is known to coincide with containment under the usual set semantics:[5]

THEOREM 3.5 ([16]). *If K is a distributive lattice then for any UCQs $\bar{P}, \bar{Q}$*

$$\bar{P} \sqsubseteq_K \bar{Q} \quad \textit{iff} \quad \bar{P} \sqsubseteq_{\mathbb{B}} \bar{Q}$$

$\mathsf{PosBool}(X)$ is a distributive lattice, so Theorem 3.5 justifies the "$\Leftrightarrow$" between $\mathbb{B}$ and $\mathsf{PosBool}(X)$ in the diagrams in Figure 2. Other interesting examples of annotations from distributive lattices include the semiring of full Boolean expressions (including negation), the *fuzzy semiring* [19], and finite total orders such as the semiring of *security clearances* proposed in [14].

Taking again the same query $\bar{Q}$ and applying it to the source table in Figure 4(a) viewed as a $\mathsf{PosBool}(X)$-relation, we obtain the $\mathsf{PosBool}(X)$-relation shown in Figure 4(f).

The last and simplest form of provenance information we consider is the data warehousing *lineage* of [13]. In this scheme, a tuple t in a query output is annotated with the *set* of all contributing source tuples (its *lineage*). This can be captured using the following semiring [3]:

DEFINITION 3.6 (LINEAGE SEMIRING). *The* lineage semiring *for X is $(\mathcal{P}(X) \cup \{\bot\}, +, \cdot, \bot, \emptyset)$ where X is a set of variables, $\bot + S = S + \bot = S$, $\bot \cdot S = S \cdot \bot = \bot$, and $S + T = S \cdot T = S \cup T$ if $S, T \neq \bot$.*

We can interpret the source relation in Figure 4(a) as a lineage annotated relation by nesting the annotations, e.g., p becomes $\{p\}$. Applying the same query $\bar{Q}$ as before to this relation, we obtain the lineage annotated relation shown in Figure 4(g). Note that the lineage for an output tuple can be obtained from the why-provenance of the tuple by *flattening*

[5]This result was claimed in [19], but Gösta Grahne recently pointed out to the author that [16] had already proved this in a more general form, for queries on relations annotated with elements of a *distributive bilattice*. Related results have also been established in the contexts of *parametric databases* [26] and *deterministic XML* [4].

the set of sets, i.e., applying the function $h : \mathsf{Why}(X) \to \mathsf{Lin}(X)$ defined by $h(I) = \bigcup_{S \in I} S$. Once again, we can show that h is a surjective semiring homomorphism.

4. THE NATURAL ORDER

We define containment of K-relations and queries over K-instances in terms of the *natural order.* Let $(K, +, \cdot, 0, 1)$ be a semiring and define $a \leq b \overset{\text{def}}{\iff} \exists c\ a + c = b$. When $\leq$ is a partial order we say that K is *naturally-ordered.* $\mathbb{B}, \mathbb{N}$, $\mathsf{PosBool}(X)$, and all of the semirings for provenance from Section 3 are naturally ordered. For $\mathsf{PosBool}(X)$ the natural order corresponds to logical entailment: $\varphi \leq \psi$ iff $\varphi \models \psi$. For $\mathbb{B}[X]$ we have $a \leq b$ iff every monomial in a also appears in b. For $\mathbb{N}[X]$ we have $a \leq b$ iff every monomial in a also appears in b with an equal or greater coefficient. Thus $2x^2y \leq 5x^2y + 2z$ but $x + 2y \not\leq 5x + 3y^2$. For lineage and why-provenance the natural order corresponds to set inclusion (n.b. for why-provenance, this is only set inclusion "at the outer level" – e.g., $\{\{x\}\} \leq \{\{x\}, \{y, z\}\}$ but $\{\{x\}, \{y, z\}\} \not\leq \{\{x, y\}, \{y, z\}\}$).

Definition 4.1. *Let K be a naturally-ordered semiring and let R_1, R_2 be two K-relations. We define containment of R_1 in R_2 by*

$$R_1 \leq_K R_2 \overset{def}{\iff} \forall t\ R_1(t) \leq R_2(t)$$

We define containment of queries P, Q with respect to K-relation semantics by

$$P \sqsubseteq_K Q \overset{def}{\iff} \forall I\ P(I) \leq_K Q(I)$$

When K is $\mathbb{B}$ ($\mathbb{N}$) we get the usual notion of query containment with respect to set (bag) semantics. For $\mathsf{PosBool}(X)$, we get the *structural containment* and *structural equivalence* of [31].[6]

5. CONTAINMENT MAPPINGS

In characterizing K-containment of CQs we will use variations on the notion of *containment mappings.* Let P, Q be conjunctive queries, and let h be a mapping $h : \mathsf{vars}(Q) \to \mathsf{vars}(P) \cup \mathsf{consts}(P)$ extended to be the identity on constants (we will typically use the shorthand $h : Q \to P$). We define h to operate component-wise on tuples, atoms, and CQs by replacing each occurrence of a variable x with $h(x)$. We say that $h : Q \to P$ is a *containment mapping* if $h(\mathsf{head}(Q)) = \mathsf{head}(P)$ and for every atom $R_i(\bar{u})$ in the body of Q the atom $R_i(h(\bar{u}))$ occurs in the body of P.

We will also make use of the notion of the *canonical database* (or *tableau*) for a query. This is the instance $\mathsf{can}(Q)$ obtained by viewing the body of a CQ Q as a database; i.e., $\mathsf{can}(Q) \models R(\bar{x})$ iff $R(\bar{x}) \in \mathsf{body}(Q)$. In doing this we blur the distinction between variables and domain values. When a query has duplicate atoms in the body, this does not result in duplicate tuples in the canonical database.

[6]There are reasonable alternatives to the natural order for incomplete databases, such as considering various orders on the sets of *possible worlds* they represent.

The classical result of [6] relates containment mappings, canonical databases, and containment of CQs under set semantics:

Theorem 5.1 ([6]). *For CQs P, Q the following are equivalent:*

1. *$P \sqsubseteq_{\mathbb{B}} Q$*
2. *$P(\mathsf{can}(P)) \leq_{\mathbb{B}} Q(\mathsf{can}(P))$*
3. *there is a containment mapping $h : Q \to P$*

We will also exploit the device of canonical databases, but for the provenance models we will use various *abstractly-tagged* versions. The *abstractly-tagged version* $\mathsf{ab}_K(R)$ of a K-relation R is obtained by annotating each tuple in the support of R with its own tuple id from X. For $\mathbb{N}[X]$, $\mathbb{B}[X]$, and $\mathsf{Trio}(X)$ this is simply a fresh variable x from X. For lineage the variable is nested in a singleton set, $\{x\}$, and for why-provenance the variable is doubly-nested, $\{\{x\}\}$. We will use the shorthand $\mathsf{can}_K(Q)$ to mean $\mathsf{ab}_K(\mathsf{can}(Q))$. Abstractly-tagged instances will also play a role outside of the context of canonical databases (cf. Lemma 7.14).

6. BOUNDS FROM SEMIRING HOMOMORPHISMS

In this section we establish some initial bounds on the "relative behavior" of the various provenance models w.r.t. query containment and equivalence, based on surjective semiring homomorphisms.

A function $h : K \to K'$ can be made to transform a K-relation R into a K'-relation $h(R)$ by applying h to each tuple annotation in R. Performing this transformation component-wise on the K-relations of a K-instance I transforms it into a K'-instance $h(I)$. It was shown in [19] that semiring homomorphisms work nicely with UCQs on K-relations:

Proposition 6.1 ([19]). *Let $h : K \to K'$ and assume that K, K' are commutative semirings. Then $\bar{Q}(h(I)) = h(\bar{Q}(I))$ for all $\bar{Q} \in$ UCQ and K-instances I iff h is a semiring homomorphism.*

The observations we have made in Section 3 about the existence of surjective semiring homomorphisms relating the various provenance models turn out to yield some easy bounds on their "relative behavior" with respect to query containment (and therefore also equivalence). We write $K_1 \Rightarrow K_2$ to mean that for all UCQs $\bar{Q}_1, \bar{Q}_2$, if $\bar{Q}_1 \sqsubseteq_{K_1} \bar{Q}_2$ then $\bar{Q}_1 \sqsubseteq_{K_2} \bar{Q}_2$. Then we have the following:

Lemma 6.2. *For naturally-ordered semirings K_1, K_2, if there exists a surjective homomorphism $h : K_1 \to K_2$, then $K_1 \Rightarrow K_2$.*

The proof is in the Appendix. Based on our previous observations, we can conclude the following about the "relative behavior" of the semirings for provenance w.r.t. containment (and therefore also equivalence) of UCQs:

THEOREM 6.3. *If there is a path downward from K_1 to K_2 in Figure 3, then $K_1 \Rightarrow K_2$.*

We shall see in Section 7 which of the implications are strict (as indicated by the ticked arrows "$\not\Rightarrow$" in Figure 2).

Finally, we note that using similar reasoning, it is possible to establish bounds for containment/equivalence of UCQs for *arbitrary* semirings:

THEOREM 6.4. *For all K, $\mathbb{N}[X] \Rightarrow K$. For all* positive *$K$, $K \Rightarrow \mathbb{B}$.*

PROOF. See Appendix. □

The definition of positive semiring is given in the Appendix. This is a large class of semirings: $\mathbb{B}$, $\mathbb{N}$, $\mathsf{PosBool}(X)$, and all of the semirings for provenance we have considered in this paper are positive. For the special case of CQs containing no self-joins, the bounds of Theorem 6.4 collapse to a uniform condition for equivalence:

COROLLARY 6.5. *If CQs P, Q contain no self-joins, then for any positive K, we have $P \equiv_K Q$ iff $P \cong Q$.*

Therefore, for conjunctive queries without self-joins, every "$\not\Rightarrow$" in Figure 2(b) becomes a "$\Leftrightarrow$".

7. MAIN RESULTS

We are now ready to present our main results on containment and equivalence.

For all but the provenance polynomials, the decision procedures for containment of CQs (and the accompanying complexity results) extend easily to UCQs because of the following general fact which was first noted for the case of set semantics in [29]:

PROPOSITION 7.1. *If a semiring K is idempotent (i.e., addition in K is idempotent), then for all UCQs $\bar{P}, \bar{Q}$, we have $\bar{P} \sqsubseteq_K \bar{Q}$ iff for every CQ P in $\bar{P}$ there is a CQ Q in $\bar{Q}$ such that $P \sqsubseteq_K Q$. As a consequence, checking K-containment of UCQs is polynomially equivalent to checking K-containment of CQs.*

The semirings for lineage, why-provenance, minimal witness provenance, and $\mathbb{B}[X]$-provenance are all idempotent. $\mathbb{N}[X]$ and $\mathsf{Trio}(X)$ are not idempotent, nor is the semiring of natural numbers used for bag semantics (and the failure of Proposition 7.1 for bag semantics was noted in [8]).

We also note that for idempotent semirings, containment and equivalence of UCQs are easily inter-reducible (and polynomially equivalent). This again generalizes a well-known fact for set semantics [29]:

PROPOSITION 7.2. *For UCQs $\bar{Q}_1, \bar{Q}_2$ and idempotent K we have*

1. $\bar{Q}_1 \sqsubseteq_K \bar{Q}_2$ *iff* $\bar{Q}_1 \cup \bar{Q}_2 \equiv_K \bar{Q}_2$
2. $\bar{Q}_1 \equiv_K \bar{Q}_2$ *iff* $\bar{Q}_1 \sqsubseteq_K \bar{Q}_2$ *and* $\bar{Q}_2 \sqsubseteq_K \bar{Q}_1$

(The second item is just the definition of K-equivalence of UCQs.)

7.1 Lineage

THEOREM 7.3. *For CQs P, Q the following are equivalent:*

1. $P \sqsubseteq_{\mathsf{Lin}(X)} Q$
2. $P(\mathsf{can}_{\mathsf{Lin}(X)}(P)) \leq_{\mathsf{Lin}(X)} Q(\mathsf{can}_{\mathsf{Lin}(X)}(P))$
3. *for every atom $A(\bar{x}) \in \mathsf{body}(P)$ there is a containment mapping $h : Q \to P$ with $A(\bar{x})$ in the image of h*

PROOF. (Sketch) similar to the proof of Theorem 7.11. □

It is easy to find examples of CQs P, Q such that there is a containment mapping $h : Q \to P$, but condition (3) above is not satisfied, e.g.:

$$P(x, y) \text{ :- } R(x, y) \qquad Q(x, y) \text{ :- } R(x, y), R(x, z)$$

There is no containment mapping $h : P \to Q$ with $R(x, z)$ in the image of h, so $P \not\sqsubseteq_{\mathsf{Lin}(X)} Q$. However, one can find containment mappings $h' : P \to Q$ and $h'' : Q \to P$ in both directions, so by Theorem 5.1, $P \equiv_{\mathbb{B}} Q$. This justifies the "$\not\Rightarrow$" between lineage and $\mathsf{PosBool}(X)/\mathbb{B}$ in Figures 2(a)–(d).

Note that the above example seems to contradict[7] Theorem 4.8 of [13] which claims that $P \equiv_{\mathsf{Lin}(X)} Q$ iff $P \equiv_{\mathbb{B}} Q$. In fact, the contradiction is explained by the fact that the definition of lineage given in that paper only makes sense for CQs without self-joins. We have already seen (Corollary 6.5) that for this class of queries, K-equivalence is the same as isomorphism, for any positive K (including the lineage semiring).

Also, condition (3) of Theorem 7.3 was identified previously in [8] as a necessary (but not sufficient) condition for bag containment of CQs. This justifies the "$\not\Rightarrow$" between $\mathbb{N}$ and lineage in Figure 2(a).

While the conditions for checking lineage containment and set containment of CQs/UCQs are different, the complexity turns out to be the same:

COROLLARY 7.4. *Checking $\mathsf{Lin}(X)$-containment or $\mathsf{Lin}(X)$-equivalence of CQs or UCQs is* NP*-complete.*

7.2 Why-Provenance

To characterize $\mathsf{Why}(X)$-containment of CQs, we define the concept of *onto containment mappings.* A mapping $h : Q \to P$ is an *onto containment mapping* if it is a containment mapping and $\mathsf{body}(P) \leq_{\mathbb{N}} h(\mathsf{body}(Q))$.

[7]The example and this observation are due to James Cheney and Wang-Chiew Tan.

THEOREM 7.5. *For CQs P, Q the following are equivalent:*

1. $P \sqsubseteq_{\mathsf{Why}(X)} Q$
2. $P(\mathit{can}_{\mathsf{Why}(X)}(P)) \leq_{\mathsf{Why}(X)} Q(\mathit{can}_{\mathsf{Why}(X)}(P))$
3. *there is an onto containment mapping* $h : Q \to P$

PROOF. (Sketch) Similar to the proof of Theorem 7.11. □

The existence of an onto containment mapping is a strictly stronger requirement than condition (3) of Theorem 7.3. For example, consider the queries

$$\begin{array}{lcl} P(x) & :- & R(x,y), R(x,x) \\ Q(u) & :- & R(u,v) \end{array}$$

There is no onto containment mapping from Q to P, hence $P \not\sqsubseteq_{\mathsf{Why}(X)} Q$, but one can find containment mappings satisfying condition (3) of Theorem 7.3 in both directions, so $P \equiv_{\mathsf{Lin}(X)} Q$. This justifies the "$\not\Rightarrow$" between why-prov. and lineage in Figure 2(a)-(d).

We note that the existence of onto containment mappings was identified in [8] as a sufficient (but not necessary) condition for bag containment of CQs. This justifies the "$\not\Rightarrow$" between $\mathsf{Why}(X)$ and $\mathbb{N}$ in Figure 2(a).

The existence of onto containment mappings in both directions leads to a simple characterization of $\mathsf{Why}(X)$-equivalence of CQs:

THEOREM 7.6. *For CQs P, Q, $P \equiv_{\mathsf{Why}(X)} Q$ iff $P \cong Q$.*

PROOF. See Appendix. □

It was shown in [8] that bag equivalence of CQs is also the same as isomorphism, hence the "$\Leftrightarrow$" between $\mathbb{N}$ and $\mathsf{Why}(X)$ in Figure 2(b). Also, note that there are $\mathsf{Lin}(X)$-equivalent CQs which are not isomorphic, for example:

$$P(x) \ :- \ R(x,y) \qquad Q(x) \ :- \ R(x,y), R(x,z)$$

Thus we have the "$\not\Rightarrow$" between $\mathsf{Why}(X)$ and $\mathsf{Lin}(X)$ in Figure 2(b).

For UCQs $\bar{P}, \bar{Q}$, we note that Theorem 7.6 does not imply that for UCQs $\bar{P} \equiv_{\mathsf{Why}(X)} \bar{Q}$ iff $\bar{P} \cong \bar{Q}$ (and indeed this is not the case).

COROLLARY 7.7. *Checking $\mathsf{Why}(X)$-containment for CQs or UCQs and $\mathsf{Why}(X)$-equivalence for UCQs is* NP*-complete. Checking $\mathsf{Why}(X)$-equivalence for CQs is* GI*-complete.*

7.3 $\mathbb{B}[X]$-Provenance

To characterize $\mathbb{B}[X]$-containment of CQs we will need another variation on containment mappings, which we call *exact containment mappings.* A mapping $h : Q \to P$ is an *exact containment mapping* if $h(Q) = P$, i.e., $h(\mathsf{head}(Q)) = \mathsf{head}(P)$, and the bag of atoms $h(\mathsf{body}(Q))$ is identical to the bag of atoms $\mathsf{body}(P)$. Note that there is an exact containment mapping from Q to P iff P can be obtained from Q (up to isomorphism) by *unifying* variables in Q.

THEOREM 7.8. *For CQs P, Q, the following are equivalent:*

1. $P \sqsubseteq_{\mathbb{B}[X]} Q$
2. $P(\mathit{can}_{\mathbb{B}[X]}(P)) \leq_{\mathbb{B}[X]} Q(\mathit{can}_{\mathbb{B}[X]}(P))$
3. *there is an exact containment mapping* $h : Q \to P$

PROOF. See Appendix. □

Every exact containment mapping is also an onto containment mapping, but the converse is not true. For example, the mapping $h : Q \to P$ which sends w to u, z to v, and everything else to itself in

$$\begin{array}{lcl} P(x,y) & :- & R(x,y), S(u,v) \\ Q(x,y) & :- & R(x,y), S(u,v), S(w,z) \end{array}$$

is an onto containment mapping, but not an exact containment mapping. This justifies the "$\not\Rightarrow$" between $\mathbb{B}[X]$ and $\mathsf{Why}(X)$ in Figure 2(a),(c). To justify the "$\not\Rightarrow$" between $\mathbb{B}[X]$ and $\mathsf{Why}(X)$ in Figure 2(d), consider P, Q as above and define the UCQs $\bar{P} = (P)$ and $\bar{Q} = (P, Q)$. Then $\bar{P} \equiv_{\mathsf{Why}(X)} \bar{Q}$ but $\bar{P} \not\equiv_{\mathbb{B}[X]} \bar{Q}$.

Like $\mathsf{Why}(X)$-equivalence, $\mathbb{B}[X]$-equivalence of CQs turns out to be the same as isomorphism:

THEOREM 7.9. *For CQs P, Q, $P \equiv_{\mathbb{B}[X]} Q$ iff $P \cong Q$.*

This justifies the "$\Leftrightarrow$" between $\mathsf{Why}(X)$ and $\mathbb{B}[X]$ in Figure 2(b).

Checking for the existence of an exact containment mapping turns out to have the same complexity as checking for the existence of a containment mapping:

COROLLARY 7.10. *Checking $\mathbb{B}[X]$-containment of CQs or UCQs, or $\mathbb{B}[X]$-equivalence of UCQs, is* NP*-complete. Checking $\mathbb{B}[X]$-equivalence of CQs is* GI*-complete.*

PROOF. See Appendix. □

7.4 Provenance Polynomials

We now prove the results for $\mathbb{N}[X]$-containment. For CQs, this turns out to be the same as for $\mathbb{B}[X]$-containment (thus justifying the "$\Leftrightarrow$" between $\mathbb{N}[X]$ and $\mathbb{B}[X]$ in Figure 2(a)):

THEOREM 7.11. *For CQs P, Q the following are equivalent:*

1. $P \sqsubseteq_{\mathbb{N}[X]} Q$

2. $P(\mathsf{can}_{\mathbb{N}[X]}(P)) \leq_{\mathbb{N}[X]} Q(\mathsf{can}_{\mathbb{N}[X]}(P))$

3. *there is an exact containment mapping* $h : Q \to P$

Proof. See Appendix. $\square$

Since $\mathbb{N}[X]$-containment of CQs holds exactly when $\mathbb{B}[X]$-containment holds, the same is true for $\mathbb{N}[X]$-equivalence:

Theorem 7.12. *Let P, Q be two CQs. Then $P \equiv_{\mathbb{N}[X]} Q$ iff $P \cong Q$.*

This justifies the "$\Leftrightarrow$" between $\mathbb{N}[X]$ and $\mathbb{B}[X]$ (and therefore also $\mathsf{Why}(X)$ and $\mathbb{N}$) in Figure 2(b).

Next we consider $\mathbb{N}[X]$-containment of UCQs. Using similar reasoning as in Theorem 7.11, it is not hard to see that a weaker version of the Sagiv-Yannakakis property for set-containment of UCQs [29] holds for $\mathbb{N}[X]$:

Lemma 7.13. *For UCQs $\bar{P}, \bar{Q}$, if $\bar{P} \sqsubseteq_{\mathbb{N}[X]} \bar{Q}$, then for every $P_i \in \bar{P}$ there exists $Q_j \in \bar{Q}$ s.t. $P_i \sqsubseteq_{\mathbb{N}[X]} Q_j$.*

Proof. (Sketch) Similar reasoning as in Theorem 7.11, using the abstractly-tagged canonical database for $\bar{P}$. $\square$

A natural question to ask is whether the lemma above can be strengthened to require that each $P_i \in \bar{P}$ correspond to a *unique* $Q_j \in \bar{Q}$; as this is clearly also a sufficient condition for containment, this would therefore yield a decision procedure for containment. However, the strengthened version is not true: consider the UCQs $\bar{P} = (P_1, P_2)$ and $\bar{Q} = (Q_1, Q_2)$ where

$$\begin{array}{lll} P_1 & \text{:-} & R(x,y), R(x,x) \quad Q_1 \text{ :- } R(x,y), R(u,u) \\ P_2 & \text{:-} & R(x,y), R(y,y) \quad Q_2 \text{ :- } R(x,x), R(x,x) \end{array}$$

Both P_1 and P_2 are $\mathbb{N}[X]$-contained in Q_1, but neither is $\mathbb{N}[X]$-contained in Q_2; nevertheless, one can show that $\bar{P} \sqsubseteq_{\mathbb{N}[X]} \bar{Q}$.

Another natural idea is to check containment of $\bar{P}$ in $\bar{Q}$ by evaluating both queries on the canonical database for $\bar{P}$, in analogy with Theorem 7.11; unfortunately, one can easily find counterexamples showing that this procedure is unsound.

However, we are able to show that $\mathbb{N}[X]$-containment of UCQs is decidable, at least, by establishing a "small counterexample" property. In particular we show that if $\bar{P} \not\sqsubseteq_{\mathbb{N}[X]} \bar{Q}$, then $\bar{P}(I) \not\leq_{\mathbb{N}[X]} \bar{Q}(I)$ for some I whose size is bounded by the size of $\bar{P}$ and $\bar{Q}$.

When looking for such counterexamples, it is helpful to know that it suffices to consider only abstractly-tagged instances:

Lemma 7.14. *For any naturally-ordered semiring K, if $\bar{P}, \bar{Q} \in UCQ$ and $\bar{P}(I) \not\leq_K \bar{Q}(I)$ for some K-instance I, then $\bar{P}(\mathsf{ab}_K(I)) \not\leq_{\mathbb{N}[X]} \bar{Q}(\mathsf{ab}_K(I))$.*

Proof. Straightforward argument using Proposition 6.1, the universality property of $\mathbb{N}[X]$, and Proposition A.2. $\square$

Of course, the lemma holds in particular for $K = \mathbb{N}[X]$. We now state our "small counterexample" result:

Theorem 7.15. *$\bar{P} \not\sqsubseteq_{\mathbb{N}[X]} \bar{Q}$ iff $\bar{P}(I) \not\leq_{\mathbb{N}[X]} \bar{Q}(I)$ for some abstractly-tagged instance I containing at most k tuples, where k is the maximum number of atoms in the body of a CQ in $\bar{Q}$.*

Proof. See Appendix. $\square$

Theorem 7.15 leads immediately to a decision procedure for checking $\mathbb{N}[X]$-containment of UCQs: simply test $\bar{P}(I) \leq_{\mathbb{N}[X]} \bar{Q}(I)$ for all instances I containing at most k tuples over, say, the first nk values of the domain, where n is the maximum arity of a relation in the schema. (If $\bar{P}$ and $\bar{Q}$ contain constants, these must be included among the values considered as well.) Moreover, one can check that this can be done using only polynomial space:

Corollary 7.16. *For UCQs $\bar{P}, \bar{Q}$, checking $\bar{P} \sqsubseteq_{\mathbb{N}[X]} \bar{Q}$ is in* PSPACE.

The exact complexity of the problem remains open.

Finally, what about $\mathbb{N}[X]$-equivalence of UCQs? Theorem 7.15 tells us that it is decidable, but not much else. However, it turns out we can use Theorem 7.11 along with Lemma 7.13 to show that, as with CQs, $\mathbb{N}[X]$-equivalence of UCQs is the same as isomorphism.

Theorem 7.17. *For UCQs $\bar{P}, \bar{Q}$, we have $\bar{P} \equiv_{\mathbb{N}[X]} \bar{Q}$ iff $\bar{P} \cong \bar{Q}$.*

In the proof we make use of the following simple proposition which states that removing $\mathbb{N}[X]$-equivalent CQs from $\mathbb{N}[X]$-equivalent UCQs yields $\mathbb{N}[X]$-equivalent UCQs:

Proposition 7.18. *Let $\bar{P}, \bar{Q} \in UCQ$ and suppose $\bar{P} \equiv_{\mathbb{N}[X]} \bar{Q}$. Then for all $P \in \bar{P}, Q \in \bar{Q}$, if $P \cong Q$, then $\bar{P}' \equiv_{\mathbb{N}[X]} \bar{Q}'$, where $\bar{P}'$ ($\bar{Q}'$) is the UCQ obtained from $\bar{P}$ ($\bar{Q}$) by removing P (Q).*

Proof. (of Theorem 7.17) "$\Leftarrow$" is trivial. For "$\Rightarrow$" we argue by induction on $|\bar{P}| + |\bar{Q}|$. In the base case, $|\bar{P}| + |\bar{Q}| = 0$, and the queries are trivially $\mathbb{N}[X]$-equivalent and isomorphic. In the inductive case, consider $\bar{P} = (P_1, \ldots, P_n)$ and $\bar{Q} = (Q_1, \ldots, Q_m)$ with $n + m > 0$, and assume inductively that for all $\bar{P}', \bar{Q}'$ s.t. $|\bar{P}'| + |\bar{Q}'| < n + m$, if $\bar{P}' \equiv_{\mathbb{N}} \bar{Q}'$ then $\bar{P}' \cong \bar{Q}'$. If $\bar{P} \equiv_{\mathbb{N}} \bar{Q}$, then using Lemma 7.13, one can show that there exists some non-empty sequence $i_1, \ldots, i_{2k}$ such that $P_{i_1} \sqsubseteq_{\mathbb{N}[X]} Q_{i_2} \sqsubseteq_{\mathbb{N}[X]} \cdots \sqsubseteq_{\mathbb{N}[X]} P_{i_{2k-1}} \sqsubseteq_{\mathbb{N}[X]} Q_{i_{2k}}$ and $Q_{i_{2k}} \sqsubseteq_{\mathbb{N}[X]} P_{i_1}$. It follows that all the CQs in the sequence are $\mathbb{N}[X]$-equivalent, and hence (by Theorem 7.15) isomorphic. In particular, we have $P_{i_1} \cong Q_{i_1}$. Denote by $\bar{P}'$ the UCQ obtained by removing P_{i_1} from $\bar{P}$, and denote by $\bar{Q}'$ the UCQ obtained by removing Q_{i_1} from $\bar{Q}$. By Proposition 7.18, we have $\bar{P}' \equiv_{\mathbb{N}} \bar{Q}'$. Using the induction hypothesis, this implies $\bar{P}' \cong \bar{Q}'$. Since $\bar{P}' \cong \bar{Q}'$ and $P_{i_1} \cong Q_{i_1}$, it follows that $\bar{P} \cong \bar{Q}$ as required. $\square$

Since $\mathbb{B}[X]$ is idempotent, but $\mathbb{N}[X]$ is not, it is easy to find examples of $\bar{P}, \bar{Q}$ where $\bar{P} \equiv_{\mathbb{B}[X]} \bar{Q}$ but $\bar{P} \not\equiv_{\mathbb{N}[X]} \bar{Q}$, e.g., $\bar{P} = (P)$ and $\bar{Q} = (P, P)$ where P is an arbitrary CQ. This justifies the "$\not\Rightarrow$" between $\mathbb{N}[X]$ and $\mathbb{B}[X]$ in Figure 2(c) and Figure 2(d).

7.5 Bag Semantics

In this section, we discuss some further connections between provenance annotations and bag semantics.

We note that by Theorem 6.4, $\mathbb{N}[X]$-containment of UCQs implies bag-containment. Since the former is decidable and the latter is not, it follows that there exist UCQs for which bag-containment holds but $\mathbb{N}[X]$-containment does not. This justifies the "$\not\Rightarrow$" between $\mathbb{N}[X]$ and $\mathbb{N}$ in Figure 2(d). Also, we can show that:

PROPOSITION 7.19. *For containment of UCQs, we have*

1. $\mathbb{N} \not\Rightarrow \mathbb{B}[X]$ *and* $\mathbb{B}[X] \not\Rightarrow \mathbb{N}$
2. $\mathbb{N} \not\Rightarrow \mathsf{Why}(X)$ *and* $\mathsf{Why}(X) \not\Rightarrow \mathbb{N}$
3. $\mathbb{N} \not\Rightarrow \mathsf{Lin}(X)$

This justifies the "$\not\Rightarrow$" between $\mathbb{N}$ and lineage in Figure 2(c) and shows that $\mathbb{N}$ is incomparable there with $\mathbb{B}[X]$ and $\mathsf{Why}(X)$.

Next, the "$\Leftrightarrow$" between $\mathbb{N}$ and $\mathbb{N}[X]$ in Figure 2(d) follows from the following result:

THEOREM 7.20. *For UCQs $\bar{P}, \bar{Q}$ we have $\bar{P} \equiv_{\mathbb{N}} \bar{Q}$ iff $\bar{P} \equiv_{\mathbb{N}[X]} \bar{Q}$*

PROOF. $\mathbb{N}[X] \Rightarrow \mathbb{N}$ follows from Theorem 6.4. We prove $\mathbb{N} \Rightarrow \mathbb{N}[X]$ by contrapositive. Suppose $\bar{P} \not\equiv_{\mathbb{N}[X]} \bar{Q}$. Then for some $\mathbb{N}[X]$-instance I and tuple t, we have $\bar{P}(I)(t) = A$ and $\bar{Q}(I)(t) = B$ and $A \neq B$. Since A and B are non-identical polynomials, one can always find a valuation $\nu : X \to \mathbb{N}$ such that $\mathsf{Eval}_\nu(A) \neq \mathsf{Eval}_\nu(B)$. By Proposition 6.1, we have $\bar{P}(\nu(I)(t)) \neq \bar{Q}(\nu(I)(t))$. Since $\nu(I)$ is an $\mathbb{N}$-instance, it follows that $\bar{P} \not\equiv_{\mathbb{N}} \bar{Q}$. Therefore $\mathbb{N}[X] \not\Rightarrow \mathbb{N}$, as required. $\square$

By Theorem 7.17 it follows from the above that bag equivalence of UCQs is also the same as isomorphism. Prior to receiving the reviews of this paper it seemed to us that the community considers the decidability of equivalence of UCQs under bag semantics an open problem. However, one of the referees pointed out (as related work on bag-set semantics) the papers [11, 12]. [11] stated the result that bag-set equivalence of UCQs (called *disjunctive queries* there) is the same as isomorphism, and added as an observation that this also holds for bag semantics. The outline of the proof of the bag-set semantics result is provided in [9] and although bag semantics is not discussed further there we have observed that, in fact, results on bag-set semantics do correspond to results on bag semantics via the following *transfer lemma*:

LEMMA 7.21. *There exists a mapping $\varphi : CQ \to CQ$ (which we extend to UCQs by applying it componentwise on CQs), a mapping f from bag instances to set instances, and a mapping g from set instances to bag instances, such that for any UCQ $\bar{Q}$, bag instance I, and set instance J, we have:*

1. $\bar{Q}(I) = \varphi(\bar{Q})(f(I))$
2. $\varphi(\bar{Q})(J) = \bar{P}(g(J))$

PROOF. See Appendix. $\square$

Lemma 7.21 implies that bag-containment (bag-equivalence) of CQs/UCQs is polynomial time reducible to bag-set containment (bag-set equivalence). Moreover, for UCQs $\bar{P}, \bar{Q}$, the transformation φ defined in the proof of Lemma 7.21 satisfies $\bar{P} \cong \bar{Q}$ iff $\varphi(\bar{P}) \cong \varphi(\bar{Q})$. Thus Lemma 7.21 transfers to bag semantics the isomomorphism results for equivalence under bag-set semantics.

7.6 Trio

For CQs, $\mathsf{Trio}(X)$-containment turns out to coincide with $\mathsf{Why}(X)$-containment:

THEOREM 7.22. *For CQs P, Q we have $P \sqsubseteq_{\mathsf{Trio}(X)} Q$ iff $P \sqsubseteq_{\mathsf{Why}(X)} Q$.*

Therefore, Theorem 7.5 (Theorem 7.6) applies to $\mathsf{Trio}(X)$-containment ($\mathsf{Trio}(X)$-equivalence) as well, and we have a "$\Leftrightarrow$" between $\mathsf{Trio}(X)$ and $\mathsf{Why}(X)$ in Figure 2(a) and Figure 2(b).

To establish the decidability of $\mathsf{Trio}(X)$-equivalence of UCQs, we note that:

PROPOSITION 7.23. $\mathsf{Trio}(X) \Rightarrow \mathbb{N}$

Combined with Theorem 7.20 this implies:

THEOREM 7.24. *For UCQs $\bar{P}, \bar{Q}$ we have $\bar{P} \equiv_{\mathsf{Trio}(X)} \bar{Q}$ iff $\bar{P} \cong \bar{Q}$.*

This justifies the "$\Leftrightarrow$" between $\mathsf{Trio}(X)$ and $\mathbb{N}$ in Figure 2(d).

Finally, we note that one can find examples of UCQs showing that $\mathbb{N}[X] \not\Rightarrow \mathsf{Trio}(X)$ and $\mathsf{Trio}(X) \not\Rightarrow \mathbb{N}$, as indicated in Figure 2(d). We leave open the decidability of $\mathsf{Trio}(X)$-containment of UCQs.

8. RELATED WORK

The seminal paper by Chandra and Merlin [6] introduced the fundamental concepts of containment mappings and canonical databases in showing the decidability of containment of CQs under set semantics and identifying its complexity as NP-complete. The extension to UCQs is due to Sagiv and Yannakakis [29]. We have built upon the techniques from these papers.

The papers by Ioannidis and Ramakrishnan [23] and Chaudhuri and Vardi [8] initiated the study of query containment under bag semantics. Chaudhuri and Vardi showed that bag-equivalence of CQs is the same as isomorphism, established the Π_2^p-hardness of checking bag-containment of CQs, and gave partial conditions for checking bag-containment (see Section 7 for further connections with our results)[8]. Ioannidis and Ramakrishnan showed that bag-containment of UCQs is undecidable and introduced a framework of annotations from algebraic structures similar in spirit to the semiring annotations we consider.

In Section 7.5 we have discussed the results of Cohen et al. [11] and Cohen [9] on bag equivalence and bag-set equivalence of UCQs. The decidability of bag-containment of CQs remains open. Recent progress was made on the problem by Jayram et al. [24] who established the undecidability of checking bag-containment of CQs with inequalities.

Semiring-annotated relations are also related to the lattice-annotated relations used in *parametric databases* by Lakshmanan and Shiri [26]. That paper also studied query containment and equivalence, giving a number of positive decidability reults. None of our provenance models fall into this framework (with the exception of $\mathsf{PosBool}(X)$, cf. Theorem 3.5).

We have already mentioned in Section 3 the paper by Grahne et al. [16], which studied containment and equivalence of positive relational queries on bilattice-annotated relations.

Green et al. [17] proposes $\mathbb{Z}$-relations, which are relations whose tuples are annotated with integer counts (positive or negative), and shows that $\mathbb{Z}$-equivalence is decidable for the full relational algebra (including difference). The proof makes essential use of the earlier results for bag semantics [8, 11].

Tan [33] showed that query containment is decidable for CQs on relations with *where-provenance* information. Our results here on why-provenance complement the where-provenance results (why-provenance and where-provenance were introduced together in [4]).

[8]Chaudhuri and Vardi [8] also introduced the study of *bag-set semantics*, and showed that bag-set equivalence of CQs (without repeated atoms in the body) is the same as isomorphism. This was essentially a rediscovery of a well-known result in graph theory due to Lovász [27] (see also [21]), who showed that for finite relational structures F, G, if $|\mathrm{Hom}(F,H)| = |\mathrm{Hom}(G,H)|$ for all finite relational structures H, where $\mathrm{Hom}(A,B)$ is the set of homomorphisms $h : A \to B$, then $F \cong G$. In database terminology, this says that bag-set equivalence of Boolean CQs (without repeated atoms in the body) is the same as isomorphism.

Green et al. [19] showed that when K is a distributive lattice, K-containment of UCQs is the same as set containment of UCQs. This was essentially a rediscovery of an earlier result due to Buneman et al. [4] presented there in the context of queries over tree-structured data with *minimal witness why-provenance* (see Section 3). The result was generalized to complex values and XML trees in [14].

Cohen [10] recently initiated the study of query optimization under *combined semantics*, which generalizes bag semantics and bag-set semantics by enriching the relational algebra with a *duplicate elimination* operator. “Duplicate elimination” also makes sense for K-relations in the form of the *support* operator:

$$\mathsf{supp}(R) \overset{\text{def}}{=} \lambda t. \begin{cases} 0 & \text{if } R(t) = 0 \\ 1 & \text{otherwise} \end{cases}$$

For $K = \mathbb{N}$, this is duplicate elimination; for $K = \mathsf{PosBool}(X)$ it corresponds to the poss operator of [1] which returns the “possible” tuples of an incomplete relation. It would be interesting to see whether the decidability results presented here can be extended to queries using supp.

Finally, the work in AI on *soft constraint satisfaction problems* [2] is closely related to the framework of K-relations. Their constraints over semirings are in fact the same as our K-relations and the two operations on constraints correspond indeed to relational join and projection. The semirings used in [2] are such that $+$ is idempotent and 1 is a top element in the resulting order. This rules out $\mathbb{N}$, $\mathbb{B}[X]$, $\mathbb{N}[X]$, and $\mathsf{Trio}(X)$.

9. CONCLUSION

We have mapped out some of the foundations of query optimization for databases with provenance information, by giving positive decidability results and complexity characterizations for checking K-containment/equivalence for CQs/UCQs, for various semirings K used to track provenance information. We also used these results to establish some necessary and some sufficient conditions for K-containment of CQs for *any* semiring K, and we showed that for the special case of CQs without self-joins and positive K, K-equivalence is the same as isomorphism. We also highlighted connections between query containment under set and bag semantics and containment under the various provenance semantics.

Moving beyond UCQs, it would be interesting to consider the same questions for Datalog programs on K-relations [19]. Unlike with UCQs, it is easy to see that $\mathbb{N}[X]$-equivalence of Datalog programs does not reduce to isomorphism, and it seems likely that the undecidability results for set semantics [32] will carry over to the forms of provenance information we have considered here. On the other hand, the positive decidability results concerning containment/equivalence of a Datalog program and a UCQ [7] might also carry over. We conjecture that when K is a distributive lattice, K-containment of Datalog programs holds exactly when the same holds for ordinary set semantics.

We assumed a Datalog-style representation for UCQs, which is expressively equivalent to the positive relational algebra ($\mathcal{RA}^+$) on K-relations, but exponentially less concise. Under set semantics, it is well-known [29] that checking containment of $\mathcal{RA}^+$ queries is correspondingly harder (Π_2^p-complete rather than NP-complete). An obvious question is how the move to an algebraic representation affects the results presented here.

Finally, semiring annotations also make sense for a positive version of XQuery on unordered XML data, as shown in [14]. It would be worthwhile to investigate how the same issues of query containment and equivalence considered here play out for annotated XML.

Acknowledgments

James Cheney, Zack Ives, Grigoris Karvounarakis, and Stijn Vansummeren offered useful comments on earlier revisions of this paper. Val Tannen suggested many of the semirings and their constructions described in Section 3 and offered guidance and encouragement in preparing this paper. We thank the anonymous referees for bringing the papers [11, 12] and [27] to our attention, and we thank Gösta Grahne for pointing out [16] and [26]. Our work is supported by the National Science Foundation under grants IIS-0447972, 0513778, and 0629846.

10. REFERENCES

[1] L. Antova, C. Koch, and D. Olteanu. From complete to incomplete information and back. In *SIGMOD*, 2007.

[2] S. Bistarelli. *Semirings for Soft Constraint Solving and Programming.* Springer, 2004.

[3] P. Buneman, J. Cheney, W.-C. Tan, and S. Vansummeren. Curated databases. In *PODS*, 2008.

[4] P. Buneman, S. Khanna, and W.-C. Tan. Why and where: A characterization of data provenance. In *ICDT*, 2001.

[5] P. Buneman, S. Khanna, and W. C. Tan. On propagation of deletions and annotations through views. In *PODS*, 2002.

[6] A. K. Chandra and P. M. Merlin. Optimal implementation of conjunctive queries in relational data bases. In *STOC*, pages 77–90, 1977.

[7] S. Chaudhuri and M. Y. Vardi. On the equivalence of recursive and nonrecursive datalog programs. In *PODS*, 1992.

[8] S. Chaudhuri and M. Y. Vardi. Optimization of *real* conjunctive queries. In *PODS*, 1993.

[9] S. Cohen. Containment of aggregate queries. *SIGMOD Record*, 34(1):77–85, 2005.

[10] S. Cohen. Equivalence of queries combining set and bag-set semantics. In *PODS*, 2006.

[11] S. Cohen, W. Nutt, and A. Serebrenik. Rewriting aggregate queries using views. In *PODS*, 1999.

[12] S. Cohen, Y. Sagiv, and W. Nutt. Equivalences among aggregate queries with negation. *ACM TOCL*, 6(2):328–360, April 2005.

[13] Y. Cui, J. Widom, and J. L. Wiener. Tracing the lineage of view data in a warehousing environment. *TODS*, 25(2), 2000.

[14] J. N. Foster, T. J. Green, and V. Tannen. Annotated XML: Queries and provenance. In *PODS*, 2008.

[15] N. Fuhr and T. Rölleke. A probabilistic relational algebra for the integration of information retrieval and database systems. *TOIS*, 14(1):32–66, 1997.

[16] G. Grahne, N. Spyratos, and D. Stamate. Semantics and containment of queries with internal and external conjunctions. In *ICDT*, 1997.

[17] T. J. Green, Z. G. Ives, and V. Tannen. Reconcilable differences. In *ICDT*, 2009.

[18] T. J. Green, G. Karvounarakis, Z. G. Ives, and V. Tannen. Update exchange with mappings and provenance. In *VLDB*, 2007.

[19] T. J. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. In *PODS*, 2007.

[20] T. J. Green and V. Tannen. Models for incomplete and probabilistic information. In *IIDB*, March 2006.

[21] P. Hell and J. Nešetřil. *Graphs and Homomorphisms.* Oxford University Press, 2004.

[22] T. Imieliński and J. Witold Lipski. Incomplete information in relational databases. *J. ACM*, 31(4), 1984.

[23] Y. E. Ioannidis and R. Ramakrishnan. Containment of conjunctive queries: Beyond relations as sets. *TODS*, 20(3):288–324, 1995.

[24] T. S. Jayram, P. G. Kolaitis, and E. Vee. The containment problem for *real* conjunctive queries with inequalities. In *PODS*, 2006.

[25] J. Köbler, U. Schöning, and J. Torán. *The Graph Isomorphism Problem: its Structural Complexity.* Birkhäuser Verlag, 1993.

[26] L. V. S. Lakshmanan and N. Shiri. A parametric approach to deductive databases with uncertainty. *IEEE Trans. Knowl. Data Eng.*, 13(4):554–570, 2001.

[27] L. Lovász. Operations with structures. *Acta Mathematica Hungarica*, 18(3–4):321–328, 1967.

[28] W. Nutt, Y. Sagiv, and S. Shurin. Deciding equivalences among aggregate queries. In *PODS*, 1998.

[29] Y. Sagiv and M. Yannakakis. Equivalences among relational expressions with the union and difference operators. *J. ACM*, 27(4):633–655, 1980.

[30] A. D. Sarma, M. Theobald, and J. Widom. Exploiting lineage for confidence computation in uncertain and probabilistic databases. In *ICDE*, 2008.

[31] P. Senellart and S. Abiteboul. On the complexity of managing probabilistic XML data. In *PODS*, 2007.

[32] O. Shmueli. Equivalence of datalog queries is undecidable. *J. of Logic Programming*, 15, 1993.

[33] W.-C. Tan. Containment of relational queries with annotation propagation. In *DBPL*, September 2003.

[34] E. Zimányi. Query evaluation in probabilistic relational databases. *TCS*, 171(1-2), 1997.

APPENDIX

A. BACKGROUND

Definition A.1 (Semiring homomorphism). *Let K_1, K_2 be semirings. A mapping $h : K_1 \to K_2$ is called a* semiring homomorphism *if $h(0) = 0$, $h(1) = 1$, and for all $a, b \in K_1$, we have $h(a + b) = h(a) + h(b)$ and $h(a \cdot b) = h(a) \cdot h(b)$.*

Proposition A.2. *Let K_1, K_2 be naturally-ordered commutative semirings. If $h : K_1 \to K_2$ is a semiring homomorphism then for all $a, b \in K_1$, $a \leq_{K_1} b \implies h(a) \leq_{K_2} h(b)$. If h is also surjective, then for all $a, b \in K_1$, $a \leq_{K_1} b \iff h(a) \leq_{K_2} h(b)$.*

PROOF. Straightforward calculation. □

Given a semiring K define $\dagger : K \to \mathbb{B}$ as follows:

$$\dagger(0) \overset{\text{def}}{=} \mathsf{false}$$
$$\dagger(a) \overset{\text{def}}{=} \mathsf{true} \text{ when } a \neq 0$$

PROPOSITION A.3. *The following are equivalent:*

1. $\dagger$ *is a semiring homomorphism*
2. K *satisfies*
   (a) $0 \neq 1$
   (b) $a + b = 0$ *implies* $a = 0$ *or* $b = 0$
   (c) $ab = 0$ *implies* $a = 0$ *or* $b = 0$

A semiring K is called *positive* if it satisfies either of the (equivalent) statements in Proposition A.3.

DEFINITION A.4 (CONGRUENCE RELATION). *If K is a semiring and $\approx$ is an equivalence relation on K, then we say that $\approx$ is a* congruence relation *on K if $a \approx a'$ and $b \approx b'$ implies $a + b \approx a' + b'$ and $a \cdot b \approx a' \cdot b'$.*

DEFINITION A.5 (QUOTIENT SEMIRING). *Let K be a semiring and let $\approx$ be a congruence relation on K. If $a \in K$ then denote the equivalence class of a in $\approx$ by $a/\approx$. Then the* quotient of K by $\approx$ *is the semiring whose domain is the set $K/\approx$ of equivalence classes of $\approx$, $0 \overset{def}{=} 0_K/\approx$, $1 \overset{def}{=} 1_K/\approx$, $(a/\approx) + (b/\approx) = (a + b)/\approx$, and $(a/\approx) \cdot (b/\approx) \overset{def}{=} (a \cdot b)/\approx$.*

B. PROOFS

PROOF. (of Lemma 6.2) Suppose that $h : K_1 \to K_2$ is a surjective semiring homomorphism and that $\bar{Q}_1 \sqsubseteq_{K_1} \bar{Q}_2$. Consider an arbitrary K_2-instance I. We want to show that $\bar{Q}_1(I) \leq_{K_2} \bar{Q}_2(I)$. Since h is surjective, there exists a K_1-instance J such that $I = h(J)$. Since $\bar{Q}_1 \sqsubseteq_{K_1} \bar{Q}_2$ we have that $\bar{Q}_1(J) \leq_{K_1} \bar{Q}_2(J)$. By Proposition A.2, this implies $h(\bar{Q}_1(J)) \leq_{K_2} h(\bar{Q}_2(J))$. But by Proposition 6.1, $h(\bar{Q}_1(J)) = \bar{Q}_1(h(J)) = \bar{Q}_1(I)$, and likewise, $h(\bar{Q}_2(J)) = \bar{Q}_2(h(J)) = \bar{Q}_2(I)$. It follows that $\bar{Q}_1(I) \leq_{K_2} \bar{Q}_2(I)$, as required. □

PROOF. (of Theorem 6.4) $\mathbb{N}[X] \Rightarrow K$ follows from similar reasoning as in Proposition 6.2, but using the universality of the provenance polynomials rather than the existence of surjective semiring homomorphisms to establish the relationship. $K \Rightarrow \mathbb{B}$ follows immediately from Proposition 6.2 using the definition of positive semiring. □

PROOF. (of Theorem 7.6) Clearly isomorphism implies K-equivalence for any K, in particular for why-provenance. In the other direction, if $P \equiv_{\mathsf{Why}(X)} Q$ by Theorem 7.5 there must exist onto containment mappings $h : Q \to P$ and $g : P \to Q$. But since both mappings are surjective they must also be injective. It follows that $P \cong Q$. □

PROOF. (of Theorem 7.8) (1) $\Rightarrow$ (2) is trivial, and (3) $\Rightarrow$ (2) is straightforward to check.

For (2) $\Rightarrow$ (3), we assume for simplicity that $\mathsf{body}(P)$ contains no duplicate atoms (the argument can be extended to work without this assumption). Now suppose

$$P(\mathsf{can}_{\mathbb{B}[X]}(P)) \leq_{\mathbb{B}[X]} Q(\mathsf{can}_{\mathbb{B}[X]}(P)).$$

Then in particular

$$P(\mathsf{can}_{\mathbb{B}[X]}(P))(\bar{u}) \leq Q(\mathsf{can}_{\mathbb{B}[X]}(P))(\bar{u}),$$

where $\bar{u}$ is the tuple of distinguished variables in $\mathsf{head}(P)$. Also, the polynomial $P(\mathsf{can}_{\mathbb{B}[X]}(P)))(\bar{u})$ contains as a term (i.e., with Boolean coefficient true) the product $x_1 \cdots x_n$ of all tuple ids $x_1, \ldots, x_n$ in $\mathsf{can}_{\mathbb{B}[X]}(P))$. Since containment holds, the polynomial $Q(\mathsf{can}_{\mathbb{B}[X]}(P)))(\bar{u})$ must also contain the same term. Working backwards, there must be some valuation $\nu : \mathsf{vars}(Q) \to \mathbb{D}$ justifying the term. Moreover, in order to yield all variables $x_1, \ldots, x_n$ in the term, ν must map the atoms of $\mathsf{body}(Q)$ surjectively onto the tuples of $\mathsf{can}_{\mathbb{B}[X]}(P))$; and in order for all the exponents in the term to equal one, the mapping of atoms to tuples must be injective. It follows that ν is an exact containment mapping from Q to P. □

PROOF. (of Corollary 7.10) (Sketch) It is clear that checking for exact containment mappings is in NP. As with containment mappings, the NP-hardness of the problem can be shown via a reduction from the graph 3-coloring problem. The main difference is that instead of reducing an instance of the 3-coloring problem for a graph (V, E) to one instance of the exact containment mapping problem, we reduce it to $\leq n^3$ instances of the exact containment problem (where $n = |E|$), one for each possible multiplicity of red, green, and blue edges, and observe that there is a 3-coloring of the graph iff there is an exact containment mapping for one of the instances. □

PROOF. (of Theorem 7.11) (2) "$\Rightarrow$" (3) is exactly the same as in Theorem 7.8. For (3) "$\Rightarrow$" (1) some additional care is required because addition in $\mathbb{N}[X]$ is not idempotent. We need to make sure that the coefficient of an arbitrary term in the polynomial $Q(I)(t)$, for some arbitrary $\mathbb{N}[X]$-instance I and tuple t, is at least as large as the coefficient of the same term in the polynomial $P(I)(t)$. To check this, it suffices to observe that for any valuations $\nu, \nu' : \mathsf{vars}(P) \to \mathbb{D}$ justifying a term in $P(I)(t)$, the valuations $\nu \circ h$ and $\nu' \circ h$ justify the same monomial in $Q(I)(t)$; and moreover (this is the important part) if $\nu \neq \nu'$ then $\nu \circ h \neq \nu' \circ h$. Hence every justification for $P(I)(t)$ corresponds to a unique justification for $Q(I)(t)$. Since addition in $\mathbb{N}$ is monotone this implies the required inequality for the term coefficients. □

PROOF. (of Theorem 7.15) "$\Leftarrow$" is trivial. For "$\Rightarrow$", suppose $\bar{P} \not\sqsubseteq_{\mathbb{N}[X]} \bar{Q}$. Then for some $\mathbb{N}[X]$-instance J, we have $\bar{P}(J) \not\leq_{\mathbb{N}[X]} \bar{Q}(J)$. By Lemma 7.14, we may assume that J is abstractly-tagged. Choose some tuple t such that $\bar{P}(J)(t) \not\leq \bar{Q}(J)(t)$. There must be some term α in the polynomial $\bar{P}(J)(t)$ with coefficient m such that the same term α in the polynomial $\bar{Q}(J)(t)$ has coefficient n and $m > n$. Now restrict J to contain only the source tuples identified in that term (call the resulting instance I). I has at most k tuples. Moreover, the coefficients for α in the polynomials for $\bar{P}(I)(t)$ and $\bar{Q}(I)(t)$ are unchanged. Hence $\bar{P}(I)(t) \not\leq \bar{Q}(I)(t)$, and therefore $\bar{P}(I) \not\leq_{\mathbb{N}[X]} \bar{Q}(I)$. $\square$

PROOF. (of Lemma 7.21) We first define $\varphi : \mathrm{CQ} \to \mathrm{CQ}$, as follows. Let Q be a CQ over schema Σ, and let Σ' be the schema obtained from Σ by replacing each n-ary relational predicate R with an $n+1$-ary relational predicate R'. Then φ maps Q to the CQ over schema Σ' obtained from Q by replacing each join atom $R(x_1, \ldots, x_k)$ with an atom $R'(x_1, \ldots, x_k, u)$ where u is a fresh variable. For example, if Q is the CQ

$$Q(x, y) \text{:-} R(x, y), R(y, z)$$

then $\varphi(Q)$ is the CQ

$$\varphi(Q)(x, y) \text{:-} R(x, y, u), R(y, z, v)$$

We extend φ to map UCQs to UCQs by applying it componentwise on CQs.

Next we define an encoding f of bag-instances over schema Σ to bag-set instances over schema Σ', and an encoding g of bag-set instances over Σ' to bag-instances over Σ, as follows:

- f maps a tuple t in relation R with multiplicity k to k distinct tuples $t'_1, \ldots, t'_k$ in R' each obtained from t by appending a fresh constant in the last column
- g assigns to tuple t in R the multiplicity k where k is the number of tuples t' in R' such that t and t' agree on all columns of t

It is straightforward to verify that φ, f, and g satisfy the conditions required by the Lemma. $\square$