Provenance and Probabilities in Relational Databases: From Theory to Practice

Pierre Senellart

To cite this version:

Pierre Senellart. Provenance and Probabilities in Relational Databases: From Theory to Practice. SIGMOD record, 2017, pp.1-11. hal-01672566

HAL Id: hal-01672566
https://inria.hal.science/hal-01672566
Submitted on 26 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ABSTRACT
We review the basics of data provenance in relational databases. We describe different provenance formalisms, from Boolean provenance to provenance semirings and beyond, that can be used for a wide variety of purposes, to obtain additional information on the output of a query. We discuss representation systems for data provenance, circuits in particular, with a focus on practical implementation. Finally, we explain how provenance is practically used for probabilistic query evaluation in probabilistic databases.

1. INTRODUCTION
The central task in data management is query evaluation: given a database instance and a query, compute the results of that query on that instance. But what if we want something more than just the query result? We might want to know:
• why this specific result was obtained;
• where values in the result come from;
• how the result was produced from the query;
• how the result would change if some of the input tuples were missing;
• how many times each query result was obtained;
• what probability the result has, given a probability distribution on the input data;
• what minimal security clearance is needed to see the result, given some security information on the input data;
• what the most economical way to obtain this result is, in terms of number of data accesses.

All these questions, and more, can be answered using the tool of data provenance [11,12], which is some additional bookkeeping information maintained during query evaluation, that allows answering a large number of such meta-questions on the output of a query. The precise nature and form this provenance information should take depends on the question we want to answer, and on the query language we consider.

We focus in this paper on the setting of relational databases, though provenance and its applications apply as well and are equally important in other settings, such as scientific workflows [19], knowledge graphs [30], or hierarchical data [11].

There has been a large amount of work on the foundations of data provenance in relational database systems: early definitions and implementations of data provenance (called lineage at the time) for specific applications [13,48]; the influential framework of where- and why-provenance introduced by Buneman, Khanna, and Tan [11]; the seminal paper on provenance semirings [27] by Green, Karvounarakis, and Tannen; and further extensions thereof [7,8,22,24].

Provenance has also been a particularly useful tool in the area of probabilistic databases [3,23,32], where the use of provenance is dubbed the intensional approach to probabilistic query evaluation.

The goal of this paper is to review some of the most important definitions of provenance, unifying them in a single framework as much as possible, and to address some of the concrete issues that arise in building a modern provenance-aware database management, and in implementing the intensional approach to probabilistic query evaluation.

Example 1. We will use throughout this paper a very simple running example of a single-table database, in Table 1, that contains information on the personal of a (fictitious) intelligence agency, to illustrate various aspects of data provenance. Note in Table 1 the $t_i$ annotation on each tuple. These will be used in the following as provenance tokens associated with the tuple, which means that they will contain the elementary provenance information associated to the tuple. You can think of them for now as simple tuple identifiers.

We will consider the following two example queries
on this database. The first query, $Q_1$, asks for the cities referenced in the $Personal$ table which host at least two different employees of the agency:

**SELECT DISTINCT** $P1.city$
FROM $Personal$ $P1$, $Personal$ $P2$
WHERE $P1.city = P2.city$ AND $P1.id < P2.id$

Obviously, the answer to this query, disregarding provenance, is a single-attribute table, containing the three cities New York, Paris, and Berlin. Our second query, $Q_2$, asks for the cities that host exactly one employee of the agency:

**SELECT DISTINCT** city FROM $Personal$
**EXCEPT**
**SELECT DISTINCT** $P1.city$
FROM $Personal$ $P1$, $Personal$ $P2$
WHERE $P1.city = P2.city$ AND $P1.id < P2.id$

Its output, once again disregarding provenance, is the empty table.

The paper is organized as follows: in Section 2, we first introduce the simple setting of **Boolean provenance** that is in particular used in probabilistic databases and has the advantage of being definable for an arbitrary query language. We move in Section 3 to **semiring provenance**, which captures more information than Boolean provenance but is defined for a specific, monotone, query language. We explore formalisms that go beyond semiring provenance in Section 4. In Section 5, we address the concrete problem of which formalism to use to represent data provenance, and settle on circuits as compact formalism. We explain in Section 6 how probabilistic query evaluation in probabilistic databases can be solved using provenance, and which tools can be used to do this efficiently.

2. BOOLEAN PROVENANCE

**Boolean provenance** is one of the simplest forms of provenance, while having a major conceptual advantage: it can be defined independently of a specific query language. The notion of Boolean provenance is implicit in the work of Imelinski and Lipski on conditional tables [31], although this predates the notion of provenance itself and the query language was limited to the relational algebra. The term **Boolean provenance** was introduced specifically in the setting of provenance semirings [26], though this was restricted to the monotone case. The generalization to non-monotone queries we give below is straightforward, and was made, e.g., in [3].

We fix a finite set $X = \{x_1, \ldots, x_n\}$, the elements of which we call **Boolean events** (i.e., variables that can be either $\top$ or $\bot$).

As in [31], we let provenance tokens (the annotations attached to tuples of the input databases) be Boolean functions over $X$, that is, functions of the form $\varphi : (X \rightarrow \{\top, \bot\}) \rightarrow \{\top, \bot\}$. They are interpreted under a possible-world semantics: every valuation $\nu : X \rightarrow \{\top, \bot\}$ denotes a possible world of the database; in this possible world, a tuple with annotation $\varphi$ exists if and only if $\varphi(\nu) = \top$. For a given database $D$, we denote $\nu(D)$ the set of tuples $t$ with annotation $\varphi_t$ such that $\varphi_t(\nu) = \top$. It is a subset of the database $D$.

**Example 2.** Consider again Table 1 and assume that, for each $i$, $t_i$ is the indicator function of a distinct Boolean event $x_i$, i.e., the function that maps a valuation $\nu$ to $\top$ if and only if $\nu(x_i) = \top$. Then the set of possible worlds of $Personal$ is the set of all subrelations of $Personal$. For instance, if $\nu$ maps $x_1$, $x_3$, $x_5$, and $x_7$ to $\top$ and all other variables to $\bot$, then $\nu(Personal)$ is the subrelation of $Personal$ where only the tuples of $id$ 1, 3, 5, and 7 survive.

Let $Q$ be an arbitrary query, i.e., a function that takes as input a finite relational database over a fixed database schema, and produces as output a finite relation over a fixed relational schema. Then the **Boolean provenance** of $Q$ over a database $D$, denoted $\text{prov}_{Q,D}$, is a function that maps a tuple $t$ of the

---

**Table 1: Table Personal for the personal of an intelligence agency, used as a running example**

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>position</th>
<th>city</th>
<th>classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>John</td>
<td>Director</td>
<td>New York</td>
<td>unclassified</td>
</tr>
<tr>
<td>2</td>
<td>Paul</td>
<td>Janitor</td>
<td>New York</td>
<td>restricted</td>
</tr>
<tr>
<td>3</td>
<td>Dave</td>
<td>Analyst</td>
<td>Paris</td>
<td>confidential</td>
</tr>
<tr>
<td>4</td>
<td>Ellen</td>
<td>Field agent</td>
<td>Berlin</td>
<td>secret</td>
</tr>
<tr>
<td>5</td>
<td>Magdalen</td>
<td>Double agent</td>
<td>Paris</td>
<td>top_secret</td>
</tr>
<tr>
<td>6</td>
<td>Nancy</td>
<td>HR</td>
<td>Paris</td>
<td>restricted</td>
</tr>
<tr>
<td>7</td>
<td>Susan</td>
<td>Analyst</td>
<td>Berlin</td>
<td>secret</td>
</tr>
</tbody>
</table>

**Example 1.** Consider the first query, $Q_1$, which asks for the cities that host at least two different employees of the agency.
output relational schema to the Boolean function that maps a valuation \( \nu \) to \( \top \) if \( t \in Q(\nu(D)) \), and to \( \bot \) otherwise. In other words, given provenance annotations on input database tuples, we obtain as output of a query a new database, namely,

\[
\bigcup_{\nu : X \to \{\top, \bot\}} Q(\nu(D)),
\]

with new provenance annotations on each tuple \( t \), namely \( \text{prov}_Q(D, t) \).

Note that if \( Q \) is monotone, i.e., if \( D \subseteq D' \Rightarrow Q(D) \subseteq Q(D') \), then \( \bigcup_{\nu : X \to \{\top, \bot\}} Q(\nu(D)) \subseteq Q(D) \), but this is not true for arbitrary queries.

**Example 3.** We will denote in this example Boolean functions using propositional formulas. See Section 5 for an alternate representation. To simplify, we assume that, once again, every \( t_i \) is an indicator function of a different Boolean event, and we write simply \( t_i \) for this event instead of \( x_i \).

One can check that the Boolean provenances of \( Q_1 \) and \( Q_2 \) on Personal are, respectively:

<table>
<thead>
<tr>
<th>city</th>
<th>( t_1 \land t_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>( (t_3 \land t_6) \lor (t_3 \land t_5) \lor (t_4 \land t_6) )</td>
</tr>
<tr>
<td>Paris</td>
<td>( t_4 \land \lnot t_7 )</td>
</tr>
<tr>
<td>Berlin</td>
<td>( (t_4 \land \lnot t_7) \lor (t_5 \land \lnot t_7) )</td>
</tr>
</tbody>
</table>

and:

<table>
<thead>
<tr>
<th>city</th>
<th>( (t_1 \land \lnot t_2) \lor (t_2 \land \lnot t_1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>( (t_3 \land \lnot t_5 \lor t_5) \lor (t_3 \land \lnot t_6 \lor t_6) \lor (t_5 \land \lnot t_6 \lor t_6) \lor (t_6 \land \lnot t_5 \lor t_5) )</td>
</tr>
<tr>
<td>Paris</td>
<td>( (t_4 \land \lnot t_7) \lor (t_7 \land \lnot t_4) )</td>
</tr>
<tr>
<td>Berlin</td>
<td>( (t_4 \land \lnot t_7) \lor (t_5 \land \lnot t_7) )</td>
</tr>
</tbody>
</table>

One of the major applications of Boolean provenance is query evaluation in probabilistic databases. Assume that each Boolean event \( x_i \) comes with an independent probability \( \text{Pr}(x_i) \) of being true. Then we can define the probability of a valuation \( \nu : X \to \{\top, \bot\} \) as:

\[
\text{Pr}(\nu) = \prod_{\nu(x_i) = \top} \text{Pr}(x_i) \prod_{\nu(x_i) = \bot} (1 - \text{Pr}(x_i)).
\]

From there, it is natural to define, for any given Boolean function \( \varphi \) over \( X \):

\[
\text{Pr}(\varphi) = \sum_{\nu : X \to \{\top, \bot\}} \text{Pr}(\nu).
\]

In particular, this defines a probability distribution \( \text{Pr}(\text{prov}_Q(D, t)) \) on provenance annotations of output tuples given a probability distribution on the provenance annotations of input tuples. This observation was first made by Green and Tannen in [28].

When \( t_i \)'s are indicator functions, one gets the simple model of tuple-independent databases [14, 23, 36] that has been widely studied.

**Example 4.** Assume again every \( t_i \) is an indicator function of a different Boolean function, and comes with an independent probability \( \text{Pr}(t_i) \) of being true. Then \( \text{Pr}(\text{New York} \in Q_1(\text{Personal})) = \text{Pr}(t_1 \land t_2) = \text{Pr}(t_1) \times \text{Pr}(t_2) \).

Note that we were able to compute the probability this way because the Boolean formula was simple enough. We discuss in Section 6 options when Boolean functions are more complex.

Though Boolean provenance can be formally defined for any query, what we need in practice is efficient algorithms for computing the provenance of a query in a given query language. We also want to capture more with provenance than what Boolean provenance can do. This is what provenance semirings, and extensions thereof, offer.

### 3. SEMIRING PROVENANCE

*Provenance semirings* have been introduced in [27] as a formalism for data provenance that has been shown [34] to cover and generalize, using a clean mathematical framework, previous formalisms such as why-provenance [11], lineages used in view maintenance [13], or the lineage used by the TRIO uncertain management system [9].

A semiring \((K, 0, 1, \oplus, \otimes)\) is a set \( K \) with distinguished elements \( 0 \) and \( 1 \), along with two binary operators:

- \( \oplus \), an associative and commutative operator, with identity \( 0 \);
- \( \otimes \), an associative and commutative\(^1\) operator, with identity \( 1 \).

We further require \( \otimes \) to distribute over \( \oplus \), and \( 0 \) to be annihilating for \( \otimes \).

Examples of semirings include [27, 28, 34]:

- \((\mathbb{N}, 0, 1, +, \times)\): the *counting* semiring;
- \((\{\bot, \top\}, \bot, \top, \lor, \land)\): the *Boolean* semiring;
- \((\{\text{unclassified}, \text{restricted}, \text{confidential}, \text{secret}, \text{top secret}\}, \text{top secret}, \text{unclassified}, \min, \max)\): the *security* semiring;
- \((\mathbb{N} \cup \{\infty\}, \infty, 0, \min, +)\): the *tropical* semiring;

\(^1\)It is almost always required in the literature [27, 34] that the semiring be commutative, which means that \( \otimes \) must be commutative. Note that this is only necessary if the cross product operator of the relational algebra is assumed to be commutative, which is the case in the named perspective, but not in the unnamed one [1]. This assumption has some technical impact, e.g., on the universality of \( \mathcal{N}[X] \), but is actually not critical to implement provenance support.
• \(\{\text{positive Boolean funct. over } X\}, \bot, \top, \lor, \land\): the semiring of positive Boolean functions over \(X\);
• \((\mathbb{N}[X], 0, 1, +, \times)\): the semiring of integer polynomials with variables in \(X\) (also called how-semiring or universal semiring, see further);
• \((\mathcal{P}(\mathcal{P}(X)), \emptyset, \{\emptyset\}, \cup, \uplus)\): why-semiring over \(X\) \((A \uplus B := \{a \cup b \mid a \in A, b \in B\})\).

Now, given a fixed semiring \((K, 0, 1, \oplus, \otimes)\), semiring provenance works as follows: we assume provenance tokens are all in \(K\). We consider a query \(Q\) from the positive relational algebra [1] (selection, projection, renaming, cross product, union). We define a semantics for the provenance of a tuple \(t \in Q(D)\) inductively on the structure of \(Q\), informally as follows (formal definitions can be found in [27]):

- selection and renaming do not affect provenance annotations;
- in the set semantics, the provenance annotations of tuples that are identical after projection are \(\oplus\)-ed; in the bag semantics [29] that more closely models SQL, projection does not affect provenance annotations, but duplicate elimination \(\oplus\)-es the annotations of merged tuples;
- the provenance annotations of unioned tuples are \(\oplus\)-ed;
- the provenance annotations of tuples combined in a cross product are \(\otimes\)-ed.

**Example 5.** Consider the security semiring and query \(Q_1\), which can be rewritten in the relational algebra as:

\[
\Pi_{\text{city}}(\sigma_{\text{id2} < \text{id2}}(\Pi_{\text{id, city}}(\text{Personal}) \Join \\
\rho_{\text{id} \rightarrow \text{id2}}(\pi_{\text{id, city}}(\text{Personal}))))
\]

(the join operator \(\Join\) being a combination of a cross product, selection, and projection). Using the inductive definition of the provenance of a tuple in a query result, and assuming that the initial provenance tokens \(t_i\) are equal to the classification attribute of the tuple, one can compute the provenance of the output of the query as:

<table>
<thead>
<tr>
<th>city</th>
<th>New York</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Paris</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Berlin</td>
<td>28</td>
</tr>
</tbody>
</table>

Indeed, simply using this inductive definition of semiring provenance, one can use different semirings to compute different meta-information on the output of a query, with polynomial-time overhead in data complexity:

- **counting semiring:** the number of times a tuple can be derived;
- **Boolean semiring:** if a tuple exists when a sub-database is selected;
- **security semiring:** the minimum clearance level required to get a tuple as a result;
- **tropical semiring:** minimum-weight way of deriving a tuple (as when computing shortest paths in a graph);
- **positive Boolean functions:** Boolean provenance, as previously defined;
- **integer polynomials:** universal provenance, see further;
- **why-semiring:** why-provenance of [11], set of combinations of tuples needed for a tuple to exist.

However, [27] makes two important observations that lead to a different way to compute provenance annotations, instead of doing it one semiring at a time. First, semiring homomorphisms commute with provenance computation: if there is a homomorphism from semiring \(K\) to semiring \(K'\), then one can compute the provenance in \(K\), apply the homomorphism, and obtain the same result as when computing provenance in \(K'\). Second, the integer polynomial semiring \(\mathbb{N}[X]\) is the unique universal semiring in the following sense: there exists a unique homomorphism to any other commutative semiring that respects a given valuation of the variables. Combining these two facts, we have that all computations can be performed in the universal semiring, with homomorphisms only applied when the provenance for a given semiring is required. This suggests a way to implement provenance computation in a DBMS, discussed in Section 5.

Note that two queries that are equivalent in the usual sense [1] can have different semiring provenance, as semiring provenance captures more than logical equivalence. Indeed, two queries are logically equivalent if and only if they have the same Boolean provenance on every database.

Provenance semirings only capture the positive relational algebra, a relatively small fragment of SQL. We next discuss how to go beyond this fragment,
and investigate if all interesting forms of provenance are captured by provenance semirings.

4. BEYOND SEMIRING PROVENANCE

We now discuss some extensions of the provenance semiring framework.

Semirings with monus. Semiring provenance can only be defined for the positive fragment of the relational algebra, excluding non-monotone operations such as difference. However, some semirings can be straightforwardly equipped with a monus operator \( \ominus \) \([6,24]\), that captures non-monotone behavior. Such an operator must verify the following properties, for all \( a, b, c \in K \):

- \( a \oplus (b \ominus a) = b \oplus (a \ominus b) \);
- \( (a \ominus b) \ominus c = a \ominus (b \ominus c) \);
- \( a \ominus a = 0 - a = 0 \).

This is the case for the Boolean function semiring, which, equipped with the monus operator \( a \ominus b = a \wedge \neg b \), forms a semiring with monus, or \( m \)-semiring for short. This is also the case \([7]\) for the why-semiring with set difference, the integer polynomial and counting semirings with truncated difference on scalar values \( (a \ominus b = \max(0, a - b)) \), etc. Indeed, most natural semirings (though not all \([5]\)) can be extended to \( m \)-semirings.

Once such an \( m \)-semiring is defined, provenance of the full relational algebra can be captured in that \( m \)-semiring. For Boolean functions, it coincides with the Boolean provenance introduced in Section 4.

Note, however, that sometimes some seemingly natural axioms, such as distributivity of \( \otimes \) over \( \ominus \), fail over \( m \)-semirings \([7]\), which implies that two very similar queries may return different provenances.

Another important difference between \( m \)-semirings and semirings is that \( \mathbb{N}[X] \) is not a universal \( m \)-semiring \([24]\). There does indeed exist a unique universal \( m \)-semiring \([24]\), but it is simply the free \( m \)-semiring, i.e., the \( m \)-semiring of free terms constructed using \( \oplus, \otimes, \ominus \), quotiented by the equivalence relations imposed by the \( m \)-semiring structure.

We will illustrate in Section 5 how computation is performed using \( m \)-semirings.

Provenance for aggregates. One of the most natural ways to extend the relational algebra is to add aggregation capabilities \([37]\). There have been attempts at defining provenance formalisms for aggregate queries \([8,22]\). This is feasible in the case of associative and commutative aggregation, though it requires moving from annotations at the level of tuples to annotations at the level of values, as elements of a semimodule that combines provenance semiring annotation with scalar values from the range of the aggregation function. We believe the representation systems of Section 5 could be extended to provide somewhat compact representations of these semimodule elements, but leave this for future work.

Where-provenance. One notable provenance formalism that was introduced early on \([11]\) is where-provenance. The where-provenance is a bipartite graph that connects values in the output relation to values in the input relation to indicate where a specific value may come from in the input. It was shown \([12]\) that where-provenance cannot be captured by semiring provenance: there is no semiring for which semiring provenance allows reconstructing the where-provenance of a query. This is, intuitively, for two reasons:

- since where-provenance is assigned to individual values instead of tuples, it is affected either by renaming (in the named perspective) or by projection (in the unnamed perspective), as there needs to be a way to keep track of which value of a given tuple has which where-provenance;

- where-provenance is affected by joins differently as by a combination of cross product, selection, and projection: a value that results from a join of two relations has where-provenance pointing to the joined value in both relations.

However, a system keeping track of semiring provenance could be relatively straightforwardly extended to keep track of where-provenance: instead of dealing with values in a semiring (or in an \( m \)-semiring), just maintain value in a free algebra of terms, whose operator includes, in addition to \( \oplus, \otimes, \ominus \), operators to record projections (or renaming) and joined values.

Recursive queries. Query languages considered so far are unable to express recursive queries, such as shortest distance in a graph. It is also possible to define provenance notions for such queries as extension to semiring provenance, as long as the provenance formalism can express cycles. This was done in the original work on provenance semirings \([27]\) for \( \omega \)-continuous semirings, showing that the semiring \( \mathbb{N}[\mathbb{Z}[X]] \) of formal power series with integer coefficients is a universal \( \omega \)-continuous semiring. An alternative approach is the use of semirings with Kleene stars \([21]\), such as \( k \)-closed semirings \([39]\), for which efficient algorithms for provenance computation can be designed \([39]\). We leave details for further work, though we note that the representation systems we are introducing next – circuits – need to
be amended in the case of recursive queries, using either equation systems as in [27] or cycluits (cyclic circuits) as in [2].

5. PROVENANCE CIRCUITS

We now discuss concrete representations for provenance annotations. As we have seen, leaving the case of provenance for aggregates, where-provenance, and recursive queries, to future work, provenance annotations can be Boolean functions (see Section 2) useful for probabilistic databases, semiring values (see Section 3), or m-semiring values (see Section 4).

In addition, positive Boolean provenance is a special case of semiring provenance, and non-monotone Boolean provenance is a special case of m-semiring provenance. Finally, there exist a universal semiring and a universal m-semiring.

In some semirings (the Boolean, counting, and security semirings, for instance), provenance annotations are elementary, i.e., they are easily representable with enumerations or native types. Other semirings, such as \( \mathbb{N}[X] \) or the Boolean function semirings have complex annotations, for which a compact representation needs to be found.

In many previous works [27, 31, 45], provenance annotations have been represented as formulas, e.g., propositional formulas for Boolean provenance. But this leads to suboptimal representations, as (Boolean) formulas can be less compact than (Boolean) circuits [4, 47]. We therefore argue in favor of using provenance circuits, arithmetic circuits whose gates are the operators of the (m-)semiring as in [3, 20], as a compact representation system for provenance.

Example 6. Consider queries \( Q_1 \) and \( Q_2 \) on Personal. We can represent the provenance annotations of their output as references to gates in the universal m-semiring circuit shown in Figure 1. The output of \( Q_1 \) is as follows: _______ while that of \( Q_2 \) is: _______

\[
\begin{array}{ccc}
\text{city} & g_1 & g_2 \\
\text{New York} & 1 & \oplus \\
\text{Paris} & g_2 & \ominus \\
\text{Berlin} & g_3 & \otimes \\
\end{array}
\]

Indeed, by developing the circuit, one can for instance verify that the provenance of “New York” for \( Q_2 \) on Personal is \( (t_1 \oplus t_2) \ominus (t_1 \otimes t_2) \). As can be seen on Figure 1, a significant amount of sharing can be obtained for provenance within and across queries by using provenance circuits.

This suggests a practical way for computing provenance of queries over a relational database: inductively construct a provenance circuit over input tuples for each operation performed in a query, reusing parts of the circuit that have been constructed by subqueries. By constructing this circuit in the universal m-semiring, it then becomes easy to instantiate it to a wide variety of semirings and m-semirings.

Example 7. Consider the query \( Q_3 \) on Personal for which we want to compute the security and counting semiring annotations. Since we have already computed in Figure 1 a circuit for this query in the universal m-semiring, we can directly obtain a circuit whose evaluation returns the provenance of \( Q_3 \) in either of these semirings by applying the appropriate semiring homomorphisms. This is what is shown in Figures 2 and 3. One can verify that the evaluation of these queries returns the provenance annotations already computed in Example 5.

Similarly, one can compute the Boolean circuit of Figure 4 by applying the m-semiring homomorphism from the universal m-semiring of Figure 1 to the Boolean function semiring.

This approach of incremental provenance computation in the universal m-semiring, with specialization to arbitrary semirings and m-semirings on demand, is that taken by PROVSQL [43], a lightweight add-on to the PostgreSQL database management system for support of (m-)semiring provenance computation on relational databases. To our knowledge, this is the only publicly available system for management of data with semiring provenance, with support of a large subset of the SQL query language. The closest such software may be ORCHESTRA [25], which is unfortunately unavailable.

6. PROBABILITY EVALUATION

As discussed in Section 2, Boolean provenance is a very important tool to compute the probability of a query in probabilistic databases, an intractable (#P-hard) problem in general [45]. It allows separating the concerns between query evaluation on the one hand, which produces a Boolean provenance annotation in polynomial time, and probability evaluation of the provenance annotation on the other hand, itself a #P-hard problem.

Let us thus assume we have obtained a Boolean circuit of the data provenance of a query over a probabilistic database. What can, then, be done to evaluate the probability of the provenance annotation, given the intractability of the problem?

Brute-force algorithms. The first possible way is to resort to an exponential-time enumera-
Figure 1: Provenance circuit for queries $Q_1$ and $Q_2$ in the universal m-semiring

Figure 2: Provenance circuit for query $Q_1$ in the security semiring

Figure 3: Provenance circuit for query $Q_1$ in the counting semiring
tion of all possible valuations of the Boolean events occurring in the provenance annotation, and summing the probabilities of valuations mapped to \( \top \) by the provenance annotation to compute the overall probability. This is rarely feasible, but note that it is at least more efficient than enumerating all possible worlds of the initial database.

**Approximations.** One can always resort to approximating the probability of the query, either by Monte-Carlo sampling, which is always feasible but fairly slow, or by more refined approximation techniques, as in [42, 44].

**Exploiting the query structure.** Jha and Suciu [33] have shown that, when queries have specific forms, it is possible to construct Boolean provenance circuits of certain types, that allow for efficient probability evaluation. In particular, if a union of conjunctive queries (UCQ) is inversion-free, an ordered binary diagram (OBDD [10]) for it can be obtained. If some more general property is satisfied, then it admits a deterministic decomposable negation normal form (d-DNNF [16]). Both OBDDs and the more general class of d-DNNFs allow for efficient (linear-time) probabilistic query evaluation. Of course, this approach is inapplicable if the query is not of the specific form required. Note also that this approach is specific to Boolean provenance, which means it precludes computation of provenance in a more general (m-)semiring before specializing to the Boolean function case.

**Exploiting the data structure.** An alternative is to exploit the fact that the structure of the data is not arbitrary. Indeed, if the data has the structure of a tree, or has a low treewidth, meaning that its structure is close to that of a tree, it has been shown [2, 3] that a bounded-treewidth provenance circuit can be constructed, which in turn supports tractable query evaluation. This line of technique has been successfully applied to synthetic [40] and real-world [38] data, for specific kinds of queries.

When none of this is feasible, one can resort to general knowledge compilation techniques [18]. Knowledge compilation is the problem of transforming Boolean functions of a certain form into another, more tractable, form. Over the years, a wide variety of techniques, results, heuristics, and tools have emerged from the knowledge compilation community. In particular, tools such as c2d [17], Dsharp [41], and d4 [35] compile arbitrary formulas in conjunctive normal form into d-DNNFs.

One practical approach for probabilistic query evaluation is therefore to produce a Boolean provenance circuit, transform it into a conjunctive normal form in linear time using the standard Tseitin transformation [46], and feed it to a knowledge compiler. This approach is used in ProvSQL.

**Example 8.** Consider the middle connected component of the Boolean circuit of Figure 4, and, in particular, gate \( g_5 \) which yields the Boolean provenance of “Paris” in the output of query \( Q_2 \) on **Personal**. One can transform this part of the circuit into the following equivalent conjunctive normal form, where variables are inputs of the circuit along with its
Figure 5: d-DNNF for tuple “Paris” in the output of query $Q_2$

internal gates, using Tseitin transformation:

\[
\begin{align*}
g_5 \lor g_8 \lor \overline{g}_7 & \land g_5 \lor g_8 \\
\land \overline{g}_5 \lor g_7 & \land \overline{g}_8 \lor t_3 \lor t_5 \lor t_6 \\
\land g_8 \lor \overline{t}_3 & \land g_8 \lor \overline{t}_5 \\
\land g_8 \lor \overline{t}_6 & \land g_7 \lor \overline{g}_2 \\
\land \overline{g}_7 \lor \overline{g}_2 & \land \overline{g}_2 \lor t_9 \lor t_{10} \lor t_{11} \\
\land g_2 \lor t_9 & \land g_2 \lor \overline{t}_{10} \\
\land g_2 \lor \overline{t}_{11} & \land g_9 \lor \overline{t}_3 \lor \overline{t}_6 \\
\land \overline{g}_9 \lor t_3 & \land \overline{g}_9 \lor t_6 \\
\land g_{10} \lor \overline{t}_3 \lor \overline{t}_5 & \land \overline{g}_{10} \lor t_3 \\
\land \overline{g}_{10} \lor t_5 & \land g_{11} \lor \overline{t}_6 \lor \overline{t}_6 \\
\land g_{11} \lor t_6 & \land g_{11} \lor t_6 \\
\land g_5 &
\end{align*}
\]

Once such a formula obtained, it can be given as input to a knowledge compiler. For example, $d4$ outputs the d-DNNF in Figure 5. A d-DNNF is a special case of a Boolean circuit, where every $\neg$-gate is directly connected to an input, every $\land$-gate has children with disjoint sets of descendant leaves, and every $\lor$-gate is such that only one of its child can be true in any possible world. These restrictions make it possible to compute the probability of a gate in linear-time, given a probability distribution on input gate: $\land$-gates become products, while $\lor$-gates become sums. The computation of $\Pr(g_5)$ from the d-DNNF in Figure 4 is shown in Figure 6.

Note that this intensional (provenance-based) approach to probabilistic query evaluation is not the only one. The extensional approach, which directly manipulates probabilities as the query is evaluated, without an intermediate provenance representation, has also been successfully used [14,15]. It is an open problem [15,33] whether there are cases where the extensional approach succeeds but no compact d-DNNF is obtainable.

7. CONCLUSION

Data provenance is a major tool to obtain additional information on query output. It allows answering questions about:

- **why** in the why-semiring;
- **where** using where-provenance;
- **how** using integer polynomials;
- **missing tuples** using Boolean provenance;
- **how many times** in the counting semiring;
- **probability** using probability evaluation of Boolean provenance;
- **minimal security clearance** in the security semiring;
- **most economical way** in the tropical semiring.

Practical implementation of provenance management is very much possible, since it introduces a relatively low overhead. One avenue for practical implementations is to perform all computations in a universal structure, such as the universal m-semiring, and only specialize when needed. Using these provenance representations, it is also possible to perform query evaluation on probabilistic databases, for instance using knowledge compilation to obtain Boolean provenance representations on which probabilistic evaluation is efficient.

8. REFERENCES


Figure 6: Probability computation following the d-DNNF of Figure 5


