

On Optimizing the Trade-off between Privacy and Utility in Data Provenance

Daniel Deutch
Tel Aviv University
danielde@tauex.tau.ac.il

Ariel Frankenthal
Tel Aviv University
frankenthal@mail.tau.ac.il

Amir Gilad
Duke University
agilad@cs.duke.edu

Yuval Moskovitch
University of Michigan
yuvalm@umich.edu

Abstract

Organizations that collect and analyze data may wish or be mandated by regulation to justify and explain their analysis results. At the same time, the *logic* that they have followed to analyze the data, i.e., their queries, may be proprietary and confidential. Data provenance, a record of the transformations that data underwent, was extensively studied as means of explanations. In contrast, only a few works have studied the tension between disclosing provenance and hiding the underlying query.

This tension is the focus of the present paper, where we formalize and explore for the first time the tradeoff between the utility of presenting provenance information and the breach of privacy it poses with respect to the underlying query. Intuitively, our formalization is based on the notion of provenance abstraction, where the representation of some tuples in the provenance expressions is abstracted in a way that makes multiple tuples indistinguishable. The privacy of a chosen abstraction is then measured based on how many queries match the obfuscated provenance, in the same vein as k -anonymity. The utility is measured based on the entropy of the abstraction, intuitively how much information is lost with respect to the actual tuples participating in the provenance. Our formalization yields a novel optimization problem of choosing the best abstraction in terms of this tradeoff. We show that the problem is intractable in general, but design greedy heuristics that exploit the provenance structure towards a practically efficient exploration of the search space. We experimentally prove the effectiveness of our solution using the TPC-H benchmark and the IMDB dataset.

1 Introduction

Data provenance, namely a record of the transformations that pieces of data underwent when processed by a query, has been the subject of extensive investigation in recent years [44, 33, 28, 17, 8, 27, 49]. Most of these works focus on the utility of provenance, showing that it is highly effective for applications such as hypothetical reasoning [3, 4, 24], explaining and justifying query results [22, 9, 12], and others. The cost of provenance tracking is typically measured in terms of the execution time / memory overhead it incurs, and significant research effort has been ded-

icated to optimizing such computational aspects. In this paper, we shed light on a different kind of cost incurred by publishing provenance: the exposure of the *query* that has been executed and for which provenance has been tracked. We ask: *can we obfuscate provenance so that it remains useful, while hiding the underlying query?*

This aspect of provenance has become increasingly important as more and more agencies and organization aim to provide explanations for their decisions [31, 26] while governmental bodies and research communities stress the need for privacy-aware mechanisms [47, 34, 41].

Interests			
	PID	Interest	Source
i_1	1	Music	WikiLeaks
i_2	2	Music	Facebook
i_3	3	Music	LinkedIn
i_4	1	Parties	WikiLeaks
i_5	2	Parties	Facebook
i_6	4	Movies	WikiLeaks

Hobbies			
	PID	Hobby	Source
h_1	1	Dance	Facebook
h_2	2	Dance	LinkedIn
h_3	4	Dance	Facebook
h_4	1	Trips	Facebook
h_5	2	Trips	LinkedIn
h_6	3	Trips	WikiLeaks

Persons		
	PID	Name
p_1	1	James T
p_2	2	Brenda P

Figure 1: Partial Database instance of hobbies and interests of people collected from different sources

Example 1.1. Consider an online advertising company that wishes to match ads to people. Their database contains information about people, their hobbies and interests, a sample of which appears in Figure 1. Each tuple has an identifier, appearing to its left. The company may run queries such as Q_{real} appearing in Table 1 looking for people that like dancing and music. The query output includes James and Brenda, and relevant advertisements may then be presented to them. Upon request, Brenda may receive an explanation of why the advertisement was shown to her (see e.g., [31, 26]). In the case where James and Brenda are friends, they may obtain each other explanation in addition to their own. However, the company may wish to avoid disclosing the general criteria (i.e., the query Q_{real}),

Table 1: Queries for the running example. Q_{real} is the original, Q_{false1} , Q_{false2} are similar but not identical, and $Q_{general}$ is a generalization of the original

Name	Query
Q_{real}	$Q(id) \text{ :- } Person(id, name, age), \text{ Hobbies}(id, 'Dance', src1), \text{ Interests}(id, 'Music', src2)$
Q_{false1}	$Q(id) \text{ :- } Person(id, name, age), \text{ Hobbies}(id, 'Trips', src1), \text{ Interests}(id, 'Music', src2)$
Q_{false2}	$Q(id) \text{ :- } Person(id, name, age), \text{ Hobbies}(id, 'Dance', src1), \text{ Interests}(id, 'Parties', src2)$
$Q_{general}$	$Q(id) \text{ :- } Person(id, name, age), \text{ Hobbies}(id, 'Dance', src1), \text{ Interests}(id, interest, src2)$

since these criteria are part of the company’s confidential business strategy.

The provenance of a given query result describes the tuples used by the query to derive the result and the manner in which they were used. We use here the well-established model of *provenance semirings* [33].

Output	Provenance
1	$p_1 \cdot h_1 \cdot i_1$
2	$p_2 \cdot h_2 \cdot i_2$

(a) Ex_{real}

Output	Provenance
1	$p_1 \cdot h_4 \cdot i_1$
2	$p_2 \cdot h_5 \cdot i_2$

(b) Ex_{false1}

Output	Provenance
1	$p_1 \cdot h_1 \cdot i_4$
2	$p_2 \cdot h_2 \cdot i_5$

(c) Ex_{false2}

Figure 2: K -examples. Ex_{real} , Ex_{false1} and Ex_{false2} are the outputs of Q_{real} , Q_{false1} and Q_{false2} , respectively

Example 1.2. The provenance of the output tuple (1) according to the query Q_{real} shown in Table 1 is presented in the first row of Figure 2a. The expression, formulated as a product of the annotations p_1, h_1, i_1 , intuitively means that the three tuples with these annotations in the database (Figure 1) have jointly participated in an assignment to Q_{real} that yielded this result.

We denote by K -example a subset (“example”) of the results of a (hidden) query and an explanation for each result, formulated as its provenance (e.g., Figure 2a shows K -example derived by Q_{real} , modeling the explanations for James and Brenda). Given a K -example, the problem we address is *how to modify the provenance in a way that still allows users to gain information from it, but without divulging the underlying query that produced it?*

We next detail the main components of our solution.

Obfuscating provenance through abstraction. We propose a simple way to obfuscate provenance, based on *provenance abstraction*. The main idea is to allow identification of multiple provenance annotations, replacing them with a common “meta-annotation”. Not all such identifications make sense in general, and so their

choice is constrained by a tree whose leaves correspond to actual annotations and ancestors can be used as abstractions of their descendants. This technique has recently been proposed in [24], where it was used in a different context of reducing the provenance size.

Quantifying loss of information. We use entropy [45] to quantify the loss of information incurred by a choice of provenance abstraction. Information entropy expresses the level of uncertainty of a given data. In our context, we wish to measure “how uncertain” is a viewer of the abstracted provenance expression, with respect to the actual one (each possibility for the actual provenance, given an abstraction, is called a concretization). We assume a given distribution over the concretizations. Lacking additional knowledge, this distribution may simply be taken as uniform. The entropy for an abstraction is then defined with respect to a tree and a distribution.

Model for provenance privacy. Recall that our goal is to show an abstraction of a given K -example, while hiding the query that yielded the K -example. To measure the privacy of an abstraction, we may thus look at the set of its possible concretizations, and then at the set of queries that would have yielded each concretization. In fact, not all such queries are “interesting”: we may restrict attention to connected inclusion-minimal queries [23], i.e., queries whose join graph is connected and are not included in any other query in this set. These queries are representative of the viable options for the hidden query. We then define the privacy incurred by an abstraction as the cardinality of this set (i.e., how many connected inclusion-minimal queries match some concretization).

The problem of optimizing abstractions. The last two components are then combined to define the problem introduced and studied in this paper: given an example of query results and their provenance Ex , a provenance abstraction tree T , and a privacy threshold k , we aim at finding an abstraction that has at least k connected inclusion minimal queries that ‘can fit’ it, and minimizes the loss of information among all such abstractions.

Example 1.3. Consider the K -example Ex_{real} presented in Figure 2a showing two outputs of the query Q_{real} and their provenance. The allowed abstractions are defined based on the tree T depicted in Figure 3. The leaves of T are annotations (identifiers) of the tuples in Figure 1, and its inner nodes are abstracted forms of these annotations. An abstraction of the provenance in Ex_{real} w.r.t. T may, e.g., replace the annotation h_1 with its ancestors Facebook or Social Network. Other tuple annotations may be abstracted as well. A choice of abstraction dictates a certain amount of information loss since the annotation Facebook can stand for any one of the annotations h_1, h_3, h_4, i_2, i_5 , and when viewing the annotation Facebook we cannot be sure which annotation is the original. At the same time, it may obfuscate the underlying query Q_{real} , as more queries become consistent with the observable provenance information.

We study the complexity of the problem and show that

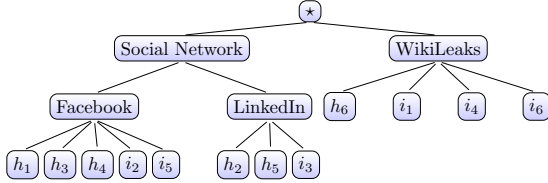


Figure 3: Abstraction tree containing a subset of tuple annotations in the database in Figure 1 as leaves, and inner nodes that are abstractions of the leaves

it is intractable in general. Namely, deciding the existence of an abstraction with privacy at least k and loss of information of at most l , is NP-hard. Bearing this bound in mind, we provide novel heuristic algorithms for computing optimal abstractions in practically efficient ways. Our approach revolves around several key ideas. First, we optimize the order of traversal over the possible abstractions, by examining “simpler” abstractions first. We further prioritize the computation of loss of information over privacy, as the former can be done significantly more efficiently. Additionally, privacy computation is performed in a greedy fashion, relying on the properties of the K -example. Finally, caching is used in order to avoid repetitive computations. Our heuristics and optimizations render our approach scalable even for large databases and complex queries, as observed in our experiments overviewed next.

Experimental evaluation. We have conducted an experimental study using the TPC-H [5] and the IMDB [37] datasets in which we examined the scalability and usability of our solution for different settings. We study the performance in terms of varying data, tree sizes, query complexity, K -example size, and privacy thresholds. We show that thanks to our optimizations, our solution is efficient even in complex settings that involve queries with many joins, large volumes of data and a large space of abstractions. We have also compared our solution with the provenance compression-based method presented in [24]. Finally, we performed a user study, showing that abstracted K -examples provide the desired privacy while still being informative and useful.

2 Preliminaries

We now define the background needed for our model. A summary of the notations used throughout the paper is shown in Table 2.

2.1 Query Language and Provenance

We give a brief review of the concepts of Union of Conjunctive Queries and Provenance Polynomials.

Union of conjunctive queries. We recall the concept of Unions of Conjunctive Queries. Fix a database schema \mathcal{S} with relation names $\{R_1, \dots, R_n\}$ over a domain \mathcal{C} of constants. Further fix a domain \mathcal{V} of variables. A CQ Q over \mathcal{S} is an expression of the form

Table 2: Notations

Q	Union conjunctive query
Ex	K -example
T	Abstraction tree
A_T	Abstraction function
\tilde{Ex}	Abstracted K -example
$Var(Ex)$	Set of variables in Ex
V_T	Set of nodes of tree T
L_T	Set of leaves of tree T
$L_T(v)$	Set of leaves of the subtree of T rooted in v
$C(\tilde{Ex})$	Concretization set of \tilde{Ex}

$T(\vec{u}) : -R_1(\vec{v}_1), \dots, R_l(\vec{v}_l)$ where T is a relation name not in \mathcal{S} . For each $1 \leq i \leq n$, \vec{v}_i is a vector of the form (x_1, \dots, x_k) where $\forall 1 \leq j \leq k. x_j \in \mathcal{V} \cup \mathcal{C}$. $T(\vec{u})$ is the query head, denoted $head(Q)$, and $R_1(\vec{v}_1), \dots, R_l(\vec{v}_l)$ is the query body and is denoted $body(Q)$. The variables appearing in \vec{u} are called the *head variables* of Q , and each of them must also appear in the body. A union of such queries is a UCQ. We use UCQ to denote the class of all UCQs, omitting details of the schema when clear from the context.

Next, we define the notion of *derivations* for UCQs. A derivation α for a query $Q \in UCQ$ with respect to a database instance D is a mapping of the relational atoms of Q to tuples in D that respects relation names and induces a mapping over arguments, i.e., if a relational atom $R(x_1, \dots, x_n)$ is mapped to a tuple $R(a_1, \dots, a_n)$ then we say that x_i is mapped to a_i (denoted $\alpha(x_i) = a_i$). We require that a variable x_i will not be mapped to multiple distinct values, and a constant x_i will be mapped to itself. For a CQ $q \in Q$, we define $\alpha(head(q))$ as the tuple obtained from $head(q)$ by replacing each occurrence of a variable x_i by $\alpha(x_i)$.

Example 2.1. Reconsider the CQ Q_{real} depicted in Table 1 and the output tuple (1) in the first row of Figure 2a. It is derived using the tuples with annotations p_1, h_1, i_1 (Figure 1) that are mapped to the first, second and third atom of Q_{real} respectively.

Provenance semirings. We focus on databases whose tuples are associated (“annotated”) with elements of a set X , or polynomials (with positive coefficients) thereof [33]. X may be thought of as a set of identifiers each attached to a single input tuple.

A *commutative monoid* (from [23]) is an algebraic structure $(M, +_M, 0_M)$ where $+_M$ is an associative and commutative binary operation and 0_M is an identity for $+_M$. A *commutative semiring* is then a structure $(K, +_K, \cdot_K, 0_K, 1_K)$ where $(K, +_K, 0_K)$ and $(K, \cdot_K, 1_K)$ are commutative monoids, \cdot_K is distributive over $+_K$, and $a \cdot_K 0_K = 0 \cdot_K a = 0_K$. A K -relation is a mapping between tuples and elements of K . A K -database D over a schema $\{R_1, \dots, R_n\}$ is then a collection of K -relations, over each R_i . Unless stated otherwise, we will assume that in databases used as input to queries, all relations

are abstractly-tagged: namely, each tuple is annotated by a distinct element of X (intuitively, its identifier).

We then define UCQs as mappings from K -databases to K -relations. Intuitively, we define the annotation (provenance) of an output tuple as a combination of annotations of input tuples. The idea is that given a set of basic annotations X (elements of which may be assigned to input tuples), the provenance of an output is represented by a sum of products, i.e., a polynomial. Coefficients serve in a sense as a “shorthand” for multiple derivations using the same tuples, and exponents as a “shorthand” for multiple uses of a tuple in a derivation.

Definition 2.2 (adapted from [33]). *Let D be a K -database and let $Q \in UCQ$, with T_i being the relation name in $\text{head}(q_i)$ where $q_i \in Q$ is a CQ in Q . For every tuple $t \in T_i$, let α_t be the set of derivations of q_i w.r.t. D that yield t . $q_i(D)$ is defined to be a K -relation T_i s.t. for every t , $T_i(t) = \sum_{\text{head}(q_i)=T_i} \sum_{\alpha \in \alpha_t} \prod_{t' \in \text{Im}(\alpha)} \text{Ann}(t')$, where $\text{Im}(\alpha)$ is the image of α , and $\text{Ann}(t')$ is the annotation of t' according to its K -relation.*

Example 2.3. *In Example 2.1, we showed that the output tuple (1) of Q_{real} (Table 1) is derived from the tuples annotated by p_1, h_1, i_1 . As a provenance polynomial, this corresponds to the monomial $p_1 \cdot h_1 \cdot i_1$.*

Provenance examples. We now define the notion of a K -example, which intuitively captures output examples and their explanations as provenance.

Definition 2.4 (adapted from [23]). *A K -example is a pair (I, O) where I is an abstractly-tagged K -database called the input and O is a K -relation called the output.*

In words, O denotes an output example and I its provenance.

Example 2.5. *A K -example is depicted in Figure 2a where the left column shows two output examples, O_1 and O_2 , and the right column shows the provenance of each of them, I_1 and I_2 , respectively.*

For a K -example $Ex = (I, O)$, we denote by $\text{Var}(Ex)$ the set of tuple annotations in I (see Table 2).

2.2 Provenance Abstraction Tree

We define an *abstraction tree* over the provenance variables, drawing on [24]. Intuitively, this defines groupings of different variables with a single value as a generalized representation of all of them. The tree is structured so that the labels associated with tuples of the input examples are at the leaf level; inner nodes stand for abstractions of the labels associated with leaves of their sub-trees.

Definition 2.6. *An abstraction tree T is a rooted labeled tree, where each node has a unique label (we thus use “node” and “label” interchangeably). V_T is used to denote the set of labels in T and L_T is the set of labels of the leaves in T . Given a K -database D , we say that T is compatible with D if $(V_T \setminus L_T) \cap (\cup_{t \in D} \text{Ann}(t)) = \emptyset$.*

We say that an abstraction tree T is *compatible* with a K -example (I, O) if T is compatible with I . If T is not compatible with a K -example then it cannot be used as an abstraction tree for this particular K -example. We will discuss ways of constructing abstraction trees at the end of Section 4.

Example 2.7. *Reconsider the K -example Ex_{real} presented in Figure 2a. The abstraction tree T shown in Figure 3 is compatible with Ex_{real} since none of the inner nodes of T (e.g., Facebook) are labeled by the variables of Ex_{real} .*

3 Model

We define our novel model for the problem of provenance privacy.

3.1 Abstractions and Concretizations

Let T be an abstraction tree. For $v, v' \in V_T$, we say that $v \leq_T v'$ if v is a descendant of v' in T (or $v' = v$).

Definition 3.1 (Abstraction Function). *Given an abstraction tree T that is compatible with a K -example Ex and an ordering over the variables of Ex where each variable occurrence is assigned an index $i \in \mathbb{N}$, an abstraction function over T is a function $A_T : \text{Var}(Ex) \times \mathbb{N} \rightarrow (V_T \cup \text{Var}(Ex))$ that maps each occurrence of a variable $v \in \text{Var}(Ex)$ at index i such that $v \in L_T$ to $v' \in V_T$, where $v \leq_T v'$. If $v \notin L_T$, $A_T(v, i) = v$.*

Note that A_T may map different occurrences of the same variable v to different nodes in T , namely, it is possible to have $A_T(v, i) \neq A_T(v, j)$, where $A_T(v, i)$ ($A_T(v, j)$) is the mapping of the i -th (resp. j) occurrence of v . To simplify notations, in the rest of the paper we assume each variable appears once, and omit the index from A_T . Overloading notation, we use $A_T(Ex)$ to denote the K -example \bar{Ex} obtained by replacing each $v \in \text{Var}(Ex)$ by $A_T(v)$ for all $v \in L_T$.

We next demonstrate the notion of abstraction function. In practice, these functions are generated automatically by the algorithm given in Section 4. In the rest of the paper, we will use the term *abstraction* interchangeably for the concepts of an abstraction function and its output, an abstracted K -example.

Example 3.2. *Reconsider the K -example Ex_{real} given in Figure 2a and the abstraction function A_T^1 depicted in Figure 4. Using A_T^1 on Ex_{real} will create the abstracted K -example \bar{Ex}_{abs1} shown in Figure 5. Formally, $A_T^1(Ex_{\text{real}}) = \bar{Ex}_{\text{abs1}}$.*

A concretization is then the ‘reverse’ operation of abstraction.

Definition 3.3 (Concretization). *Given an abstracted K -example \bar{Ex} and an abstraction tree T , a K -example Ex is*

$$A_T^1(v) = \begin{cases} \text{Facebook}, & \text{if } v = h_1, h_4 \\ \text{LinkedIn}, & \text{if } v = h_2, h_5 \\ v, & \text{otherwise} \end{cases}$$

$$A_T^2(v) = \begin{cases} \text{WikiLeaks}, & \text{if } v = i_1, i_4 \\ \text{Facebook}, & \text{if } v = i_2, i_5 \\ v, & \text{otherwise} \end{cases}$$

$$A_T^3(v) = \begin{cases} \text{WikiLeaks}, & \text{if } v = i_1 \\ v, & \text{otherwise} \end{cases}$$

Figure 4: Abstraction Functions

$$\widetilde{Ex}_{abs1} = A_T^1(Ex_{real}) = A_T^1(Ex_{false1}) =$$

Output	Provenance
1	$p_1 \cdot \text{Facebook} \cdot i_1$
2	$p_2 \cdot \text{LinkedIn} \cdot i_2$

$$\widetilde{Ex}_{abs2} = A_T^2(Ex_{real}) = A_T^2(Ex_{false2}) =$$

Output	Provenance
1	$p_1 \cdot h_1 \cdot \text{WikiLeaks}$
2	$p_2 \cdot h_2 \cdot \text{Facebook}$

$$\widetilde{Ex}_{abs3} = A_T^3(Ex_{real}) =$$

Output	Provenance
1	$p_1 \cdot h_1 \cdot \text{WikiLeaks}$
2	$p_2 \cdot h_2 \cdot i_2$

Figure 5: Abstracted K -examples

a concretization of \widetilde{Ex} if there exists an abstraction function A_T such that $A_T(Ex) = \widetilde{Ex}$. The concretization set of \widetilde{Ex} is $C(\widetilde{Ex}) = \{Ex \mid \exists A_T. A_T(Ex) = \widetilde{Ex}\}$

Since sub-trees in the abstraction tree may have multiple leaves, an abstracted K -example can have more than one concretization. Therefore, we have defined the *concretization set* containing all options for concretizations.

Example 3.4. Consider again the abstracted K -example \widetilde{Ex}_{abs1} presented in Figure 5, the K -example Ex_{real} shown in Figure 2a and the abstraction function A_T^1 given in Figure 4. From Example 3.2, we have $Ex_{real} \in C(\widetilde{Ex}_{abs1})$ since $A_T^1(Ex_{real}) = \widetilde{Ex}_{abs1}$. Now consider the K -example Ex_{false1} shown in Figure 2b. It also holds that $A_T^1(Ex_{false1}) = \widetilde{Ex}_{abs1}$, and thus $Ex_{false1} \in C(\widetilde{Ex}_{abs1})$, i.e., Ex_{false1} is also in the concretization set of \widetilde{Ex}_{abs1} . $C(\widetilde{Ex}_{abs1})$ also contains other K -examples beside Ex_{real} and Ex_{false1} .

The following are simple observations regarding the size of a concretization set that will be useful in the sequel. Note that L_T is the set of leaves of the abstraction tree T and $L_T(v)$ is the set of leaves of the subtree of T rooted in v .

Proposition 3.5. Given an abstraction tree T that is compatible with a K -example Ex and an abstraction function A_T , it holds that:

$$1. |C(A_T(Ex))| = \prod_{v \in Var(Ex)} |L_T(A_T(v))|$$

$$2. 1 \leq |C(A_T(Ex))| \leq |L_T|^n, \text{ where } n = |\{v \in Var(Ex) \mid v \neq A_T(v)\}|, \text{ and these bounds are tight.}$$

Proof. 1. In induction on the number of abstracted values $n = |\{v \in Var(A_T(Ex)) \mid v \neq A_T(v)\}|$. If $n = 0$ it holds that $\forall v \in Var(A_T(Ex)), v = A_T(v)$. Thus, $\forall v$ it holds that

$$\begin{aligned} \prod_{v \in Var(Ex)} |L_T(A_T(v))| &= \prod_{v \in Var(Ex)} |L_T(v)| \\ &= \prod_{v \in Var(Ex)} |\{v\}| = 1 \end{aligned}$$

It is also clear that $|C(A_T(Ex))| = 1$ since the abstraction function is the identity function, so Ex itself is the only concretization that holds $Id(Ex) = Ex$ and the base case is true.

About the inductive step, let's assume the proposition holds for n and we will prove it for $n + 1$. Let's denote $Var(Ex) = \{v_1, \dots, v_m\}, n < m$. Now, w.l.o.g, assume that if $i \in \{1, \dots, n\}$ then $v_i \neq A_T(v_i)$ and if $i \in \{n + 1, \dots, m\}$ then $v_i = A_T(v_i)$. Now, by the inductive assumption it holds that:

$$\begin{aligned} |C(A_T(Ex))| &= \prod_{v \in Var(Ex)} |L_T(A_T(v))| \\ &= \prod_{i=1}^n |L_T(A_T(v_i))| \end{aligned}$$

The last equality holds since if $v = A_T(v)$ then

$$|L_T(A_T(v))| = |L_T(v)| = 1$$

so it is not effect the product.

Now, for the $n + 1$ case, we changed A_T s.t.

$$v_{n+1} \in Var(A_T(Ex)), v_{n+1} \neq A_T(v_{n+1})$$

The concretization set contains only K -examples Ex' that holds $\exists A_T. A_T(Ex') = Ex$. By definition, an abstraction function $A_T : L_T \rightarrow V_T$ is a function that transform each leaf v to a single ancestor v' in the tree. From that we know that if $Ex' \in C(Ex)$ it holds that

$$\forall v \in V(Ex'), v \neq A_T(v) \Rightarrow v \in L_T(v)$$

We also know that $v_{n+1} \neq A_T(v_{n+1})$ so it holds that $A_T(v_{n+1})$ can be any $l \in L_T(A_T(v_{n+1}))$, and since there are $|L_T(A_T(v_{n+1}))|$ options for that value, we multiple all the previous concretization with every new option. Thus,

$$\begin{aligned} |C(A_T(Ex))| &= |L_T(A_T(v_{n+1}))| \cdot \prod_{i=1}^n |L_T(A_T(v_i))| \\ &= \prod_{i=1}^{n+1} |L_T(A_T(v_i))| \end{aligned}$$

and we are done.

2. It is clear that $1 \leq |C(A_T(Ex))|$ since if the abstraction function is the identity function it is always true that $Id(Ex) = Ex$, so Ex itself is a concretization.

Now, since $\forall v, L_T(A_T(v)) \leq L_T$, from the previous part we get that:

$$|C(A_T(Ex))| = \prod_{i=1}^n |L_T(A_T(v_i))| \leq \prod_{i=1}^n |L_T| = |L_T|^n$$

and we are done.

3. For the first equality, choosing A'_T to be the identity function, i.e., $A'_T(v) = v$, the only concretization is Ex itself, so $|C(A'_T(Ex))| = 1$.

For the second equality, we denote by r the abstraction tree T 's root. Consider the following abstraction function: $A''_T(v) = r, \forall v \in Var(Ex)$. With this abstraction tree, $|L_T(A''_T(v_i))| = |L_T|, \forall v \in Var(Ex)$, so it holds that:

$$|C(A''_T(Ex))| = \prod_{i=1}^n |L_T(A''_T(v_i))| = \prod_{i=1}^n |L_T| = |L_T|^n$$

and we are done. \square

3.2 Loss of Information

Each abstraction entails a loss of information. We measure the loss of information of an abstracted K -example \widetilde{Ex} via the notion of *Entropy*. Entropy is the average level of ‘information’ or ‘uncertainty’ inherent in the possible outcomes of a random variable [45]. Given a random variable X , with possible outcomes x_i , each with probability $P_X(x_i)$, the entropy $H(X)$ of X is as follows: $H(X) = -\sum_i P_X(x_i) \ln P_X(x_i)$. The entropy quantifies how ‘informative’ or ‘surprising’ the random variable is, averaged over all of its possible outcomes. Next, we define the entropy induced by abstraction, as follows:

Definition 3.6. *Given an abstraction tree T that is compatible with a K -example Ex , an abstraction function A_T and a probability space on $X = C(A_T(Ex))$ (the concretization set of $A_T(Ex)$) we define the loss of information by $LOI(A_T(Ex)) = -\sum_{i=1}^n P_X(x_i) \ln P_X(x_i)$ where $X = C(A_T(Ex)) = \{x_1, \dots, x_n\}$ and $P_X(x_i)$ is the probability of the concretization x_i .*

The probabilities may be determined using statistical properties of the database or external information. Note that for a finite probability space X with a discrete uniform distribution over n states, the entropy is $H(X) = \ln(n)$. Since $C(A_T(Ex))$ is a finite set (Proposition 3.5), if the probabilities of all concretizations in $C(A_T(Ex))$ are equal then $LOI(A_T(Ex)) = \ln(|C(A_T(Ex))|)$.

Example 3.7. *Reconsider the abstracted K -example Ex_{real} presented in Figure 2a, the abstracted tree T shown in Figure 3 and the abstraction function A_T^3 depicted in Figure 4. The output of $A_T^3(Ex_{real})$ is the abstracted K -example \widetilde{Ex}_{abs3} shown in Figure 5. The concretization set*

of \widetilde{Ex}_{abs3} is given in Figure 6. Assuming the probabilities of the concretizations are $P_X(c_1) = 0.1$, $P_X(c_2) = 0.2$, $P_X(c_3) = 0.3$ and $P_X(c_4) = 0.4$. the loss of information of \widetilde{Ex}_{abs3} is $-\sum_{i=1}^4 P_X(c_i) \ln P_X(c_i) = -(0.1 \cdot \ln 0.1 + \dots + 0.4 \cdot \ln 0.4) \approx 1.279$

3.3 Privacy

We next define our privacy measure.

Consistent and CIM queries. Next, we define the concepts of consistent and connected inclusion-minimal queries with respect to a K -example. Our definitions are inspired by [23] and extend them. As a preliminary step, we define subsumption of K -relations.

Definition 3.8 (from [23]). *Let $(K, +_K, \cdot_K, 0, 1)$ be a semiring and define $a \leq_K b$ iff $\exists c. a +_K c = b$. If \leq_K is a (partial) order relation then we say that K is naturally ordered. Given two K -relations R_1, R_2 we say that $R_1 \subseteq_K R_2$ iff $\forall t. R_1(t) \leq_K R_2(t)$.*

We now define a consistent query w.r.t. an abstracted example. Intuitively, a query Q is consistent w.r.t. \widetilde{Ex} if there exists a concretization of \widetilde{Ex} for which Q generates the output tuples when given the provenance, and the provenance generated by Q matches the one specified in the concretization.

Definition 3.9. *[consistent query] Given an abstracted K -example \widetilde{Ex} and a CQ Q we say that Q is consistent with respect to the example \widetilde{Ex} if there exists $(I, O) \in C(\widetilde{Ex})$ such that $O \subseteq_K Q(I)$.*

To define privacy, we use the concept of *connected inclusion-minimal queries* (CIM queries). Intuitively, we define the privacy criterion by the number of the most ‘focused’ queries. We draw on previous works in the field of query-by-example [38] that looks for connected queries and on [23] that looks for minimality in terms of inclusion. Recall that the join graph for a CQ is defined by the set of relations in its body $\{R_1, \dots, R_m\}$ with an edge (R_i, R_j) iff R_i and R_j share at least one variable. We say that a query is connected if its join graph is connected.

Definition 3.10 (CIM query). *A consistent query Q with respect to a given abstracted K -example \widetilde{Ex} is a CIM query if it is connected and for every query Q' such that $Q' \subsetneq_K Q$, (i.e., for every K -database D it holds that $Q'(D) \subseteq_K Q(D)$, but not vice-versa), Q' is not consistent with respect to \widetilde{Ex} . Namely, $\forall Ex \in C(\widetilde{Ex}), Q'$ is not a consistent query of Ex .*

Example 3.11. *Consider the abstracted K -example \widetilde{Ex}_{abs3} in Figure 5 and its concretization set given in Figure 6. There is only one CIM query w.r.t. \widetilde{Ex}_{abs3} which is Q_{real} (shown in Table 1) since it is consistent w.r.t. the concretization c_2 , connected and minimal w.r.t. all other consistent connected queries. Now consider the query $Q_{general}$ (shown in Table 1). It is consistent w.r.t.*

the concretization c_3 and connected. However, $Q_{general}$ is not CIM since $Q_{real} \subseteq Q_{general}$ (both queries have the same structure but Q_{real} contains an extra constant).

Definition 3.10 may consider trivial queries as CIM if we allow for union. For example, in \tilde{Ex}_{abs3} in Figure 5, the concretization c_1 in Figure 6 leads to the trivial CIM query $Q = q_1 \cup q_2$ where $q_1(1) : -p_1, h_1, h_6$ and $q_2(1) : -p_2, h_2, i_2$. Naturally, these types of UCQs do not generalize the K -example and therefore are not likely queries. In Section 4, we discuss a version of our solution that disqualifies such trivial queries.

$$C(\tilde{Ex}_{abs3}) = \left\{ \begin{array}{l} c_1 = \begin{array}{|c|c|} \hline \text{Output} & \text{Provenance} \\ \hline 1 & p_1 \cdot h_1 \cdot h_6 \\ \hline 2 & p_2 \cdot h_2 \cdot i_2 \\ \hline \end{array} \quad c_3 = \begin{array}{|c|c|} \hline \text{Output} & \text{Provenance} \\ \hline 1 & p_1 \cdot h_1 \cdot i_4 \\ \hline 2 & p_2 \cdot h_2 \cdot i_2 \\ \hline \end{array} \\ c_2 = \begin{array}{|c|c|} \hline \text{Output} & \text{Provenance} \\ \hline 1 & p_1 \cdot h_1 \cdot i_1 \\ \hline 2 & p_2 \cdot h_2 \cdot i_2 \\ \hline \end{array} \quad c_4 = \begin{array}{|c|c|} \hline \text{Output} & \text{Provenance} \\ \hline 1 & p_1 \cdot h_1 \cdot i_6 \\ \hline 2 & p_2 \cdot h_2 \cdot i_2 \\ \hline \end{array} \end{array} \right\}$$

Figure 6: Concretization Set of \tilde{Ex}_{abs3} (from Figure 5)

Privacy of an abstracted K -example. We are now ready to define the privacy of a K -example. Our definition is similar in spirit to the k -anonymity criterion in data privacy [48].

Definition 3.12 (Privacy). *The privacy of an abstracted K -example \tilde{Ex} is the number of unique CIM queries w.r.t. \tilde{Ex} .*

As with k -anonymity, a higher number of unique CIM queries w.r.t. an abstracted K -example indicates that this abstracted K -example is more private. Even an abstracted K -example can reveal some information about the query structure. In particular, the tables participating in the query and possibly also the join structure can be inferred from the combination of the schema and the K -example.

Example 3.13. *Reconsider the abstracted K -example \tilde{Ex}_{abs1} presented in Figure 5. We now detail the CIM queries w.r.t. \tilde{Ex}_{abs1} . First, we note that the consistent queries w.r.t. \tilde{Ex}_{abs1} are depicted in Table 3. We choose only the queries that are connected (the queries marked by ‘con’). From these, we choose only the queries that are inclusion-minimal w.r.t. \tilde{Ex}_{abs1} . Those are the queries marked with ‘min’ as well. Therefore, the CIM queries are annotated with ‘con, min’. There are only 2 queries that fulfill these terms, Q_{real} and Q_{false1} (shown in Table 1). Thus, the privacy of \tilde{Ex}_{abs1} is 2.*

Note that in Example 3.13, all disconnected queries are missing the logic expressed by the connected queries.

3.4 Problem Definition

We are now ready to define the problem of provenance abstraction. In short, given a K -example and a privacy

Table 3: Some of the consistent queries w.r.t. \tilde{Ex}_{abs1} from Figure 5. There is a total of 14 consistent queries. From those, 3 are connected (labeled ‘con’), and from those 2 are CIM (labeled ‘con, min’). This shows that the privacy of \tilde{Ex}_{abs1} is 2

Class	Query
con, min	$Q(a) :- \text{Person}(a,b,c), \text{Hobbies}(a,\text{‘Dance’},d), \text{Interests}(a,\text{‘Music’},e)$
	$Q(a) :- \text{Person}(a,p,q), \text{Hobbies}(r,s,t), \text{Interests}(u,v,w)$
	$Q(a) :- \text{Person}(a,b,c), \text{Hobbies}(d,\text{‘Dance’},e), \text{Interests}(a,\text{‘Music’},f)$
con	$Q(a) :- \text{Person}(a,b,c), \text{Hobbies}(a,d,e), \text{Interests}(a,\text{‘Music’},f)$
con, min	$Q(a) :- \text{Person}(a,b,c), \text{Hobbies}(a,\text{‘Trips’},d), \text{Interests}(a,\text{‘Music’},e)$
	$Q(a) :- \text{Person}(a,b,c), \text{Interests}(d,\text{‘Music’},e), \text{Interests}(a,\text{‘Music’},f)$

threshold, we want to find an abstraction that satisfies this threshold but also minimizes the loss of information.

Definition 3.14. [Problem Definition] *Given an abstraction tree T that is compatible with a K -example Ex and $k \in \mathbb{N}$ a privacy threshold, our goal is to find an abstraction function A_T where $A_T(Ex)$ has privacy $\geq k$, and A_T minimizes $A_T(Ex)$ ’s loss of information out of all the abstraction functions that guarantee privacy $\geq k$. We call this abstraction an optimal abstraction.*

Example 3.15. *Reconsider the database depicted in Figure 1, the query Q_{real} shown in Table 1, its output Ex_{real} given in Figure 2a and the abstraction tree T presented in Figure 3. Assume that the privacy threshold is 2 (i.e., we want our privacy to be at least 2) and the loss of information is entropy with discrete uniform distribution. We can use the abstraction function A_T^2 (detailed in Figure 4) so that $A_T^2(Ex_{real})$ yields \tilde{Ex}_{abs2} (depicted in Figure 5). Since the queries Q_{real} and Q_{false2} (shown in Table 1) are CIM w.r.t. \tilde{Ex}_{abs2} , its privacy is 2. In addition, $\ln|C(\tilde{Ex}_{abs2})| = \ln(5 \cdot 4) = \ln 20 \approx 2.996$, thus the loss of information incurred by $A_T^2(Ex_{real})$ is 2.996. On the other hand, we can use the abstraction function A_T^1 (detailed in Figure 4) so that $A_T^1(Ex_{real})$ yields \tilde{Ex}_{abs1} (depicted in Figure 5). In Example 3.13 we have seen that the privacy of \tilde{Ex}_{abs1} is 2. In addition, $\ln|C(\tilde{Ex}_{abs1})| = \ln(5 \cdot 3) = \ln 15 \approx 2.708$, thus the loss of information incurred by $A_T^1(Ex_{real})$ is 2.708. Since the loss of information of A_T^1 is smaller than all possible abstraction functions that guarantee privacy ≥ 2 (in particular, A_T^2), it is an optimal abstraction.*

Aggregate queries. A model for provenance for aggregation queries was defined in [1]. In a nutshell, the aggregation result is represented as a semimodule, that couples, using a tensor product, values from the aggregate domain and the tuple annotations. For example, consider an aggregate query with a similar structure to that of Q_{real} (shown in Table 1), that performs a MAX aggregation on the age attribute, i.e., instead of the people ids it returns

the maximal age of all people that like dancing and music. In this case the resulting aggregate value would be $(p_1 \cdot h_1 \cdot i_1) \otimes 27 +_{MAX} (p_2 \cdot h_2 \cdot i_2) \otimes 31$. Our model can support queries with aggregation over the head variables, where abstraction functions operate on the tuple's annotation part in the semimodule. For instance, the result of applying A_T^1 (shown in Figure 4) on the aforementioned aggregate result is $(p_1 \cdot \text{Facebook} \cdot i_1) \otimes 27 +_{MAX} (p_2 \cdot \text{LinkedIn} \cdot i_2) \otimes 31$.

4 Hardness and Solution

We first note that the optimal abstraction problem is intractable. To this end we define the decision problem version of the optimal abstraction: given an abstraction tree compatible with a K -example and integers k, l , determine whether there is an abstraction function that gives a privacy of at least k with at most l loss of information. This decision problem is NP-hard in the size of the intersection of the provenance variables with the leaves of the abstraction tree.

Proposition 4.1. *The decision problem version of the optimal abstraction is NP-hard.*

prov.	V	E	N
VC_1	v_1	e_i	2
\vdots	\vdots	\vdots	\vdots
VC_{2m}	v_n	e_j	2
yes	0	1	3

(a) Relation VC

prov.	J
E_1	e_1
\vdots	\vdots
E_m	e_m
ec	1

(b) Relation E

Figure 7: Database instance for the proof of Proposition 4.1

Output	Provenance
e_1, \dots, e_m	$VC_1 \dots VC_k \cdot E_1 \dots E_m$
$1, \dots, 1$	$yes^k \cdot ec^m$

Figure 8: K -example for the proof of Proposition 4.1

Proof. We show that the problem is NP-hard by reduction from the decision problem version of Vertex Cover. The input to Vertex Cover is a graph $G = (V, E)$, where $|V| = n$, $|E| = m$ and an integer $k \in \mathbb{N}$, and the solution is a set of vertices $C \subseteq V$ such that $\forall e \in E. C \cap e \neq \emptyset$.

Given such an input, we define the relation $VC(V, E, N)$, where $(v_i, e_j, 2) \in VC$ iff $v_i \in e_j$ (these tuples are denoted by VC_1, \dots, VC_{2m}). VC also contains the additional tuple $(0, 1, 3)$, denoted by yes . The relation is shown in Figure 7a. Next, we define the relation E depicted in Figure 7b: for each edge $e_j \in E$, we have a tuple $E(e_j)$ denoted by E_j and an additional tuple $E(1)$ denoted by ec .

We then define a K -example Ex with two rows as seen in Figure 8, where VC_1, \dots, VC_k are chosen at random. Clearly, Ex has a consistent query w.r.t. it which just

projects the attributes of the atoms with relation E to the output. We also define the abstraction tree to be T where $L(T) = VC_1, \dots, VC_N$ and each VC_i is connected to a node $VC(\tilde{v}, \tilde{e}, 2)$ (denoted by \widetilde{VC}), where the weight of each edge is 1.

Now, we claim that G has a cover of size at most k iff there is an abstraction function that gives privacy at least 1 with at most k loss of information.

(\Leftarrow) Suppose we have an abstraction function A_T that gives privacy at least 1 with at most k loss of information. Let Q be a CIM query w.r.t. $A_T(Ex)$. In particular, Q is consistent w.r.t. a certain concretization of $A_T(Ex)$ (Definition 3.9). Assume that the monomial in the first row of this concretization is $VC'_1 \dots VC'_k \cdot E_1 \dots E_m$, like in this illustration:

Output	Provenance
e_1, \dots, e_m	$VC'_1 \dots VC'_k \cdot E_1 \dots E_m$
$1, \dots, 1$	$yes^k \cdot ec^m$

Given this concretization, Q should be connected, i.e., all atoms should have at least one join to another atom. Note that every atom with relation E has to be connected to an atom with relation VC (as they cannot be connected to each other). Suppose Q is of the form:

$$Q(x_1, \dots, x_m) : -VC(y_1, x_1, z), \dots, VC(y_k, x_m, z), \\ E(x_1), \dots, E(x_m)$$

This structure is necessary because each $E(x_j)$ has to be connected to some $VC(y_i, x_j, z)$ as it is the only option to create a connected query. Thus, given the provenance of the first row, Q maps x_1, \dots, x_m to e_1, \dots, e_m . We choose the vertices represented by $VC'_1 \dots VC'_k$ as the vertex cover for the graph G , since each tuple $VC'_i = VC(v_i, e_j, 2)$ represents a cover of e_j by v_i and all edges e_1, \dots, e_m appear in $VC'_1 \dots VC'_k$.

(\Rightarrow) Suppose we have a cover of size $\leq k$, $\{v_1, \dots, v_k\}$. We show how to generate an abstraction function from this cover. A_T would abstract all VC_1, \dots, VC_k to \widetilde{VC} :

Output	Provenance
e_1, \dots, e_m	$\widetilde{VC}^k \cdot E_1 \dots E_m$
$1, \dots, 1$	$yes^k \cdot ec^m$

Clearly, this gives k loss of information. Next, we show that there is at least one consistent connected query, which, in particular, shows that there exists a CIM query (it may be contained in the query we show, but its existence will be proved).

First, we generate the concretization that our connected query will be consistent with. For every $v_i \in \{v_1, \dots, v_k\}$, and the edge covered by it e_j , we replace an instance of \widetilde{VC} with $VC(v_i, e_j, 2)$. This creates the concretization shown in the previous part of the proof. We now claim that the following query is connected and consistent w.r.t. this concretization:

$$Q(x_1, \dots, x_m) : -VC(y_1, x_1, z), \dots, VC(y_k, x_m, z), \\ E(x_1), \dots, E(x_m)$$

Q is clearly connected. To see consistency, assign the VC'_i tuples to the VC atoms and the E_j tuples to the E atoms. \square

We have defined the problem for general semirings and UCQs (with aggregation). Now, we discuss the solution, starting from $\mathbb{N}[X]$ and CQs. At the end of this section we consider other versions of the problem, where the provenance is given in a different model and the query class is more general. As shown above, the problem is intractable, and our algorithms incur exponential time in the worst case – yet we design heuristics that significantly improve the performance in practice. We first give a high-level description of our solution and then introduce our algorithms.

4.1 High Level Description

The brute force approach for solving the problem would go over all possible abstractions, compute the privacy and the loss of information of each and return the one with minimal loss of information among the ones that meet the privacy threshold. We next overview of how each of these components may be improved. The observed improvement over the brute force solution is reported in Section 5.2.

Efficiently computing privacy. The privacy computation is the most time consuming part of the solution (see Section 4.2). We next give an overview of how the privacy induced by a given abstraction may be efficiently computed.

1. *Computing privacy row by row.* Consistency with a K -example is monotone in the sense that each consistent query must be consistent with each subset of the rows in K -example. We use this fact to effectively compute privacy. For every abstracted K -example \widetilde{Ex} , we first check whether the K -example containing only the first two rows of \widetilde{Ex} has at least k CIM queries w.r.t. it, where k is the privacy threshold. We store only concretizations of \widetilde{Ex} that admit consistent connected queries by storing which concretization creates each query. Then, we add the next row of \widetilde{Ex} to the stored concretizations from the previous step and repeat these steps.

2. *Concretizations connectivity.* We say that a K -example Ex is connected if every provenance monomial in Ex defines a connected graph where the nodes are the tuples and there is an edge between two tuples if they share a constant (e.g., $R(1, 2), R(2, 3)$ are connected). Observe that a connected consistent query cannot be obtained from a disconnected K -example; therefore, disconnected concretizations can be filtered out.

3. *Caching information about concretizations and queries.* Given two abstractions $\widetilde{Ex}, \widetilde{Ex}'$, it is common that $C(\widetilde{Ex}) \cap C(\widetilde{Ex}')$ contains multiple shared concretizations. Therefore, we use caching to store the consistent connected queries w.r.t. each concretization, to avoid repetitive computations (we do not store the CIM

queries since the minimality of a query is measured w.r.t. the concretization set, which varies between different abstractions). Additionally, for each concretization, we store whether it is connected or not and use it in the following computations that involve this concretization.

Efficiently finding an optimal abstraction. Our next goal is to improve the naïve iteration over all abstractions. If we cleverly choose the *order in which we iterate over the abstractions* and avoid complicated calculations for irrelevant abstractions we can find a solution quickly. To do so, we use the following components. In Section 5.2, we will show that these components have improved performance by a factor of over $500\times$.

1. *Sorting abstractions.* When we iterate over all the abstractions, we sort them in increasing order according to the number of tree edges they use, prioritizing abstractions with small loss of information. In this manner, abstractions that use fewer edges of the abstraction tree appear first (these are the easiest to compute privacy for since they have fewer concretizations). Practically, such abstractions often meet the privacy threshold.

2. *Prioritizing loss of information over privacy computation.* Unlike the loss of information that can be quickly and efficiently computed, computing the privacy of an abstracted K -example is a complex and pricey procedure (see Section 4.2). Therefore, given an abstracted K -example, we first compute the loss of information for each abstraction and only then compute the privacy. After finding the first abstraction that satisfies the privacy threshold, we only have to compute privacy for abstractions that incur less information loss.

4.2 Algorithm Details

We next detail the implementation of the ideas we have described.

Privacy computation. We use the following components:

1. *Finding consistent queries.* To find all consistent queries w.r.t. a concretization we recall the algorithm *FindConsistentQuery* from [23] that finds one consistent query for a given K -example by modeling the two provenance monomials of the first two rows in the K -example as a bipartite graph and finding partial matchings that ‘cover’ the output attributes. The algorithm returns the first consistent query that is generated by such a matching. We adjust this algorithm to output all the consistent queries from all matchings instead of returning the first one we find. We then minimize each query using the lattice algorithm described in the paper.

2. *Finding minimal queries.* Given a set of queries Q , $q \in Q$ is minimal if there is no query $q' \in Q$ such that $q' \subsetneq q$. We iterate over all the queries $q \in Q$, and for every $q' \in Q, q' \neq q$ we check whether $q' \subsetneq q$ using the

procedure *QueryContainment* that checks query containment (adapted from [15]).

Algorithm 1 computes the privacy of a given abstracted K -example \widetilde{Ex} . The input is an abstracted K -example \widetilde{Ex} with n rows, an abstraction tree T and the privacy threshold k . The output is the privacy guaranteed by \widetilde{Ex} , or -1 if the privacy is smaller than k . The algorithm initializes a set of good concretizations *GoodConc* (concretizations that create consistent connected queries, as described in the ‘Computing privacy row by row’ component in Section 4.1) with the first row of \widetilde{Ex} (line 1). Then, it iterates over the rows in \widetilde{Ex} (lines 2–22), and for each row preforms the following operations. First, it collects the concretization sets of each abstracted K -example in *GoodConc* combined with the current row from \widetilde{Ex} (lines 3–5). Second, it removes all the disconnected concretizations (line 6) while for each concretization it uses caching to store whether it is connected or not, to avoid redundant computations. Third, it collects all consistent queries w.r.t. every connected concretization and adds them to a set Q_{cons} and to a map *QueriesToConc* that stores, for each concretization, the queries that were created from it (lines 7–12). Then, it removes all the disconnected queries from Q_{cons} (line 13) and also uses caching to store whether it is connected or not. After that, it checks whether the number of connected queries is lower than our privacy threshold, and if so it returns -1 as the privacy does not satisfy the threshold (lines 14–15). Then, the algorithm re-sets the good concretization set *GoodConc* with all the concretizations that create consistent connected queries using *QueriesToConc* (lines 16–19). These concretizations will continue to the next iteration. Finally, the algorithm selects only minimal queries (lines 20–22) and checks again whether their number satisfies the privacy threshold (line 21). After the algorithm iterates over all rows, it returns the number of CIM queries (line 23).

Example 4.2. Consider the K -example Ex_{real} , the tree T , the abstraction function A_T^3 , and the abstracted K -example $\widetilde{Ex}_{abs3} = A_T^3(Ex_{real})$ (depicted in Figures 2a, 3, 4, and 5, respectively). Assume our privacy threshold is 2 (i.e., we want our privacy to be at least 2). First, the algorithm generates the concretization set $C(\widetilde{Ex}_{abs3})$ (shown in Figure 6) and removes the disconnected concretizations (which are c_1 and c_4). For each of the remaining concretization, the algorithm finds the consistent queries and amongst these, finds the CIM queries. As we saw in Example 3.11, after removing the disconnected queries we are left with Q_{real} and $Q_{general}$ (shown in Table 1) and since $Q_{real} \subseteq Q_{general}$ there is only one CIM query which is Q_{real} so the algorithm will return -1 .

Loss of information computation. The loss of information can be easily computed given the abstracted K -example \widetilde{Ex} and the abstraction tree. If we use entropy with discrete uniform distribution, then the loss of information is equal to $\ln(|C(\widetilde{Ex})|)$, i.e., the size of the concretization set. For other distributions, we can find the

Algorithm 1: Compute Privacy

input: Abstracted K -example \widetilde{Ex} , abstraction tree T , privacy threshold k
output: The privacy of \widetilde{Ex} if it's at least k or -1 otherwise

Let \widetilde{Ex}_i be the i th row of \widetilde{Ex} and n be the number of rows of \widetilde{Ex} ;

```

1 GoodConc  $\leftarrow \{\widetilde{Ex}_1\}$ ;
2 for  $i \in \{2, \dots, n\}$  do
3    $C \leftarrow \emptyset$ ;
4   for  $gc \in \textit{GoodConc}$  do
5      $gc + \widetilde{Ex}_i$  denotes appending the  $i$ 'th row of  $\widetilde{Ex}$  to  $gc$ ;
6      $C \leftarrow C \cup \textit{GetConcretizationSet}(gc + \widetilde{Ex}_i, T)$ ;
7    $C_{connect} \leftarrow \textit{RemoveDisconnected}(C)$ ;
8    $Q_{cons} \leftarrow \emptyset$ ;  $\textit{QueriesToConc} \leftarrow (\emptyset, \emptyset)$ ;
9   for  $c \in C_{connect}$  do
10     $Q_{cur} \leftarrow \textit{GetConsistentQueries}(c)$ ;
11     $Q_{cons} \leftarrow Q_{cons} \cup Q_{cur}$ ;
12    for  $q \in Q_{cur}$  do
13       $\textit{QueriesToConc} \leftarrow \textit{QueriesToConc} \cup (q, c)$ ;
14   $Q_{conn} \leftarrow \textit{GetConnectedQueries}(Q_{cons})$ ;
15  if  $|Q_{conn}| < k$  then
16    return  $-1$ ;
17   $\textit{GoodConc} \leftarrow \emptyset$ ;
18  for  $q \in Q_{conn}$  do
19    for  $c \in \textit{QueriesToConc}(q)$  do
20       $\textit{GoodConc} \leftarrow \textit{GoodConc} \cup \{c\}$ ;
21   $Q_{cim} \leftarrow \textit{GetMinimalQueries}(Q_{conn})$ ;
22  if  $|Q_{cim}| < k$  then
23    return  $-1$ ;
24 return  $|Q_{cim}|$ ;
```

concretization set with the abstraction tree and calculate the entropy using the given distribution.

Optimal abstraction algorithm. Given a K -example, an abstraction tree and a privacy threshold, Algorithm 2 finds the optimal abstraction which guarantees the threshold with minimal loss of information. First, the algorithm creates a set of all possible abstraction (line 1) and sorts it in increasing order by the number of edges in the abstraction tree used by each of the abstractions (ties are broken by their loss of information, line 2). Then it initializes the optimal abstraction to be *null* and the optimal loss of information to be ∞ (line 3). For each abstraction, the algorithm first computes the loss of information (line 5). If the loss of information is lower than the optimal loss observed, it computes the privacy (line 7), otherwise, it continues to the next abstraction. If the computed privacy meets the privacy threshold, the algorithm updates the optimal abstraction to be the current one, and updates the current optimal loss of information (lines 8–9). Finally, it returns the abstraction that meets the privacy threshold and incurred the minimum loss of information (or \emptyset if no abstraction has been found).

Example 4.3. Reconsider the K -example Ex_{real} and the abstraction tree T (shown in Figures 2a and 3 resp.). Assume that the privacy threshold is 2 and the loss of information is entropy with discrete uniform distribution.

Algorithm 2: Find Optimal Abstraction

input: K -example Ex , abstraction tree T , privacy threshold k

output: Optimal abstraction

```

1  $A \leftarrow \text{AllPossibleAbstractions}(Ex, T);$ 
2  $A_{\text{sort}} \leftarrow \text{SortAbstractions}(A, T);$ 
3  $a_{\text{best}} \leftarrow \emptyset; l_{\text{best}} \leftarrow \infty;$ 
4 for  $a \in A_{\text{sort}}$  do
5    $l \leftarrow \text{GetLossOfInformation}(a);$ 
6   if  $l < l_{\text{best}}$  then
7      $p \leftarrow \text{ComputePrivacy}(a);$ 
8     if  $p \geq k$  then
9        $a_{\text{best}} \leftarrow a; l_{\text{best}} \leftarrow l;$ 
10 return  $a_{\text{best}};$ 

```

First, the algorithm creates a set of all possible abstracted K -examples of Ex_{real} . Among these we have \tilde{Ex}_{abs1} and \tilde{Ex}_{abs3} (depicted in Figure 5). The corresponding abstraction functions are A_T^1 and A_T^3 (shown in Figure 4). The algorithm starts iterating all abstractions until it gets to $\tilde{Ex}_{\text{abs3}} = A_T^3(Ex_{\text{real}})$, which does not meet the threshold (as shown in Example 4.2). Then, the algorithm gets to $\tilde{Ex}_{\text{abs1}} = A_T^1(Ex_{\text{real}})$. Its privacy is 2 (see Example 3.13), satisfying the threshold. The loss of information is $\ln |C(\tilde{Ex}_{\text{abs1}})| = \ln(5 \cdot 3) = \ln 15 \approx 2.708$ (Proposition 3.5). Since this is the first abstraction that meets the threshold, we keep it as the current optimal one. The algorithm continues to iterate over all other abstractions for which the loss of information is smaller than the current optimal one. Since all of them do not satisfy the privacy threshold, it returns \tilde{Ex}_{abs1} as the optimal abstraction.

Complexity. Given a K -example Ex and an abstraction tree T , the complexity of Algorithm 2 for finding the optimal abstraction is $O((hl)^n q)$ where h is the height of T , $l = |L_T|$ is the number of leaves in T , $n = |\text{Var}(Ex) \cap L_T|$ is the number of variables in Ex that appears in T and q is an exponential expression in the arity (all considered queries have the same arity) which involves the consistency checks [23], connectivity check and containment checks [11]. First, the number of abstractions is $O(h^n)$ since there are n variables that can be abstracted, and for each one of them we have h options of abstracted values. Thus, for each abstraction we compute the concretization set which is of size $O(l^n)$ (since $|C(A_T(Ex))| \leq |L_T|^n$ from Proposition 3.5). Finally, for each concretization we check for consistency, connectivity and containment in $O(q)$ where q is exponential in the query arity. Our experimental evaluation that follows shows the practical efficiency of our solution.

Extending the solution. Table 4 summarizes the augmentations needed for Algorithm 1 when the provenance in the K -example is given in different semirings (table columns) and the query is permitted to be CQ, UCQ or aggregate query as specified in Section 2.1 (table rows).

Gray cell. First, for the $\mathbb{N}[X]$ and $\mathbb{B}[X]$ semirings, Algorithm 1 does not need to be modified for CQs, as the $\mathbb{B}[X]$ semiring simply drops coefficients from the polynomials

Table 4: Privacy computation for the semirings (or semi-modules) from [32, 1] and different query classes. The approach we have detailed so far is designed for the scenario given in the gray cell and the modifications needed to adjust it to the other scenarios are given in the corresponding cells. The $\text{Lin}(X)$ semiring is discussed in the text

	$\mathbb{N}[X], \mathbb{B}[X]$	$\text{Trio}(X), \text{PosBool}(X), \text{Why}(X)$
CQ	Alg. 1	Change line 9 to Alg. 2 in [23]
UCQ, AGG	Change lines 13 and 20	Change lines 9, 13 and 20

and coefficients do not have an impact on the algorithm.

Orange cell. For UCQs (and aggregate queries), line 13 needs to be adjusted to account for the definition of disconnected UCQ (a UCQ containing a disconnected CQ). Moreover, in line 20 we may get CIM queries that are trivial, i.e., the simple union of the tuples that participate in the provenance of a concretization is a CIM query. Therefore, we can augment this procedure by eliminating such trivial queries by, e.g., changing Definition 3.10 that every CIM query has to have at least one variable.

Red cell. The semirings $\text{Trio}(X)$, $\text{PosBool}(X)$ and $\text{Why}(X)$ drop coefficients as well as powers and even monomials subsumed by other monomials ($\text{PosBool}(X)$). The procedure for finding consistent queries in line 9, therefore, needs to be adjusted to Algorithm 2 from [23] that finds consistent queries when given the provenance in these semirings. The algorithm accounts for the missing powers by expanding the provenance as much as needed until a consistent query is found. The algorithm proposed in [23] needs to be augmented as specified in Bullet (1) at the beginning of Section 4.2.

Green cell. Similarly, for UCQs and aggregate queries, lines 9, 13, and 20 have to change in the aforementioned manners.

The $\text{Lin}(X)$ semiring. For the $\text{Lin}(X)$ semiring, adapting our solution is more challenging. This semiring incurs a significant loss of information about the query structure [32], both due to the nature of the semiring and due to the order relation in Definition 3.8. For example, the provenance represented in the $\mathbb{N}[X]$ semiring $2ab^2$ is represented as $\{a, b\}$. Furthermore, the order relation is translated to set containment, and thus, the provenance shown in the K -example can be any subset of the original set, i.e., the empty subset is also valid as provenance. If only part of the provenance set is given (i.e., there are missing tuples in the provenance set), we may employ an approach that ‘completes’ the provenance in the most reasonable way for every concretization [29] and then apply our solution as a subsequent step. If no provenance is given, we may be able to utilize methods from the field of query-by-example and query reverse-engineering [46, 38, 53, 52, 50] to find the query structure strictly from the output, such as column mappings and candidate query generation. This will be the subject of future work.

The dual problem. The dual problem is defined as searching for the optimal abstraction whose loss of information does not exceed a certain threshold l_{max} . Algorithm 2 can be adjusted to solve this problem using the following changes: (1) initializing $p_{best} \leftarrow 0$ in line 3 (p_{best} will store the current optimal privacy), (2) changing the condition in line 6 to be $l < \min(l_{best}, l_{max})$ (this will limit the abstraction we scan to those which do not exceed the given threshold l_{max}), (3) changing the condition in line 8 to be $p \geq p_{best}$ (this will optimize the privacy of the output abstraction) and (4) adding $p_{best} \leftarrow p$ to line 9 (this will update the current best privacy for the next abstractions we scan). With those changes, the algorithm terminates if the loss of information exceeds l_{max} . This reduces the number of abstractions considered, thus the dual problem is more efficiently solvable.

Constructing abstraction trees. Domain experts who know the database structure may be able to phrase rules that place annotations of similar tuples in proximity in the tree. For example, tuples containing the same values in the same attributes (e.g., Figure 3), or are included in the same relation, etc. Another possible manner of constructing abstraction trees is based on *ontologies* that encode abstractions for the different tuples by grouping tuples with similar meaning. Existing methods for identifying semantic relationships between tuples may be used [39, 36]. To further hone the constructed tree in terms of height and size, users could input the relevant queries and database to our system and try to adjust those parameters so that the system incurs the fastest runtime (see Figures 12 and 14 in Section 5). The height can be adjusted, e.g., by adding or removing sub-categories in the ontology. The size can be modified by adding more tuples from the database to the tree. If the tree contains more tuple annotations, more abstractions are possible, which affects the possibility of finding an abstraction that meets the privacy threshold using less edges in the abstraction tree.

5 Experiments

We next detail the settings of our experimental study and its results. We further show end-to-end use cases of our framework.

The algorithms were implemented in Java 13 using the `TreeNode` interface implementation to represent the abstraction trees. All experiments were performed on Mac OS 10.15, 64-bit, with 16GB of RAM and Intel Quad-Core i7 2.2 GHz processor.

5.1 Settings and Summary of the Results

We next review the settings and the summary of our experiments.

Settings. We study the scalability of our solution in terms of runtimes and the size of the optimal abstraction, i.e., the output of the algorithm (we measure the size as

the number of edges in the abstraction tree that were used to get the optimal abstraction). For runtime experiments and the size of optimal abstraction experiments, we use the settings shown in Table 5. To our knowledge, there is no comparable solution in previous work. We thus use the brute force approach as a baseline, studying the effects of each of our algorithm components described in Section 4.1. We have used the TPC-H dataset [5] which consists of a suite of business oriented queries and the IMDB movies dataset [37]. We have randomly sampled a database of 1GB for all experiments. Our basic settings is a privacy threshold of 5; 5-levels abstraction tree with 10000 leaves (10244 nodes); 2 rows in K -example; and discrete uniform distribution for the loss of information measure.

Abstraction trees. The TPC-H abstraction tree consists of a single relation ‘lineitem’, randomly divided into subcategories evenly throughout the tree. The IMDB abstraction tree was created as follows: (1) Directors and actors were categorized by their year of birth, which were further categorized by ranges of years. (2) Tables that connect actors and directors to movies were categorized similarly. (3) Genres were categorized by the genre type. (4) Movies were categorized by their released year, which were further categorized by ranges. (5) Each one of the previous was categorized under a main category and all of those were categorized under the root.

Queries. We have used the TPC-H queries whose details appear in Table 6. We have adapted those queries to our setting, i.e., we have converted them to CQs by dropping aggregation and arithmetics. The queries are relatively complex (e.g., Q21 includes a triple self-join, i.e., a relation name occurring in 3 atoms). We also use the following IMDB queries: (Q1) All the actors starring in a movie from 1995, (Q2) All the actors who starred in a drama movie directed by an american director, (Q3) All the actors which have a bacon number of 1 (actors who act in a movie with Kevin Bacon), (Q4) All the directors which created an action movie and a comedy movie, (Q5) All the comedy movies starred by an actor born in 1978, (Q6) All the directors who directed a movie starring Tom Cruise, and (Q7) All the actors who act in at least two action movies. All experiments were performed with all the queries. However, to avoid visual overloading in graphs and since the results of queries TPCH-Q5, TPCH-Q9, IMDB-Q3 and IMDB-Q4 were very similar to the results of queries TPCH-Q3, TPCH-Q7, IMDB-Q6 and IMDB-Q7 respectively, we omit their curves from the graphs.

Summary of the results.

1. Our solution scales well with the first three parameters in Table 5, due to the components presented in Section 4.1.
2. An increase in the number of rows in the K -example causes a significant runtime increase compared to the other parameters since Algorithm 2 often has to iterate and analyze all possible abstractions, as in the brute force approach.

Table 5: Scalability experiments settings for Figures 9–17

Figures	Privacy threshold	Abst. tree size	Abst. tree height	# rows in K -example
9, 10, 11	varying	10244	5	2
12, 13	5	varying	5	2
14, 15	5	10244	varying	2
16	5	10244	5	2
17	5	10244	5	varying

Table 6: TPC-H and IMDB queries for the experiments

Query	# Atoms	# Joins
TPCH-Q3	3	2
TPCH-Q4	2	1
TPCH-Q5	7	6
TPCH-Q7	6	5
TPCH-Q9	6	5
TPCH-Q10	4	3
TPCH-Q21	6	5

Query	# Atoms	# Joins
IMDB-Q1	3	2
IMDB-Q2	6	5
IMDB-Q3	5	4
IMDB-Q4	7	6
IMDB-Q5	4	3
IMDB-Q6	5	4
IMDB-Q7	7	6

3. The tree height that yields minimum runtime for finding an optimal abstraction varies according to the query structure, though the number of required tree edges used steadily increases.

4. As the size of the tree increases, the time for finding the optimal abstractions also increases, however, the number of required tree edges used for the abstraction decreases.

5. Our solution is not sensitive to the loss of information distribution used, i.e., changing this parameter will not significantly change the runtime. However, the optimal abstraction may change since the distributions has changed, so another abstraction can now incur a smaller loss of information.

6. The effect of the components described in Section 4.1 was dramatic in improving the scalability of our solution.

7. Compared to a provenance compression approach that also utilizes abstraction trees [24], our solution is able to output abstractions with a significantly lower loss of information.

8. We conducted a comprehensive user study, showing that users are unable to infer the original query from the abstracted K -example, while still being able to use the provenance to answer hypothetical questions about the data.

5.2 Results

We next detail our scalability results for the different settings.

Privacy threshold. For this experiment we have increased the privacy threshold while fixing the other parameters (first row in Table 5). There are no strong and clear criteria on how to choose the privacy threshold exactly. For example, in the healthcare world when medical data is shared with k -anonymity property with a small number

of people (typically for research purposes), k is often chosen between 5 and 15. Thus, we have increased the privacy threshold from 2 to 20. For privacy thresholds larger than 20, we noticed that the optimal abstraction returned had a significantly larger privacy than requested. For example, for a privacy threshold of 23, in 90% of the runs the algorithm returned an optimal abstraction with at least $2\times$ privacy than requested (i.e., the number of CIM queries of the optimal abstraction was at least $2\times$ larger than the threshold). We have performed the following experiments:

(a) *Runtime.* The results are shown in Figure 9 and indicate that our solution remains scalable even for a large privacy threshold.

(b) *Optimal abstraction size.* We use ‘Optimal abstraction size’ to represent the number of abstraction tree edges used in the optimal abstraction. The results are shown in Figure 10 and indicate that we do not need a much larger abstraction to get larger privacy. We can see here that for TPC-H-Q21 whose runtime was the slowest, we need fewer edges than for the other queries.

(c) *Loss of information.* We study the loss of information as a function of varying privacy threshold. The results are shown in Figure 11 and indicate that the loss of information increases as privacy increases, as expected.

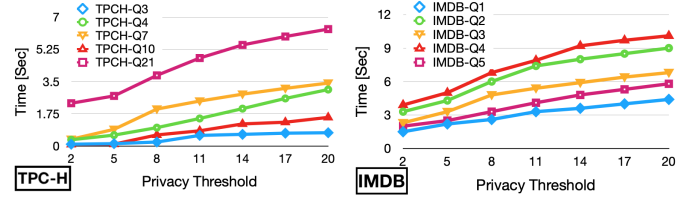


Figure 9: Runtime for varying number of privacy thresholds

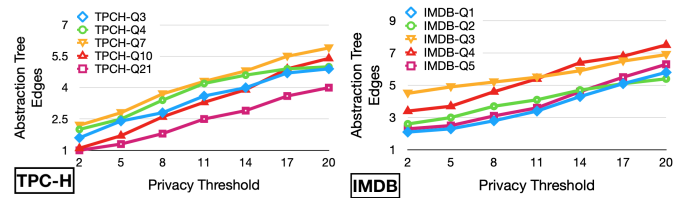


Figure 10: Optimal abstraction size for varying number of privacy thresholds

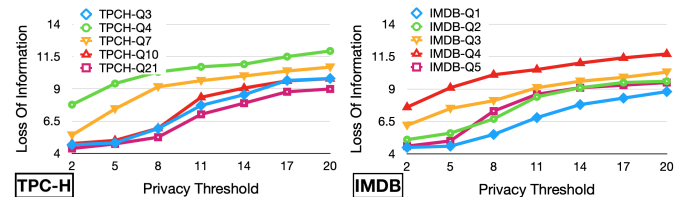


Figure 11: Loss of information for varying privacy thresholds

Abstraction tree size. For this experiment we have increased the number of leaves in the tree from 10K to 810K. We have performed the following experiments:

(a) *Runtime.* The results are shown in Figure 12. Our solution remains scalable even when the size of the abstraction tree nears the size of the data. We observed a similar trend when the tree size reached the data size. TPC-H queries Q3, Q5 and Q10 were faster than the rest since they have one ‘lineitem’ atom which is connected to the rest of the query by a single attribute, as opposed to the other queries. Hence, there are fewer restrictions on these queries in terms of connectivity, making it easier to find CIM queries.

(b) *Optimal abstraction size.* The results are shown in Figure 13 and indicate that when the abstraction tree is larger, the optimal abstraction requires fewer edges. The reason for this is that when the abstraction tree is larger there are more concretizations for each abstraction, and then the privacy can be larger for such abstractions. Here we have not directly measured Loss of Information since it depends on the tree structure which is varied here.

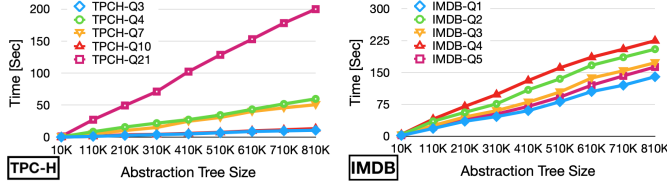


Figure 12: Runtime for varying abstraction tree size

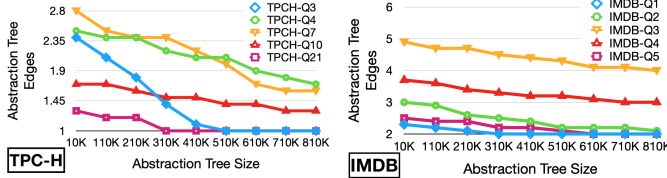


Figure 13: Optimal abstraction size for varying tree size

Abstraction tree height. We next examined the abstraction tree height. We have performed the following experiments:

(a) *Runtime.* The results are shown in Figure 14. Interestingly, we noticed that every query has an optimal height for which the runtimes are the fastest (e.g., for TPC-H-Q7, the optimal height is 5). Particularly, there is no trend of the sort “higher tree implies longer runtime to find the optimal abstraction”. Instead, the tree height that yields the fastest runtime is dependent on the query structure.

(b) *Optimal abstraction size.* The results are shown in Figure 15 and indicate that the optimal abstraction size increases when the tree height increases.

We have observed that different queries require traversing a different number of concretizations to achieve the desired privacy. If the query is relatively simple (e.g., TPC-H-Q4) it needs less and if the query is relatively complicated (e.g., TPC-H-Q21) it needs more. On the one hand, if the tree is not sufficiently high, every abstraction has more concretizations than we need, so the runtime will be slower.

On the other hand, if the tree is too high, every abstraction has fewer concretizations than we need, so we have to scan more abstractions to find a solution and the runtime will also be slower.

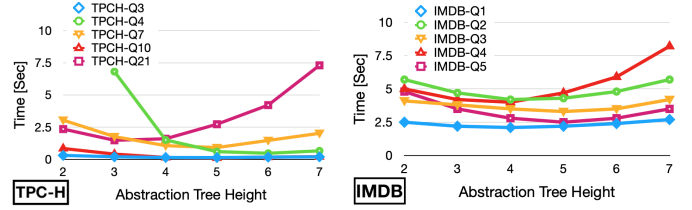


Figure 14: Runtime for varying abstraction tree height

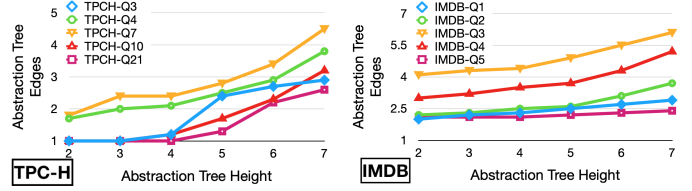


Figure 15: Optimal abstraction size for varying tree height

Number of query joins (query complexity). In this experiment we used TPC-H queries Q5, Q7, Q9, Q21 and IMDB queries Q2, Q4, Q7 (as this is the subset of queries with at least 6 joins) and examined the change in runtime as we increase the number of joins in each. We do so by starting with a version of the queries with only 3 joins and adding an atom for each tick on the X axis. The results (depicted in Figure 16) show that the runtime is not significantly affected by the increase in the number of joins.

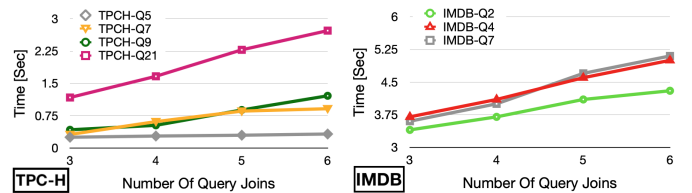


Figure 16: Runtime for varying number of joins

K-example rows. We examine our scalability in terms of increased the number of rows in the K -example. The results (shown in Figure 17) indicate that the number of rows is a determining factor in the runtime of our algorithm. This is because a large number of rows implies fewer CIM queries for each concretization (since each row must be connected). Therefore, the algorithm was forced to try all possible (exponentially many) abstractions, similarly to the brute force approach, which significantly worsened the runtime. In particular, for TPC-H-Q21, the algorithm had to examine a large number of abstractions since this query includes three joined atoms with the ‘lineitem’ relation, where each of them can be abstracted.

Loss of information distribution. We have conducted all of the experiments for two loss of information distributions. The first is entropy with discrete uniform distribution and the second is entropy with random distribution (Section 3.2). We found that on average, the

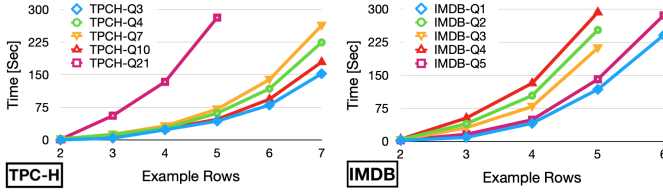


Figure 17: Runtime for varying K -example rows

runtimes are not affected by different distributions. As the probabilities change, the optimal abstraction for one distribution may not necessarily be the optimal for the other one. For example, if there is another abstraction with the same privacy, it may now have a smaller loss of information and will be the new optimal one.

Comparing to a different abstraction approach. The notion of abstraction trees was presented in [24], where the goal of the abstraction was reducing the provenance size. We used this approach to construct an alternative algorithm for our problem. Since the framework of [24] was not designed to achieve privacy, we used it as a black-box, which we executed multiple times with a decreasing target provenance size, until we met the desired privacy threshold. We compared the loss of information incurred by our algorithm to that of [24]. The results are shown in Figure 18. The compression-based approach of [24] unnecessarily increases the loss of information by approximately $2\times$ to $3\times$ to achieve the same privacy as our approach.

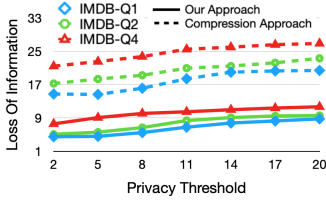


Figure 18: Loss of information for varying privacy thresholds, for our approach and the approach from [24]

Effect of each algorithm component. We now present the effects on the execution time of the five algorithm components we have detailed in Section 4.1, compared to a brute-force approach. The effect of each component is measured as a standalone optimization. Figure 19 shows the results for each component. Referring to the names of the components in Section 4.1, ‘Sorting the abstractions’ and ‘Prioritizing loss of information over privacy computation’ have improved performance by a factor of over $500\times$. The third component of ‘Computing privacy row by row’ has improved performance by approximately $2\times$ to $4\times$ for a K -example with three rows. For a K -example with four rows, it improved performance by approximately $10\times$ to $100\times$. For K -example with more than five rows we were unable to find a solution to the problem in a reasonable time using the brute force approach, in contrast to our approach. The fourth component, ‘Concretizations connectivity’, has improved performance by approximately $1.5\times$ to $1.8\times$ when we filtered out about 60% of the concretizations. The last component, ‘Caching information about concretizations and queries’, has im-

proved performance by approximately $1.5\times$ to $4\times$.

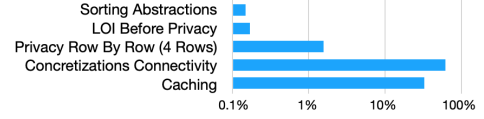


Figure 19: Effect of each of algorithm component from Section 4.1 as compared to the brute force approach (brute force execution time is marked by 100%)

Table 7: User Study Results Summary

	Group A	Group B
Number of group members that were able to find the original query	6/6 (100%)	0/6 (0%)
Number of correct answers in hypothetical questions (on average)	9.6/10 (96%)	8.5/10 (85%)

User Study: We have conducted a user study, involving 12 users with knowledge of databases. The users were randomly divided into two groups of equal size: control group (Group A) and treatment group (Group B). We used IMDB-Q3 (all the actors who played in a movie with the actor Kevin Bacon), the IMDB abstraction tree, 2 rows of output, and a privacy threshold of 2. Then, with Algorithm 2 we found the optimal abstraction. Group A was given the output with the original provenance while group B received the output with the abstracted provenance and the abstraction tree. The users were given two tasks: (1) Infer the underlying query from the original (Group A)/abstracted (Group B) provenance and (2) Answer 10 hypothetical questions regarding the effect of deleting rows (e.g., regarding action movies) from the database on the query result. The study results are summarized in Table 7.

For the first task, all members of group A and none of the members of group B were able to identify the original query. For the second task, the members of group A were able to answer on average 9.6 out of 10 questions correctly, while the members of group B were able to answer on average 8.5 out of 10 questions. This shows a reasonable loss of information. The breakdown of correct answers is shown in Figure 20 and indicates the following conclusions. In most cases, the abstracted provenance has provided enough information to answer the question. For example, for question Q6, which considers the effect of the removal of all comedy movies released after 1980, the abstracted provenance could be used to determine the correct answer. This is because the abstracted value that replaced the relevant tuple was “comedy movie released in 1990–2000”. In some cases, there were a few mistakes due to misunderstandings or lack of concentration. In contrast, naturally, there were cases where the abstracted provenance was not detailed enough to answer the question. For instance, question Q9, that refers to a case where directors born before 1970 are removed from the database. The abstracted provenance indicated that the output is related to a person born between 1950 to 1960, but not

to the person’s role in the movie (actor or director), thus the members of group B were unable to answer the question. Overall, our user study indicates that our method was successful in hiding the original query and incurred a reasonable loss of information in terms of using provenance.

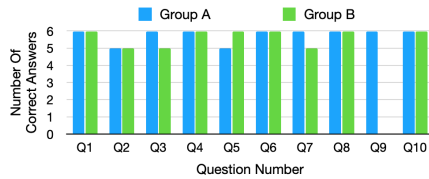


Figure 20: Breakdown of correct answers in hypothetical questions of the user study

6 related work

We next review previous work in the fields of provenance and privacy, highlighting our novelty.

There is a wealth of works on data provenance and its uses, including relational algebra, XML query languages, Nested Relational Calculus, and functional programs (see e.g., [44, 33, 28, 17, 8, 27, 30, 49] and a survey [35]). These works have generally focused on provenance modeling, efficient tracking and storage, and algorithms that use provenance for different applications. As such, they are orthogonal to our work: extending our solutions to additional query and provenance formalisms proposed in these works is an important challenge for future work.

The area of privacy and security in the context of provenance has been explored by various works [19, 20, 21, 6, 43, 42, 2, 51, 16]. These works have focused on privacy and security in different settings than ours such as IoT [43], Blockchain [42] and workflows [19, 20, 21], while our focus was the relational setting. The difference in the setting is reflected in the provenance models (we focus on provenance polynomials whereas, e.g., [19] focuses on workflow provenance in the form of input-output relationship between modules). In turn, the technical problems and solutions are inherently different.

A recent work on fine-grained provenance privacy [23] has focused on learning queries from K -examples where the provenance is given in different semirings [33, 32]. It showed that reducing the granularity of the provenance by using less detailed semirings (which may be seen as an alternative to our approach of abstracting provenance expressions) is inadequate for privacy purposes: it does not introduce significant added difficulty when attempting to reverse-engineer the underlying query.

In [19, 20, 21] the authors studied workflow privacy, with a privacy criterion inspired by l -diversity [40] and k -anonymity [48]. This model achieves privacy by obfuscating entire attributes of a relation that represents a workflow. In contrast, we do not focus on black-box modules, but rather on detailed fine-grained provenance obtained from queries. This makes the technical results of these works inapplicable to our setting. The work of [16] has described an abstract framework for provenance secu-

rity and defines the notions of the disclosure and obfuscation properties of provenance. Given a query and two traces, the problem is then to determine whether the output of the query is equal on these two traces, if they have the same provenance view. A prominent difference from our model is the assumption that the underlying query is known which makes the problem definition and solution fundamentally different.

Previous work on abstracting provenance has primarily focused on workflow provenance abstractions and graph abstractions [18, 10, 25, 13, 7, 14, 24], mainly for the purpose of reducing the provenance size and/or optimizing its generation. Security Views [13] is a framework for access control where users can specify the desired security of the components of a scientific workflow. The framework then omits the inaccessible components from the provenance view. ZOOM [7] abstracts the provenance view by grouping models together allowing users to focus only on the relevant part of the workflow, and ProPub [25] allows users to publish provenance while anonymizing, abstracting, or hiding parts of the provenance graph. Here again, the models (coarse-grained workflow provenance models) and problems that are studied in these works significantly differ from those of the present work.

Query reverse-engineering from output examples [46, 38, 53, 52, 50] attempts to assist users who lost access to the original query or want an automatic system to infer a query based on output examples. In the context of our work, such systems may be of use in the computation of privacy when the provenance is given in the $Lin(X)$ semiring, as mentioned in Section 4. This is an intriguing subject of future work.

7 Conclusion and Limitations

We have proposed in this paper a novel framework for striking a balance between utility and privacy when releasing data provenance. The framework is based on obfuscating provenance by identifying annotations appearing in it, thereby hiding to some extent the query whose execution has yielded the provenance. This kind of obfuscation may be done in many ways, and we aim at choosing the optimal one. The resulting problem is NP-hard, yet we have provided practically effective heuristics.

There are many important directions for future work. First, our work assumes an abstraction tree as input, which may not be readily given. (Semi-)automatic inference of abstraction trees, as briefly discussed in the paper, is an important complementary problem. Second, our loss-of-information model relies on a probability distribution over the leaves, and in our experiments, we have mostly assumed a uniform distribution; we intend to study means for inferring probabilities, as well as other weight-based models for loss of information. Third, our model is tailored to the provenance semiring model; studying provenance obfuscation in the context of other provenance models is another intriguing goal for future research.

References

- [1] Y. Amsterdamer, D. Deutch, and V. Tannen. Provenance for aggregate queries. In *PODS*, pages 153–164, 2011.
- [2] P. Anderson and J. Cheney. Toward provenance-based security for configuration languages. In U. A. Acar and T. J. Green, editors, *4th Workshop on the Theory and Practice of Provenance, TaPP*, 2012.
- [3] B. S. Arab, D. Gawlick, V. Krishnaswamy, V. Radhakrishnan, and B. Glavic. Reenactment for read-committed snapshot isolation. In *CIKM*, pages 841–850, 2016.
- [4] S. Assadi, S. Khanna, Y. Li, and V. Tannen. Algorithms for provisioning queries and analytics. In *ICDT*, volume 48, pages 18:1–18:18, 2016.
- [5] T. Benchmark. <http://www.tpc.org/tpch>.
- [6] E. Bertino, G. Ghinita, M. Kantarcioglu, D. Nguyen, J. Park, R. S. Sandhu, S. Sultana, B. M. Thuraisingham, and S. Xu. A roadmap for privacy-enhanced secure data provenance. *J. Intell. Inf. Syst.*, 43(3):481–501, 2014.
- [7] O. Biton, S. C. Boulakia, S. B. Davidson, and C. S. Hara. Querying and managing provenance through user views in scientific workflows. In *ICDE*, pages 1072–1081, 2008.
- [8] P. Buneman, J. Cheney, and S. Vansumneren. On the expressiveness of implicit provenance in query and update languages. *ACM Trans. Database Syst.*, pages 28:1–28:47, 2008.
- [9] P. Buneman, S. Khanna, and W. Tan. Why and where: A characterization of data provenance. In *ICDT*, pages 316–330, 2001.
- [10] T. Cadenhead, V. Khadilkar, M. Kantarcioglu, and B. M. Thuraisingham. Transforming provenance using redaction. In *SACMAT*, pages 93–102, 2011.
- [11] A. K. Chandra and P. M. Merlin. Optimal implementation of conjunctive queries in relational data bases. In *STOC*, pages 77–90, 1977.
- [12] A. Chapman and H. V. Jagadish. Why not? In *SIGMOD*, pages 523–534, 2009.
- [13] A. Chebotko, S. Chang, S. Lu, F. Fotouhi, and P. Yang. Scientific workflow provenance querying with security views. In *WAIM*, pages 349–356, 2008.
- [14] A. Chebotko, S. Lu, S. Chang, F. Fotouhi, and P. Yang. Secure abstraction views for scientific workflow provenance querying. *IEEE Trans. Serv. Comput.*, 3(4):322–337, 2010.
- [15] C. Chekuri and A. Rajaraman. Conjunctive query containment revisited. *Theoretical Computer Science*, 239(2):211 – 229, 2000.
- [16] J. Cheney. A formal framework for provenance security. In *CSF*, pages 281–293, 2011.
- [17] J. Cheney, L. Chiticariu, and W. C. Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, pages 379–474, 2009.
- [18] J. Cheney and R. Perera. An analytical survey of provenance sanitization. In *IPAW*, volume 8628, pages 113–126, 2014.
- [19] S. B. Davidson, S. Khanna, T. Milo, D. Panigrahi, and S. Roy. Provenance views for module privacy. In *PODS*, pages 175–186, 2011.
- [20] S. B. Davidson, S. Khanna, S. Roy, J. Stoyanovich, V. Tannen, and Y. Chen. On provenance and privacy. In *ICDT*, pages 3–10, 2011.
- [21] S. B. Davidson, S. Khanna, V. Tannen, S. Roy, Y. Chen, T. Milo, and J. Stoyanovich. Enabling privacy in provenance-aware workflow systems. In *CIDR*, pages 215–218, 2011.
- [22] D. Deutch, N. Frost, and A. Gilad. Explaining natural language query results. *VLDB J.*, 29(1):485–508, 2020.
- [23] D. Deutch and A. Gilad. Reverse-engineering conjunctive queries from provenance examples. In *EDBT*, pages 277–288, 2019.
- [24] D. Deutch, Y. Moskovitch, and N. Rinetzky. Hypothetical reasoning via provenance abstraction. In *SIGMOD*, pages 537–554, 2019.
- [25] S. C. Dey, D. Zinn, and B. Ludäscher. Propub: Towards a declarative approach for publishing customized, policy-aware provenance. In *SSDBM*, volume 6809, pages 225–243, 2011.
- [26] Facebook. Understand why you’re seeing certain ads and how you can adjust your ad experience. <https://about.fb.com/news/2019/07/understand-why-youre-seeing-ads/>.
- [27] R. Fink, L. Han, and D. Olteanu. Aggregation in probabilistic databases via knowledge compilation. *PVLDB*, 5(5):490–501, 2012.
- [28] F. Geerts and A. Poggi. On database query languages for k-relations. *J. Applied Logic*, pages 173–185, 2010.
- [29] A. Gilad and Y. Moskovitch. Towards inferring queries from simple and partial provenance examples. In *CIKM*, pages 3273–3276, 2020.
- [30] B. Glavic, J. Siddique, P. Andritsos, and R. J. Miller. Provenance for data mining. In *TaPP*, 2013.
- [31] Google. Why you’re seeing an ad. <https://support.google.com/accounts/answer/1634057>.
- [32] T. J. Green. Containment of conjunctive queries on annotated relations. In *ICDT*, pages 296–309, 2009.
- [33] T. J. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. In *PODS*, pages 31–40, 2007.
- [34] B. D. U. G. W. Group. Un handbook on privacy-preserving computation techniques. <http://publications.officialstatistics.org/handbooks/privacy-preserving-techniques-handbook/UN%20Handbook%20for%20Privacy-Preserving%20Techniques.pdf>, 2019.
- [35] M. Herschel, R. Diestelkämper, and H. Ben Lahmar. A survey on provenance: What for? what form? what from? *VLDB J.*, 26(6):881–906, 2017.

- [36] B. E. Idrissi, S. Baïna, and K. Baïna. Ontology learning from relational database: How to label the relationships between concepts? In *BDAS*, volume 521, pages 235–244, 2015.
- [37] IMDB. <https://www.imdb.com/interfaces>.
- [38] D. V. Kalashnikov, L. V. S. Lakshmanan, and D. Srivastava. Fastqre: Fast query reverse engineering. In *SIGMOD*, pages 337–350, 2018.
- [39] M. Li, X.-Y. Du, and S. Wang. Learning ontology from relational database. In *2005 International Conference on Machine Learning and Cybernetics*, volume 6, pages 3410–3415. IEEE, 2005.
- [40] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE*, page 24, 2006.
- [41] F. Ricciato, A. Wirthmann, K. Giannakouris, M. Skaliotis, et al. Trusted smart statistics: Motivations and principles. *Statistical Journal of the IAOS*, (Preprint):1–15, 2019.
- [42] P. Ruan, G. Chen, A. Dinh, Q. Lin, B. C. Ooi, and M. Zhang. Fine-grained, secure and efficient data provenance for blockchain. *Proc. VLDB Endow.*, 12(9):975–988, 2019.
- [43] J. L. C. Sanchez, J. B. Bernabé, and A. F. Skarmeta. Towards privacy preserving data provenance for the internet of things. In *WF-IoT*, pages 41–46, 2018.
- [44] A. D. Sarma, M. Theobald, and J. Widom. Exploiting lineage for confidence computation in uncertain and probabilistic databases. In *ICDE*, pages 1023–1032, 2008.
- [45] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- [46] Y. Shen, K. Chakrabarti, S. Chaudhuri, B. Ding, and L. Novik. Discovering queries based on example tuples. In *SIGMOD*, pages 493–504, 2014.
- [47] E. D. P. Supervisor. Preliminary opinion on privacy by design. https://edps.europa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_privacy_by_design_en_0.pdf, 2018.
- [48] L. Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.
- [49] W. C. Tan. Containment of relational queries with annotation propagation. In *DBPL*, pages 37–53, 2003.
- [50] W. C. Tan, M. Zhang, H. Elmeleegy, and D. Srivastava. Reverse engineering aggregation queries. *Proc. VLDB Endow.*, 10(11):1394–1405, 2017.
- [51] Y. S. Tan, R. K. L. Ko, and G. Holmes. Security and data accountability in distributed systems: A provenance survey. In *HPCC/EUC*, pages 1571–1578, 2013.
- [52] Q. T. Tran, C.-Y. Chan, and S. Parthasarathy. Query reverse engineering. *The VLDB Journal*, 23(5):721–746, 2014.
- [53] M. Zhang, H. Elmeleegy, C. M. Procopiuc, and D. Srivastava. Reverse engineering complex join queries. In *SIGMOD*, pages 809–820, 2014.