Provenance-aware Discovery of Functional Dependencies on Integrated Views
Ugo Comignani, Laure Berti-Équille, Noël Novelli, Angela Bonifati

To cite this version:

HAL Id: hal-03474955
https://hal.science/hal-03474955
Submitted on 10 Dec 2021

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

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Provenance-aware Discovery of Functional Dependencies on Integrated Views

Ugo Comignani
Tyrex team, Grenoble INP,
INRIA, France
ugo.comignani@inria.fr

Laure Berti-Equille
IRD, ESPACE-DEV
Montpellier, France
laure.berti@ird.fr

Noël Novelli
Aix-Marseille Univ.
LIS CNRS, France
noel.novelli@lis-lab.fr

Angela Bonifati
Lyon 1 University
Lyon, France
angela.bonifati@univ-lyon1.fr

Abstract—The automatic discovery of functional dependencies (FDs) has been widely studied as one of the hardest problems in data profiling. Existing approaches have focused on making the FD computation efficient while inspecting single relations at a time. In this paper, for the first time we address the problem of inferring FDs for multiple relations as they occur in integrated views by solely using the functional dependencies of the base relations of the view itself. To this purpose, we leverage logical inference and selective mining and show that we can discover most of the exact FDs from the base relations and avoid the full computation of the FDs for the integrated view itself, while at the same time preserving the lineage of FDs of base relations. We propose algorithms to speedup the inferred FD discovery process and mine FDs on-the-fly only from necessary data partitions. We present InFine (Inferred Functional Dependency), an end-to-end solution to discover inferred FDs on integrated views by leveraging provenance information of base relations. Our experiments on a range of real-world and synthetic datasets demonstrate the benefits of our method over existing FD discovery methods that need to rerun the discovery process on the view from scratch and cannot exploit lineage information on the FDs. We show that InFine outperforms traditional methods necessitating the full integrated view computation by one to two order of magnitude in terms of runtime. It is also the most memory efficient method while preserving FD provenance information using mainly inference from base table with negligible execution time.

I. INTRODUCTION

The automatic discovery of all functional dependencies (FDs) holding in a single relation is amongst the hardest problems in data profiling [11], [24]. Typically, an FD $X \rightarrow Y$ with attribute sets $X$ and $Y$ in a given table allows to enforce that the combination of values in the set $X$ uniquely determines the values of every attribute in the set $Y$. Functional dependencies are key ingredients in database design, table decomposition, database normalization, and for several other data management tasks, such as data cleaning [23] and query optimization [13], [26]. Due to the high complexity of FD discovery [19] (exponential in the number of attributes and quadratic in the number of records of a relation), a wealth of algorithms have been proposed by relying on aggressive pruning and sophisticated validation. Many of these algorithms [11], [12], [29], [21], [25] have the same theoretical complexity and are able to compute all minimal FDs in less runtime.

Despite the vast literature on the topic, all existing algorithms focus on the FD discovery problem for one relation at a time and do not explore reuse of computation when dealing with multiple relations. However, relations are oftentimes involved in view computation operations, among which the notable class of SPJ (Select-Project-Join) view operations. In this case, the existing algorithms addressing FD discovery would have to entirely recompute the set of FDs for the obtained integrated view without being able to reuse any of the previous computations on the individual base tables of the view. In this paper, for the first time we tackle the problem of inferring FDs on integrated views (also called Inferred FD discovery problem) by solely relying on the FDs available from the base tables of the views. To do so, we employ the well-known concept of why-provenance [3], which allows us to identify the lineage of the FDs on the integrated view and throughout the view computation operations.

In our work, we address the Inferred FD discovery problem, which consists of considering a SPJ view on top of a set of base tables and computing the FDs holding on the view by reusing as much as possible the discovered FDs on the base tables, and, in turn, reduce the SPJ view computation with only the needed attributes.

Explicitly relating the FDs of the base table to the FDs of the view has interest on its own as it allows for instance to understand whether FDs on the base tables are persistent on the view. It can also be beneficial whenever the user needs help while debugging the FDs on the base tables by choosing the most relevant ones if they also apply to the view. Since SPJ views might be obtained as results of data integration and ETL scenarios [10], our method allows to understand how constraints (namely the FDs) are actually affected by the integration process. New FDs that hold on the view but not on the base tables may help the user better understand and explain the result of the data integration process.

Our method allows us to obtain time savings in terms of FD discovery from integrated views. By leveraging logical inference and provenance triples, we can avoid recomputing the FDs holding on the view from scratch and only focus on the new FDs that are not holding on the base tables.

To address the Inferred FD discovery problem, we use a multi-step pipeline encompassing the discovery of different types of FDs and leveraging logical inference. To this end, we propose InFine, an efficient solution for automatically discovering multi-relation FDs starting from the FDs of the...
We design five algorithms to compute the provenance of base tables. Our main contributions are as follows:

1. We develop a system that computes the provenance of base tables. Our main contributions are as follows:
   - We develop a system that computes the provenance of base tables.
   - We present an illustrative example. Section III presents the necessary background and notations. In Section IV, we formalize the Inferred FD discovery problem and provide an overview of InFine. We also present our main contributions and the algorithms at the core of InFine. We describe our performance experiments evaluating the efficiency and accuracy of InFine in Section V. Finally, we discuss related work in Section VI and conclude in Section VII.

II. Motivating Example

Let us consider a real-world clinical database MIMIC-III, out of which we extract the PATIENT table containing information about patients: their identifier (subject_id), gender, date of birth (dob), date of death (dod), and a boolean expire_flag indicating whether the patient passed away, and the table ADMISSION containing administrative and clinical information about patients such as the hospital admission time (admittime), the admission location (admission_location), the insurance, the diagnosis, and a boolean h_expire_flag indicating whether the patient died at the hospital. Let us consider V, the SPJ view (illustrated in Fig. 1) that computes a join between the two above tables: SELECT * FROM PATIENT, ADMISSION WHERE PATIENT.subject_id=ADMISSION.subject_id. We are interested in discovering the set of FDs that hold over the SPJ view by reusing as much as possible the FDs of the base tables PATIENT and ADMISSION. For ease of exposition, we focus on FDs with RHS limited to one single attribute (as it is the case in a canonical cover of FDs). The LHS might consist of multiple attributes as those mined from the join in Fig. 1. A total of 42 minimal and canonical FDs is considered in this running example thus making the example far from being trivial. These FDs can be obtained by FD discovery methods running on the view. However, their lineage is also important as they carry the information of whether they are valid in the base tables or they solely hold on the view. Precisely, we use color coding to encode the FD provenance in the figure. We can observe that the 9 exact FDs from the base table PATIENT (highlighted in green), as well as 9 (out of the 14) exact FDs from the base table ADMISSION (highlighted in pink) are preserved in the view. Better informed, s/he may choose a subset of the integrated view. For example, the query fragment SELECT * FROM PATIENT, ADMISSION WHERE PATIENT.subject_id=ADMISSION.subject_id AND subject_id=252 out of which we extract the PATIENT table containing information about patient 252, gender F, date of birth 31/03/33, and a boolean expire_flag indicating whether the patient passed away, and the table ADMISSION containing administrative and clinical information about patient 252 such as the hospital admission time (admittime), the admission location (admission_location), the insurance, the diagnosis, and a boolean h_expire_flag indicating whether the patient died at the hospital. Let us consider V, the SPJ view (illustrated in Fig. 1) that computes a join between the two above tables: SELECT * FROM PATIENT, ADMISSION WHERE PATIENT.subject_id=ADMISSION.subject_id. We are interested in discovering the set of FDs that hold over the SPJ view by reusing as much as possible the FDs of the base tables PATIENT and ADMISSION. For ease of exposition, we focus on FDs with RHS limited to one single attribute (as it is the case in a canonical cover of FDs). The LHS might consist of multiple attributes as those mined from the join in Fig. 1. A total of 42 minimal and canonical FDs is considered in this running example thus making the example far from being trivial. These FDs can be obtained by FD discovery methods running on the view. However, their lineage is also important as they carry the information of whether they are valid in the base tables or they solely hold on the view. Precisely, we use color coding to encode the FD provenance in the figure. We can observe that the 9 exact FDs from the base table PATIENT (highlighted in green), as well as 9 (out of the 14) exact FDs from the base table ADMISSION (highlighted in pink) are preserved in the result of the integrated view. Imagine that a data steward would like to investigate why a few valid FDs in the base tables are no longer valid in the integrated view result or, conversely, why some FDs that are not valid in the base tables become valid in the view. Better informed, s/he may...
change and adapt the constraint (FD) enforcement strategies when curating the data using the FDs. Moreover, 10 FDs (highlighted in blue) from the view result can be obtained by logical inference over the sets of exact FDs discovered from each base table: for instance, diagnosis → dod is obtained from insurance, diagnosis → subject_id and subject_id → insurance in ADMISSION and subject_id → dod in PATIENT. As a side note, we are not addressing here the problem of meaningfulness of the FDs, which is orthogonal and of independent interest. Indeed, a valid FD may not be semantically meaningful and judging whether a valid FD is relevant typically needs human intervention.

When the two tables are joined, patient #257 is removed due to the absence of the corresponding subject_id value in the other table and approximate FDs (AFD) (such as expire_flag → dod in PATIENT) become exact in the SPJ view. These FDs are highlighted in grey in the figure.

Finally, only 10 exact FDs (highlighted in orange) that hold in the case of a join operation are performed: for instance, (#249,#252) or (#249,#251) and (#250,#251) or (#250,#252), we can obtain the remaining 10 join FDs without having to compute the entire view beforehand.

This example show that with existing FD discovery approaches, the FDs would have to be computed on both the base tables and the integrated view result to preserve the FD provenance information. Furthermore, to identify the provenance of the FDs, a comparison among the two FD sets would have to be performed. We will show the overhead of this process in our experimental study in Section V.

Hence, a better understanding of the mechanisms underlying the provenance of FDs from integrated views is highly needed. Based on these observations, in our work, we address the following key questions: How can we preserve the provenance of FDs discovered from integrated views? Instead of executing FD discovery over each base table and a SPJ view result independently, can we infer most of the FDs on the view as well as reuse the FDs from the base tables and achieve a non negligible speedup? We answer these key questions in the rest of the paper and propose an efficient solution for discovering FDs from integrated views.

III. Preliminaries

Next, we recall the necessary definitions of FDs and join operators with their application to our problem.

Definition 1 (Functional dependency satisfaction): Let I be an instance over a relation schema R, and X, Y be two sets of attributes from R. I satisfies a functional dependency d : X → Y, denoted by I |= d, if and only if:

\[ \forall t_1, t_2 \in I, t_1[X] = t_2[X] \Rightarrow t_1[Y] = t_2[Y]. \]  \hspace{1cm} (1)

Note that the above definition holds regardless of the selected null semantics. This implies that our approach is not depending on the underlying semantics of null values, in line with other approaches in the literature [2], [8], [13].

Following the convention, we use uppercase letters for attribute sets and lowercase letters for single attributes. Moreover, in the rest of the paper we use canonical FDs, i.e., minimal FDs with only one attribute in their right-hand part. We recall that for any set of FDs, it can be computed a logically equivalent set of canonical FDs, thus this is done without loss of generality. We define the SPJ view specification and the set of projected attributes over this view as follows:

Definition 2 (SPJ view specification): Let \( R \) = \( \{R_1; \ldots; R_n\} \) be a set of relational instances. We define a view specification \( V_R \) as a relational algebra formula over relations in \( R \) and limited to the following set of operators:

\[ \{\pi; \sigma; \Join; \bowtie; \bowtie \times; \bowtie \times; \bowtie \times; \bowtie \times \} \]

Definition 3 (Projected attributes set): Let \( V, V_1, \) and \( V_2 \) be view specifications. Let \( R \) be a relational instance. Let \( X \) be a set of attributes. Let \( \rho \) be a set of constraints. Let \( \text{atts}(S) \) denotes the set of attributes over a relational instance \( S \). Let \( \circ \) be a join operator in \( \{\bowtie; \bowtie \times; \bowtie \times\} \). Then the set of projected attributes, denoted by \( \text{proj}(\cdot) \) is defined as follows:

\[ \text{proj}(R) = \text{atts}(R) \quad \text{proj}(V_1 \bowtie V_2) = \text{proj}(V_1) \]

\[ \text{proj}(\pi_X(V)) = X \quad \text{proj}(V_1 \bowtie V_2) = \text{proj}(V_1) \cup \text{proj}(V_2) \]

Finally, we define the types of FDs and the provenance triples derived in our framework and used to maintain the provenance information of every discovered FD. The base FDs are defined as follows:

Definition 4 (Base FD): Let \( R = \{R_1; \ldots; R_n\} \) be a set of relational instances. Let \( V_R \) be a view specification over relations in \( R \). Let \( \text{fds}(R_i) \) denotes the set of minimal FDs over the relation \( R_i \in R \). Let \( \text{fds}(V_R) \) denotes the set of minimal FDs over the view specified by \( V_R \). An FD \( d \) is a base FD if \( d \in \text{fds}(V_R) \) and \( \exists R_i \in R \) s.t. \( d \in \text{fds}(R_i) \). We now define the notion of upstaged FD, which can occur either in the case of a selection or of a join operation:

Definition 5 (Upstaged FD): Let \( R_1, R_2 \) be two relational instances. Let \( \text{fds}(R_i) \) denotes the set of minimal FDs over a relational instance \( R_i \). An FD \( d \) is an upstaged FD for \( R_1 \) if \( d \) is defined over attributes from \( R_1 \) and:

- in the case of a selection \( \sigma_{R_1}(R_1) \): \( d \in \text{fds}(\sigma_{R_1}(R_1)) \) and \( d \notin \text{fds}(R_1) \)
- in the case of a join \( R_1 \bowtie R_2 \): \( d \in \text{fds}(R_1 \bowtie R_2) \) and \( d \notin \text{fds}(R_1) \)

In the following, we distinguish the upstaged FDs by considering if they come from a selection operation or a join operation. We also define two types of FDs that arise specifically after a join operation is performed:

Definition 6 (Inferred FD): Let \( R_1, R_2 \) be two relational instances. Let \( \text{fds}(R_i) \) denotes the set of minimal FDs over a relational instance \( R_i \). Let \( \text{atts}(R_i) \) denotes the set of attributes of \( R_i \). Let \( d : X \rightarrow y \) be a FD such that \( d \in \text{fds}(R \bowtie R') \). Then \( d \) is an inferred FD if \( \{X \cup \{y\}\} \cap \text{atts}(R_1) \neq \emptyset \land (X \cup \{y\}) \cap \text{atts}(R_2) \neq \emptyset \)
and either:

- $d'$ can be inferred through the use of Armstrong’s axioms;
- $\exists x' \ni X \rightarrow y$ such that $X \subseteq X'$ and $d'$ can be inferred through Armstrong’s axioms.

**Definition 7 (Join FD):** Let $d : X \rightarrow y$ be a FD such that $d \in fds(R \bowtie R')$. Then $d$ is a join FD if

$$(X \cup \{y\}) \cap \text{atts}(R_1) \neq \emptyset \land (X \cup \{y\}) \cap \text{atts}(R_2) \neq \emptyset$$

and $d$ did not belong to the set of inferred FDs.

From these types of FDs, we define the provenance triples as follows:

**Definition 8 (FD Provenance Triple):** Let $R = \{R_1; \ldots; R_n\}$ be a set of relational instances. Let $V_R$ be a view specification over relations in $R$. A provenance triple $(d,t,s)$ for an FD $d$ over the view specified by $V_R$ is a triple composed of:

- the FD $d$ whose provenance is described;
- the type $t$ of $d$, taking one of the following values: “base”, “upstaged selection”, “upstaged left”, “upstaged right”, “inferred” or “join FD”;
- the first sub-query $s$ over the view specification $V_R$ in which $d$ holds during the view computation.

We illustrate the provenance triples in the following example:

**Example 1:** Following the running example of Figure 1, the provenance triple for FDs $\text{subject}_\text{id} \rightarrow \text{dob}; \text{expire}_\text{flag} \rightarrow \text{dob}$ and $\text{gender, h_expire}_\text{flag} \rightarrow \text{insurance}$ will be the following:

$$(\text{subject}_\text{id} \rightarrow \text{dob}, \text{“base”, } \text{ADMISSION})$$
$$(\text{expire}_\text{flag} \rightarrow \text{dob}, \text{“upstaged left”, } \text{PATIENT} \bowtie_{\text{subject}_\text{id}=\text{subject}_\text{id}, \text{ADMISSION}})$$
$$(\text{gender, h_expire}_\text{flag} \rightarrow \text{insurance}, \text{“join FD”, } \text{PATIENT} \bowtie_{\text{subject}_\text{id}=\text{subject}_\text{id}} \text{ADMISSION})$$

**IV. MINING FDs PROVENANCE TRIPLES OVER VIEWS**

In this section, we describe the workflow of InFine to compute FDs and their provenance information from a view specification and the base relations used in the view specification. In this workflow, after discovering FDs from the base tables, we mine the FDs appearing during the view computation by relying as much as possible on inference methods as well as on efficient methods to discover the remaining FDs that cannot be inferred.

**Problem statement.** Let $R$ be a set of base tables; $V_R$ be a view specification over $R$. The Inferred FD discovery problem consists of efficiently inferring functional dependencies over the view specified by $V_R$ by reusing functional dependencies from $R$ and annotating each FD with its provenance information under the form of a provenance triple.

In order to compute the functional dependencies for a given view, one needs to tackle the complexity of the FD mining problem and thus to tame the size of explored FDs lattices.

In the following theorem, we state that at each computation step of a view without projection, the FDs previously mined continue to be valid. Conversely, we show that the application of a projection can only lead to the suppression of previously valid FDs, and never to the discovery of new FDs.

**Theorem 1:** Let $V, V_1$, and $V_2$ be view specifications. Let $D_1, D_2$, and $D_3$ be the sets of FDs over views specified by $V, V_1$ and $V_2$, respectively. Let $fds(V)$ denotes the set of FDs over a view specified by $V$.

Then:

$$fds(\pi_X(V)) \subseteq D_1, \quad fds(\rho_r(V)) \supseteq D_2,$$

and

$$fds(V_1 \circ V_2) \supseteq D_1 \cup D_2, \text{ with } \circ \in \{\bowtie, \bowtie, \bowtie, \bowtie, \bowtie\}.$$
Algorithm 1: InFine

Input: $R$: a set of relational instances; $V_R$: a view specification over instances in $R$.
Result: the set of provenance triples of the FDs over $V_R$.

1. $\mathcal{E}_P \leftarrow \emptyset$;
2. $A_V \leftarrow \text{compute the set of attributes $\text{proj}(V_R)$};$
3. for each relation $R_i \in R$ do
   4. $\mathcal{D}_{R_i} \leftarrow \text{compute FDs in } R_i \text{ limited to attributes in } A_V$;
   5. add $\mathcal{D}_{R_i}$ to $\mathcal{E}_P$
6. return $\text{provFDs}(R, V_R, \mathcal{E}_P, A_V)$;

Subroutine $\text{provFDs}(R, V_R, \mathcal{E}_D, A_V)$

switch $V_R$ do
   case $R \in R$ do
      $\mathcal{P} \leftarrow \emptyset$;
      $\mathcal{D} \leftarrow \text{get } \mathcal{D}_D \text{ in } \mathcal{E}_D$;
      for each $FD$ $d \in \mathcal{D}$ do
         add triple $(d, \text{"base"}, R)$ to $\mathcal{P}$
   14. return $\mathcal{P}$;
   case $\pi_x(V_R)$ do
      return $\text{provFDs}(R, V_R, \mathcal{E}_D, A_V)$;
   case $\sigma_{\rho}(V_R)$ do
      $\mathcal{P} \leftarrow \text{provFDs}(R, V_R, \mathcal{E}_D, A_V)$;
      return $\mathcal{P} \cup \text{selectionFDs}(R, V_R, \mathcal{E}_D, A_V)$;
   case $V_R \Join V'_R$ do
      $\mathcal{P}_{rec} \leftarrow \text{provFDs}(R, V'_R, \mathcal{E}_D, A_V)$;
      $\mathcal{P}_{rec} \leftarrow \text{provFDs}(R, V'_R, \mathcal{E}_D, A_V)$;
      $\mathcal{P}_{up} \leftarrow \text{joinUpFDs}(A_V)$;
      $\mathcal{P}_{inf} \leftarrow \text{inferFDs}()$;
      $\mathcal{P}_{join} \leftarrow \text{joinFDs}()$;
      return $\mathcal{P}_{rec} \cup \mathcal{P}_{up} \cup \mathcal{P}_{inf} \cup \mathcal{P}_{join}$;
switch $V_R$ do

A. Reducing Space of Explored Candidates using Projections

As our goal is to discover the FDs occurring on top of an integrated view, we focus on mining FDs containing only attributes that are retrieved in the output SPJ view.

The number of considered attributes greatly influences the size of the explored lattices during FD mining. Thus, we make use of the knowledge of the projected attribute set at the very first step of our framework, in order to efficiently reduce the cost of mining FDs from the base relations. This can be seen in the main algorithm of our framework, Algorithm 1 during the mining of the FDs on the base relations (lines #3-5). After this step, each following step of our framework relies on the FDs mined at the previous steps. Thus, additionally to reducing the FD discovery complexity over the base relations, removing unwanted FDs at this step also prevents useless computation during the following steps of our framework.

B. Mining of FDs Through View Selections

New FDs may appear due to selections in the view definition. This will occur when the selection operation leads to filter the tuples violating an FD, making this FD valid in the resulting instance. In such a case, our framework aims at producing provenance triples for these new FDs with Algorithm 2. This algorithm checks if the selection operation leads to some tuple filtering (from line #4) to avoid unnecessary mining step if the filter is not applicable. If some tuples are filtered by the selection, then upstaged FDs need to be mined and, to do so, we rely on a level-wise approach for FD mining. The intuition behind such an approach is to explore the possible sets of attributes involved in the candidate FDs, by organizing them into a lattice beginning from an infimum with FDs with singleton as lhs (line #5). Then, FD validity is tested and next level candidates are generated (line #11) while taking into account the FDs discovered at previous levels to prune the candidates FDs (lines #12-13). The algorithm stops when no candidates are left, and then the resulting set of minimal FDs are labelled with their provenance triples, the latter being returned by the algorithm.

Algorithm 2: selectionFDs

Input: $R$: a set of relational instances; $V_R$: a view specification over instances in $R$; $\rho$: the selection condition; $\mathcal{P}_V$: the sets of provenance triples of $V_R$; $A_V$: the set of projected attributes to explore.
Result: upstaged FDs provenance triples over $V_R$.

1. $\mathcal{P} \leftarrow \emptyset$;
2. $\mathcal{D}_V \leftarrow$ get the set of FDs in the triples in $\mathcal{P}_V$;
3. $\mathcal{V}_\text{set} \leftarrow \pi_x(V_R)$;
4. if size($\mathcal{V}_\text{set}$) < size($V_R$) then
   5. $\mathcal{D}_\text{cand} \leftarrow$ generate candidate FDs for first level of $V_R$;
   6. prune FDs in $\mathcal{D}_\text{set}$ with attributes not in $A_V$;
   7. repeat
      8. prune non-minimal FDs in $\mathcal{D}_\text{cand}$ knowing $\mathcal{D}_\text{out}$;
      9. prune non-minimal FDs in $\mathcal{D}_\text{cand}$ knowing $\mathcal{D}_V$;
     10. add to $\mathcal{D}_\text{out}$ the FDs from $\mathcal{D}_\text{cand}$ holding in $V$;
     11. $\mathcal{D}_\text{cand} \leftarrow$ generate candidate FDs for next level;
     12. until $\mathcal{D}_\text{cand} = \emptyset$;
   13. for each $FD$ $d \in \mathcal{D}_\text{out}$ do
      14. add triple $(d, \text{"upstaged selection"}, V_R)$ to $\mathcal{P}$
   15. return $\mathcal{P}$;
   16. else
      17. return $\emptyset$;

C. Mining of FDs from Joined Tables

In the next sections, we will show how we can mine the upstaged, inferred and join FDs generated during a join operation. In the next theorem, we state that the join order can only affect the type of upstaged FDs which can switch between left or right upstaged FDs:

Lemma 1 (FD types preservation though join operations): Let $R = \{R_1; \ldots; R_n\}$ be a set of relational instances. Let $\mathcal{T}(R_i)$ denotes the set of minimal FDs over the relational instance $R_i$. Let $\Sigma_{\text{left up}}, \Sigma_{\text{right up}}, \Sigma_{\text{inf}}$ and $\Sigma_{\text{join}}$ denote the sets of left upstaged, right upstaged, inferred and join FDs over $R_1 \circ \ldots \circ R_n$, respectively. Then, the sets $\Sigma_{\text{left up}} \cup \Sigma_{\text{right up}}, \Sigma_{\text{inf}}$ and $\Sigma_{\text{join}}$ are equal regardless of the join order. At the opposite, equality between sets $\Sigma_{\text{left up}}$ and $\Sigma_{\text{right up}}$ cannot be guaranteed.

Next, we show how upstaged FDs can be mined over joins.

1) Mining of Upstaged FDs from Joined Tables: Upstaged FDs may appear due to the join operations in the SPJ view when tuples from one base relation cannot be joined with their counterpart in the other base relation, i.e., when some join at-
tribute values are missing in one of the tables. This mechanism is expressed more formally in the following lemma:

**Lemma 2 (Upstaged join FDs):** Let $L$ and $R$ be two instances over relations $S$ and $T$, respectively, and $D_L$ and $D_R$ be the two sets of all FDs such that $L \models D_L$ and $R \models D_R$, respectively. Then the sets of upstaged join FDs denoted $D_L^{\text{new}}$ and $D_R^{\text{new}}$ are the sets:

$$D_L^{\text{new}} = \{ d \mid d \notin D_L \land (L \odot_X Y(\pi_Y(R)) \models \pi_Y(R)) \}$$

$$D_R^{\text{new}} = \{ d \mid d \notin D_R \land (\pi_X(L) \odot_X Y(R) \models \pi_Y(R)) \}$$

**Example 2:** To illustrate the case of upstaged join FDs we use the example in Figure 1. The FD $\text{expire}_\text{flag} \rightarrow \text{dob}$ in Table $\text{PATIENT}$ is violated by the presence of two tuples for patient (#257). However, in the join result of $\text{PATIENT} \bowtie \text{subject}_\text{id} \text{ADMISSION}$, the violating tuple #257 has no counterpart in the $\text{ADMISSION}$ table and it disappears from the join result. Consequently, the FD $\text{expire}_\text{flag} \rightarrow \text{dob}$ becomes valid in the join result.

**Algorithm 3: joinUpFDSs**

**Input:** $V_L, V_R$: the join views; $X, Y$: the join attributes; $\bowtie$: a join operator in $\{\bowtie_\text{left}, \bowtie_\text{right}\}$; $\mathcal{P}$: the provenance triples set; $\mathcal{P}_{V_L}, \mathcal{P}_{V_R}$: the provenance triples sets of $V_L$ and $V_R$; $\mathcal{A}_V$: the set of projected attributes to explore.

**Result:** upstaged FDs provenance triples over $V_L$ and $V_R$

1. $\mathcal{P} \leftarrow \emptyset$
2. $\mathcal{D}_{V_L} \leftarrow \{ \text{set of FDs in the triples in } \mathcal{P}_{V_L} \}$
3. $\mathcal{D}_{\text{left}} \leftarrow \text{upstagedFDSs}(V_L, V_R, X, Y, \mathcal{D}_{V_L}, \mathcal{A}_V)$
4. for each FD $d \in \mathcal{D}_{\text{left}}$
5.   add triple $(d, "\text{leftstaged left"}, V_L \odot V_R)$ to $\mathcal{P}$
6. $\mathcal{D}_{V_R} \leftarrow \{ \text{set of FDs in the triples in } \mathcal{D}_{V_R} \}$
7. $\mathcal{D}_{\text{right}} \leftarrow \text{upstagedFDSs}(V_R, V_L, Y, X, \mathcal{D}_{V_R}, \mathcal{A}_V)$
8. for each FD $d \in \mathcal{D}_{\text{right}}$
9.   add triple $(d, "\text{upstaged right"}, V_L \odot V_R)$ to $\mathcal{P}$
10. return $\mathcal{P}$

**Subroutine upstagedFDSs**($I, J, X, Y, D, \mathcal{A}_V$)

11. $D_{\text{out}} \leftarrow \emptyset$
12. $I_{\text{join}} \leftarrow \{ X = Y(\pi_Y(J)) \}$
13. if size($I_{\text{join}}$) < size($I$) then
14.   $D_{\text{cand}} \leftarrow \text{generate candidate FDs for first level of } I_{\text{join}}$
15.   prunes FDs in $D_{\text{cand}}$ with attributes not in $\mathcal{A}_V$
16. repeat
17.   prunes non-minimal FDs in $D_{\text{cand}}$ knowing $D_{\text{out}}$
18.   prunes non-minimal FDs in $D_{\text{cand}}$ knowing $D_{\text{out}}$
19. add to $D_{\text{out}}$ the FDs from $D_{\text{cand}}$ holding in $I$
20. $D_{\text{cand}} \leftarrow \text{generate candidate FDs for next level}$
21. until $D_{\text{cand}} = \emptyset$
22. return $D_{\text{out}}$

To compute upstaged FDs, we propose Algorithm 3. Lines 12–17 handle the inputs of the SPJ view for each table participating in the join operation. For each side of the join, the subroutine upstagedFDSs is executed (lines 18–21) and computes partially the join only with the join attributes from the left side table (line 18) to check the assumption of the join value set preservation [9]. If the assumption is violated (i.e., if some tuples have been deleted through the join operation, as checked at line 14), some upstaged join FDs are produced. The subroutine works analogously to Algorithm 2 to discover the FDs in the input instance by using a level-wise approach where the previously discovered FDs are used to improve the pruning of candidates (lines 18–19).

In this algorithm, the computation is performed over upstaged FDs only. Next, we discover the inferred FDs by relying on the characteristics of the join and on the FDs discovered previously.

2) Mining of Inferred FDs from Joined Tables: We will now show how the inferred FDs can be deduced from the sets of FDs discovered from the tables involved in the join operation. In the next lemma, we show that if the right-hand side (rhs) of an FD is not functionally defined by the set of join attributes, then this FD cannot be an inferred FD in a join result. This property of the join operation is formally defined as follows:

**Lemma 3:** Let $L$ and $R$ be two instances over relations $S$ and $T$, respectively. Let $L \odot_X Y = R$ be a join result with $X \subseteq \text{atts}(L)$, $Y \subseteq \text{atts}(R)$. For all $A \subseteq \text{atts}(L) \setminus X$ and $B \subseteq \text{atts}(R) \setminus Y$:

$$L \odot_X Y \models R \nRightarrow A \rightarrow B$$

**Example 3:** To illustrate the property proved in Lemma 3, we observe that the diagnosis is not determined by the patient identifier in Figure 1, for example patient #249 has been admitted three times for a different pathology each time, i.e., $\text{PATIENT} \bowtie \text{subject}_\text{id} \text{ADMISSION} \nRightarrow \text{subject}_\text{id} \rightarrow \text{diagnosis}$. From Lemma 3, we know that diagnosis in the join result $\text{PATIENT} \bowtie \text{subject}_\text{id} \text{ADMISSION}$ cannot be determined by any set of attributes coming from $\text{PATIENT}$ table. Such similar inferences may be trivial for the user, but they usually require the knowledge of the attribute semantics. If not encoded, they are difficult to capture by a system. However, the property shown in Lemma 3 can be used to drastically reduce the set of possible FDs that can appear after a join operation.

Intuitively, from Lemma 3 follow that inferring FDs in the results of a join operation using Armstrong’s transitivity axiom can only be done if the transitivity is done through the join attributes. This is formalized in the following theorem:

**Theorem 2:** Let $L$ and $R$ be two instances over relations $S$ and $T$, respectively. Let $L \odot_X Y = R$ be a join result with $X \subseteq \text{atts}(L)$, $Y \subseteq \text{atts}(R)$. For all $A \subseteq \text{atts}(L) \setminus X$ and $B \subseteq \text{atts}(R) \setminus Y$:

$$L \odot_X Y \models R \nRightarrow A \rightarrow B$$

**Example 4:** To be determined in $\text{PATIENT} \bowtie \text{subject}_\text{id} \text{ADMISSION}$ and $\text{subject}_\text{id} \rightarrow \text{dob}$ in $\text{PATIENT}$. Since these tables do not contain any null values for the $\text{lhs}$ and $\text{rhs}$ attributes, joining them with attribute
subject_id leaves these FDs unchanged with no violation. By transitivity, we obtain: subject_id → dob.

**Algorithm 4: inferFDs**

Input: $V_L, V_R$: two view specifications;
$X, Y$: the sets of join attributes for $V_L$ and $V_R$;
∅: a join operator in $\{\cap; \times; \times\cap; \times\cap; \times\cap; \times\cap\}$;
$\mathcal{P}_{V_L}, \mathcal{P}_{V_R}$: the provenance triples sets of $V_L$ and $V_R$;
$A_U$: the set of projected attributes to explore.

Result: inferred FDs provenance triples over $\mathcal{L}_{X\leftarrow Y}$.

To compute the set of FDs with lhs attributes coming from a single instance, as described in Theorem 4, we propose Algorithm 4. First, the subroutine infer extracts the FDs that can be retrieved by transitivity (lines 12 and 13 in subroutine infer). Note that in the case of equijoins, equality of values might be enforced between sets of attributes with different names (i.e., $X$ and $Y$ might be different), thus, for the general case, the FD (line 13) cannot be simplified into an FD $X \rightarrow b$. Conversely, if we restrict the join operations to natural joins only, such a simplification can be made.

Then, for each FD returned by infer, the subroutine refine checks whether the FD is minimal or if a subset of its lhs leads to a minimal FD. To do so, subroutine refine uses an horizontal partition of the joined instances in which only the necessary attributes to perform the verification are considered (line 19). These necessary attributes are the join attributes (to perform the join operation), and the rhs attributes of the refined FD $A \rightarrow b$ (line 18) as refine only considers candidates with subsets of $A$ as lhs and $b$ as rhs (lines 20 and 21).

3) Mining of Join FDs from Joined Tables: Now, we consider the set of join FDs which hold on a join result. Contrarily to the inferred FDs that can be deduced directly using a simple logical reasoning, these join FDs need to be discovered and validated from the data. For example, gender, expire_flag→ insurance of our example has attributes from PATIENT in lhs and attributes from ADMISSION in rhs and it cannot be inferred logically. Other FDs with the same properties are illustrated in orange in Figure 1. In the following theorem, we show that if lhs attributes of an FD come from the instances participating in the join, then we cannot predict their validity without checking them directly with some representative (if not all) tuples of the join result:

Theorem 3: Let $L$ and $R$ be two instances over relations $S$ and $T$, respectively. Let $L \times_{X\leftarrow Y} R$ be a join result with $X \subseteq \text{atts}(L)$, $Y \subseteq \text{atts}(R)$. We cannot guarantee that all FDs over $L \times_{X\leftarrow Y} R$ can be inferred from Armstrong’s axioms over the FDs over $L$ and $R$ taken separately. Such FDs are illustrated in the following example:

Example 5: In our example of Figure 1, FDs that cannot be inferred are highlighted in orange. For example, gender, expire_flag→ insurance is specific to the join of PATIENT and ADMISSION. It holds in PATIENT but not in ADMISSION. Attributes gender and expire_flag come from PATIENT and attribute insurance comes from ADMISSION.

Theorem 5 motivates the need for designing a new method for computing FDs from partial join results, as we cannot always infer all the FDs only using logical reasoning. However, we can rely on the following theorem to greatly reduce number of remaining FDs to check from the data:

Theorem 4: Let $L$ and $R$ be two instances over relations $S$ and $T$, respectively. Let $L \times_{X\leftarrow Y} R$ be a join result with $X \subseteq \text{atts}(L)$, $Y \subseteq \text{atts}(R)$. For all $A \subseteq \text{atts}(L)$, $A' \subseteq \text{atts}(R)$ and $b \in \text{atts}(R)$: If $L \times_{X\leftarrow Y} R \models AA' \rightarrow b$, Then $L \times_{X\leftarrow Y} R \not\models Y A' \rightarrow b$.

In-line with Theorem 4, we propose Algorithm 5 for selective mining and use the FDs previously discovered with Algorithms 3 and 4 to compute the remaining join FDs. Intuitively, Theorem 4 shows that a given attribute $b$ can be a rhs of a remaining join FDs only if we have previously found an FD of the form $YA \rightarrow b$ with $Y$ being the join attributes of the instance containing $b$. Thus, it allows us to focus only on the plausible rhs (lines 11 and 15 in subroutine discover) and explore their candidate lhs (lines 12 and 16) and 17. In practice, there is no need to generate every candidate FDs initially. Instead, candidate FDs can be explored by generating a first level containing only the smallest candidates and by generating upper levels only when currently evaluated candidates are not valid. Moreover, we can avoid the computation of the full join by deleting a given lhs attribute $a$ if $a$ is not a possible rhs and, for every FD candidate $d : A \rightarrow b$ such that $a \in A$, $d$ is logically implied by previously discovered FDs.

D. Completeness, Correctness, and Complexity

In this section we show that the set of FDs retrieved by lnFine is both complete and correct, and present our solution
complexity. First, we show that InFine always retrieve every functional dependency in the specified SPJ view.

Theorem 5 (InFine completeness): Let \( R = \{R_1; \ldots; R_n\} \) be a set of relational instances. Let \( V_R \) be a view specification over relations in \( R \). Let \( fds(V_R) \) denote the set of minimal FDs over the view specified by \( V_R \). Let \( fds(inFine) \) denotes the set of FDs computed by InFine over the view specified by \( V_R \). Then:

\[
\forall d \in fds(V_R), \exists d' \in fds(inFine) \text{ s.t. } d \equiv d'
\]

This establish the completeness of our approach. Now, in the next theorem, we show that every FD retrieved by InFine is valid in the specified SPJ view:

Theorem 6 (InFine correctness): Let \( R = \{R_1; \ldots; R_n\} \) be a set of relational instances. Let \( V_R \) be a view specification over relations in \( R \). Let \( I_{V_R} \) be the view specified by \( V_R \). Let \( fds(inFine) \) denotes the set of FDs computed by InFine over the view specified by \( V_R \). Then:

\[
\forall d \in fds(inFine), I_{V_R} \models d
\]

We now detail the complexity of our algorithms. Algorithms 2 and 3 are based on level-wise algorithms through the attributes lattices. Their complexity is exponential in the number of attributes of the considered table. They prune candidates at each level when it is possible. In terms of memory, only two levels are required. The memory size is bounded by \( O(k^{1/2}) \) where \( k \) is the number of attributes. Algorithm 4 infers and refines FDs coming from the previous step with complexity \( O(n \cdot f) \), where \( f \) is the number of validated FDs and \( n \) the maximal number of tuples in the left or right instance. The complexity of Algorithm 5 is \( \mathcal{O}(f : f_j) \) where \( f \) is the maximal number of validated FDs in the left or right instance and \( f_j \) the number of validated FDs in the join instance.

V. Experiments

Evaluation Goals. The two main points we seek to validate in each experiment are as follows: (1) Does our approach enable us to discover all FDs with their provenance information in an efficient manner and faster than the straight-forward approach? (2) What is the impact of different data and SPJ view characteristics on the InFine performance?

Setup. We perform all experiments on a laptop HP ZBook 15 machine with an Intel Core i7-4900MQ, 2.8 GHz, 32 GB RAM, powered by Windows 10 pro 64-bit. Our implementation in C++ uses only one thread. Sharable datasets, SPJ queries, scripts, and code are available at https://github.com/ucognani/InFine

Data. We use three real-world datasets and one synthetic dataset in our experiments: (1) MIMIC-3, a clinical database[6]; (2) PTE[7]: a database for predictive toxicology evaluation, used to predict whether a compound is carcinogenic, and (3) PTC[8]: the dataset from the Predictive Toxicology Challenge that consists of more than three hundreds of organic molecules marked according to their carcinogenicity on male and female mice and rats; and (4) the TPC-H Benchmark[9] with scale-factor 1. The datasets characteristics are given in Table I and the characteristics of the queries corresponding to the views on the above datasets are provided in Table II. To the best of our knowledge, all the datasets used in the literature for

<table>
<thead>
<tr>
<th>DB</th>
<th>Table</th>
<th>(Attr# : Tuple#)</th>
<th>FD#</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMIC3</td>
<td>Patients</td>
<td>(7 ; 46.52k)</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Admissions</td>
<td>(18 ; 58.976k)</td>
<td>631</td>
</tr>
<tr>
<td></td>
<td>Diagnoses_icd</td>
<td>(4 ; 651.047k)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Diag_Diagnoses</td>
<td>(3 ; 14.710k)</td>
<td>2</td>
</tr>
<tr>
<td>PTE</td>
<td>active</td>
<td>(2 ; 300)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>bond</td>
<td>(4 ; 9.317k)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>atm</td>
<td>(5 ; 9.189k)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>drug</td>
<td>(1 ; 340)</td>
<td>0</td>
</tr>
<tr>
<td>TPC-H</td>
<td>atom</td>
<td>(3 ; 12.333k)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>connected</td>
<td>(3 ; 24.758k)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>bond</td>
<td>(3 ; 12.379k)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>molecule</td>
<td>(2 ; 343)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Supplier</td>
<td>(7 ; 10k)</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Customer</td>
<td>(8 ; 150k)</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Orders</td>
<td>(9 ; 1.5M)</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>LineItem</td>
<td>(16 ; 6M)</td>
<td>3946</td>
</tr>
<tr>
<td></td>
<td>Nation</td>
<td>(4 ; 23)</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>(3 ; 5)</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Part</td>
<td>(7 ; 200k)</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Partsupp</td>
<td>(5 ; 800k)</td>
<td>11</td>
</tr>
</tbody>
</table>

TABLE I: Data characteristics.
Table II: SPJ queries considered in our experiments.

<table>
<thead>
<tr>
<th>DB</th>
<th>SPJ View</th>
<th>Tuple#</th>
<th>FD#</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMICS</td>
<td>Q(patients ⊙ admissions)</td>
<td>58,976</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>diagnoses⊙ patients</td>
<td>58,798</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>dicddiagnoses ⊙ diagnosesid</td>
<td>658,498</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>[dicddiagnoses⊙patients]⊙ dicddiagnoses</td>
<td>658,498</td>
<td>44</td>
</tr>
<tr>
<td>PTC</td>
<td>atom ⊙ molecule</td>
<td>9,111</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>connected ⊙ bond</td>
<td>24,758</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>[connected ⊙ bond] ⊙ molecule</td>
<td>18,312</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>connected ⊙ atm</td>
<td>18,312</td>
<td>12</td>
</tr>
<tr>
<td>PTE</td>
<td>atm ⊙ drug</td>
<td>9,189</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>active ⊙ drug</td>
<td>299</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>[bond ⊙ drug] ⊙ active</td>
<td>7,994</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>[atom ⊙ bond ⊙ atm] ⊙ drug</td>
<td>9,317</td>
<td>24</td>
</tr>
<tr>
<td>TPC-H</td>
<td>Q2' (P ∞ PS ⊙ S ⊙ N ⊙ R)</td>
<td>21,696</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Q3' (C ∞ O ∞ L)</td>
<td>60,150</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Q9' (P ∞ PS ⊙ S ⊙ L ⊙ O ⊙ N)</td>
<td>3,735,632</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Q11' (P ∞ S ⊙ N)</td>
<td>284,160</td>
<td>151</td>
</tr>
</tbody>
</table>

For \( Coverage(R\bowtie L) > 1 \), there are more tuples in the join result than in tables \( L \) or \( R \) as some tuples may be repeated through the join. The coverage values of the corresponding views are given in Table III.

Methods. We compare the InFine algorithms applied to base tables against four state-of-the-art FD discovery methods, i.e.: (1) TANE [11, 12], (2) Fast_FDs [29], and (3) FUN [21], and (4) HyFD [23] with Java implementation of Metanome [24] using command line. The datasets employed in our study are stored in a PostgreSQL DBMS (version 12.7). Join attributes are indexed with both B-Tree and hash indexes. The reported results correspond to the average over 10 runs for each query.

Metrics. We compute accuracy, average runtime, and maximal memory consumption for each view. In our setting, accuracy corresponds to precision and is defined as the fraction of the number of FDs correctly discovered by InFine applied only to the base tables over the total number of FDs found by baseline methods classically applied to the SPJ view results. Accuracy values are given in Table III and represented as percentages in pie charts of Fig. 5. All methods (including ours) reach accuracy of 1 at the end of their execution.

Comparison Setup. Our goal is to compare our method which discovers FDs from a SPJ view specification and base tables and provides FD provenance information against the straightforward approach which consists of discovering all the FDs from each base table, computing the view result, and discovering the FDs on top of the latter. Classical methods do not provide provenance information. Then, a fair comparison requires the preservation of FD provenance information from the base tables to the view. To know the origin of each FD, both the baselines and our approach have to discover FDs from the base tables first. Since these execution times are the same in both cases, we don’t include them. On the one hand, the competing methods are applied to each SPJ view result and we report their average execution time to which we added the execution time of the full SPJ view computation. On the other hand, InFine is applied to the base tables only and it computes a partial SPJ view depending on the view specification (query sub-tree) and generate FD provenance triples; we report InFine average execution time with time breakdown per algorithm. The time of the partial SPJ view computation is included in mineFDs. The generation of provenance triples is included in the execution of each InFine algorithm in charge of annotating each FD accordingly. Total I/O times of InFine are also reported in Table III.

A. Efficiency Evaluation

In a first set of experiments, we evaluate the runtime and memory consumption of the InFine algorithms compared to the state-of-the-art FD discovery methods that follow the straightforward approach over the 16 SPJ queries on the real-world and synthetic datasets with a wide range of coverage values. The selected queries are representative in terms of coverage and size (number of tables, tuples, and attributes). Results for other queries follow the same trend and are available as supplementary material on Github.

For \( Coverage(R\bowtie L) > 1 \), there are more tuples in the join result than in tables \( L \) or \( R \) as some tuples may be repeated through the join. The coverage values of the corresponding views are given in Table III.
1) Runtime: Fig. 3 presents, for each method, the average total runtime in seconds for FD discovery (including data loading) in log scale. For the competing methods, we included the average execution time of the SPJ view over the indexed data. Similarly, we included the partial SPJ view computation time included in mineFDs. For all methods including ours, we did not include the time for discovering FDs from base tables since these costs are the same. Since our method does not require the full SPJ view computation to discover FDs and does not operate over the query result but only on the single base tables, it is much faster than the traditional methods with one order of magnitude on average up to two orders of magnitude compared to FastFDs (> 2,000 seconds). A provenance triple is generated each time an FD is validated by one of the InFine algorithms and is appended to the corresponding FD data structures with negligible time (included in each algorithm execution time). For MIMIC3 and TPC-H SPJ queries with the highest values for coverage and number of tuples (e.g., Q9* and Q11*), InFine is still outperforming the other methods. The time difference between full and partial SPJ computations shown in Fig. 3 also explains the advantage of our approach.

2) Memory Consumption: As shown in Fig. 4, the average maximal memory consumption of InFine (for accuracy equals 1) is the lowest for all queries across the datasets. HyFD has the second position in terms of memory consumption efficiency and TANE is the worst one. For all methods, Q9* has the highest memory consumption due to its large size of
3.73 millions of tuples, very high coverage (25.8k), and the number of joins between 6 tables. This case is an example of a worst case scenario where many tuples are repeated through the join of the base tables. FD discovery from Q9* requires the mining and checking of all the valid FDs and storing all the relevant information. In this particular case, InFine time performance becomes comparable to HyFD over the full SPJ view computation (due to mineFDs), although all FDs were discovered before by upstageFDs and inferFDs (see Fig. 5).

B. Quality Evaluation

In Fig. 5 we report the average runtime breakdown of each algorithm upstageFDs, inferFDs, and mineFDs of InFine (as horizontal histograms) and their respective percentages of discovered FDs (in the corresponding pie charts with the same color coding). Error bars represent the standard deviation of the average total runtime of InFine algorithms. selectionFDs’s time is included in upstageFDs’s time. Various base table data distributions, SPJ views, and coverage values with very different characteristics across the datasets illustrate the behavior of our algorithms.

Noticably, upstageFDs can retrieve from 33.3% to 59% of the FDs in MIMIC3 in 0.75ms on average, from 58.3% to 75% in PTC in 1.12ms, from 66.7% to 100% in PTE in 1.5ms, and from 42.8.7% to 87.5% in TPC-H in 7.65ms, which are negligible times compared to mineFDs time and I/O time, the most time-consuming steps. Table III gives the number of FDs retrieved by each algorithm and time breakdowns.

The logical inference times are negligible (below 0.0001 ms) and are not reported in the table. inferFDs can retrieve 100% of the FDs for 3 queries over PTE. In some cases, mineFDs is executed but does not return any new FD (e.g., Q*9 in TPC-H or [atm ⊙ bond ⊙ atm]⊙drug), whereas in the other queries, mining subsets of the SPJ view using mineFDs is necessary to recover the remaining FDs. Without computing and mining the join results, upstageFDs and inferFDs can retrieve 83.01% of FDs (68.38±20.00 and 14.63±15.47, respectively) on average across all the datasets. We also observed that join ordering does not affect the total number of discovered FDs, but it changes FD provenance. The main reason is due to the difference in the query execution sub-trees triggering different FD validation by each InFine algorithm.

It should be noted that, due to the NP-completeness of the FD mining problem (cf. [7]), this exploration is necessary to ensure that no valid FD is missed.

These results clearly show the main advantages of our approach, reducing drastically the execution time of FD discovery from SPJ queries with minimal memory consumption, outperforming all the state-of-the art methods tested in our experiments over a large representative range of queries.

VI. RELATED WORK

In the last three decades, numerous approaches from the database and the data mining communities have been proposed to extract automatically valid exact and approximate FDs from single relational tables [16]. Liu et al. [19] have shown
that the complexity of FD discovery is in $O(n^2(\frac{1}{2})^{2k})$ where $k$ is the number of attributes and $n$ the number of records considered. To find FDs efficiently, existing approaches can be classified into three categories: (1) Tuple-oriented methods (e.g., FastFDs [29], DepMiner [20]) that exploit the notion of tuples agreeing on the same values to determine the combinations of attributes of an FD; (2) Attribute-oriented methods (e.g., Tane [11], Fun [21], FDMine [30]) that use pruning techniques and reduce the search space to the necessary set of attributes of the relation to discover exact and approximate FDs. HyFD [25] exploits simultaneously the tuple- and attribute-oriented approaches to outperform the previous approaches; and more recently (3) Structure learning methods relying on sparse regression [31], or on entropy-based measures [15] to score candidate constraints (not limited to FDs alone). More particularly, FDX [31] performs structure learning over a sample constructed by taking the value differences over sampled pairs of tuples from the raw data. In addition, incremental approaches (e.g., [27], [4]) have been developed to tackle data volume and velocity with updating all valid FDs when new tuples are inserted outperforming classical approaches that recalculate all FDs after each data update. Extensive evaluations of FD discovery algorithms can be found in [8, 24]. To the best of our knowledge, previous work on FD discovery did not attempt to address the problem of FD discovery from integrated views in an efficient manner while preserving data provenance. Our approach combining logical inference and selective mining from the base tables adapting FD lattice approach avoids the full computation of FDs from the integrated views and it is the first solution in this direction.

VII. CONCLUSIONS

We addressed the problem of FD discovery from integrated views starting from the FDs of multiple base tables by avoiding the full computation of the view beforehand. The salient features of our work are the following: (1) We leverage single-table approximate FDs that become exact FDs due to the join operation; (2) We leverage logical inference to discover FDs from the base tables without computing the full view result; and (3) We find new multi-table join FDs from partial join using selective mining on the necessary attributes. We empirically show that InFine outperforms, both in terms of runtime and memory consumption, the state-of-the-art FD discovery methods applied to the SPJ views that have to be computed beforehand. We hope that our work will open a new line of research for reusing the FDs discovered from multiple base tables. Various orderings of the base tables lead to different sets of potential upstaged FDs, which, in turn, may trigger different logical inferences. Future work will be to find the optimal ordering of the base tables to reduce the overall execution time and memory consumption.

TABLE III: Accuracy and time breakdowns of InFine algorithms

<table>
<thead>
<tr>
<th>DB Type</th>
<th>SPJ View</th>
<th>(Att# : Tuple#)</th>
<th>Cov.</th>
<th>UpstageFDs</th>
<th>InferFDs</th>
<th>MineFDs</th>
<th>Total Accur.</th>
<th>Total Accur.</th>
<th>I/O (s)</th>
<th>upstageFDs</th>
<th>mineFDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTE</td>
<td>atm ⊢ drug</td>
<td>(5 : 9,189)</td>
<td>14.01</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1 (5 FDs)</td>
<td>0.0024</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>active ≃ drug</td>
<td>(2 : 299)</td>
<td>0.94</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1 (1 FD)</td>
<td>0.0015</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>[bond ≃ drug] ⊢ active</td>
<td>(6 : 7,994)</td>
<td>13.83</td>
<td>0.67</td>
<td>0.33</td>
<td>0</td>
<td>1 (6 FDs)</td>
<td>0.0215</td>
<td>0.0030</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>[atom ≃ bond] ⊢ atom ⊢ drug</td>
<td>(14 : 9,317)</td>
<td>14.20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1 (24 FDs)</td>
<td>0.0879</td>
<td>0.0030</td>
<td>0.0492</td>
<td>0.0492</td>
</tr>
<tr>
<td></td>
<td>atom ≃ molecule</td>
<td>(4 : 9,111)</td>
<td>13.67</td>
<td>0.75</td>
<td>0.25</td>
<td>0</td>
<td>1 (4 FDs)</td>
<td>0.0231</td>
<td>0.0000</td>
<td>0.0015</td>
<td>0.0015</td>
</tr>
<tr>
<td></td>
<td>connected ≃ bond</td>
<td>(5 : 24,758)</td>
<td>1.50</td>
<td>0.625</td>
<td>0.375</td>
<td>0</td>
<td>1 (8 FDs)</td>
<td>0.1012</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0015</td>
</tr>
<tr>
<td></td>
<td>[connected ≃ bond] ⊢ molecule</td>
<td>(6 : 18,312)</td>
<td>27.08</td>
<td>0.75</td>
<td>0.25</td>
<td>0</td>
<td>12 (12FDs)</td>
<td>0.0868</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>connected ≃ [atom ≃ molecule]</td>
<td>(6 : 18,312)</td>
<td>27.08</td>
<td>0.583</td>
<td>0.417</td>
<td>0</td>
<td>1 (12 FDs)</td>
<td>0.0903</td>
<td>0.0030</td>
<td>0.0045</td>
<td>0.0045</td>
</tr>
<tr>
<td>MIMIC3</td>
<td>diagnosed ≃ patients</td>
<td>(12 : 651,047)</td>
<td>7.50</td>
<td>0.591</td>
<td>0.273</td>
<td>0.136</td>
<td>1 (22 FDs)</td>
<td>2.1876</td>
<td>0.0015</td>
<td>2.3120</td>
<td>2.3120</td>
</tr>
<tr>
<td></td>
<td>discriminator ≃ discriminator</td>
<td>(7 : 658,498)</td>
<td>22.84</td>
<td>0.333</td>
<td>0</td>
<td>0.667</td>
<td>1 (12 FDs)</td>
<td>1.7200</td>
<td>0.0000</td>
<td>0.3497</td>
<td>0.3497</td>
</tr>
<tr>
<td></td>
<td>[diagnosed ≃ patients] ≃ discriminator</td>
<td>(14 : 658,498)</td>
<td>22.84</td>
<td>0.545</td>
<td>0</td>
<td>0.455</td>
<td>1 (44 FDs)</td>
<td>5.1323</td>
<td>0.0000</td>
<td>0.1325</td>
<td>0.1325</td>
</tr>
<tr>
<td></td>
<td>[patients ≃ admissions]</td>
<td>(10 : 6,736)</td>
<td>0.79</td>
<td>0.563</td>
<td>0</td>
<td>0.037</td>
<td>1 (16 FDs)</td>
<td>0.2360</td>
<td>0.0015</td>
<td>0.0230</td>
<td>0.0230</td>
</tr>
<tr>
<td></td>
<td>Q2′ ≃ P ⊢ (S ⊸ S ⊸ S ∋ N) ⊸ (S ∋ N)</td>
<td>(10 : 21,096)</td>
<td>1.50</td>
<td>0.594</td>
<td>0.087</td>
<td>0.319</td>
<td>1 (69 FDs)</td>
<td>0.1299</td>
<td>0.0045</td>
<td>0.0120</td>
<td>0.0120</td>
</tr>
<tr>
<td></td>
<td>Q3′ (C ⊸ O ⊸ O)</td>
<td>(6 : 60,150)</td>
<td>0.12</td>
<td>0.429</td>
<td>0</td>
<td>0.571</td>
<td>1 (14 FDs)</td>
<td>13.036</td>
<td>0.0000</td>
<td>0.0198</td>
<td>0.0198</td>
</tr>
<tr>
<td></td>
<td>Q9∗ (P ⊢ PS ⊢ S ⊸ (L ⊢ O ⊸ O) ⊸ N)</td>
<td>(9 : 3,735,632)</td>
<td>25.813</td>
<td>0.875</td>
<td>0.125</td>
<td>0</td>
<td>1 (8 FDs)</td>
<td>16.967</td>
<td>0.0015</td>
<td>12.1261</td>
<td>12.1261</td>
</tr>
<tr>
<td></td>
<td>Q11∗ (P ⊢ S ⊸ S ⊸ N)</td>
<td>(15 : 284,160)</td>
<td>80.09</td>
<td>0.636</td>
<td>0.232</td>
<td>0.132</td>
<td>1 (151 FDs)</td>
<td>13.771</td>
<td>0.0246</td>
<td>0.5777</td>
<td>0.5777</td>
</tr>
</tbody>
</table>

REFERENCES


APPENDIX

Theorem and sketches of proofs:

Theorem 1: Let \( V, V_1, \) and \( V_2 \) be view specifications. Let \( \mathcal{D}, \mathcal{D}_1, \) and \( \mathcal{D}_2 \) be the sets of FDs over views specified by \( V, V_1 \), and \( V_2, \) respectively. Let \( fds(V) \) denote the set of FDs over a view specified by \( V \). Then:

\[
fds(\pi_X(V)) \subseteq \mathcal{D}, \quad fds(\sigma_\rho(V)) \supseteq \mathcal{D}, \quad \text{and} \quad fds(V_1 \circ V_2) \supseteq \mathcal{D}_1 \cup \mathcal{D}_2, \quad \text{with} \ \phi \in \{ \exists; \exists^+; \exists\star; \exists^\star; \text{x} \}
\]

Proof: Trivial.

Lemma 1 (FD types conservation through join operations): Let \( \mathcal{R} = \{ \mathcal{R}_1; \ldots; \mathcal{R}_n \} \) be a set of relational instances. Let \( fds(\mathcal{R}) \) denote the set of minimal FDs over the relational instance \( \mathcal{R} \). Let \( \Sigma_{left\_up}, \Sigma_{right\_up}, \Sigma_{in\_f} \) and \( \Sigma_{join} \) denote the sets of left upstaged, right upstaged, inferred and join FDs over \( \mathcal{R}_1 \circ \cdots \circ \mathcal{R}_n \), respectively. Then, the sets \( \Sigma_{left\_up} \cup \Sigma_{right\_up}, \Sigma_{in\_f} \) and \( \Sigma_{join} \) are equals regardless of the join order. At the opposite, equality between sets \( \Sigma_{left\_up} \) and \( \Sigma_{right\_up} \) cannot be guaranteed.

Proof:

Lemma 2 (Upstaged join FDs): Let \( \mathcal{L} \) and \( \mathcal{R} \) be two instances over relations \( S \) and \( T \), respectively, and \( \mathcal{D}_L \) and \( \mathcal{D}_R \) be the two sets of all FDs such that \( L = \mathcal{D}_L \) and \( R = \mathcal{D}_R \), respectively. Then the sets of upstaged FDs denoted \( \mathcal{D}^{new}_L \) and \( \mathcal{D}^{new}_R \) are the sets:

\[
\mathcal{D}^{new}_L = \{ d \mid d \notin \mathcal{D}_L \land (L \circ \pi_X(Y) \circ \mathcal{R}_L) = d \} \tag{4}
\]

\[
\mathcal{D}^{new}_R = \{ d \mid d \notin \mathcal{D}_R \land (\pi_X(L) \circ \mathcal{R}_L = d \} \tag{5}
\]

Theorem 2: Let \( L \) and \( R \) be two instances over relations \( S \) and \( T \), respectively. Let \( L \circ \pi_X = Y \) be a join result with \( X \subseteq atts(L), Y \subseteq atts(R) \). For all \( A \subseteq atts(L) \setminus X \) and \( B \subseteq atts(R) \setminus Y \):

\[
\text{if } L \circ \pi_X = Y \neq X \rightarrow B \text{ then } L \circ \pi_X = Y \neq A \rightarrow B
\]

Proof sketch: Let \( a_1, \ldots, a_n \) be values over the attributes in \( X \), and \( b_1, \ldots, b_m, b_1', \ldots, b_m' \) be values over the attributes in \( Y \). If \( L \circ \pi_X = Y \neq X \rightarrow B \) then there exist two tuples:

\[
t_R(x_1, \ldots, x_n, a_1, \ldots, a_k, b_1, \ldots, b_m)
\]

\[
t'_R(x_1, \ldots, x_n, a_1, \ldots, a_k, b_1', \ldots, b_m')
\]

in \( R \) such that: \( \exists i \in [1, \ldots, n], b_i' \neq b'_i \) and there exists a tuple \( t_L(x_1, \ldots, x_n, a_1, \ldots, a_k) \) in \( L \) with \( a_1, \ldots, a_k \), the values of the attributes in \( A \) (otherwise, tuples \( t_R \) and \( t'_R \) would have been filtered during the join operation). Thus, the join \( L \circ \pi_X = Y \) leads to the two tuples:

\[
t(x_1, \ldots, x_n, a_1, \ldots, a_k, b_1, \ldots, b_m)
\]

\[
t'(x_1, \ldots, x_n, a_1, \ldots, a_k, b_1', \ldots, b_m')
\]

which violate the FD \( A \rightarrow B \) and \( L \circ \pi_X = Y \neq A \rightarrow B \)

Proof: This is trivially derived from lemma 3 by transitivity, with the use of Armstrong’s transitivity axiom.

Theorem 3: Let \( L \) and \( R \) be two instances over relations \( S \) and \( T \), respectively. Let \( L \circ \pi_X = Y \) be a join result with \( X \subseteq atts(L), Y \subseteq atts(R) \). We cannot guarantee that all FDs over \( L \circ \pi_X = Y \) can be inferred from \( X \)’s axioms over the FDs over \( L \) and \( R \) taken separately.

Proof: In the two following instances \( L \) and \( R \), we can see that only the FDs \( YA' \rightarrow b \) and \( Yb \rightarrow A' \) hold.

In the join result, the FD \( AA' \rightarrow b \) holds but it cannot be inferred using Armstrong’s axioms over the FDs discovered from each instance \( L \) and \( R \).

Theorem 4: Let \( L \) and \( R \) be two instances over relations \( S \) and \( T \), respectively. Let \( L \circ \pi_X = Y \) be a join result with \( X \subseteq atts(L), Y \subseteq atts(R) \). For all \( A \subseteq atts(L) \setminus X \) and \( A' \subseteq atts(R) \setminus Y \):

\[
\text{if } L \circ \pi_X = Y \neq X \rightarrow B \text{ then } L \circ \pi_X = Y \neq A \rightarrow B
\]

Proof sketch: Let \( a_1, \ldots, a_n \) be values over the attributes in \( X \), and \( b_1, \ldots, b_m, b_1', \ldots, b_m' \) be values over the attributes in \( Y \). If \( L \circ \pi_X = Y \neq X \rightarrow B \) then there exist two tuples:

\[
t_R(x_1, \ldots, x_n, a_1, \ldots, a_k, b_1, \ldots, b_m)
\]

\[
t'_R(x_1, \ldots, x_n, a_1, \ldots, a_k, b_1', \ldots, b_m')
\]

in \( R \) such that: \( \exists i \in [1, \ldots, n], b_i \neq b'_i \) and there exists a tuple \( t_L(x_1, \ldots, x_n, a_1, \ldots, a_k) \) in \( L \) with \( a_1, \ldots, a_k \), the values of the attributes in \( A \) (otherwise, tuples \( t_R \) and \( t'_R \) would have been filtered during the join operation). Thus, the join \( L \circ \pi_X = Y \) leads to the two tuples:

\[
t(x_1, \ldots, x_n, a_1, \ldots, a_k, b_1, \ldots, b_m)
\]

\[
t'(x_1, \ldots, x_n, a_1, \ldots, a_k, b_1', \ldots, b_m')
\]

which violate the FD \( A \rightarrow B \) and \( L \circ \pi_X = Y \neq A \rightarrow B \)

Proof sketch: Algorithm [1] begins by extracting the attributes that are retrieved by a view in order to guide the mining process of these attributes while still guaranteeing incompleteness. Then Algorithm [1] recursively explores the view specification, thus its incompleteness can be shown inductively.

For the base case where we have a relational instance \( R \) (Algorithm [1, line# 9]), the FDs are mined using a classic level-wise algorithm allowing to retrieve every FD in \( R \). In the case of a projection (Algorithm [1, line# 15]), as the relevant attributes are extracted at the very first step of Algorithm [1], no computation is performed and the previously computed FDs are returned. In the cases of a selection or a join (Algorithm [1, line# 17 or line# 20 respectively), the tuples are filtered by the join operation and Algorithm [2] mines all exact FDs. Algorithm [4] retrieves the minimal FDs from the logically inferred FDs, thus no lhs subset of FDs remains unchecked.
Theorem 4 indicates which part of the candidate FD lattice is pruned. Then, Algorithm 5 explores the candidate FDs using a classic bottom-up approach, thus no minimal FD remains unchecked. Overall, our algorithms explore the lattice of candidate FDs until they find minimal FDs; they avoid only the parts of the lattice that do not contain valid candidate FDs, thus InFine with selective mining retrieves the complete set of minimal candidate FDs.

**Theorem 6 (InFine correctness):** Let $R = \{R_1; \ldots; R_n\}$ be a set of relational instances. Let $V_R$ be a view specification over relations in $R$. Let $I_{V_R}$ be the view specified by $V_R$. Let $fds(InFine)$ denotes the set of FDs computed by InFine over the view specified by $V_R$. Then:

$$\forall d \infds(InFine), I_{V_R} \models d$$

**Proof sketch:** Algorithm 1 recursively explores the view specification, thus its correctness can be shown inductively. For the base case where we have a relational instance $R$ (Algorithm 1, line# 9) the FDs are mined using a classic level-wise algorithm allowing to retrieve FDs in $R$.

In case of a projection (Algorithm 1, line# 15), the process ensure that only the FDs with relevant attributes (i.e., the one that are retrieved in the output view) are kept. Thus, if the FDs output by the steps are correct, the handling of projection will only lead to remove irrelevant FDs without introducing incorrect ones.

In case of a selection (Algorithm 1, line# 17), Algorithm 2 valid exact FDs are discovered from the data using the principle of the level-wise algorithms adapted to mine upstaged FDs, thus the FDs are guaranteed to hold on the joined instance and only minimal FDs are kept.

In case of a join (Algorithm 1 line 20), at first in Algorithm 3 valid exact FDs are discovered from the data using the same principle as in Algorithm 2 thus returning FDs that are guaranteed to hold on the joined instance. Theorem 2 shows the correctness of the set of FDs inferred through logical inference. Then, the subroutine refine checks the correctness of its candidates FDs holding on the data. Therefore, the set of FDs $D_2$ discovered by Algorithm 4 is such that $D_2 \models D_0$. In Algorithm 5, Theorem 4 enforces the retrieval of FDs based only on the attributes that can become rhs in the joined instance, then only plausible candidate FDs are explored. Therefore, every discovered FD holds in the joined instance. By construction, Algorithm 4 and 5 lead to the retrieval of FDs that are minimal.