Extending inclusion dependencies with conditions

Citation for published version:

Digital Object Identifier (DOI):
10.1016/j.tcs.2013.11.002

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Theoretical Computer Science

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Extending inclusion dependencies with conditions

Shuai Ma\textsuperscript{a,\*}, Wenfei Fan\textsuperscript{a,\textsuperscript{b}}, Loreto Bravo\textsuperscript{c}

\textsuperscript{a} SKLSDE Lab, Beihang University, China
\textsuperscript{b} University of Edinburgh, UK
\textsuperscript{c} University of Concepcion, Chile

ARTICLE INFO

Article history:
Received 30 January 2013
Received in revised form 25 October 2013
Accepted 2 November 2013
Communicated by D. Sannella

Keywords:
Satisfiability
Implication
Inclusion dependencies
Data cleaning
Schema matching

ABSTRACT

This paper introduces a class of conditional inclusion dependencies (CINDs), which extends inclusion dependencies (INDs) by enforcing patterns of semantically related data values. We show that CINDs are useful not only in data cleaning, but also in contextual schema matching. We give a full treatment of the static analysis of CINDs, and show that CINDs retain most desired properties of traditional INDs: (a) CINDs are always satisfiable; (b) CINDs are finitely axiomatizable, i.e., there exists a sound and complete inference system for the implication analysis of CINDs; and (c) the implication problem for CINDs has the same complexity as its traditional counterpart, namely, PSPACE-complete, in the absence of attributes with a finite domain; but it is EXPTIME-complete in the general setting. In addition, we investigate the interaction between CINDs and conditional functional dependencies (CFDs), as well as two practical fragments of CINDs, namely acyclic CINDs and unary CINDs. We show the following: (d) the satisfiability problem for the combination of CINDs and CFDs becomes undecidable, even in the absence of finite-domain attributes; (e) in the absence of finite-domain attributes, the implication problem for acyclic CINDs and for unary CINDs retains the same complexity as its traditional counterpart, namely, NP-complete and PTIME, respectively; but in the general setting, it becomes PSPACE-complete and coNP-complete, respectively; and (f) the implication problem for acyclic unary CINDs remains in PTIME in the absence of finite-domain attributes and coNP-complete in the general setting.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

A class of \textit{conditional functional dependencies} (CFDs) has recently been proposed in \cite{1} as an extension of \textit{functional dependencies} (FDs). In contrast to traditional FDs, CFDs hold conditionally on a relation, i.e., they apply only to those tuples that satisfy certain data-value patterns, rather than to the entire relation. CFDs have proven useful in data cleaning \cite{1–7}: inconsistencies and errors in the data may be captured as violations of CFDs, whereas they may not be detected by traditional FDs.

It has been recognized (e.g., \cite{8,9}) that to clean relational data, one should make use of not only FDs, but also \textit{inclusion dependencies} (INDs). Furthermore, INDs are commonly used in schema matching systems, e.g., Clio \cite{10}: INDs associate attributes in a source schema with semantically related attributes in a target schema. Nevertheless, schema matching and data cleaning often need to use inclusion dependencies that hold on a part of data that satisfies certain constant patterns, rather than on the entire relations. As illustrated by the examples below, dependencies with constant patterns cannot be expressed as traditional INDs. These suggest that we extend INDs by incorporating patterns of semantically-related data values, along the same lines as CFDs.
Example 1.1. Consider the two relational schemas, referred to as source and target:

**source**: order(asin: string, title: string, type: string, price: real)

**target**: book(isbn: string, title: string, price: real, format: string)

The source database contains a single relation `order`, specifying items of various types such as books, CDs and DVDs, ordered by customers. The target database has two relations, namely, book and CD, specifying items of books and CDs ordered by customers, respectively. In each relation the underlined attributes indicate a key (a special case of FDs), e.g., asin is a key for `order`. Example source and target instances $D_1$ are shown in Fig. 1. Note that keys (FDs) are not the focus of this work, but to make the schemas more rational.

To find schema mappings from the source to the target, or to detect errors across these databases, one might be tempted to use standard INDs such as:

$$\text{ind}_1: \text{order} (\text{title}, \text{price}) \subseteq \text{book} (\text{title}, \text{price})$$

$$\text{ind}_2: \text{order} (\text{title}, \text{price}) \subseteq \text{CD} (\text{album}, \text{price})$$

These INDs, however, do not make sense: one cannot expect the title and price of each book item in the `order` table to find a matching CD tuple (e.g., Harry Potter); similarly for the `CD` items in the `order` table. Nevertheless, there are indeed inclusion dependencies from the source to the target, as well as on the target database, but only under certain conditions:

$$\text{cind}_1: \text{order} (\text{title}, \text{price}, \text{type} = \text{‘book’}) \subseteq \text{book} (\text{title}, \text{price})$$

$$\text{cind}_2: \text{order} (\text{title}, \text{price}, \text{type} = \text{‘CD’}) \subseteq \text{CD} (\text{album}, \text{price})$$

$$\text{cind}_3: \text{CD} (\text{album}, \text{price}, \text{genre} = \text{‘a-book’}) \subseteq \text{book} (\text{title}, \text{price}, \text{format} = \text{‘audio’})$$

Here constraint `cind_1` states that for each `order` tuple $t$, if its type is ‘book’, then there must exist a book tuple $t’$ such that tuples $t$ and $t’$ agree on their `title` and `price` attributes; similarly for constraint `cind_2`. Here `type = ‘book’` specifies a condition under which the constraint can be applied. Constraint `cind_3` asserts that for each CD tuple, if its `genre` is ‘a-book’ (audio book), then there must be a book tuple $t’$ such that the `title` and `price` of $t’$ are identical to the `album` and `price` of $t$, and moreover, the `format` of $t’$ must be ‘audio’. Here `format = ‘audio’` is an additional condition on matched tuples.

Constraints `cind_1` and `cind_2` specify a form of contextual schema matching studied in [11]. As shown in [11], contextual schema matching often allows us to derive sensible schema mapping from a source to a target, which cannot be found via schema matching specified with traditional INDs.

Such constraints also allow us to detect errors across different relations. For instance, while $D_1$ of Fig. 1 satisfies `cind_1` and `cind_2`, it violates `cind_3`. Indeed, tuple $t_5$ in the CD table has an ‘a-book’ genre, but it cannot find a match in the book table with ‘audio’ format. The violation suggests that there may exist inconsistencies in the CD and book tables in the target database. Such inconsistencies cannot be detected by traditional INDs. Note that the book tuple $t_4$ is not a match for $t_5$: although $t_5$ and $t_4$ agree on their `album` (title) and `price` attributes, the format of $t_4$ is ‘paperback’ rather than ‘audio’ as required by `cind_3`.

Like CFDs, dependencies `cind_1–cind_3` are required to hold only on a subset of tuples satisfying certain patterns. In other words, they apply only conditionally to relations order, CD, and book. These dependencies are specified with constants, and hence, cannot be expressed as standard INDs. Although such dependencies are needed for schema matching and data cleaning, to the best of our knowledge, the earlier conference version [12] of this paper is the first to study these constraints.

Contributions To this end we introduce an extension of INDs, and investigate the static analyses of these constraints.
1 Our first contribution is a notion of conditional inclusion dependencies (CINDs). A CIND is defined as a pair consisting of an IND \( R_1[X] \subseteq R_2[Y] \) and a pattern tableau, where the tableau enforces binding of semantically related data values across relations \( R_1 \) and \( R_2 \). For example, \( \text{ind}_1 \), \( \text{ind}_2 \), and \( \text{ind}_1 \rightarrow \text{cind}_3 \) given above all can be expressed as CINDs. In particular, traditional INDs are a special case of CINDs. This mild extension of INDs captures a fundamental part of the semantics of the data, and suffices to express rules commonly used in data cleaning and schema matching.

2 Our second contribution consists of complexity results for fundamental problems associated with CINDs, as well as an inference system for reasoning about CINDs.

Given a set of CINDs, the first question one would ask is whether the CINDs are satisfiable, i.e., whether they are “dirty” themselves. Indeed, one does not want to enforce the CINDs on a database at running time but find, after repeated failures, that the CINDs cannot possibly be satisfied by a nonempty database. Similarly, one does not want to match schema based on CINDs that do not make sense. The satisfiability analysis helps users identify satisfiable sets of CINDs for data cleaning and schema matching.

Another important question concerns the implication analysis, which is to decide whether a set of CINDs entails another CIND. The implication analysis is useful in reducing redundant CINDs, and hence improving performance when detecting CIND violations in a database, and speeding up the derivation of schema mappings from CINDs [10].

For traditional INDs, the satisfiability analysis is not an issue: any set of INDs is satisfiable. Their implication analysis is PSPACE-complete, and furthermore, it is infinitely axiomatizable: there exists a finite, sound and complete set of axioms (see, e.g., [13]).

We show that although CINDs are more expressive than INDs, they retain most desired properties of their IND counterpart: (a) CINDs are always satisfiable; (b) the implication of CINDs is infinitely axiomatizable; (c) in the absence of finite-domain attributes, the implication problem for CINDs is PSPACE-complete.

In the real world, it is common to find finite-domain attributes, e.g., Boolean, date, etc. It is hence necessary to get the complexity of these problems right in the general setting when finite-domain attributes may be present. The implication problem of INDs remains intact in the general setting, as their inference does not involve any data values. However, in the general setting, the implication problem for CINDs becomes EXPSPACE-complete, due to the interaction between the data values in finite domains and constants in pattern tableaux.

3 Our third contribution consists of complexity bounds for reasoning about the combination of CINDs and CFDs. As remarked earlier, to clean relational data, one may need both CFDs and CINDs in practice.

We show that the presence of finite-domain attributes does not complicate the satisfiability and implication analyses when INDs and FDs (resp. CINDs and CFDs) are taken together. Nonetheless, while a set of INDs and FDs is always satisfiable [13], we show that the satisfiability problem for CINDs and CFDs taken together becomes undecidable, even in the absence of finite-domain attributes, due to the presence of data values.

4 Our fourth contribution consists of complexity bounds for reasoning about two fragments of CINDs: acyclic CINDs and unary CINDs, which extend acyclic INDs and unary INDs (see e.g., [13] for details), respectively. Many CINDs found in practice are either acyclic or unary. For instance, the set \( \{\text{cind}_1, \text{cind}_2, \text{cind}_3\} \) of CINDs that we have seen earlier is a set of acyclic CINDs.

We show that in the absence of attributes with a finite domain, the implication problem for acyclic CINDs is NP-complete, and it is in PTIME for unary CINDs, the same as their counterparts for classical acyclic and unary INDs, respectively. That is, acyclic CINDs (resp. unary CINDs) retain the same complexity as acyclic INDs [14] (resp. unary INDs [15]). Nevertheless, we also show that in the general setting, the implication problem becomes PSPACE-complete for acyclic CINDs, and it is coNP-complete for unary CINDs. This tells us that the increased expressive power of CINDs does not come for free. Nevertheless, these complexity bounds are still lower than their counterparts for general CINDs, namely, PSPACE and EXPSPACE, respectively. Therefore, when only acyclic or unary CINDs are needed, we do not have to pay the price of the complexity of the full-fledged CINDs.

We show that further constraining acyclic (or unary) CINDs does not make our lives easier. Indeed, the implication problem for acyclic unary CINDs remains coNP-complete in the general setting (while in the absence of attributes with finite domains, it of course remains in PTIME).

These results settle several fundamental problems associated with CINDs, an extension of INDs that is useful in schema matching and data cleaning. (1) We show that one can specify any CINDs without worrying about their satisfiability. (2) We develop a sound and complete inference system for the implication analysis of CINDs, which provides algorithmic insight into reasoning about CINDs. (3) We present a comprehensive picture of complexity bounds on the implication analysis of CINDs, when finite-domain attributes are present or absent, for general CINDs and for practical fragments (acyclic CINDs and unary CINDs). (4) We also show that there exists interaction between CFDs and CINDs.

Remark. (1) This paper is an extension of our earlier work [12] by including (a) the study of acyclic CINDs (Section 5.1), (b) the investigation of unary CINDs (Section 5.2), (c) the complexity bounds on the implication problem for standard INDs, AINDs and UINDs in the presence of finite-domain attributes (Sections 3, 5.1 and 5.2), and (d) the proofs for the sound and complete inference system and for the complexity bounds for the satisfiability and implication analyses of CINDs (Section 3). Some of the proofs are nontrivial and the techniques are interesting in their own right. The algorithm for the satisfiability checking of CINDs and FDs and its experimental study were reported in [12]; these are left out from this paper in order to focus on the theoretical aspects of CINDs.
(2) CINDs do not introduce a new logical formalism. Indeed, in first-order logic, they can be expressed in a form similar to tuple-generating dependencies (TGDs), which have lately generated renewed interests in data exchange (see [16] for a survey). However, (a) these simple CINDs suffice to capture data consistency and contextual schema matching commonly found in practice, without incurring the complexity of full-fledged TGDs (e.g., the undecidability of their implication problem), and (b) no prior work has studied the satisfiability, implication and finite axiomatizability of TGDs in the presence of constants or finite-domain attributes.

Organization The remainder of the paper is organized as follows. We define CINDs in Section 2, and investigate their satisfiability and implication problems in Section 3. The interaction between CFDs and CINDs is studied in Section 4. We study acyclic CINDs and unary CINDs in Section 5, followed by related work in Section 6 and topics for future work in Section 7. To improve the readability we defer the detailed proofs to Appendix A, but include their sketches in the main paper.

2. Conditional inclusion dependencies

A relational database schema \( \mathcal{R} \) is a finite collection of relation schemas \( (R_1, \ldots, R_n) \), where for each \( i \in [1, n] \), \( R_i \) is defined over a finite set of attributes, denoted as \( \text{attr}(R_i) \). For each attribute \( A \in \text{attr}(R_i) \), its domain is specified in \( R_i \), denoted as \( \text{dom}(A) \), which is either finite (e.g., bool) or infinite (e.g., string). We use \( \text{finattr} \) to denote the set of all the finite-domain attributes that appear in \( \mathcal{R} \). We refer to an attribute in terms of the name of the relation and the name of the attribute, and only use the attribute name when it is clear in the context.

An instance \( I_i \) of \( R_i \) is a finite set of tuples such that for each \( t \in I_i \), \( t[A] \in \text{dom}(A) \) for all attributes \( A \in \text{attr}(R_i) \).

A database instance \( D \) of \( \mathcal{R} \) is a collection of relation instances \((I_1, \ldots, I_n)\), where \( I_i \) is an instance of \( R_i \) for each \( i \in [1, n] \).

Conditional inclusion dependencies A conditional inclusion dependency (CIND) \( \psi \) is defined as a pair

\[
\left( \mathcal{R}_a[X \colon X_p] \subseteq \mathcal{R}_b[Y \colon Y_p], T_p \right),
\]

where

1. \( X, X_p \) and \( Y, Y_p \) are lists of attributes in \( \text{attr}(R_a) \) and \( \text{attr}(R_b) \), respectively, such that \( X \) and \( X_p \) (resp. \( Y \) and \( Y_p \)) are disjoint;
2. \( \mathcal{R}_a[X] \subseteq \mathcal{R}_b[Y] \) is a standard IND, referred to as the IND embedded in \( \psi \); and
3. \( T_p \) is a tableau, called the pattern tableau of \( \psi \), which is a set of tuples over attributes \( X_p \cup Y_p \) such that each tuple \( t_p \in T_p \) and each attribute \( A \) in \( X \cup Y, t_p[A] \) is a constant drawn from \( \text{dom}(A) \).

We require that \( X \) and \( X_p \) (resp. \( Y \) and \( Y_p \)) are disjoint to make CINDs succinct, and to help eliminate useless constraints. One can readily verify that when \( X \) and \( X_p \) (resp. \( Y \) and \( Y_p \)) are not disjoint, we can easily define an equivalent CIND that satisfies the disjointness condition.

We adopt the following conventions and notations. (1) Let \( X = [A_1, \ldots, A_m] \) and \( Y = [B_1, \ldots, B_m] \). We require that \( \text{dom}(A_i) \subseteq \text{dom}(B_i) \) for each \( i \in [1, m] \). (2) If a list \( Z \) of attributes occurs in both \( X_p \) and \( Y_p \), we use \( Z_L \) and \( Z_R \) to indicate the occurrence of \( Z \) in \( X_p \) and \( Y_p \), respectively. (3) When both \( X_p \) and \( Y_p \) are empty lists, \( T_p \) is an empty set \( \emptyset \). (4) We use \( U \cup V \) to denote the set of all attributes of \( X \) and \( Y \), and \( X \setminus Y \) to denote the list obtained from list \( X \) by removing all the elements in list \( Y \). We denote \( X \cup X_p \) as \( \text{LHS}(\psi) \) and \( Y \cup Y_p \) as \( \text{RHS}(\psi) \), and separate the LHS and RHS attributes in a pattern tuple with ‘|’.

We use \( \texttt{nil} \) to denote an empty list. (6) In addition, we adopt the common assumption that each finite or infinite domain contains at least two elements [17], which was used when studying, e.g., FDSs [13].

Example 2.1. Constraints \( \text{ind}_{1}, \text{ind}_{2}, \) and \( \text{cind}_{1} \text{--cind}_{3} \) given in Examples 1.1 can all be expressed as CINDs, as shown in Fig. 2: \( \psi_1 \) and \( \psi_2 \) for \( \text{ind}_{1} \) and \( \text{ind}_{2} \), and \( \psi_{3} \text{--} \psi_{5} \) for \( \text{cind}_{1} \text{--cind}_{3} \), respectively. Observe that \( \text{ind}_{1} \) and \( \text{ind}_{2} \) are standard INDs embedded in \( \psi_1 \) and \( \psi_2 \), respectively. In \( \psi_5 \), \( X \) is [album, price], \( Y \) is [title, price], \( X_p \) is [genre], and \( Y_p \) is [format]. The standard IND embedded in \( \psi_5 \) is CD[album, price] \subseteq book[title, price]. □
Consider CIND $\psi = (R_a[X; X_p] \subseteq R_b[Y; Y_p], T_p)$. In general, the embedded IND may not hold on the entire $R_a$ relation: it applies only to $R_a$ tuples matching certain pattern tuples in $T_p$. We say that an $R_a$ (resp. $R_b$) tuple $t_1$ (resp. $t_2$) matches a pattern tuple $t_p \in T_p$ if $t_1[X] = t_p[X]$ (resp. $t_2[Y] = t_p[Y]$).

An instance $(I_a, I_b)$ of $(R_a, R_b)$ satisfies the CIND $\psi$, denoted by $(I_a, I_b) \models \psi$, if and only if for each tuple $t_1$ in the relation $I_a$, and for each pattern tuple $t_p$ in the pattern tableau $T_p$, if $t_1[X] = t_p[X]$, then there exists a tuple $t_2$ in the relation $I_b$ such that $t_1[X] = t_2[Y]$, and moreover, $t_2[Y] = t_p[Y]$.

That is, if $t_1[X] = t_p[X]$, then the standard IND embedded in $\psi$ and the pattern specified by $t_p$ must be satisfied. More specifically, there exists a tuple $t_2$ such that (1) $t_2[Y] = t_1[X]$, and (2) $t_2[Y] = t_p[Y]$. Note that $t_1[X] = t_p[X]$ if $X_p = \text{nil}$, and $t_2[Y] = t_p[Y]$ if $Y_p = \text{nil}$. Intuitively, $X_p$ is used to identify the $R_a$ tuples over which $\psi$ is applied. The pattern on $Y_p$ enforces the matching $R_b$ tuples to have certain values in their $Y_p$ attributes.

We say that a database $D$ satisfies a set $\Sigma$ of CINDs, denoted by $D \models \Sigma$, if $D \models \psi$ for each $\psi \in \Sigma$. Two sets $\Sigma_1$ and $\Sigma_2$ of CINDs are said to be equivalent, denoted by $\Sigma_1 \equiv \Sigma_2$, if for any instance $D$, $D \models \Sigma_1$ iff $D \models \Sigma_2$.

**Example 2.2.** Database $D_1$ given in Fig. 1 satisfies CINDs $\psi_3$ and $\psi_4$. However, the INDs embedded in these CINDs do not necessarily hold. For example, while $\psi_4$ is satisfied, $\text{ind}^3_2$ in $\psi_4$ is not, since item Harry Potter in the order table cannot find a match in the CD table. The pattern $X_p$ in LHS($\psi_4$) identifies the order tuples on which $\psi_4$ has to be enforced, i.e., those CD tuples; similarly for $\psi_3$.

On the other hand, $\psi_5$ is violated by the database. Indeed, for CD tuple $t_5$, there exists a pattern tuple $t_p$ in $T_5$ such that $t_5[\text{genre}] = t_p[\text{genre}] = \text{‘book’}$ but there exists no tuple $t$ in table book such that $t[\text{format}] = \text{‘audio’}$, $t[\text{title}] = t_5[\text{title}] = \text{‘Snow White’}$, and $t[\text{price}] = t_p[\text{price}] = 7.99$. Here the genre pattern is to identify CD tuples on which $\psi_5$ is applicable, while format is a constraint on the book tuples that match those CD tuples via the IND embedded in $\psi_5$.

**Normal form** A CIND $\psi = (R_a[X; X_p] \subseteq R_b[Y; Y_p], T_p)$ is in the normal form if $T_p$ only consists of a single pattern tuple $t_p$. We write $\psi$ as $(R_a[X; X_p] \subseteq R_b[Y; Y_p], t_p)$.

It is straightforward to verify that a CIND $\psi = (R_a[X; X_p] \subseteq R_b[Y; Y_p], T_p)$ can be expressed as a set $\Sigma_\psi = \{(R_a[X; X_p] \subseteq R_b[Y; Y_p], t_p) | t_p \in T_p\}$ of CINDs in the normal form, which is equivalent to $\{\psi\}$, i.e., $\{\psi\} \equiv \Sigma_\psi$. In light of this, in the sequel we shall w.l.o.g. consider CINDs in the normal form only.

### 3. Reasoning about conditional inclusion dependencies

For any constraint language $L$, there are two fundamental problems associated with it. One is the satisfiability problem, which is to determine whether a given set of constraints in $L$ has conflicts. The other is the implication problem, to derive other constraints from a given set of constraints in $L$. As remarked in Section 1, to effectively use a constraint language in practice, it is often necessary to answer these two questions at compile time.

One might be tempted to use a constraint language that is more powerful than CINDs, e.g., full-fledged TGDs extended by allowing constants (data values). The question is whether the language allows us to effectively reason about its constraints. We need a constraint language that is powerful enough to express dependencies commonly found in schema matching and data cleaning, while at the same time well-behaved enough so that its associated decision problems are tractable or, at the very least, decidable [16]. For full-fledged TGDs, it was known 30 years ago that the implication problem is **undecidable** even in the absence of data values [18].

As found in most database textbooks, standard INDs have several desired properties. (a) INDs are always satisfiable. (b) For INDs, the implication problem is decidable (PSPACE-complete). (c) Better still, INDs are finitely axiomatizable, i.e., there exists a finite set of axioms that is sound and complete for implication of CINDs. The question is: do CINDs still have these properties (which extend INDs by incorporating data values)?

As observed in [18], if TGDs were extended by including data values, their analysis would become more intricate. Although we are not aware of previous work on the complexity of the static analyses of TGDs with constants, the study of CFDs [1] tells us that data values in the pattern tableaux of dependencies would make our lives much harder: (1) as opposed to standard FDs for which the implication problem is in linear-time, the implication analysis for CFDs is coNP-complete, and (2) while a set of FDs is always satisfiable, the satisfiability problem for CFDs is NP-complete. Moreover, for the satisfiability and implication problems, we have to consider the impact of finite-domain attributes, as a finite domain imposes an additional constraint on how a relation can be populated such that the relation observes the constant patterns and satisfies the dependencies. The interaction between data values in finite domains and constants in pattern tableaux makes the inference complicated.

In this section we study the satisfiability and implication problems for CINDs. We show that despite that CINDs contain constants and are more expressive than INDs, they retain most of the desired properties of INDs. That is, CINDs strike a balance between the expressive power and complexity. Below we first settle the satisfiability problem for CINDs in positive, in Section 3.1. We then study the implication analysis of CINDs in Section 3.2. More specifically, we develop a sound and complete inference system for CINDs in Section 3.2.1, and then establish the complexity of the implication problem in Section 3.2.2. Moreover, we also revisit the implication problem for standard INDs in the presence of finite-domain attributes, an issue that has not been studied.
3.1. The satisfiability analysis

One cannot expect to derive sensible schema matches or to effectively clean data from a set of constraints if the constraints are not satisfiable themselves. Thus before any run-time computation is conducted, we have to make sure that the constraints are satisfiable, or in other words, make sense themselves.

The satisfiability problem for a constraint language \( L \) is to determine, given a finite set \( \Sigma \) of constraints in \( L \) defined on a database schema \( \mathcal{R} \), whether there exists a nonempty instance \( D \) of \( \mathcal{R} \) such that \( D \models \Sigma \).

Traditional FDs and INDs do not contain constants, and any set of FDs or INDs is always satisfiable [19]. However, adding data values to constraints may make their satisfiability analysis much harder, e.g., CFDs [1]. To illustrate this, we first review CFDs below, which will also be needed when we study their interaction with CINDs in Section 4.

**CFDs** A conditional functional dependency (CFD) \( \phi \) on a relation \( R \) is a pair \((R : X \rightarrow Y, T_p)\), where (1) \( X \) and \( Y \) are subsets of \( \text{attr}(R) \); (2) \( R : X \rightarrow Y \) is a standard FD, referred to as the FD embedded in \( \phi \); and (3) \( T_p \) is a tableau with all attributes in \( X \) and \( Y \), referred to as the pattern tableau of \( \phi \), where for each \( A \) in \( X \) or \( Y \) and each tuple \( t_p \in T_p \), \( t_p[A] \) is either a constant \( a \in \text{dom}(A) \), or an unnamed variable \( '_' \), as defined for CINDs given earlier.

To formalize the semantics of CFDs, we define an operator \( \times \) on constants and the symbol \( '_' \): \( \eta_1 \times \eta_2 \) if either \( \eta_1 = \eta_2 \), or one of \( \eta_1 \) and \( \eta_2 \) is \( '_' \). The operator \( \times \) naturally extends to tuples. For example, \((\text{Mayfield, EDI}) \times (\_, \text{EDI})\) but \((\text{Mayfield, EDI}) \neq (\_, \text{EDI})\).

An instance \( I \) of \( R \) satisfies the CFD \( \phi \), denoted by \( I \models \phi \), if and only if for each pair of tuples \( t_1, t_2 \) in the relation \( I \), and for each tuple \( t_p \) in the pattern tableau \( T_p \), if \( t_1[X] = t_2[X] \times t_p[X] \), then \( t_1[Y] = t_2[Y] \times t_p[Y] \). That is, if \( t_1[X] \) and \( t_2[X] \) are equal to each other and match the pattern \( t_p[X] \), then \( t_1[Y] \) and \( t_2[Y] \) must also be equal to each other and match the pattern \( t_p[Y] \). Note that standard FDs are a special case of CFDs in which the pattern tableau contains a single tuple that consists of \( '_' \) only.

Along the same lines as CINDs in normal form, we say that a CFD \( \phi = (R : X \rightarrow Y, T_p) \) is in the normal form if \( T_p \) consists of a single tuple \( t_p \) and \( Y \) contains a single attribute \( A \), and we write \( \phi \) as \((R : X \rightarrow A, t_p)\). We can always rewrite a CFD into an equivalent set of CFDs in the normal form. In the sequel, we only consider CFDs in the normal form.

A set of CFDs, which extend FDs by adding constant patterns, may be unsatisfiable, as illustrated by the following example [1].

**Example 3.1.** Consider a relation schema \( (A, B) \), and the CFDs below defined on \( R \), refining standard FDs \( A \rightarrow B \) and \( B \rightarrow A \):

\[
\begin{align*}
\phi_1 &: (A \rightarrow B, (\text{true} \parallel b_1)), \\
\phi_2 &: (A \rightarrow B, (\text{false} \parallel b_2)), \\
\phi_3 &: (B \rightarrow A, (b_1 \parallel \text{false})), \\
\phi_4 &: (B \rightarrow A, (b_2 \parallel \text{true})),
\end{align*}
\]

where \( \text{dom}(A) \) is bool, and \( b_1, b_2 \) are two distinct constants in \( \text{dom}(B) \). The CFD \( \phi_1 \) (resp. \( \phi_2 \)) asserts that for any \( R \) tuple \( t \), if \( t[A] = \text{true} \) (resp. \( \text{false} \)), then \( t[B] \) must be \( b_1 \) (resp. \( b_2 \)). On the other hand, \( \phi_3 \) (resp. \( \phi_4 \)) requires that if \( t[B] = b_1 \) (resp. \( b_2 \)), then \( t[A] \) must be \( \text{false} \) (resp. \( \text{true} \)). Observe that there exists no nonempty instance of \( R \) that satisfies all these CFDs. Indeed, for any \( R \) tuple \( t \), no matter what Boolean value \( t[A] \) is, these CFDs together force \( t[A] \) to take the other value from the finite domain bool.

Note that if \( \text{dom}(A) \) and \( \text{dom}(B) \) were infinite, one could find a tuple \( t \) such that \( t[A] \) is neither true nor false, and \( t[B] \) is not \( b_1 \) or \( b_2 \); then the \( R \) instance \( I \) satisfies these CFDs. This tells us that attributes with a finite domain complicate the satisfiability analysis. □

It is known [1] that the satisfiability problem for CFDs is NP-complete. As opposed to CFDs, the satisfiability analysis of CINDs is as trivial as their standard INDs counterpart, despite the increased expressive power of CINDs. Intuitively, this is because CFDs are a subset of EGDs [13] (universally quantified formulas), while CINDs are TGDs in which the existential quantification allows us to add tuples that match the pattern and satisfy the embedded IND of a CIND. Moreover, the requirement that \( X \) and \( X_p \) (resp. \( Y \) and \( Y_p \)) are disjoint in a CIND \( (R_p[X, X_p] \subseteq R_p[Y, Y_p], t_p) \) simplify the analysis.

**Theorem 1.** Any set of CINDs is satisfiable.

**Proof.** We show that given a set \( \Sigma \) of CINDs over a database schema \( \mathcal{R} = (R_1, \ldots, R_n) \), we can always construct a nonempty finite instance \( D \) of \( \mathcal{R} \) such that \( D \models \Sigma \).

We build such an instance \( D \) as follows.\( \square \)

(1) We start with the construction of the active domains. For each attribute \( A \), we first collect in \( \text{dom}(A) \) all those constants that appear in some pattern of \( A \) in the CINDs. We then propagate these constants from \( \text{dom}(A) \) to \( \text{dom}(B) \) for each attribute \( B \) that is connected to \( A \) via a CIND of \( \Sigma \).

More specifically, for each attribute \( A \) in some relation of \( \mathcal{R} \), we start with \( \text{dom}(A) = \emptyset \). For every CIND \((R_p[X, X_p] \subseteq R_p[Y, Y_p], t_p)\) in \( \Sigma \), if \( A \in X_p \) or \( A \in Y_p \), we include in \( \text{dom}(A) \) the constant \( t_p[A] \), i.e., we let \( \text{dom}(A) = \text{dom}(A) \cup \{t_p[A]\} \).
IR1: (reflexivity) If X is a list of distinct attributes of R, then \( R[X; nil] \subseteq R[X; nil], t_p \), where \( t_p = \mathbb{N} \).

IR2: (projection and permutation) If \( R[A_1, \ldots, A_m; X_p] \subseteq R[B_1, \ldots, B_m; Y_p], t_p \), then \( R[A_j, \ldots, A_n; X_p] \subseteq R[B_j, \ldots, B_n; Y_p], t_p \), where \( (1 \leq j, \ldots, n) \) is a list of distinct integers in \( \{1, \ldots, m\} \), or \( A_1, \ldots, A_m = B_1, \ldots, B_m = \text{nil}; \) \( X_p \) and \( Y_p \) are permutations of \( X_p \) and \( Y_p \), respectively.

IR3: (transitivity) If \( R[A_1; X_p] \subseteq R[Y_p], t_p \), and \( R[A_2; Y_p] \subseteq R[Z_p], t_p \), then \( R[A_1; Z_p], t_p \).

IR4: (instanation) If \( R[X; X_p] \subseteq R[Y; Y_p], t_p \), then \( R[X[A_1]; X_p] \subseteq R[Y[A_1]; Y_p], t_p \), where \( X = (A_1, \ldots, A_n) \), \( Y = (B_1, \ldots, B_m) \), \( t_p[A_1] \in \text{adom}(A) \), \( t_p[A] \in \text{adom}(A) \), and \( t_p[A_1] = t_p[A] \).

IR5: (LHS expansion) If \( R[X; X_p] \subseteq R[Y; Y_p], t_p \), then \( R[A; X_p] \subseteq R[B; Y_p], t_p \), where \( A \in \text{attr}(A) \) \( \cup \text{attr}(X) \) \( \cup \text{attr}(Y) \), \( X_p[A] \in \text{adom}(A) \), and \( t_p[A] = t_p[A] \).

IR6: ( RHS reduction) If \( R[X; X_p] \subseteq R[Y; Y_p], t_p \), then \( R[A; X_p] \subseteq R[B; Y_p], t_p \), where \( t_p[A] = t_p[A] \).

IR7: (finite-domain attribute elimination) If \( R[A_1; X_p] \subseteq R[Y; Y_p], t_p \), then \( R[A_2; X_p] \subseteq R[Y; Y_p], t_p \), where \( t_p[A_1] = t_p[A_2] \).

IR8: (finite-domain attribute abstraction) If \( R[A_1; X_p] \subseteq R[B; Y_p], t_p \), then \( R[A_2; X_p] \subseteq R[B; Y_p], t_p \), where \( t_p[A_1] = t_p[A_2] \).

If do\( \text{dom}(A) = c \) is still empty after all the CINDs in \( \Sigma \) have been inspected, we let adom\( (A) = c \) for an arbitrary constant \( c \in \text{dom}(A) \).

We propagate these initial values as follows. For each CIND \( R[A_1, A_2, \ldots, A_m; X_p] \subseteq R[B_1, B_2, \ldots, B_m; Y_p], t_p \) in \( \Sigma \), we expand adom\( (B_1) \) by letting adom\( (B_1) = \text{dom}(B_1) \cup \text{adom}(A) \) for each \( i \in \{1, \ldots, m\} \). The propagation process is recursively applied to all attributes, and proceeds until no further changes can be made to adom\( (A) \) of any attribute \( A \). Since it starts with a finite set of values for each adom\( (A) \), it is easy to verify that the process always terminates.

(2) We construct \( D \) as follows. For each \( R[A_1, \ldots, A_k] \in \mathcal{R} \), we define \( I = \text{adom}(A_1) \times \cdots \times \text{adom}(A_k) \), where \( \times \) is the Cartesian product operation. And we define \( D = \{t_1, \ldots, t_k\} \).

It is easy to verify that the database instance \( D \) is nonempty, finite, and that \( D \models \Sigma \).

3.2. The implication analysis

As remarked earlier, the implication analysis allows us to remove redundancies from data quality rules to improve performance, and it is also critical to deriving schema mappings from schema matchings [10,11]. Recall that (1) a schema mapping is data transformation from instances of a source schema to instances of a target schema while preserving the appropriate information of the source instances; and (2) a schema matching is a pairing of attributes (or groups of attributes) of a source schema and attributes of a target schema such that the pairs are likely to be semantically related. In practice, finding a schema matching is often an early step in building a schema mapping, which is a common task in a variety of data exchange and integration scenarios [11]. It is known that contextual schema matching needs to make use of (contextual) foreign keys, a primitive and special case of CINDs.

The implication problem for a constraint language \( L \) is to determine, given a finite set \( \Sigma \) of constraints in \( L \) and another \( \psi \) of \( L \), whether \( \psi \) can be derived from \( \Sigma \) using rules of \( \mathcal{I} \). As will be seen shortly, these rules are both sound, i.e., if \( \Sigma \models \psi \) then \( \Sigma \models \psi \). Complete, i.e., if \( \Sigma \models \psi \) then \( \Sigma \models \psi \). We refer to the inference system for CINDs as \( \mathcal{I} \) as the inference system for CINDs.

Example 3.2. Consider the CINDs given in Fig. 2. Let \( \Sigma = \{\psi_1, \psi_2, \psi_3\} \). One can verify that \( \Sigma \models \psi_3 \) and \( \Sigma \models \psi_4 \). That is, CINDs \( \psi_3 \) and \( \psi_4 \) are redundant, and we only need to focus on CINDs in \( \Sigma \), and ignore \( \psi_3 \) and \( \psi_4 \).

3.2.1. An inference system

As remarked earlier, for standard INDs the implication problem is not only decidable but also finitely axiomatizable. We show that CINDs are also finitely axiomatizable, by providing an inference system for CINDs, denoted by \( \mathcal{I} \). Given a finite set \( \Sigma \) of CINDs and another CIND \( \psi \), we denote by \( \Sigma \vdash \psi \) if \( \psi \) is provable from \( \Sigma \) using rules of \( \mathcal{I} \). As will be seen shortly, these rules are both sound, i.e., if \( \Sigma \vdash \psi \) then \( \Sigma \models \psi \). Complete, i.e., if \( \Sigma \models \psi \) then \( \Sigma \vdash \psi \).

Recall that for standard INDs, the inference system consists of three rules: reflexivity, projection–permutation and transitivity [13,20]. To cope with the richer semantics of CINDs, the inference system \( \mathcal{I} \) is more complicated than the inference system for INDs.

The inference system \( \mathcal{I} \) is shown in Fig. 3. We briefly illustrate the inference rules (axioms) in \( \mathcal{I} \) as follows. Intuitively, rules IR1–IR3 correspond to the inference rules for INDs. IR1 is the reflexivity rule. IR2 shows that the patterns, i.e., \( X_p \) and \( Y_p \), can also be permuted, in addition to permutation and projection of the embedded IND. IR3 extends the transitivity rule. It requires only the RHS of the first CIND to match the LHS of the second CIND, but also their pattern tuples to be matched, i.e., \( t_p [Y_p] = t_p [Y_p] \).

Observe that rules IR1–IR3 do not consider the interaction between the pattern tuples and the embedded INDs, or the complication introduced by finite-domain attributes. Hence, in contrast to classical INDs, these three rules are not enough to characterize the implication analysis of CINDs. In light of these, we need rules IR4–IR8, which do not find a counterpart in the inference system for INDs.
IR4 allows us to instantiate attributes in X and their corresponding attributes in Y. Given a CIND \((R_a[X; X_p] \subseteq R_b[Y; Y_p], tp)\), we can take an attribute \(A_j\) from X and the corresponding \(B_j\) in Y, assign a value in \(\text{dom}(A_j)\) to them, and move the attribute \(A_j\) (resp. \(B_j\)) to the pattern tuple \(X_p\) (resp. \(Y_p\)) of the CIND.

IR5 enhances the LHS pattern of a CIND by adding an attribute to the pattern \(X_p\). Consider a CIND \((R_a[X; X_p] \subseteq R_b[Y; Y_p], tp)\). For any attribute \(A \in \text{attr}(R_a)\) that is in neither X nor \(X_p\), we can add \(A\) to \(X_p\) with an arbitrary value from \(\text{dom}(A)\). Intuitively, if \(\psi\) holds for all data values in \(\text{dom}(A)\), then it also holds for a specific value in \(\text{dom}(A)\).

IR6 weakens the RHS pattern of a CIND by removing an attribute from \(Y_p\). Consider a CIND \((R_a[X; X_p] \subseteq R_b[Y; Y_p], B, tp)\) holds, then for each tuple \(t_b\) in \(R_b\) that satisfies the pattern \(tp\), there exists a matching tuple \(t_p\) in \(R_a\) that satisfies the pattern \(tp, B\). If an attribute is removed from \(Y_p\), the CIND certainly still holds since the same tuple \(t_b\) satisfies \(tp\).

Finally, IR7 and IR8 are only needed in the presence of attributes with a finite domain. IR7 says that if (1) they differ only in the value of \(t_p[A]\) of a finite-domain attribute \(A\), and the union of all those \(t_p[A]\) values covers the entire domain \(\text{dom}(A)\), then we can replace the set of CINDs by a single CIND in which attribute \(A\) is removed from \(X_p\). That is, the presence of \(A\) in the LHS pattern has no effect at all, and hence, we can just exclude it from the CINDs.

IR8 is the “inverse” of IR4. If IR4 is applied to a CIND \(\psi\) to instantiate attributes \(A\) and \(B\) in the pattern tuple, when \(t_p[A]\) ranges over all the values of \(\text{dom}(A)\), then IR8 can take all those CINDs and restore \(\psi\). In short, IR8 merges a set of CINDs if (1) they differ only in the value of \(t_p[A]\), (2) \(t_p[A]\) ranges over all the values in the domain \(\text{dom}(A)\), and (3) there exists an attribute \(B\) in the RHS of each CIND such that \(t_p[A] = t_p[B]\).

**Example 3.3.** Recall \(\psi_1\) and \(\psi_3\) from **Example 2.1.** From \(\psi_1\), we can derive \(\psi_3\) by using rule IR5, i.e., \(\{\psi_1\} \vdash_{I} \psi_3\).

The next example demonstrates that finite-domain attributes make the inference process more intricate. Consider a database that consists of three relations \(R_1(ABC), R_2(EGF)\), and \(R_3(GBK)\) such that \(\text{dom}(B) = \{b_1, b_2\}\) is finite and all the other attributes have an infinite domain, e.g., string. Let \(\{\psi_1, \psi_2, \psi_3, \psi'\}\) be a set of CINDs, where

\[
\psi_1 = (R_1[A; B] \subseteq R_2[E; G], (b_1 \parallel g)),
\]

\[
\psi_2 = (R_1[A; B] \subseteq R_3[H; G], (b_2 \parallel g)),
\]

\[
\psi_3 = (R_2[E; nil] \subseteq R_3[H; G], (nil \parallel g)),
\]

and

\[
\psi' = (R_1[A; nil] \subseteq R_3[H; nil], \emptyset).
\]

Then \(\psi'\) can be derived from \(\{\psi_1, \psi_2, \psi_3\}\):

\(1\) \(R_2[E; G] \subseteq R_3[H; G], (g \parallel g)\), \(\psi_3', \text{ IR5}\)

\(2\) \(R_1[A; B] \subseteq R_3[H; G], (b_1 \parallel g)\), \(\psi_1', \text{ IR3}\)

\(3\) \(R_1[A; nil] \subseteq R_3[H; G], (nil \parallel g)\), \(\psi_2', \text{ IR7}\)

\(4\) \(\psi' = (R_1[A; nil] \subseteq R_3[H; nil], \emptyset)\), \(\psi_3', \text{ IR6}\)

That is, \(\{\psi_1, \psi_2, \psi_3\} \vdash_{I} \psi'\). □

The soundness and completeness of \(I\) We next show that \(I\) is sound and complete for the implication analysis of CINDs. That is, for any set \(\Sigma\) of CINDs and another CIND \(\psi\) defined over the same schema \(R\), \(\Sigma \models \psi\) iff \(\Sigma \vdash_{I} \psi\).

As we have seen from **Example 3.1.** the presence of finite-domain attributes complicates the implication analysis of CFDs. This is also the case for CINDs: when some attributes in \(\Sigma\) or \(\psi\) have a finite domain, the implication analysis is more intricate, as indicated by **Example 3.3** and the rules IR7 and IR8 in \(I\). In light of this, we distinguish two settings: (1) in the absence of finite-domain attributes, i.e., when none of the attributes in \(\Sigma\) or \(\psi\) has a finite domain, and (2) the general setting, when some attributes in \(\Sigma\) or \(\psi\) may have a finite domain. We next show that a set of the rules of \(I\) is sound and complete in both settings.

In the absence of finite-domain attributes This is the setting under which the inference system for standard INDs was developed [20] (see Section 6 for a detailed discussion). In this context, our main result about CIND inference is that the inference rules IR1–IR6 of \(I\) make a sound and complete inference system for CINDs. For a set \(\Sigma\) of CINDs and another CIND \(\psi\), we use \(\Sigma \vdash_{I} \psi\) to denote that \(\psi\) can be proved from \(\Sigma\) using rules IR1–IR6 of \(I\).

**Theorem 2.** In the absence of finite-domain attributes, the inference rules IR1–IR6 of \(I\) are sound and complete for the implication analysis of CINDs.

**Proof sketch.** For the soundness, we show that given a set \(\Sigma \cup \{\psi\}\) of CINDs, if \(\Sigma \vdash_{I} \psi\) by using IR1–IR6, then \(\Sigma \models \psi\). This can be readily verified by induction on the length of proofs with IR1–IR6, by showing that the application of each of IR1–IR6 is sound, by the definition of CINDs.
For the completeness, we show that given a set of CINDs $\Sigma \cup \{\psi\}$ defined over a relational schema $R$, if $\Sigma \vdash \psi$ then $\Sigma \vdash_{\mathcal{I}(1-6)} \psi$, i.e., $\psi$ is provable from $\Sigma$ by using rules IR1–IR6. The proof consists of three parts. (1) We first develop a chase procedure to characterize $\Sigma \vdash \psi$. The chase process starts with a single-tuple instance of $R$, and repeatedly adds tuples (one at a time) to the instance by applying CINDs in $\Sigma$ until no more CINDs can be applied. (2) We then show that the chase process always terminates, and moreover, that if $\Sigma \vdash \psi$, then the resulting instance satisfies $\psi$. (3) Based on (1) and (2), we finally show that if $\Sigma \vdash \psi$ then $\Sigma \vdash_{\mathcal{I}(1-6)} \psi$, by showing that the chase process can be simulated by the applications of rules IR1–IR6. The detailed proof can be found in Appendix A. □

Remark. The proof of Theorem 2 is inspired by the proof of their IND counterpart given in [20]. Nonetheless, our proof has to deal with six rules with constant patterns, whereas the inference system for INDs consists of three rules and does not need to consider constant patterns [20].

In the general setting As indicated by Example 3.3, the presence of finite-domain attributes complicates the inference process for the implication analysis of CINDs, for which rules IR7 and IR8 are used to deal with finite-domain attributes. Indeed, while one can verify that the inference system of [20] is also complete for standard INDs in the presence of finite-domain attributes, whereas the rules IR1–IR6 are no longer complete for CIND implication here, as shown below.

Example 3.4. Consider a database schema $R$ with relations $R_1(ABCD)$, $R_2(EGF)$, and $R_3(IJKLM)$ such that $\text{dom}(A) = \{a, b, c\}$, $\text{dom}(D) = \{d, e\}$, $\text{dom}(G) = \{d, e, f\}$, $\text{dom}(I) = \{g, h\}$ and all the other attributes have an infinite domain, e.g., string. Consider a set $\Sigma = \{\psi_1, \psi_2, \psi_3, \psi_4, \psi_5\}$ of CINDs and another CIND $\psi$, all defined on $R$, where

\[
\begin{align*}
\psi_1 &= (R_1[BC; AD] \subseteq R_2[EF; GH], (a, d \parallel d, f)), \\
\psi_2 &= (R_1[BC; AD] \subseteq R_3[IJK], (b, d \parallel k)), \\
\psi_3 &= (R_3[IJK] \subseteq R_2[EG; GH], (k \parallel d, g)), \\
\psi_4 &= (R_1[BC; A] \subseteq R_2[EF; GH], (c \parallel d, h)), \\
\psi_5 &= (R_1[BC; D] \subseteq R_2[EG; GH], (e \parallel e, k)), \quad \text{and} \\
\psi &= (R_1[BCD; nil] \subseteq R_2[EGF; nil], \emptyset).
\end{align*}
\]

Although $\Sigma \vdash \psi$, one can verify that $\Sigma \vdash_{\mathcal{I}(1-6)} \psi$, by showing that the existence of a tuple $t_1$ of schema $R_1$ does not guarantee the existence of a tuple $t_2$ of schema $R_2$ such that $t_2[EGF] = t_1[BCD]$, using the chase process for rules IR1–IR6 developed in the detailed proof of Theorem 2. In other words, IR1–IR6 are no longer complete for proving $\psi$ from $\Sigma$. In contrast, $\psi$ can be proved from $\Sigma$ by using IR1–IR8, as follows:

1. $(R_1[BC; AD] \subseteq R_2[EF; GH], (b, d \parallel d, g))$, $\psi_2$, $\psi_3$, IR3
2. $(R_1[BC; AD] \subseteq R_2[EF; GH], (c, d \parallel d, h))$, $\psi_4$, IR5
3. $(R_1[BC; AD] \subseteq R_2[EF; G], (a, d \parallel d))$, $\psi_1$, IR6
4. $(R_1[BC; AD] \subseteq R_2[EF; G], (b, d \parallel d))$, (1), IR6
5. $(R_1[BC; AD] \subseteq R_2[EF; G], (c, d \parallel d))$, (2), IR6
6. $(R_1[BC; D] \subseteq R_2[EF; G], (d \parallel d))$, (3), (4), (5), IR7
7. $(R_1[BC; D] \subseteq R_2[EF; G], (e \parallel e))$, $\psi_5$, IR6
8. $\psi = (R_1[BCD; nil] \subseteq R_2[EGF; nil], \emptyset)$, (6), (7), IR8

That is, $\Sigma \vdash_{\mathcal{I}} \psi$. □

In the general setting, we show that IR1–IR8 are sound and complete.

Theorem 3. The inference system $\mathcal{I}$ is sound and complete for the implication of CINDs in the general case, where finite-domain attributes may be present.

Proof sketch. The soundness of $\mathcal{I}$ can be verified by induction on the length of $\mathcal{I}$-proofs.

For the completeness of $\mathcal{I}$, consider a set $\Sigma \cup \{\psi\}$ of CINDs defined on a database schema $R$. We show that if $\Sigma \vdash \psi$, then $\Sigma \vdash_{\mathcal{I}} \psi$. The proof consists of five parts. (1) We first show that it suffices to consider a special form of CINDs. (2) We then develop a chase procedure for the special form of CINDs, which extends the one given in the proof of Theorem 2 to further deal with finite-domain attributes, and (3) we show that the chase process always terminates. (4) In addition, we establish a certain property of the chase procedure, by characterizing conditions under which $\Sigma \cup \{\psi\}$ holds and showing that the chase ensures the satisfaction of the condition. Finally, (5) we show that if $\Sigma \vdash \psi$ then $\Sigma \vdash_{\mathcal{I}} \psi$ by using rules IR1–IR8 and the property. We refer the interested reader to Appendix A for a detailed proof. □
Unrestricted implication We have so far only considered finite implication, when finite databases are considered, i.e., for database instances in which each relation is a finite set of tuples. A CIND $\psi$ is finitely implied by $\Sigma$ if for every finite database $D$, if $D \models \Sigma$, then $D \models \psi$. For theoretical interest one may also want to consider unrestricted implication, where $\psi$ is implied by $\Sigma$ if for every database $D$, either finite or infinite, if $D \models \Sigma$, then $D \vdash \psi$. To distinguish these, we denote finite implication and unrestricted implication by $\Sigma \models_{\text{fin}} \psi$ and $\Sigma \models_{\text{unr}} \psi$, respectively.

The result below tells us that, however, for CINDs these notions are equivalent. As a result, we can focus on finite implication for CINDs, and use $\Sigma \models \psi$ to denote both $\Sigma \models_{\text{fin}} \psi$ and $\Sigma \models_{\text{unr}} \psi$.

Proposition 4. Finite implication and unrestricted implication coincide for CINDs.

Proof. By definition, unrestricted implication entails finite implication. That is, if $\Sigma \models_{\text{unr}} \psi$, then $\Sigma \models_{\text{fin}} \psi$. Conversely, it is easy to verify that the inference system $I$ is sound for unrestricted implication, i.e., if $\Sigma \vdash_{I} \psi$, then $\Sigma \models_{\text{unr}} \psi$. Moreover, by Theorem 3, if $\Sigma \models_{\text{fin}} \psi$, then $\Sigma \vdash_{I} \psi$. From these it follows that finite implication entails unrestricted implication, i.e., if $\Sigma \models_{\text{fin}} \psi$, then $\Sigma \models_{\text{unr}} \psi$. \qed

3.2.2. The complexity of the implication analysis

We next establish the computational complexity bounds for the implication analysis of CINDs. We investigate the problem again in two settings, namely, in the absence of finite-domain attributes and in the general setting when finite-domain attributes may be present.

In the absence of finite-domain attributes It is known that for standard INDs in this setting, the implication problem is PSPACE-complete [20]. Below we show that the implication problem for CINDs retains the same complexity as their standard counterpart for INDs in this setting.

Theorem 5. The implication problem for CINDs is PSPACE-complete in the absence of attributes with finite domains.

Proof sketch. It is known that the implication problem for INDs is PSPACE-complete in the absence of finite-domain attributes [20]. Since CINDs subsume INDs, the implication problem for CINDs is also PSPACE-hard.

To show that the implication problem for CINDs is in PSPACE in the absence of finite-domain attributes, it suffices to give a linear space non-deterministic algorithm for deciding whether $\Sigma \models \psi$, along the same lines as its counterpart for INDs (see [13,20]). For if it holds, then by Savitch’s theorem [21], there is a deterministic quadratic-space algorithm for checking whether $\Sigma \models \psi$, and hence, the implication problem is in PSPACE. We give such an algorithm based on the chase procedure developed in the proof of Theorem 2, in which only a single tuple is kept at all time, and old tuples are replaced by new tuples derived by no-deterministically picking CINDs in $\Sigma$ (see Appendix A for the details of the algorithm). \qed

In the general setting When finite-domain attributes are present, the implication analysis of CINDs becomes more involved. Nevertheless, the increased complexity is not incurred only by the presence of finite-domain attributes. Indeed, a close examination of the proofs of [13,20] reveals that while the PSPACE-completeness for the implication analysis of standard INDs was established for relations with infinite domains only, the proofs remain intact when finite-domain attributes are present. That is, the following result holds.

Proposition 6. In the presence of finite-domain attributes, the implication problem for INDs remains PSPACE-complete.

Theorem 5 and Proposition 6 together tell us that neither finite-domain attributes nor constant patterns alone complicate the implication problem. However, below we show that when they are taken together, the implication analysis becomes more intriguing. That is, their interaction makes the implication problem for CINDs harder.

Theorem 7. In the general setting, the implication problem for CINDs is EXPTIME-complete.

Proof sketch. To show that the problem is in EXPTIME, we develop an algorithm that, given a set $\Sigma \cup \{\psi\}$ of CINDs defined on instances of a relational schema $\mathcal{R}$, returns ‘yes’ if and only if $\Sigma \models \psi$. The algorithm is in $O(2^{kn})$ time, where $k$ is a constant and $n$ is the size of $\mathcal{R}$, $\Sigma$ and $\psi$. For the lower bound, we show that the problem is EXPTIME-hard by reduction from the two-player game of corridor tiling problem (TPG-CT), which is EXPTIME-complete [22,23]. We refer the interested reader to Appendix A for a detailed proof. \qed

4. Interaction between CINDs and CFDs

We have seen that CINDs do not make their satisfiability and implication problems much harder than their traditional counterpart INDs. At the very least, these problems remain decidable for CINDs. In contrast, we show in this section that
when CINDs and CFDs are taken together, the static analyses become far more intriguing. As remarked earlier, in schema matching and data cleaning it is often necessary to use both CINDs and CFDs.

Recall the definition of CFDs given in [1] and presented in Section 3.1. For CFDs the following have been established in [1], (a) The satisfiability problem for CFDs is NP-complete. (b) The implication problem of CFDs is finitely axiomatizable. (c) The implication problem for CFDs is coNP-complete. (d) The satisfiability and implication problems are in $O(n^2)$ time, where $n$ is the size of the given CFDs, when the CFDs do not involve attributes with a finite domain.

While CFDs alone already complicate the static analyses, we next show that CFDs and CINDs together make our lives much harder.

**Implication analysis.** It is not surprising that the implication problem for CINDs and CFDs is undecidable and is not finitely axiomatizable, since the problem is already undecidable for standard INDs and FDs (see, e.g., [13]), and CINDs and CFDs subsume INDs and FDs, respectively. The result remains intact even in the absence of attributes with a finite domain, since the undecidability proof of the implication problem for FDs and INDs use attributes with infinite domains only (see [13]).

**Proposition 8.** The implication problem for CINDs and CFDs is undecidable, and is not finitely axiomatizable, even in the absence of finite-domain attributes.

**Satisfiability analysis.** Given a set of CFDs and a set of CINDs that are separately satisfiable, when they are put together, there may be conflicts among them, which make the set of CFDs and CINDs unsatisfiable, as illustrated by the following example.

**Example 4.1.** Consider a relation $R$ with $\text{attr}(R) = \{A, B\}$, on which we define a CFD $\phi = (R : A \rightarrow B, (\_ \| a))$ and a CIND $\psi = (R[\text{nil}; B] \subseteq R[\text{nil}; B], (\text{nil} || b))$, where $a$ and $b$ are distinct constants. Obviously, there exists a nonempty instance of $R$ that satisfies $\phi$ and there is an instance satisfying $\psi$. However, there exists no nonempty instance of $R$ that satisfies both $\phi$ and $\psi$. To see this, assume that such an instance $D$ exists. Then $\psi$ tells us that as long as $D$ is nonempty, there is a tuple $t$ in $D$ such that $t[B] = b$. In contrast, $\phi$ requires that $t[B] = a$, violating $\psi$. □

While the undecidability of the implication problem for CINDs and CFDs is expected, the following result is a little surprising. The undecidability can be verified by reduction from the implication problem for standard FDs and INDs. The undecidability remains intact in the absence of attributes with a finite domain.

**Theorem 9.** The satisfiability problem for CFDs and CINDs is undecidable, even in the absence of finite-domain attributes.

**Proof.** We show that the satisfiability problem for CINDs and CFDs is undecidable by reduction from the implication problem for standard FDs and INDs, which is undecidable even in the absence of attributes with finite domains (see, e.g., [13]).

More specifically, given an instance of the implication problem, namely, a set $\Sigma$ of FDs and INDs and a single FD $\varphi = (R_g : X \rightarrow A)$ over a database $\mathcal{R}(R_1, \ldots, R_g, \ldots, R_n)$, we define another set $\Sigma'$ of CINDs and CFDs over $\mathcal{R}$ such that $\Sigma'$ is satisfiable if and only if $\Sigma \not\models \varphi$.

Let $\Sigma'$ contain every $\phi \in \Sigma$ as well as the following CINDs:

- $(R_i[\text{nil}; \text{nil}] \subseteq R_g[\text{nil}; A], (\text{nil} || c_i))$ for $i = 1, \ldots, n$; that is, if any $R_i$ is nonempty, then there exists an $R_g$ tuple $t$ such that $t[A] = c_i$;
- $(R_g[X; A] \subseteq R_g[X; A], (c_1 || c_2))$ with $c_1 \neq c_2$ and $c_1, c_2 \in \text{dom}(A)$; that is, if there exists an $R_g$ tuple $t_1$ with $t_1[A] = c_1$, then there exists an $R_g$ tuple $t_2$ with $t_2[A] = c_2$ whereas $t_1[X] = t_2[X]$.

We next show that $\Sigma'$ is satisfiable if and only if $\Sigma \not\models \varphi$.

Assume first that $\Sigma'$ is satisfiable, then obviously there exists a nonempty instance $D$ of $\mathcal{R}$ such that $D \models \Sigma'$. From this it follows that $D \models \Sigma$ and moreover, that $D \not\models \varphi$, since two tuples $t_1, t_2$ described above must appear in $D$.

Conversely, assume that $\Sigma' \not\models \varphi$. Then there exists a database $D$ such that $D \models \Sigma$ but $D \not\models \varphi$. Since $D \not\models \varphi$, there must exist two $R_g$ tuples $t_1, t_2$ such that $t_1[X] = t_2[X]$, $t_1[A] = a$, $t_2[A] = d$, but $a \neq d$. Assume w.l.o.g. that $c_1, c_2$ are two constants that do not appear in $D$, which always exist since $\text{dom}(A)$ is infinite. Define a mapping $\rho$ such that $\rho(a) = c_1$, $\rho(d') = c_2$ and $\rho(d) = d$ for all other constants $d$ in $D$. Denote the database obtained this way as $D'$. Then it is easy to verify that $D'$ is a witness that satisfies $\Sigma'$, i.e., $\Sigma'$ is satisfiable.

Putting these together, we have shown that $\Sigma'$ is satisfiable if and only if $\Sigma \not\models \varphi$. □

5. Two special classes of CINDs

**Proposition 6** and **Theorem 7** tell us that it is rather expensive to reason about INDs or CINDs. Indeed, the implication analyses of INDs and CINDs are beyond reach in practice.
However, not all is lost: in practice one often needs only a fragment of INDs or CINDs with a lower complexity. That is, we do not have to pay the price of the complexity of full-fledged INDs or CINDs when we only need certain special cases of the dependencies.

We next investigate two special cases of INDs and CINDs, namely, acyclic and unary inclusion dependencies.

### 5.1. Acyclic CINDs

One of the well-studied fragments of INDs is identified as sets of acyclic INDs [24] (see e.g., [13]). A set $\Sigma$ of INDs over a relational schema $\mathcal{R}$ is acyclic if there exists no sequence $R_i[X_i] \subseteq S_i[Y_i]$ ($i \in [1, n]$) of INDs in $\Sigma$, where $R_{i+1} = S_i$ for each $i \in [1, n-1]$, and $R_1 = S_n$. We refer to such a set of INDs as a set of AINDs.

**Acyclic CINDs** Along the same lines as AINDs, we define acyclic CINDs.

A set $\Sigma$ of CINDs is acyclic if there exists no sequence $(R_i[X; X_{pi}] \subseteq S_i[Y; Y_{pi}], t_{pi})$ ($i \in [1, n]$) of CINDs in $\Sigma$ such that $R_{i+1} = S_i$ for each $i \in [1, n-1]$, and $R_1 = S_n$. We refer to such a set of CINDs as a set of ACINDs.

It is common to find a set of CINDs acyclic in practice. For example, the CINDs in a Web database [25] and a drug database [26] are both acyclic. Moreover, the set of CINDs given in Example 2.1 and the set of CINDs in Example 3.3 are both acyclic as well.

We next show that ACINDs indeed make our lives easier, i.e., they allow more efficient static analysis. From Theorem 1, it follows that any set of ACINDs is satisfiable. Hence we shall focus on the implication analysis of ACINDs, i.e., the problem for determining, given a set $\Sigma$ of ACINDs and a CIND $\varphi$, whether $\Sigma \models \varphi$. We study the implication problem for ACINDs in the absence of finite-domain attributes and in the general setting when finite-domain attributes may be present. Note that while $\Sigma$ is acyclic, $\Sigma \cup \{\varphi\}$ may be cyclic.

**Corollary 10.** The implication problem for ACINDs is NP-complete in the absence of attributes with a finite domain.

**Proof.** For the lower bounds, note that the implication problem for acyclic INDs is already NP-hard [14]. Since acyclic INDs are a special case of acyclic CINDs, the implication problem for ACINDs is also NP-hard.

We next show that the problem is in NP. Consider the non-deterministic algorithm given in the proof of Theorem 5 for determining whether a set $\Sigma$ of CINDs implies another CIND $\psi$ (see Appendix A). When $\Sigma$ is a set of ACINDs without finite-domain attributes, the initial tuple $t_0$ can be replaced by other tuples for at most $n$ times, where $n$ is the number of relations in the database schema $\mathcal{R}$. That is, the linear space non-deterministic algorithm runs in polynomial time. Hence, it is indeed an NP algorithm for deciding whether $\Sigma \models \psi$. Thus, the implication problem for ACIND is in NP without finite-domain attributes. ⊓⊔

**In the absence of finite-domain attributes** It has been shown that in this setting, the implication problem for AINDs is NP-complete [14]. We next show that the implication problem for ACINDs retains the same complexity as its standard counterpart, i.e., ACINDs do not complicate the implication analysis in the absence of finite-domain attributes.

**Corollary 11.** The implication problem for AINDs remains NP-complete when finite-domain attributes may be present.

In contrast, the presence of finite-domain does complicate the implication analysis of ACINDs: the implication problem for ACINDs in this setting becomes PSPACE-complete, up from NP-complete for ACINDs in the absence of finite-domain attributes.

**Theorem 12.** In the general setting, the implication problem for ACINDs is PSPACE-complete.

**Proof sketch.** We show that the implication problem for ACINDs is in PSPACE in the general setting, by giving a linear space non-deterministic algorithm for determining whether $\Sigma \not\models \psi$, i.e., the complement of $\Sigma \models \psi$. This suffices. For it holds, then (a) by the Immerman–Szelepcsényi theorem [27,28], there exists a linear space non-deterministic algorithm for deciding whether $\Sigma \models \psi$; and (b) by Savitch’s theorem [21], there is a deterministic quadratic-space algorithm for checking whether $\Sigma \models \psi$. Therefore, the problem is in PSPACE. For the lower bound, we show that the problem is PSPACE-hard by reduction from the Q3SAT problem, which is known to be PSPACE-complete (cf. [29]). A detailed proof is given in Appendix A. ⊓⊔
5.2. Unary CINDs

Another well-studied fragment of INDs is the class of unary inclusion dependencies, defined as follows [13,15]. A unary IND (UIND) is an IND of the form $R_a[A] \subseteq R_b[B]$, where $A \in \text{attr}(R_a)$ and $B \in \text{attr}(R_b)$. That is, a UIND is an IND in which exactly one attribute appears on each side.

**Unary CINDs** We next define and investigate unary CINDs along the same lines as UINDs.

A unary CIND (UCIND) is a CIND of the form $(R_a[A; X_p] \subseteq R_b[B; Y_p], t_p)$, where $A$ and $B$ are attributes in $R_a$ and $R_b$, respectively.

UCINDs are an extension of UINDs by incorporating patterns of data values. Observe that patterns $X_p$ and $Y_p$ in a UCIND may have more than one attribute. It is common to find UCINDs in practice, just like our familiar unary primary keys and their corresponding foreign keys [13]. Below we give another example.

**Example 5.1.** Consider a database consisting of three relations: student(SSN, name, dept), course(cno, title, dept), and enroll(SSN, cno, grade). The student relation collects all the student records in a university, and course consists of all the courses offered by the university. In contrast, the enroll relation aims to maintain a complete record of the courses registered by students in the cs department.

One would naturally want to design the following CINDs:

- $(\text{student}[SSN; dept] \subseteq \text{enroll}[SSN; nil], ('cs' \parallel \text{nil}))$
- $(\text{course}[cno; dept] \subseteq \text{enroll}[cno; nil], ('cs' \parallel \text{nil}))$
- $(\text{enroll}[SSN; nil] \subseteq \text{student}[SSN; nil], \emptyset)$, and
- $(\text{enroll}[cno; nil] \subseteq \text{course}[cno; nil], \emptyset)$.

All these CINDs are UCINDs.

We next investigate the static analysis of UCINDs. By Theorem 1 we do not have to worry about the satisfiability problem for UCINDs. Hence below we focus on the implication problem for UCINDs, i.e., the problem for determining, given a set $\Sigma$ of UCINDs and another UCIND $\psi$, whether $\Sigma \models \psi$. We study the problem in the absence of finite-domain attributes and in the general setting.

**Theorem 13.** The implication problem for UCINDs is in polynomial time in the absence of finite-domain attributes.

**Proof sketch.** We give a PTIME algorithm for checking whether $\Sigma \models \psi$ when $\psi$ and the CINDs in $\Sigma$ are unary. The algorithm can be found in Appendix A. □

**In the general setting** For UINDs, the presence of finite-domain attributes does not complicate the implication analysis. Indeed, a close examination of the PTIME algorithm of [15] for checking traditional UIND implication reveals that it still works in the general setting. Hence we have the following.

**Corollary 14.** The implication problem for UINDs remains in polynomial time in the general setting.

This is, however, no longer the case for UCINDs in this setting.

**Theorem 15.** The implication problem for UCINDs is coNP-complete in the general setting.

**Proof sketch.** We show that the problem is in coNP, by developing an NP algorithm that, given a set $\Sigma \cup \{\psi\}$ of UCINDs, checks whether $\Sigma \not\models \psi$. We verify that the problem is coNP-hard by reduction from the 3SAT problem to the complement of the problem (i.e., to decide whether $\Sigma \not\models \psi$). It is known that 3SAT is NP-complete (cf. [30]). We defer a detailed proof to Appendix A. □

**Acyclic UCINDs** One might be tempted to think that it would simplify the implication analysis if we further restrict UCINDs to be acyclic. Unfortunately, this is not the case. Indeed, in the lower bound proof of Theorem 15, the UCINDs used are acyclic. From this it follows that the implication problem for acyclic UCINDs remains coNP-complete.

**Corollary 16.** The implication problem for acyclic UCINDs is coNP-complete in the general setting.
6. Related work

Data dependencies have been studied for relational databases since the introduction of FDs by Codd [31] in 1972 (see, e.g., [13,32] for details). Recently, data dependencies have generated renewed interests for improving data quality [1–7,33–37] and for schema mapping [10,11,16].

The theory of INDs was established in [20], which developed a sound and complete inference system and the PSPACE-completeness for the implication analysis of INDs. Acyclic INDs were introduced in [24], and their implication problem was shown to be NP-complete in [14]. Unary INDs were studied in [15], which provided a sound and complete inference system for UIINDs and FDs, and proved the PTIME bound of the implication problem for UIINDs and FDs put together (see [13] for a survey on INDs, AINDs and UIINDs). While not explicitly stated, the proofs of these results indicate that the implication analysis was conducted in the absence of finite-domain attributes. In this paper we verify that the complexity bounds for INDs, AINDs and UIINDs remain intact in the presence of finite-domain attributes.

CINDs, ACINDs and UCINDs extend INDs, AINDs and UIINDs, respectively, by incorporating patterns of data values. For the implication problem in the absence of finite-domain attributes, the lower bounds for CINDs, AINDs and UIINDs are inherited from their traditional counterparts, but not the upper bounds. When finite-domain attributes may be present, however, none of the results of [14,15,18,20,24] holds on CINDs. Indeed, the implication problems for CINDs, ACINDs and UCINDs in the general setting have a higher complexity bound than their traditional counterparts.

INDs are a special case of TGDs, which can be expressed as first-order logic sentences of the form:

\[ \forall x_1 \ldots \forall x_n \left[ \varphi(x_1, \ldots, x_n) \rightarrow \exists z_1 \ldots \exists z_k \phi(y_1, \ldots, y_m) \right], \]

where (a) \( z_1, \ldots, z_k = \{ y_1, \ldots, y_m \} \setminus \{ x_1, \ldots, x_n \} \), (b) \( \varphi \) and \( \phi \) are conjunctions of relation atoms of the form \( R(w_1, \ldots, w_l) \) in which \( w_1, \ldots, w_l \) are variables (see, e.g., [13] for details). In contrast to CINDs, TGDs do not allow constants, and their implication problem is undecidable [18]. There have been extensions of TGDs [38] developed for constraint databases, notably constrained tuple-generating dependencies (cTGDs) of the form:

\[ \forall \bar{x} \left( R_1(\bar{x}) \land \cdots \land R_k(\bar{x}) \land \xi \rightarrow \exists \bar{y} \left( R'_1(\bar{x}, \bar{y}) \land \cdots \land R'_k(\bar{x}, \bar{y}) \land \xi'(\bar{x}, \bar{y}) \right) \right), \]

where \( R_i, R'_j \) are relation atoms, and \( \xi, \xi' \) are arbitrary constraints. While cTGDs support constants and can express CINDs, the increased expressive power comes at a price for static analysis. Indeed, the satisfiability and implication problems are both undecidable for cTGDs.

Closer to our work is the study of CFDs [1]. CFDs extend FDs with pattern tableaux, along the same lines as CINDs. It was shown in [1] that the satisfiability and implication problems for CFDs are NP-complete and coNP-complete, respectively, in the general setting, and they are in PTIME in the absence of finite-domain attributes. Extensions of CFDs have been proposed to support disjunction and negation [39], cardinality constraints and synonym rules [40], built-in predicates (\( \neq, <, \leq, \geq \)) [41], and to specify patterns in terms of value ranges [4]. However, CFDs and their extensions are defined on a single relation and are universally quantified. They cannot express CINDs, and neither CINDs nor their static analyses were studied in [1,4,39–41]. In addition, as we have seen earlier, the satisfiability and implication analysis of CINDs is far more intriguing than their CFD counterparts. An extension of CINDs was recently proposed to support built-in predicates [41], which was based on the results of this work. Discovering CINDs has been studied in [25].

Research on constraint-based data cleaning has mostly focused on two topics, both proposed by [33]: repairing is to find another database that is consistent and minimally differs from the original database (e.g., [8,9,42]); and consistent query answering is to find an answer to a given query in every repair of the original database (e.g., [33,43]). A variety of constraint formalisms have been used in data cleaning, ranging from standard FDs and INDs [8,9,33], denial constraints (full dependencies) [44], to logic programs (see [34–36] for surveys). To our knowledge, no prior work has considered CINDs for data cleaning albeit our work [12,41], and the recent work [26] that makes use of CINDs and CFDs to clean drug data. Moreover, previous work on data cleaning did not study inference, satisfiability and implication analyses of constraints, which are the focus of this paper.

Constraints used in schema matching in practice are typically standard INDs and keys (see, e.g., [10]), Contextual schema matching [11] investigated the applications of contextual foreign keys, a primitive and special case of CINDs, in deriving schema mapping from schema matches. While [11] partly motivated this work, it neither formalized the notion of CINDs nor considered the static analysis of CINDs. There has also been recent work on data exchange (schema mapping) and data integration based on TGDs (see [16,45,46] for surveys). However, inference systems and static analyses of constraints are not the focus of the prior work on data exchange and data integration, and none of the results of this work has been established in those lines of research.

The chase technique is widely used in implication analysis and query optimization, and has been studied for a variety of dependencies (see, e.g., [13,15,18,47]). Recently it was extended for query reformulation and schema mapping, and a number of sufficient conditions were identified for its termination (see [16,48] for recent surveys). This work extends the chase technique to study the implication analysis of CINDs, for which the chase process always terminates.
7. Conclusion

We have proposed CINDs, a mild extension of INDs that is important in both contextual schema matching and data cleaning. We have also settled several fundamental problems associated with static analysis of CINDs, from the satisfiability to the finite axiomatizability, to the complexity of the implication problem. Tables 1 and 2 summarize the main results of this work on CINDs, compared with their counterparts for standard INDs. While some proofs for CINDs in the absence of finite-domain attributes are inspired by the proofs of their IND counterparts, most proofs here are more involved. For instance, we have to deal with six inference rules, whereas the inference system for INDs has three rules and does not need to consider constant patterns [20]. In the general setting, our proofs are much more complicated, since we need to cope with the impact of finite-domain attributes and constant patterns together, as indicated by eight inference rules. The techniques for dealing with finite-domain attributes are interesting in their own right.

For the satisfiability analysis, the presence of finite-domain attributes does not complicate the problem of INDs, CINDs, INDs + FDs and CINDs + CFDs. However, while a set of INDs and FDs is always satisfiable [13], the problem of CINDs and CFDs put together is undecidable, because of the presence of data values.

For the implication analysis, in particular, we have developed a sound and complete inference system for CINDs. We have also provided a complete picture of complexity bounds for the implication analysis of CINDs and INDs, focusing on the following dichotomies:

- with constant patterns (CINDs) vs. their absence (INDs);
We have investigated the impact of these factors on the implication analysis of inclusion dependencies. (a) For traditional INDs, AINDs, UINDs and acyclic UINDs, the presence of finite-domain attributes does not complicate the implication analysis. (b) The presence of constant patterns alone does not increase the complexity. Indeed, the implication problem for CINDs, ACINDs, UCINDs and acyclic UCINDs in the absence of finite-domain attributes retains the same complexity as its counterpart for INDs, AINDs, UINDs, and acyclic UINDs, respectively. Nevertheless, (c) the presence of both constant patterns and finite-domain attributes makes our lives harder. Indeed, the implication problem for CINDs, ACINDs, UCINDs and acyclic UCINDs in the general setting has a higher complexity than its counterpart for INDs, AINDs, UINDs and acyclic UINDs, respectively. These tell us that it is the interaction between constant patterns and finite-domain attributes that complicates the implication analysis.

There is naturally much more to be done. First, we have shown that when CINDs and CFDs are taken together, both the satisfiability problem and the implication problem become undecidable. Nevertheless, it is not known yet whether the problems are decidable for CFDs (or FDs) and fragments of CINDs put together, e.g., AINDs, ACINDs, UINDs, acyclic UINDs and acyclic UCINDs. Second, an extension of CINDs was proposed in [41] by supporting built-in predicates. However, effective and efficient data repairing and schema matching algorithms are yet to be developed when both CINDs and CFDs are brought into the play.

Acknowledgements

Ma is supported in part by NSFC grant 61322207, NCFR 973 grant 2014CB340304 and MOST grant 2012BAH46804. Fan is supported in part by 973 Programs 2012CB316200 and 2014CB340302, NSFC 61133002, Guangdong Innovative Research Team Program 2011ID005 and Shenzhen Peacock Program 110510030834361 of China, as well as EPSRC EP/J015377/1, UK.

Appendix A. Proofs

A.1. Proof of Theorem 2

Theorem 2. In the absence of finite-domain attributes, the inference rules IR1–IR6 of Σ are sound and complete for the implication analysis of CINDs.

Proof. For the soundness, we show that given a set Σ ∪ {ψ} of CINDs, if Σ ⊢ IR1–IR6 ψ, then Σ |= ψ. This can be readily verified by induction on the length of proofs with IR1–IR6, by showing that the application of each of IR1–IR6 is sound, by the definition of CINDs.

For the completeness, we show that given a set of CINDs Σ ∪ {ψ} over a relational schema R, if Σ |= ψ then Σ ⊢ IR1–IR6 ψ. i.e., ψ is provable from Σ by using IR1–IR6. We assume w.l.o.g. that R = (R1, . . ., Rn), and that ψ is (R1[A1], . . ., Am[Xp]) ⊆ R1B1 . . . BmYp tpAp, in which R1 and Rm are in R.

The proof consists of three parts. (1) We first develop a chase procedure to determine whether Σ |= ψ. The chase process starts with a single-tuple instance of R, and repeatedly adds tuples (one at a time) to the instance by applying CINDs in Σ until no more CINDs can be applied. (2) We then show that the chase process always terminates, and moreover, that if Σ |= ψ, then the resulting instance satisfies ψ. (3) Based on these, we finally show that if Σ |= ψ then Σ ⊢ IR1–IR6 ψ.

(1) We first introduce the chase procedure.

We construct a tuple t0 of schema Ra such that (a) t0[Aa] = va for i ∈ [1, m]; (b) t0[Ap] = tpAp[Ap] for each attribute Ap ∈ Xp of the CIND ψ; and (c) t0[A] = v0 for the rest of the attributes A ∈ attr(Ra) \ (∪Ai∈[1,m]∪Xp). Here v0, v1, . . . , vm are m + 1 distinct variables that represent data values not appearing in Σ ∪ {ψ}.

The chase process starts with an instance D0 := (I1, . . ., Ia, . . ., In) of R such that the instance Ia of schema Ra contains the single tuple t0, and for each i ∈ [1, n] with i a, the instance Ii of schema Ra is empty.

The chase adds tuples to the database D0, one at a time, by making use of a chase operation APPLY. More specifically, given a CIND ψ ′ = (R1[U; Up] ⊆ R1[V; Vp]; tpAp) in Σ and an instance D, APPLY(D, ψ ′) transforms the instance D into a new one D ′ by applying CIND ψ ′ to D as follows.

− For each tuple ti ∈ Ii with ti[Up] = tpAp[Up], if there exists no tuple tj ∈ Ij such that tj[V] = ti[U] and tj[Vp] = tpAp[Vp], we say that the CIND ψ ′ is applicable to D. If so then the chase adds a tuple tj to Ij such that (a) tj[V] = ti[U], (b) tj[Vp] = tpAp[Vp], and (c) tj[E] := v0 for all the other attributes E ∈ attr(Rj) \ (V ∪ Vp).

− If there exists no such tuple ti, the instance D remains the same, i.e., D = APPLY(D, ψ ′).
The chase process stops when it reaches an instance $D_f$ such that no more CINDs in $\Sigma$ are applicable to $D_f$, i.e., when $D_f = \text{APPLY}(D, \psi')$ for all CINDs $\psi'$ in $\Sigma$. We refer to such $D_f$ as a result of $\text{APPLY}(D, \Sigma)$.

(2) Next we show that the chase process always terminates, and that all results $D_f = \text{APPLY}(D, \Sigma)$ satisfies $\psi$ if $\Sigma \models \psi$.

Observe that in any step of the chase process, the database is defined in terms of a finite set of elements, i.e., $m+1$ variables $\{v_0, v_1, \ldots, v_m\}$, and the constants appearing in the constant patterns of CINDs in $\Sigma \cup \{\psi\}$. There are only finitely many distinct databases with these elements. From this it follows that the chase process always terminates.

Intuitively, the chase process yields a sequence of databases $D_0, D_1, \ldots, D_f$ such that (a) $D_0$ is the initial instance, and (b) for each $i \in [0, f-1]$, $D_{i+1} := \text{APPLY}(D_i, \psi')$ by applying a CIND $\psi'$ in $\Sigma$ to $D_i$. In addition, $D_i \subseteq D_{i+1}$ for all $0 \leq i < f$, and $\text{APPLY}(D_f, \psi') = D_f$ for all $\psi'$ in $\Sigma$. That is, the instance $D_f$ is a fixpoint. When it is clear from the context, we use $\text{Chase}(\Sigma, \psi)$ to denote such a $D_f$; in other words, $\text{Chase}(\Sigma, \psi)$ denotes an arbitrary fixpoint $D_f$ obtained by such a chase process.

It is easy to see that $\text{Chase}(\Sigma, \psi) \models \Sigma$, no matter what fixpoint is obtained by the chase process and is denoted by $\text{Chase}(\Sigma, \psi)$. Thus if $\Sigma \models \psi$, then $\text{Chase}(\Sigma, \psi) \models \psi$. That is, the initial tuple $t_0 \in I_0$ enforces the existence of another tuple $t_0 \in I_0$ such that (a) $t_0[B_i] = v_i$ for $i \in [1, m]$, and (b) $t_0[Y_p] = t_p[Z_p]$.

**Example 1.** Consider a database $D_0$ with three relations shown below, and the CINDs $\psi_1'$ and $\psi_2'$ given in **Example 3.3.** Let $\Sigma$ be $\{\psi_1', \psi_2'\}$ and $\psi = (R_1[A; B] \subseteq R_3[H; G], (b_1 \equiv g))$. The chase process works on the database $D_0$ as follows.

1. $I_1$: $A \ B \ C$
   $t_1$: $A \ B \ C$
   $v_1 \ b_1 \ v_0$
2. $I_2$: $E \ F \ G$
   $t_2$: $E \ F \ G$
   $v_1 \ v_0 \ g$
3. $I_3$: $G \ H \ K$
   $t_3$: $G \ H \ K$

(1) Initial instance $D_0$

$D_1$ after executing $\text{APPLY}(D_0, \psi_1')$

1. $I_1$: $A \ B \ C$
   $t_1$: $A \ B \ C$
   $v_1 \ b_1 \ v_0$
2. $I_2$: $E \ F \ G$
   $t_2$: $E \ F \ G$
   $v_1 \ v_0 \ g$
3. $I_3$: $G \ H \ K$
   $t_3$: $G \ H \ K$

(2) Instance $D_1$ after executing $\text{APPLY}(D_0, \psi_1')$

$D_2$ after executing $\text{APPLY}(D_1, \psi_1')$

1. $I_1$: $A \ B \ C$
   $t_1$: $A \ B \ C$
   $v_1 \ b_1 \ v_0$
2. $I_2$: $E \ F \ G$
   $t_2$: $E \ F \ G$
   $v_1 \ v_0 \ g$
3. $I_3$: $G \ H \ K$
   $t_3$: $G \ H \ K$

(3) Final instance $D_2$ after executing $\text{APPLY}(D_1, \psi_1')$

Note that $\Sigma \models \psi$, and that there exists a tuple $t_3$ such that $t_3[H] = t_1[A]$ and $t_3[G] = g$.

(3) Finally show that if $\Sigma \models \psi$, then $\Sigma \vdash_{\Sigma \cup \{\psi\}} \psi$. To prove this, it suffices to show the following claim.

**Claim 1.** Assume that the instance $I_j$ of $R_j$ in the chase process contains a tuple $t$, where (a) for each $E_i \in \text{attr}(R_j)$ ($i \in [1, |k|]$), $t[E_i] = v_{ji}$ and $v_{ji}$ is in $\{v_1, \ldots, v_m\}$, and (b) $t[Z_p]$ consists of only constants for a list $Z_p$ of distinct attributes in $\text{attr}(R_j)$. Then $\Sigma \vdash_{\Sigma \cup \{\psi\}} \psi$, where $\psi$ is $(R_0[A_1], \ldots, A_k; X_p) \subseteq R_1[E_1, \ldots, E_k; Z_p]$. $t'_p[X_p] = t_{p'}[X_p]$ and $t'_p[Z_p] = t[Z_p]$.

For if $\Sigma \vdash_{\Sigma \cup \{\psi\}} \psi$, we can conclude that if $\Sigma \models \psi$, then $\Sigma \vdash_{\Sigma \cup \{\psi\}} \psi$. Indeed, as remarked earlier, $\text{Chase}(\Sigma, \psi) \models \psi$ for all fixpoints obtained by the chase and denoted by $\text{Chase}(\Sigma, \psi)$. Since the initial $D_0$ contains the tuple $t_0$, by the definition of chase, there must be a tuple $t_0 \in I_0$ in $\text{Chase}(\Sigma, \psi)$ such that (a) $t_0[B_i] = v_i$ for $i \in [1, m]$, and (b) $t_0[Y_p] = t_{p'}[Z_p]$. Hence by **Claim 1**, $\Sigma \vdash_{\Sigma \cup \{\psi\}} \psi$.

**Base case.** When the length is 1, the sequence consists of the initial instance $D_0$, which contains a single tuple $t_0$ in the instance $I_0$ of schema $R_0$. We show that $(R_0[X'_p; X_p]) \subseteq R_1[X'_p; X_p], (t_0[X_p]) \models t_0(X_p))$ for all $X' \subseteq A_1, \ldots, A_m$ and $X_p \subseteq X_p$, by repeatedly using IR4 as follows.

1. $(R_0[X'_p; X_p]) \subseteq R_1[X'_p; X_p], (t_0[X_p]) \models t_0(X_p))$, IR1
2. $(R_0[X'_p; X_p]) \subseteq R_1[X'_p; X_p], (t_0[X_p]) \models t_0(X_p))$, IR4
Inductive case. Assume that Claim 1 holds for the first $i+1$ instances $D_0, \ldots, D_i$. We next show that Claim 1 also holds on $D_{i+1} := \text{apply} (D_i, \psi')$.

If $D_{i+1} = D_i$, the instance $D_i$ is not changed by $\text{apply}(D_i, \psi')$, and the claim obviously holds on $D_{i+1}$ since it holds on $D_i$.

Assume that $D_{i+1} \neq D_i$. Then there must be a single tuple $w$ inserted into the instance $I_j$ of some schema $R_j$, i.e., $I_j = I_j \cup \{w\}$, as the result of $\text{apply}(D_i, \psi')$ for some $\psi' \in \Sigma$. Assume w.l.o.g. that $\psi' = (R_i[C_1, \ldots, C_k; U_p]) \subseteq R_j[F_1, \ldots, F_k; V_p], t_p, \psi)$. Since $\psi'$ is applicable to $D_i$, there exists a tuple $u$ in $I_i$ such that $u[U_p] = t_p, \psi, [U_p]$. By the induction hypothesis, there is a CIND $\psi_u = (R_u[A_{p_1}, \ldots, A_{p_h}; X_p] \subseteq R_j[C_{p_1}, \ldots, C_{p_h}; U_p], t_{p_u})$ such that (a) for each attribute $C \in \text{attr}(R_i)$, if $u[C]$ is a constant, then $C \in U_p$, and if $u[C]$ is a variable in $\{v_1, \ldots, v_m\}$, then $C \in \{C_{p_1}, \ldots, C_{p_h}\}$. (b) $t_{p_u}[U_p'] = u[U_p']$, and (c) $\Sigma \vdash \langle I \setminus 6 \rangle \psi_u$. Based on the tuples $w, u$ and the CINDs $\psi'$ and $\psi_u$, we can derive the following.

(a) $t_{p_u}[U_p'] = u[U_p']$, $t_{p_u}[U_p] = u[U_p]$, and $U_p \subseteq U_p'$.

(b) Let $\{C_{q_1}, \ldots, C_{q_g}\} = \{C_1, \ldots, C_k \cap \{C_{p_1}, \ldots, C_{p_h}\}\}$, where $0 \leq g \leq \min(k, h)$. Then $\{A_{q_1}, \ldots, A_{q_h}\} \subseteq \{A_{p_1}, \ldots, A_{p_h}\}$, and for the tuple $w, w[F] \in \{v_1, \ldots, v_m\}$ iff $F \in \{F_{q_1}, \ldots, F_{q_g}\}$.

(c) For each attribute $F \in \text{attr}(R_i)$, $w[F]$ is a constant iff $F \in U_p$ such that $w[F] = u[F]$ and $C_i \in U_c = U_p' \cap \{C_1, \ldots, C_k\}$. We use $V_{c}$ to denote the list of attributes in $\text{attr}(R_i)$ that corresponds to the list of attributes $U_c$ of $\text{attr}(R_i)$ in $\psi'$.

(d) $U_c \cap \{C_{q_1}, \ldots, C_{q_g}\} = \emptyset$ and $U_c \subseteq U_p'$.

We next show that $\Sigma \vdash \langle I \setminus 6 \rangle \psi_u$, where $\psi_u$ is the CIND $(R_u[A_{p_1}, \ldots, A_{p_h}; X_p] \subseteq R_j[F_{p_1}, \ldots, F_{p_h}; Z_p'], t_{p_u})$. Observe that (a) $\{F_{p_1}, \ldots, F_{p_h}\} \subseteq \{F_{q_1}, \ldots, F_{q_g}\}$ by (2), and (b) $Z_p' \subseteq V_p \cup U_c$ by (3).

In addition, we can derive the following.

(a) $(R_i[C_{q_1}, \ldots, C_{q_g}; U_c, V_{c}] U_p \subseteq R_j[F_{q_1}, \ldots, F_{q_g}; V_{c}, V_p], t_p)$, where $t_{p_1}[U_p; V_p] = t_{p_2}[U_p; V_p]$ and $t_{p_1}[U_c] = t_{p_2}[V_c] = w[U_p]$ (using $\psi'$, IR2, IR4).

(b) $(R_u[A_{q_1}, \ldots, A_{q_h}; X_p] \subseteq R_i[C_{q_1}, \ldots, C_{q_h}; U_p], t_{p_u})$, where $t_{p_2}[X_p; V_p] = t_{p_3}[X_p; V_p]$ (using $\psi_u$, IR2).

(c) $(R_i[C_{q_1}, \ldots, C_{q_g}; U_p'] \subseteq R_j[F_{q_1}, \ldots, F_{q_g}; V_{c}, V_p], t_p)$, where $t_{p_1}[U_p'] = t_{p_2}[U_p']$ and $t_{p_1}[V_c, V_p] = t_{p_1}[V_c V_p]$ (using (1), IR5).

(d) $(R_u[A_{q_1}, \ldots, A_{q_h}; X_p] \subseteq R_i[C_{q_1}, \ldots, C_{q_h}; U_p], t_{p_u})$, where $t_{p_{21}}[X_p; U_c; V_p] = t_{p_3}[X_p; U_c; V_p]$ (using (2), IR6).

(e) $(R_i[C_{q_1}, \ldots, C_{q_g}; X_p] \subseteq R_j[F_{q_1}, \ldots, F_{q_g}; V_c V_p], t_p)$, where $t_{p_{21}}[X_p] = t_{p_3}[X_p]$ and $t_{p_1}[V_c, V_p] = t_{p_1}[V_c V_p]$ (using (2), IR3 or alternatively, (1), IR2).

(f) $(R_u[A_{q_1}, \ldots, A_{q_h}; X_p] \subseteq R_i[F_{q_1}, \ldots, F_{q_g}; V_c, V_p], t_{p_u})$, where $t_{p_{21}}[X_p; V_c; V_p] = t_{p_3}[X_p; V_c V_p]$ (using (5), IR2).

(g) $(R_u[A_{q_1}, \ldots, A_{q_h}; X_p] \subseteq R_j[F_{q_1}, \ldots, F_{q_g}; Z_p', t_{p_u})$, where $t_{p_{21}}[X_p] = t_{p_3}[X_p] = t_{p_3}[X_p]$ and $t_{p_{21}}[Z_p'] = t_{p_3}[Z_p'] = w[Z_p']$ (using (6), IR6).

This verifies that Claim 1 holds on $D_{i+1}$. From this it follows that rules IR1–IR6 are complete for the implication of CINDs in the absence of finite-domain attributes. □

A.2. Proof of Theorem 3

Theorem 3. The inference system $I$ is sound and complete for the implication of CINDs in the general case, when finite-domain attributes may be present.

Proof. The soundness of $I$ can be verified by induction on the length of $I$-proofs.

For the completeness of $I$, consider a set $\Sigma \cup \{\psi\}$ of CINDs defined on a database schema $\mathcal{R}$. We show that if $\Sigma \vdash \psi$, then $\Sigma \vdash \psi$. Assume that $\mathcal{R} = (R_1, \ldots, R_n)$, and that $\psi = (R_i[A_{m_1}, \ldots, A_{m_m}; X_p] \subseteq R_k[B_{m_1}, \ldots, B_{m_m}; V_p], t_{p})$.

The proof consists of five parts. (1) We first consider that $\psi$ satisfies a special form of CINDs. (2) We then develop a chase procedure for the special form of CINDs, and (3) show that the chase process always terminates. (4) In addition, we establish an important property of the chase procedure, and based on which (5) we show that if $\Sigma \vdash \psi$ then $\Sigma \vdash \psi$ by using IR1–IR8.

(1) We first introduce the special form of CINDs. A CIND $(R_i[U; U_p] \subseteq R_j[V; V_p], t_p)$ is in the special form if (a) the attribute list $U$ only consists of attributes with an infinite domain, and (b) the attribute list $U_p$ contains all the finite-domain attributes of schema $R_i$.

As a result, the attribute list $V$ also contains only infinite-domain attributes of the schema $R_j$. However, it is possible that (a) there is an infinite-domain attribute $A$ of schema $R_i$ such that $A \in U_p$, but (b) there is a finite-domain attribute $B$ of schema $R_j$ such that $B \not\in V_p$.

It suffices to consider CINDs in this special form only. Indeed, given a CIND $\psi$, we show that there exists a set $\bar{\Sigma}_\psi$ of CINDs in the special form such that $\bar{\Sigma}_\psi$ is equivalent to $\psi$. Better still, $\bar{\Sigma}_\psi$ can be proved from $\psi$ by using rules in $I$ and vice versa, i.e., $\phi \vdash I \bar{\Sigma}_\psi$ and $\bar{\Sigma}_\psi \vdash I \psi$. 
First, let’s consider CIND $\phi = (R_1[U; A; U_p] \subseteq R_1[V; B; V_p], t_{p_0})$, where attribute $A$ has a finite-domain $\text{dom}(A) = \{a_1, \ldots, a_k\}$. For each $l \in [1, k]$, we define a new CIND $\phi_l = (R_1[U; A; U_p] \subseteq R_1[V; B; V_p], t_{p_0})$ such that $t_{p_0}[A] = a_l$ and $t_{p_0}[U_p \parallel V_p] = t_{p_0}[U_p \parallel V_p]$. This is justified by IR4. Let $\Sigma_\psi = \{\phi_1, \ldots, \phi_k\}$. It is easy to verify that $(\phi) = \Sigma_\psi$, i.e., $\Sigma_\psi \models (\phi)$ and $\Sigma_\psi \models (\psi)$.

Next, consider $\phi = (R_1[U; U_p] \subseteq R_1[V; V_p], t_{p_0})$, where there exists an attribute $A \in \text{attr}(R_1) \setminus (U \cup U_p)$ with a finite-domain $\text{dom}(A) = \{a_1, \ldots, a_k\}$. For each $l \in [1, k]$, we construct a CIND $\phi_l = (R_1[U; A; U_p] \subseteq R_1[V; V_p], t_{p_0})$ such that $t_{p_0}[A] = a_l$ and $t_{p_0}[U_p \parallel V_p] = t_{p_0}[U_p \parallel V_p]$. This is justified by IR5. Let $\Sigma_\psi = \{\phi_1, \ldots, \phi_k\}$. Then $(\phi) \models \Sigma_\psi$.

This shows how we can convert each CIND $\psi$ in $\Sigma \cup \{\psi\}$ into an equivalent set $\Sigma_\psi$ of CINDs in the special form. In addition, $(\phi) \models \Sigma_\psi$ by successive applications of IR4 and IR5, and moreover, $\Sigma_\psi \models (\psi)$ by successive applications of IR7 and IR8. Thus, we can assume w.l.o.g. that all the CINDs in $\Sigma \cup \{\psi\}$ are in the special form.

(2) We now give the chase procedure for determining whether $\Sigma \models \psi$, which extends the one given in the proof of Theorem 2 to further deal with finite-domain attributes.

To deal with the interaction between finite domains and constant patterns in the CINDs, the chase operation on trees as opposed to relations. In such a tree $T$, a $N_{\text{root}}$ is its root, (b) each node in $T$ is labeled with a tuple $t_i$ and its schema $R_i$, denoted by $N = R_1:t_1$, and (c) for each leaf node $N_{\text{leaf}}$ in $T$, the path from the root $N_{\text{root}}$ to $N_{\text{leaf}}$ denoted by $\text{PATH}(N_{\text{root}}, N)$, encodes an instance of $R$, such that for each relation schema $R$ in $R$, the set $I_R$ of tuples of $R$ carried by the nodes on the path is an instance of $R$.

We now give the detailed chase process.

The chase process starts with a tree $T_0$ consisting of only a root node $N_{\text{root}} = 'R_1:t_1'$, in which $t_0$ is a tuple of schema $R_a$ such that (a) for each node $N_{\text{leaf}}$ in $T$, there exists a set of nodes at a time, by applying a chase operation $\text{APPLY}_{\psi}$. To specify $\text{APPLY}_{\psi}$, we need the following notion.

A CIND $\psi' = (R_1[U; U_p] \subseteq R_1[V; V_p], t_{p_0})$ in $\Sigma$ is said to be applicable to a node $N = R_1:t'$ if (a) $t_1[U_p] = t_{p_0}[U_p]$; (b) there exists no node $N' = R_1:t'$ with $t_1[V] = t_1[U]$ and $t_1[V_p] = t_{p_0}[V_p]$ on $\text{PATH}(N_{\text{root}}, N)$; and (c) there exists at least a leaf node $N_{\text{leaf}}$ such that there is no such node $N'$ on $\text{PATH}(N, N_{\text{leaf}})$. Intuitively, let $D$ denote the database instance represented by $\text{PATH}(N, N_{\text{leaf}})$ on which $N$ appears. Then $D \not\models \psi'$, and hence, we need to enforce $\psi'$ on $D$.

Given a tree $T$ and the CIND $\psi'$, the chase operation $\text{APPLY}_{\psi'}(T, \psi')$ transforms $T$ into a new tree $T'$ as follows.

- It traverses $T$ starting from its root node $N_{\text{root}}$ in a breadth-first order, and checks whether there exists a node to which the CIND $\psi'$ is applicable.
- If such a node $N$ is found, then new nodes are added to $T$ to make $T'$, as follows.
  (a) Let $S = \{N_{\text{leaf}}, \ldots, N_{\text{leaf}}\}$ be the set of all leaf nodes in $T$ such that for each $i \in [1, k]$, $\text{PATH}(N_{\text{leaf}}, N_{\text{leaf}})$ exists and moreover, there exists no node $N' = R_i:t'$ with $t_1[V] = t_1[U]$ and $t_1[V_p] = t_{p_0}[V_p]$ on $\text{PATH}(N_{\text{leaf}}, N_{\text{leaf}})$.
  (b) Let $\rho(V_f)$ denote an instantiation of the list $V_f$ of all finite-domain attributes in $\text{attr}(R_1) \setminus (V \cup V_p)$. That is, for each attribute $C \in V_f$, $\rho(C)$ is a data value drawn from $\text{dom}(C)$.
  (c) For each possible instantiation $\rho(V_f)$, it generates a new node $N_{\text{leaf}}' = R_i:t'$ such that $t_{p_0}[V] = t_{\rho(V_f)}[V]$ and $t_{p_0}[C] = \rho(C)$ for all the other attributes $C \in \text{attr}(R_1)$. Observe that there are only a finite number of instantiations for the attribute list $V_f$, and for each attribute $C \in \text{attr}(R_1)$, $t_{p_0}[C]$ is a constant if the attribute $C$ has a finite domain.
  (d) For each leaf node $N_{\text{leaf}}'$ in $S$ and each new node $N_{\text{leaf}}'$, it adds $N_{\text{leaf}}'$ as a child of $N_{\text{leaf}}'$. That is, $\psi'$ is enforced on the database represented by $\text{PATH}(N_{\text{leaf}}, N_{\text{leaf}})$.

Let $T'$ denote the modified tree. Then the same process repeats starting from the root node of $T'$.

- If there are no nodes to which $\psi'$ is applicable, the tree $T$ remains unchanged, i.e., $T' = \text{APPLY}_{\psi'}(T, \psi') = T$.

The chase process stops if no CINDs in $\Sigma$ are applicable to any nodes in $T$, i.e., $T = \text{APPLY}_{\psi'}(T, \psi')$ for each $\psi'$ in $\Sigma$.

Intuitively, the chase process augments tree $T_0$ and generates a sequence $T_0, T_{t_1}, \ldots, T_{t_f}$ of trees such that (a) for each $l \in [0, f - 1]$, $T_{t_{l+1}} := \text{APPLY}_{\psi'}(T_l, \psi')$ for some $\psi' \in \Sigma$, and all the nodes in tree $T_l$ also appear in $T_{t_{l+1}}$, and (b) $T_f = \text{APPLY}_{\psi'}(T_f, \psi')$ for each $\psi'$ in $\Sigma$, i.e., $T_f$ is a fixpoint reached by $\text{APPLY}_{\psi'}$. We refer to $T_f$ as a result of the chase process with $\Sigma$ and $\psi'$, denoted by $\text{Chase}(\Sigma, \psi')$. Note that there may exist multiple distinct resulting trees $T_f$ depending on the orders of CINDs in $\Sigma$ used in the chase process. We use $\text{Chase}(\Sigma, \psi')$ to denote an arbitrary fixpoint obtained by the chase, and study the properties of all such fixpoints.

**Example 2.** Consider the set $\Sigma \cup \{\psi\}$ of CINDs given in **Example 3.4**. We first transform each CIND into the special form. For instance, for the CIND $\psi$, we generate a set $\Sigma_\psi = \{\psi_1, \psi_2, \psi_3, \psi_4, \psi_5, \psi_6\}$ of CINDs, where

$$\psi_1' = (R_1[BC; AD] \subseteq R_2[EF; G], (a, d \parallel d)),$$

$$\psi_2' = (R_1[BC; AD] \subseteq R_2[EF; G], (b, d \parallel d)).$$
We show the chase process for \( \Sigma \) and \( \psi_2 \) in Fig. 4, where (a) tree \( T_0 \) is the initial tree, (b) \( T_1 \) is derived by applying \( \psi_2 \) to \( T_0 \), (c) \( T_2 \) is the result of applying \( \psi_{3,1} \) to \( T_1 \), and (d) \( T_3 \) is produced by applying \( \psi_{3,2} \) to \( T_2 \). Here \( \psi_{3,1} = (R_3[IF; KL] \subseteq R_2[EF; G], (k, g \parallel d, g)) \), and \( \psi_{3,2} = (R_3[IF; KL] \subseteq R_2[EF; G], (k, h \parallel d, g)) \). These two CINDs are in the special form, derived from \( \psi_3 \) by using IR5.

Observe that \( (1) \Sigma \models \psi_2 \), and \( (2) \) for each leaf \( N_{\text{leaf}} \) in the result \( T_3 = \text{Chase}(\Sigma, \psi_2) \), there exists a node \( N = \{R_2 : t' \} \) on \( \text{PATH}(N_{\text{root}}, N_{\text{leaf}}) \) with \( t[EF] = (v_1, v_2) \) and \( t[G] = d' \). □

(3) We next verify that the chase process always terminates.

Observe that in each tree \( T \) generated in the chase process, \( \text{PATH}(N_{\text{root}}, N_{\text{leaf}}) \) from the root \( N_{\text{root}} \) to each leaf \( N_{\text{leaf}} \) of \( T \) represents a database instance \( D \) of \( R \). In addition, there exist no nodes \( N_1 \) and \( N_2 \) on the path such that they are labeled with the same tuple. Hence the depth of \( T \) is determined by the maximum instance of \( \Sigma \) constructed from the finite set \( \{v_0, \ldots, v_m\} \) of variables, the finite set of constants appearing in the constant patterns in \( \Sigma \), and the finite cardinalities \( C \) of finite domains, which is also determined by \( \Sigma \). Similarly, the maximum number of the children of a node in \( T \) is bounded by the maximum cardinality of finite domains, which is also determined by \( \Sigma \). Hence the size of \( T \) is bounded. There exist finitely many distinct trees that are constructed from those variables and constants with a bounded size. The chase process can generate at most finitely many such trees that are distinct, and hence, it must terminate.

(4) We next show a property of the chase procedure, which will be used to show that if \( \Sigma \models \psi \), then \( T \vdash_\Sigma \psi \).

We first define an operator \( \Upsilon(N) \), where \( N \) is a node in a tree \( T_l \) \( (l \in [0, f]) \) generated in the chase process. Given \( N = \{R_l : t_l \} \), we define \( \Upsilon(N) = (R_l[C_1, \ldots, C_m; U_p], t_l[U_p]) \) if

- for each \( j \in [1, m] \), \( t_l[C_j] = v_j \), i.e., \( t_l[C_1, \ldots, C_m] = (v_1, \ldots, v_m) \); and
- the list \( U_p \) consists of all those attributes \( C \in \text{attr}(R_l) \) such that \( t_l[C] \) is a constant;

whereas \( \Upsilon(N) \) is undefined if there exist no attributes \( C_1, \ldots, C_m \) in \( R_l \) such that \( t_l[C_1, \ldots, C_m] = (v_1, \ldots, v_m) \).

Observe that when the CIND \( \psi \) is enforced, \( \Upsilon(N) \) must be defined on some node. We shall use \( \Upsilon(N) \) to inspect the existence of nodes satisfying the conditions specified by \( \psi \).

The property is stated as follows.

**Claim 2.** Let \( \psi \) be a CIND \( (R_A[A_1, \ldots, A_m; X_P] \subseteq R_B[D_1, \ldots, D_m; V_P], t_{p_{\psi}}) \), \( T_l \) be a tree generated in the chase process \((l \in [0, f])\), and \( S = \{N_{\text{leaf}}, \ldots, N_{\text{leaf}}\} \) be the set of all the leaf nodes in \( T \). Then \( \Sigma \vdash_\Sigma \psi \) if \( \Sigma \vdash_\Sigma \psi_j \) for each \( j \in [1, k] \), where for a schema \( R_i \) in \( \mathcal{R} \),

(a) \( \psi_j = (R_l[C_1, \ldots, C_m; U_p], t_l[U_p]) \subseteq R_{p_{\psi}}[D_1, \ldots, D_m; V_P], t_{p_{\psi}} \) with \( t_{p_{\psi}}[V_P] = t_{p_{\psi}}[V_P] \)

(b) there exists a node \( N_j = \{R_i : t'_j \} \) on \( \text{PATH}(N_{\text{root}}, N_{\text{leaf}}) \) such that \( \Upsilon(N_j) = (R_l[C_1, \ldots, C_m; U_p], t_j[U_p]) \) and \( t_j[U_p] \).
Observe the following. (a) LHS(ϕ) is the same as LHS(ψ). (b) For each j ∈ [1, k], RHS(ϕj) is the same as RHS(ϕ), and ψj’s have the same LHS. As will be seen in part (5) of the proof, we use these to prove \( \Sigma \vdash \varphi \).

We show the claim by induction on the length of the sequence of trees \( T_0, T_1, \ldots, T_f \) generated by the chase process.

Base case. For the initial tree \( T_0 \), the only leaf node in \( T_0 \) is the root \( N_{\text{root}} = R_{g} : t'_{g} \). In this case, \( \Upsilon(N_{\text{root}}) = (R_{g}[A_{1}, \ldots, A_{m}; X_{f}], t_{p_{g}}[X_{f}]) \), and \( \varphi_{\text{root}} \) and \( \varphi \) are the same CIND \( (R_{g}[A_{1}, \ldots, A_{m}; X_{f}] \subseteq R_{g}[A_{1}, \ldots, A_{m}; X_{f}], (t_{p_{g}}[X_{f}] || t_{p_{g}}[X_{f}])) \). It is obvious that if \( \Sigma \vdash \varphi_{\text{root}} \), then \( \Sigma \vdash \varphi \).

Inductive case. Assume that Claim 2 holds for the first \( i \) trees \( T_{0}, T_{1}, \ldots, T_{i} \). We show that it also holds on \( T_{i+1} := \text{APPLY}(T_{i}, \psi') \), i.e., a result of applying some CIND \( \psi' \) in \( \Sigma \) to a node \( N \) in tree \( T_{i} \). Assume w.l.o.g. that \( N = R_{g} : t_{g} \), and that \( \psi' := (R_{g}[U_{1}; U_{p_{1}}] \subseteq R_{g}[U_{2}; U_{p_{2}}], t_{p_{g}}) \).

Here we consider two cases: (a) \( T_{i+1} = T_{i} \) and (b) \( T_{i+1} \neq T_{i} \). If \( T_{i+1} = T_{i} \), the tree \( T_{i} \) is not changed by \( \text{APPLY}(T_{i}, \psi') \), and the claim obviously holds on \( T_{i+1} \) since it holds on \( T_{i} \).

We next focus on the case where \( T_{i+1} \neq T_{i} \). Recall the chase operation \( \text{APPLY}(T_{i}, \psi') \), by applying the CIND \( \psi' \) to the node \( N \). Let \( S_{j} = [N_{\text{leaf}1}, \ldots, N_{\text{leaf}h}] \) be the set of all leaf nodes in tree \( T_{i} \), and let \( S_{\text{new}} = [N_{1}, \ldots, N_{h}] \) be the set of newly generated nodes by applying the CIND \( \psi' \) to the node \( N \). In \( T_{i+1} \), all the nodes in \( S_{\text{new}} \) appear as the children of each leaf node of the sub-tree rooted at node \( N \) in tree \( T_{i} \).

To illustrate this, an example of \( T_{i} \) and \( T_{i+1} \) is shown in Fig. 5. In \( T_{i} \), the sub-tree rooted at node \( N = R_{g} : t_{g} \) has two leaf nodes. In \( T_{i+1} \), three new nodes are added as the children of each of the two leaf nodes.

To simplify the discussion we assume w.l.o.g. that there is a single leaf node \( N_{\text{leaf}} \), in the sub-tree rooted at node \( N \); the proof for multiple such leaf nodes is similar. Thus, the set \( S_{i+1} \) of leaf nodes in tree \( T_{i+1} \) becomes \( S_{i} \cup S_{\text{new}} = [N_{1}, \ldots, N_{h}, N_{\text{leaf}}] \).

We show that the claim holds on \( T_{i+1} \), by considering the following cases.

Case 1. When the operator \( \Upsilon(N) \) is undefined on the node \( N \). Then for each node \( N_{j} \) in \( S_{\text{new}} \) (\( j \in [1, h] \)), \( \Upsilon(N_{j}) \) is also undefined by the definition of \( \text{APPLY} \), which generated those nodes in \( S_{\text{new}} \) to enforce \( \psi' \). In this case, the claim holds on \( T_{i+1} \). Indeed, those nodes \( N_{j} \) (\( j \in [1, k + h - 1] \)) required by the claim are in the tree \( T_{i} \), and so is \( N \). Hence \( \Sigma \vdash \varphi \) since the claim holds on \( T_{i} \) by the induction hypothesis.

Case 2. When \( \Upsilon(N) \) is defined on \( N \). Consider \( \Upsilon(N) = (R_{g}[U'; U_{p_{1}}], t_{g}[U_{p_{1}}]) \). Since the CIND \( \psi' \) is applicable to the node \( N \), we can derive the following: (a) \( U_{p_{1}} \subseteq U' \), (b) \( t_{g}[U_{p_{1}}] = t_{p_{g}}[U_{p_{1}}] \), (c) \( t_{g}[U'] = (v_{1}, \ldots, v_{m}) \), and (d) \( t_{g}[C] = v_{0} \) for each attribute \( C \) of \( \text{attr}(R_{g}) \) that is not in \( U' \cup U_{p}' \).

We distinguish the following cases.

Case 2(a). \( U' \not\subseteq U_{1} \), where \( U_{1} \) is in \( \text{LHS}(\psi') \). By \( \Upsilon(N) \) we have that \( t_{g}[U'] = (v_{1}, \ldots, v_{m}) \). Thus, if \( U' \not\subseteq U_{1} \), \( \Upsilon \) is not defined on those new nodes in \( S_{\text{new}} \). Along the same lines as for Case 1 above, one can show that the claim holds on \( T_{i+1} \).

Case 2(b). \( U' \subseteq U_{1} \). We show that \( \Sigma \vdash \varphi \) if \( \Sigma \vdash \psi_{j} \) for each \( j \in [1, k + h - 1] \), and for each leaf node \( N_{\text{leaf}} \) in \( S_{\text{leaf}} \), there exists a node \( N_{j} = R_{i} : t'_{j} \) on \( \text{PATH}(N_{\text{root}}, N_{\text{leaf}}) \) such that \( \Upsilon(N_{j}) = (R_{i}[C_{1}, \ldots, C_{m}; U_{p_{i}}], t_{j}[U_{p_{i}}]) \) and \( t_{j}[U_{p_{i}}] = t_{p_{g}}[U_{p_{i}}] \). It suffices to show that we only need to consider those nodes \( N_{j} \) (\( j \in [1, k + h - 1] \)) in \( T_{i} \). For this holds, then the same argument for Case 1 can verify that the claim holds on \( T_{i+1} \). We show this by distinguishing the following cases.

(a) For each leaf node \( N_{\text{leaf}} \) (\( j \in [2, k] \)), the node \( N_{j} \) on \( \text{PATH}(N_{\text{root}}, N_{\text{leaf}}) \) must be in \( T_{i} \) since \( N_{\text{leaf}} \) appears in \( T_{i} \).

(b) For each new leaf nodes \( N_{j} \) (\( j \in [1, h] \)) in \( S_{\text{new}} \), there exist \( \text{PATH}(N_{\text{root}}, N_{j}) \) and \( \text{PATH}(N_{\text{root}}, N_{\text{leaf}}) \) in tree \( T_{i+1} \), where there is an edge from \( N_{\text{leaf}} \) to \( N_{j} \). In this case, the leaf node \( N_{j} \) is the only node appearing on \( \text{PATH}(N_{\text{root}}, N_{j}) \), but not on \( \text{PATH}(N_{\text{root}}, N_{\text{leaf}}) \). That is, if the node \( N_{j} \) (\( j \in [1, h] \)) on \( \text{PATH}(N_{\text{root}}, N_{j}) \) is not in tree \( T_{i} \), it must be the leaf node \( N_{j} \) (see Fig. 5).
If there exists such a node $N_j (j \in [1, h])$ that is not in $S_{\text{new}}$, it must be on $\text{PATH}(N_{\text{root}}, N_i)$ and hence, there exists $\text{PATH}(N_i, N_j)$ for each $N_i$ in $S_{\text{new}}$. In this case, we only need to consider this $N_j$ in the tree $T_i$.

If such a node $N_j$ does not exist, we show that we can use the node $N$ instead of $N_j$, where $N$ is in $T_i$. In this case, for each $j \in [1, h]$, the node $N_j$ must be the leaf node $N_k$, and the CIND $\varphi_j$ must be in the form of $(R_g[U_2'; U_p], t_{p_g})$ such that $t_{p_g} = t_{p_1} = t_{p_2}$ and $t_{p_g} = t_{p_j}$. Here $U_j$ is the list of all finite-domain attributes in $\text{attr}(R_g[U_2') \cup \{U_1 \cup U_p\})$, and $\rho_{U_j} = \{\rho_1, \ldots, \rho_j\}$ is the set of all possible instantiations of $U_j$.

We show that we can use $N$ instead of $N_k$ and use a CIND $\varphi_g$ derived instead of $\varphi_j$ for $j \in [1, h]$.

- Since $\Sigma \vdash \varphi_j$ for each $j \in [1, h]$, we have that $\Sigma \vdash \varphi_g'$, where $\varphi_g' = (R_g[U_2'; U_p], t_{p_g'})$, and $t_{p_g'} = t_{p_1} | U_p | V_p$, by using IR7.

- By applying IR2 to the CIND $\varphi'$, we have that $\Sigma \vdash \varphi''$, where $\varphi'' = (R_g[U_1'; U_p], t_{p_{g''}})$, and $t_{p_{g''}} = t_{p_{g''}}$.

- By applying IR5 to the CIND $\varphi''$, we have that $\Sigma \vdash \varphi'''$, where $\varphi''' = (R_g[U_2'; U_p], t_{p_{g'''}})$, $t_{p_{g'''}} = t_{p_{g'''}}$.

- By applying IR3 to $\varphi'''$ and $\varphi_g'$, we can get that $\Sigma \vdash \varphi_g$, where $\varphi_g = (R_g[U_1'; U_p], t_{p_{g}})$, $t_{p_{g}} = t_{p_{g}} | U_p | V_p$, and hence, there exists $\varphi_g$.

Since $\forall (N) = (R_g[U_1'; U_p], t_{p_{g}})$, we can use $N$ and $\varphi_g$ instead of $N_k$ and $\varphi_j$ ($j \in [1, h]$), which still satisfy the conditions in the claim.

Hence we only need to use nodes in $T_i$, on which the claim holds based on the induction hypothesis.

Finally, we show that if $\Sigma \models \psi$, then $\Sigma \vdash \psi$, based on Claim 2. Let $T_f$ be an arbitrary fixpoint obtained by the chase (\(\Sigma, \psi\)).

Recall that for each leaf $N_{\text{leaf}}$ of $T_f$, $\text{PATH}(N_{\text{root}}, N_{\text{leaf}})$ represents a database instance $D$ of $\mathcal{R}$. Observe that $D \models \psi$. Indeed, $D \models \Sigma$ since no CINDs in $\Sigma$ are applicable to any nodes in $T_f$. By $\Sigma \models \psi$, we have that $D \models \psi$.

These tell us that for each leaf node $N_{\text{leaf}}$ in $T_f$, there must exist a node $N = \{R_b; b\}$ on $\text{PATH}(N_{\text{root}}, N_{\text{leaf}})$ such that $t_b[A_1, \ldots, A_m, Y_p] = t_b[X_p]$, and $t_b[Y_p] = t_b[X_p]$. Here $\forall (N) = (R_b[A_1, \ldots, A_m, Y_p], t[Y_p])$, where $Y_p \subseteq Y'$.

For each $N_{\text{leaf}}$, we can verify the following using the inference system $\mathcal{I}$:

\[
\begin{align*}
\varphi_1 &= (R_b[A_1, \ldots, A_m, Y_p], t_{p_{\varphi1}}), \\
\text{where } t_{p_{\varphi1}}[Y_p'] &= t_{p_{\varphi1}}[Y_p] \quad \text{IR1} \\
\varphi_2 &= (R_b[A_1, \ldots, A_m, Y_p], t_{p_{\varphi2}}), \\
\text{where } t_{p_{\varphi2}}[Y_p'] &= t_{p_{\varphi2}}[Y_p] \quad \text{IR6}
\end{align*}
\]

That is, for each $N_{\text{leaf}}$, $\Sigma \vdash \psi_2$. Taking this together with the existence of $N = \{R_b; b\}$, we have that $\Sigma \vdash \psi$ by Claim 2. That is, if $\Sigma \models \psi$, then $\Sigma \vdash \psi$.

This completes the proof for the completeness of $\mathcal{I}$ for CINDs when finite-domain attributes may be present. □

A.3. Proof of Theorem 5

Theorem 5. The implication problem for CINDs is PSPACE-complete in the absence of attributes with finite domains.

Proof. It is known that the implication problem for INDS is PSPACE-complete in the absence of finite-domain attributes [20]. Since CINDs subsume INDS for CINDs is also PSPACE-hard.

We next show that the implication problem for CINDs is in PSPACE in the absence of finite-domain attributes. We show this by giving a linear space non-deterministic algorithm for deciding whether $\Sigma \models \psi$, along the same lines as its counterpart for INDS (see [13,20]). If this holds, then by Savitch’s theorem [21], there is a deterministic quadratic-space algorithm for checking whether $\Sigma \models \psi$, and therefore, the implication problem is in PSPACE.

Indeed, the chase procedure developed in the proof of Theorem 2 gives such an algorithm. Consider a set $\Sigma \cup \{\psi\}$ of CINDs over a database schema $\mathcal{R} = (R_1, \ldots, R_n)$, where $\psi = (R_1[A_1, \ldots, A_n, Y_p], t_{p_1})$. Recall that the chase process starts with an initial database $D_0$, which contains a single tuple $t_0 = t_0$ such that $t_0[A_i] = v_i$ for all $i \in [1, m]$ and $t_0[X_p] = t_{p_1}[X_p]$. As we have seen in the proof of Theorem 2, Chase($\Sigma$, $\psi$) $\models \psi$, where Chase($\Sigma$, $\psi$) denotes an arbitrary fixpoint (database) generated by the chase process. Moreover, if $\Sigma \models \psi$, then there must exist a tuple $t_0 = t_0$ in Chase($\Sigma$, $\psi$) such that $t_0[B_i] = v_i$ for $i \in [1, m]$, and $t_0[Y_p] = t_{p_1}[Y_p]$. More specifically, there exists a finite sequence $(t_0, \ldots, t_i)$ of tuples such that $t_0 = t_0$, $t_i = t_i$, and for each $i \in [0, l - 1]$, $t_{i+1}$ is obtained by applying a CIND $\psi'$ in $\Sigma$ to $t_i$.

Based on the analysis above, we develop a linear space non-deterministic algorithm:

- Initialize a single tuple $t_0 := t_0$.
- Replace tuple $t_i$ with another tuple $t_{i+1}$ if $t_{i+1}$ can be derived from $t_i$ by applying a CIND $\psi'$ in $\Sigma$ to $t_i$ by using rules IR1–IR6. There are possibly multiple such $t_{i+1}$’s. The algorithm non-deterministically picks one of them.
- Repeat these steps until no more changes can be made.
- If tuple $t_b \in I_b$, return 'yes'; and return 'no' otherwise.

As shown in the proof of Theorem 2, if $\Sigma \models \psi$, then $t_b$ is in Chase($\Sigma$, $\psi$) and hence, the algorithm returns 'yes'. Conversely, if the algorithm returns 'yes', i.e., $t_b$ is in Chase($\Sigma$, $\psi$), then by Claim 1, $\Sigma \not\vdash \bot$. By Theorem 2, $\Sigma \models \psi$. Hence the algorithm correctly determines whether $\Sigma \models \psi$. The algorithm is obviously in PSPACE, as it only stores at most a single tuple at any time. As a result, the implication problem for CINDs is in PSPACE in the absence of finite-domain attributes. \hfill \square

### A.4. Proof of Theorem 7

Before we prove Theorem 7, we first examine the chase process introduced in the proof of Theorem 3. Given a set $\Sigma \cup \{\psi\}$ of CINDs, the chase procedure inspects whether $\Sigma \models \psi$. Below we give its computational complexity.

**Lemma 1.** Given a set $\Sigma \cup \{\psi\}$ of CINDs defined on a database schema $R$, the chase procedure given in the proof of Theorem 3 terminates in $O(2^{2n^2})$ time, where $n$ is the size of the input, i.e., the size of $R$, $\Sigma$ and $\psi$.

**Proof.** The chase procedure is obviously in $O(|T_f|)$ time, where $T_f$ is a result Chase($\Sigma$, $\psi$) of the chase process, and $|T_f|$ is the number of nodes in $T_f$. Recall that each root-to-leaf path of $T_f$ represents a database schema $R$. Hence the depth of $T_f$ is bounded by the maximum size $|I|$ of a database instance of $R$. Moreover, the maximum number of children of a node in $T_f$ is also bounded by $|I|$. We show that $|I|$ is in $O(2^{2n^2})$ time as follows.

- The cardinality of a finite domain in $R$ is determined by the schema $R$ and is hence bounded by $n$.
- For an infinite domain in $R$, the chase process uses only those constants appearing in the patterns in $\Sigma$ or $\psi$, and the finite set $\{v_0, \ldots, v_m\}$ of variables (bounded by the size of $\psi$). These are also bounded by the input size $n$.

Hence the size $|I|$ is bounded by $O(n^k) = O(2^{\log(n^k)}) = O(2^{2n^2})$. And therefore $|T_f|$ is in $O((2^{2n^2})^2) = O(2^{2n^2 \cdot 2n^2}) = O(2^{2n^2 \cdot 2n^2}) = O(2^{2n^4})$ time. \hfill \square

We are now ready to give the complexity bound for the implication analysis of CINDs in the general setting.

**Theorem 7.** In the general setting, the implication problem for CINDs is EXPTIME-complete.

**Proof.** (1) We first show that the problem is in EXPTIME. Given a set $\Sigma \cup \{\psi\}$ of CINDs on a relational schema $R$, we develop an algorithm in $O(2^{2n^4})$ time, where $k$ is a constant and $n$ is the size of $R$, $\Sigma$ and $\psi$, such that it returns 'yes' if and only if $\Sigma \models \psi$. Assume that $R = (R_1, \ldots, R_n)$, and that $\psi = (R_a[A_1, \ldots, A_m] \subseteq R_b[B_1, \ldots, B_m] Y_p \mathrel{\negatemark} t_p \negatemark Y_p)$. Then $\psi$ is in $\Sigma \cup \{\psi\}$.

We now develop an EXPTIME algorithm based the chase procedure. The main idea is to maintain a directed graph $G(V, E)$ and a mapping $H$. Given a node $u$ in $V \subseteq C$, $H(u)$ is the set of CINDs in $\Sigma$ that have already been applied to the node $u$ in the chase process. With these two data structures, we can avoid unnecessary computations.

Below we first present the algorithm, and then verify the correctness of the algorithm. Finally, we show that the algorithm is in exponential-time.

We first present the algorithm.

(a) It initializes the node set $V := \{N_{root}\}$, the edge set $E := \emptyset$, and $H(N_{root}) := \emptyset$. Here the node $N_{root}$ is the root node $R_a : t_a$ of a tree $T$ in the chase process.

(b) For each node $u = \langle R_i : t_i \rangle$ in $V$, it checks whether there exists a CIND $\psi' = (R_j[U \subseteq R_j[V, \rho(V_j)], \rho(V_j)]$ in $\Sigma$, but not in $H(u)$, such that $t_i[U] = \rho(V_j)[U]$.

(c) If there exists such a CIND $\psi'$ for the node $u = \langle R_i : t_i \rangle$, it first generates a set $S_{new}$ of new nodes, and then updates the graph $G$ and the mapping $H$ accordingly.

The set $S_{new}$ is generated along the same lines as the process in the proof of Theorem 3.

- Let $V_f$ be the list of all finite-domain attributes in $\text{attr}(R_j) \setminus (V \cup V_p)$, and $\rho(V_f)$ be an instantiation of $V_f$. That is, for each attribute $C \in V_f$, $\rho(C)$ is a data value drawn from the finite domain $\text{dom}(C)$. 


For each possible instantiation $\rho(V_f)$, it generates a new node $u'_{\rho} := \langle R_b : t_p \rangle$ such that $t_p[V] = t_l[U]$, $t_p[V_f] = t_p[Y_p]$, $t_p[V_f] = \rho(V_f)$, and $t_p[C] = v_0$ for all the other attributes $C$ in $\text{attr}(R_f)$.

Then, for the mapping $H$, $H(u) := H(u) \cup \{ \psi' \}$, and for each node $u'_{\rho}$ in $S_{\text{new}}$, but not in $V$, $H(u'_{\rho})$ is empty. For the graph $G(V,E)$, its node set $V$ is updated to be $V \cup S_{\text{new}}$, and its edge set $E$ is updated as follows. Let $S_{\text{nbr}} = [u_1, \ldots, u_k]$ be the set of neighboring nodes of the node $u$ such that for each node $u_l \ (l \in [1, k])$, there is an edge $(u, u_l)$ in $E$.

- If $S_{\text{nbr}}$ is empty, then for each node $u'_{\rho}$ in $S_{\text{new}}$, it simply adds an edge $(u, u'_{\rho})$ to $E$.
- Otherwise, for each node $u_l \ (l \in [1, k])$ in $S_{\text{nbr}}$, and each node $u'_{\rho} := \langle R_b : t_p \rangle$ in $S_{\text{new}}$, (a) if $R_j = R_b$, $t_p[B_1, \ldots, B_m] = (v_1, \ldots, v_m)$ and $t_p[Y_p] = t_p[Y_p]$, it replaces the edge $(u, u_l)$ with a new edge $(u, u'_{\rho})$, and (b) if not, it replaces the edge $(u, u_l)$ with two new edges $(u, u')$ and $(u', u_l)$.

(d) The above process repeats until there are no more changes to the node set $V$ and the edge set $E$ of the graph $G$. We denote the final resulting graph as $G_f$.

(e) The algorithm finally checks whether $\Sigma \models \psi$ based on the graph $G_f$.

Let $S_a$ be the set of nodes in $G_f$ that are reachable from the node $N_{\text{root}} = \langle R_a : t_a \rangle$ such that $t_a[A_1, \ldots, A_m] = (v_1, \ldots, v_m)$ and $t_a[X_p] = t_p[X_p]$.

Let $S_b$ be the set of nodes $\langle R_b : t_b \rangle$ in $G_f$ such that $t_b[B_1, \ldots, B_m] = (v_1, \ldots, v_m)$ and $t_b[Y_p] = t_p[Y_p]$.

The algorithm checks if there exists a node $u$ in $S_a$ such that no nodes $u'$ in $S_b$ are reachable from $u$, i.e., there exists no path $(u, u')$ in graph $G_f$.

If there exists such a node, it returns ‘no’, and returns ‘yes’ otherwise.

We next show that $\Sigma \models \psi$ iff the algorithm returns ‘yes’. Indeed, the algorithm simulates the chase procedure given in the proof of Theorem 3. If it returns ‘no’, one can readily expand $G_f$ into a tree, which represents a database instance $D$ (see the proof of Theorem 3) such that $D \models \Sigma$, but not $D \models \psi$, i.e., $\Sigma \not\models \psi$. If it returns ‘yes’, then $\Sigma \models \psi$ by Claim 2 given in the proof of Theorem 3.

To see that the algorithm is in exponential time, observe the following. (a) The number of nodes in the graph $G_f$ is bounded by the maximum size $|I|$ of a database instance. Therefore, the size of $G_f$ is bounded by $O(2|\Sigma|^2)$ as argued in the proof of Lemma 1. (b) For the set $\Sigma$ of CINDs, the number of equivalent CINDs in the special form is bounded by $|\Sigma| = O(2^{|\Sigma|})$ as indicated in the proof of Theorem 3, and the number of CINDs equivalent to $\psi$ in the special form is also bounded by $|I| = O(2^{|\Sigma|})$. (c) The size of $S_a$ is bounded by the number of nodes in $G_f$, i.e., in $O(2^{|\Sigma|})$; so is the size of $S_b$. From these it follows that graph $G_f$ can be constructed in $O(2^{2n^2} \cdot 2^{|\Sigma|}) = O(2^{|\Sigma|})$ time, and the reachability between nodes in $S_a$ and nodes in $S_b$ can be checked in $O(2^{|\Sigma|} \cdot 2^{n^2} \cdot (2^{|\Sigma|})^2) = O(2^{|\Sigma|})$ time [29]. Based on these one can see that the algorithm is indeed in EXPTIME.

(2) We next show that the problem is EXPTIME-hard, by reduction from the two-player game of corridor tiling problem (TPG-CT), which is EXPTIME-complete [22,23].

An instance of TPG-CT consists of a tiling system $(X, H, V, \tilde{t}, \tilde{b})$ and a positive integer $n$, where $X$ is a finite set of tiles (dominoes), $H, V \subseteq X \times X$ are two binary relations, $\tilde{t}$ and $\tilde{b}$ are two $n$-vectors of given tiles in $X$, and $n$ is the number of columns (the width of the corridor). It is to determine whether or not player I has a winning strategy for tiling the corridor.

By tiling the corridor we mean that there exists a tiling $\tau : \mathbb{N} \times \mathbb{N} \rightarrow X$ and a positive integer $m$ such that for all $x \in [1, n]$ and $y \in [1, m]$, the tiling adjacency conditions are observed, i.e.,

- if $\tau(x, y) = d$ and $\tau(x + 1, y) = d'$, then $(d, d') \in H$, i.e., horizontally adjacent tiles have matching “colors”;
- if $\tau(x, y) = d$ and $\tau(x, y + 1) = d'$, then $(d, d') \in V$, i.e., vertically adjacent tiles have matching colors; and
- $\tau(x, 1) = \tilde{t}(x)$ and $\tau(x, m) = \tilde{b}(x)$, where $\tilde{t}(x)$ (resp. $\tilde{b}(x)$) denotes the $x$-th element of the vector $\tilde{t}$ (resp. $\tilde{b}$); that is, the given tiles of $\tilde{t}$ and $\tilde{b}$ are placed on the top and the bottom rows, respectively. The given tiles $\tilde{t}$ are placed on the top row by the referee of the game.

Each player in turn places a tile from $X$ in the first free location (column by column from left to right, and row by row from top to bottom), observing the tiling adjacency conditions. Player I wins if either Player II makes an illegal move by placing a tile that violates one of the adjacent conditions, or if the bottom row $\tilde{b}$ is placed. Player I has a winning strategy iff Player I can always win no matter how Player II plays. The problem is already EXPTIME-complete when $n$ is odd [22,23], and thus below we assume that $n$ is an odd number, and that Player I makes the first move.

Given an instance of TPG-CT $(X, H, V, \tilde{t}, \tilde{b})$ and $n$, we define a relational schema $\mathcal{R}$, a set of CINDs $\Sigma$ and a CIND $\psi$ such that $\Sigma \not\models \psi$ if and only if there is a winning strategy for Player I. If this holds, then the problem is EXPTIME-hard. Indeed, this problem is the complement problem of the implication problem, from which it follows that the implication problem for CINDs is also EXPTIME-hard.
### Theorem 7

(A) The database schema $\mathcal{R}$ consists of two relation schemas $R(K, A_0, A_1, \ldots, A_n, \text{next}, P, Z)$ and $S(B)$, where for all $i \in [0, n]$, $A_i$ has a finite domain $\text{dom}(A_i) = X$, $P$ has a finite domain $\text{dom}(P) = [1, \ldots, n]$, the domains of attributes $K$ and next are positive integers, and $Z$ has a finite domain with two symbols # and !. The attribute $B$ has a finite domain consisting of two distinct values: $c$ and $b$.

Intuitively, an $R$ tuple $t$ encodes a placement of tiles in the game. More specifically, tuple $t$ is the snapshot of the game showing the last $n + 1$ plays, where (a) $t[A_n]$ is the new tile placed by a player, (b) $t[A_0, \ldots, A_{n-1}]$ consists of the $n$ tiles placed before $t[A_n]$, (c) $t[P]$ codes the horizontal position of tile $t[A_n]$ in a row, (d) $t[K]$ and $t[\text{next}]$ encode a list of such snapshots: $t[K]$ is the “identifier” of the current snapshot, while $t[\text{next}]$ is a pointer to the next one (see Fig. 6). In addition, $t[Z]$ indicates that the game continues when it is #, and that the game should stop if it is !. An illegal move by Player I is indicated by the presence of a tuple $s$ of schema $S$ with $s[B] = 'c'$. We want to show that there exists an instance $D = (R_I, I_S)$ of $\mathcal{R}$ such that $D$ satisfies $\psi$, but not $\psi'$, if and only if Player I has a winning strategy.

(B) We next define the set $\Sigma$ of CINDs that encodes the play. We use $N_{\text{odd}}$ and $N_{\text{even}}$ to denote the set of even numbers and the set of odd numbers in $[1, n]$, indicating the moves of Player I and Player II, respectively.

**Initial condition.** The top row of the corridor has to be set to $\overline{t}$. We use a CIND $\psi_1$ to ensure that if $I_S$ is nonempty, then there exists an $R$ tuple $t$ such that $t[A_0, \ldots, A_{n-1}]$ matches $\overline{t}$.

$$\psi_1 = (S[\text{nil}] \subseteq R[\text{nil}; A_0; \ldots, A_{n-1}; P, Z], t_{p_{\psi_1}})$$ where $t_{p_{\psi_1}}[A_0, \ldots, A_{n-1}] = \overline{t}$ and $t_{p_{\psi_1}}[P, Z] = (1, \#)$.

**Adjacency constraints.** The vertical and horizontal tiling conditions must hold. For each $R$ tuple $t$, $t[A_n]$ corresponds to the tile directly under tile $t[A_0]$, and $t[A_n]$ is the tile placed next to $t[A_{n-1}]$ in a row. Thus for each tuple $t \in I_R$, we must ensure that $(t[A_0], t[A_n]) \in V$ and that $(t[A_{n-1}], t[A_n]) \in H$.

The adjacency constraints are enforced by two sets $\Sigma_V$ and $\Sigma_H$ of CINDs below. These CINDs assert the following: (a) if Player I makes an illegal move, then $I_S$ contains a tuple (‘c’); and (b) if any player makes an illegal move, then the game should stop, by adding an $R$ tuple $t'$ with $t'[Z] = !$.

$$\Sigma_V: \quad (R[\text{nil}; A_0, A_n, P] \subseteq S[\text{nil}; B], \psi_{V(x,y)},)$$

where $t_{V(x,y)}[A_0, A_n, h \parallel B] = (x, y, h \parallel c)$ for all $(x, y) \in (X \times X) \setminus V$ and all $h \in N_{\text{odd}}$.

$$R[\text{next}, A_1, \ldots, A_{n-1}; A_0, A_n, P] \subseteq R[K, A_0, \ldots, A_{n-2}; P, A_{n-1}, Z], \psi_{V(x,y)},)$$

where $t_{V(x,y)}[A_1, \ldots, A_{n-1}; A_0, A_n, P] = (x, y, h \parallel h + 1, y, !)$ for all $h \in [1, n - 1]$ and all $(x, y) \in (X \times X) \setminus V$.

$$\Sigma_H: \quad (R[\text{nil}; A_{n-1}, A_n, P] \subseteq S[\text{nil}; B], \psi_{H(x,y)},)$$

where $t_{H(x,y)}[A_{n-1}, A_n, h \parallel B] = (x, y, h \parallel c)$ for all $(x, y) \in (X \times X) \setminus H$ and all $h \in N_{\text{odd}}$.

$$R[\text{next}, A_1, \ldots, A_{n-2}; A_{n-1}, A_n, P] \subseteq R[K, A_0, \ldots, A_{n-3}; P, A_{n-2}, A_{n-1}, Z], \psi_{H(x,y)},)$$

where $t_{H(x,y)}[A_1, \ldots, A_{n-2}; A_{n-1}, A_n, P] = (x, y, h \parallel h + 1, x, y, !)$ for all $h \in [1, n - 1]$ and all $(x, y) \in (X \times X) \setminus H$.

Player I has to respond to all possible legal moves of Player II. For each $R$ tuple $t_1$ that is a legal move, if $t_1[P]$ is even, i.e., if the last move $t_2[A_n]$ was made by Player II, then for each tile $x \in X$ that satisfies the horizontal constraint $(t_1[A_n], x) \in H$ and the vertical constraint $(t_2[A_0], x) \in V$, there must exist a tuple $t_2$ in $I_R$ with $t_2[K] = t_1[\text{next}], t_2[A_{n-1}] = x$ and moreover, $t_2[A_i] = t_1[A_{i+1}]$ for $i \in [0, n - 1]$. That is, all possible legal moves of Player II have to be considered. We encode this with a set $\Sigma_v$ of CINDs.

$$\Sigma_V: \quad (R[\text{next}, A_1, A_2, \ldots, A_{n-2}; P, A_0, A_{n-1}, A_n, Z] \subseteq R[K, A_0, A_1, \ldots, A_{n-2}; P, A_{n-1}, Z], \psi_{V(x,y,w),})$$

where $t_{V(x,y,w)} = (h, x, y, w, \# \parallel h + 1, w, \#)$, for all $h \in N_{\text{even}}$, and for all $(x, w) \in V$ and $(y, w) \in H$.
Play continues unless Player I has won. For each R tuple $t_1$, if $t_1[P] < n$ and $t_1[Z] = \#$, then there must exist some tuple $t_2$ such that $t_2[K] = t_1[next]$, $t_2[A_0, \ldots, A_{n-1}] = t_1[A_1, \ldots, A_{n}]$ and $t_2[P] = t_1[P] + 1$. If $t_1[P] = n$ and if the bottom vector $\hat{b}$ is not matched, i.e., if for some $i \in [1, n]$, $t_1[A_i] \neq \hat{b}[i]$, then there must exist some $t_2$ such that $t_2[K] = t_1[next]$, $t_2[A_0, \ldots, A_{n-1}] = t_1[A_1, \ldots, A_{n}]$ and $t_2[P] = 1$. We express this as a set $\Sigma_\psi$ consisting of the following CINDs:

$$\varphi_h = \{ R[next, A_1, \ldots, A_n; P, Z] \subseteq R[K, A_0, \ldots, A_{n-1}; P, Z], t_{ph} \},$$

where for each $h \in [1, n - 1]$, $t_{ph} = (h, \# || h + 1, \#)$.

$$\varphi(i, x) = \{ R[next, A_1, \ldots, A_i, A_{i+2}, \ldots, A_n; P, A_{i+1}, Z] \subseteq R[K, A_0, \ldots, A_{i-1}, A_{i+1}, \ldots, A_{n-1}; P, A_i, Z], t^{i,x}_{\psi} \},$$

where $t^{i,x}_{\psi} = (n, x, \# || 1, x, \#)$, for all $i \in [1, n]$ and all $x \in X \setminus \{ \hat{b}[i] \}$.

Observe that if an illegal move was made, a new move $t_2$ is added with $t_2[Z] = !$ and $t_2[P] = h + 1$, by $\Sigma_H$ or $\Sigma_V$.

The set $\Sigma$ consists of all the CINDs given above, i.e., $\Sigma = \{ \varphi_1 \cup \Sigma_V \cup \Sigma_H \cup \Sigma_P \cup \Sigma_Y \}$. The number of CINDs in $\Sigma$ is bounded by a polynomial of $n$ and the number of tiles in $X$.

(C) We define CIND $\psi = (S(nil, \#) \subseteq S(nil, B), (\text{nil} \parallel c))$. Intuitively, if $D \neq \psi$ then (a) $I_S$ is nonempty and (b) there exists no tuple in $I_S$ such that $t(B) = \cdot c$. That is, Player I does not make illegal move.

The reduction is obviously in polynomial time. We next verify that Player I has a winning strategy iff $\psi \neq \psi$.

First, suppose that $\Sigma \models \psi$. Then there exists an instance $D = (I_R, I_S)$ of $\mathcal{R}$ such that $D \models \Sigma$, but $D \neq \psi$. We give a winning strategy for Player I. We begin with the tuple in $R$ enforced by $\varphi_1 \in \Sigma$. Such a tuple must exist because $\Sigma$ must be nonempty (by $D \neq \psi$) and hence, $\varphi_1$ is applicable. At any step in the game, there is a tuple in $I_R$ that represents the last $n + 1$ moves of the play thus far. Each valid tile $x_j$ that Player II places as the next move, represented by $t$, the CINDs in $\Sigma_V$ ensure the existence of a tuple $t'$ with $t[next] = t'[K]$ and $t'[A_{n-1}] = x_j$ and $t'[A_i] = t[A_{i+1}]$ for $i \in [0, n - 1]$, i.e., a response from Player I. By $D \neq \psi$, and $D \models \Sigma_H \cup \Sigma_V \cup \Sigma_Y$, the move $t'$ satisfies both the horizontal and the vertical constraints, and it also correctly updates the last $n + 1$ tiles played. Furthermore, by $D \models \Sigma_P$, the play continues until Player I wins. Thus Player I has a winning strategy.

Conversely, suppose that Player I has a winning strategy. We then form an instance $D = (I_R, I_S)$ of $\mathcal{R}$ such that $I_S$ consists of all valid plays in any game, where each tuple codes the horizontal position of the last move in a row and the last $n + 1$ tiles played in the game, and $I_S$ has a single tuple $\tau$ such that $t(B) = \cdot b$ (i.e., Player I makes no illegal move). It is easy to correct that $D \models \Sigma$, but $D \neq \psi$. \qed

A.5. Proof of Theorem 12

Theorem 12. In the general setting, the implication problem for ACINDs is PSPACE-complete.

Proof. (1) We first show that the problem for ACINDs is in PSPACE in the general setting. We show this by giving a linear space non-deterministic algorithm for determining whether $\Sigma \models \psi$, i.e., the complement of $\Sigma \models \psi$. This suffices. For if it holds, then (a) by the Immerman-Szelepcsényi theorem [27, 28], there exists a linear space non-deterministic algorithm for determining whether $\Sigma \models \psi$; and (b) by Savitch’s theorem [21], there is a deterministic quadratic-space algorithm for checking whether $\Sigma \models \psi$. From these it follows that the problem is in PSPACE.

Consider a set $\Sigma \cup \{ \psi \}$ of acyclic CINDs defined over a database schema $\mathcal{R} = (R_1, \ldots, R_m)$. Assume w.l.o.g. that for any two schemas $R_i$ and $R_j$ (i.e., $i \neq j$), there exist no CINDs in the form of $(R_i[U; U_p] \subseteq R_j[V; V_p], t_p)$ such that $j < i$. We observe that we can always rearrange the schemas in $\mathcal{R}$ to satisfy this condition since the CINDs in $\Sigma$ are acyclic. We also assume w.l.o.g. that the CIND $\psi$ is $(R_1[A_1, \ldots, A_m; X_p] \subseteq R_1[B_1, \ldots, B_m; Y_p], t_{\psi};)$; the proof is similar for the cases where $\psi$ is from $R_i$ to $R_j$ when $i \neq 1$ or $j \neq n$.

The proof consists of three parts. (a) We first introduce notations to be used. (b) We then present the linear space non-deterministic algorithm. (c) Finally, we show that the algorithm is correct and that it runs in PSPACE.

(1)1) Before we present the algorithm, we first introduce the following notations to be used in the algorithm.

(a) Let $\{ \Sigma_1, \ldots, \Sigma_{n-1} \}$ be the partition of $\Sigma$ such that $\bigcup_{i=1}^{n-1} \Sigma_i = \Sigma$, and for each $i \in [1, n - 1]$, $\Sigma_i$ is the set of CINDs of the form $(R_i[U; U_p] \subseteq R_j[V; V_p], t_p)$ such that $j \in [i + 1, n]$.

(b) The number of CINDs in $\Sigma_i$ (i.e., $i \in [1, n - 1]$) is denoted as $n_i$. We assume w.l.o.g. that for each $i \in [1, n - 1]$, $n_i > 0$, i.e., there exists at least one CIND in $\Sigma_i$.

(c) All CINDs in $\Sigma_i$ (i.e., $i \in [1, n - 1]$) are sorted in an arbitrary order. We use a pointer $p_i$ to indicate the $p_i$-th CIND in $\Sigma_i$.

(1.2) We now present the linear space non-deterministic algorithm for determining whether $\Sigma \models \psi$.

(a) It guesses an instantiation $\rho_1$ for the list $X_f$ of all finite-domain attributes in $\text{attr}(R_1) \setminus (X_p)$, where for each attribute $C$ in $X_f$, $\rho_1(C)$ is a value drawn from the finite domain $\text{dom}(C)$.

(b) It initializes a database instance $D := (t_1, \ldots, t_n)$, such that for each $i \in [1, n]$, $t_i$ is an empty instance of schema $R_i$, except that $t_1$ contains a single tuple $t_0$ of schema $R_1$, where $t_1[A_1, \ldots, A_m] = (v_1, \ldots, v_m)$, $t_1[X_f] = t_p[X_p]$, and $t_1[X_f] = (1, \ldots, 1)$.
\[ \rho_1[X] \] for the list \( X \) of all finite-domain attributes in \( \text{attr}(R_1) \setminus \{X_p\} \). Here \( v_1, \ldots, v_m \) are \( m \) distinct variables. Intuitively, \( t_1 \) encodes the LHS of the CIND \( \psi \), and is to serve as a “witness” for \( D \not\models \psi \). We assume w.l.o.g. that \( X \setminus \{A_1, \ldots, A_m\} \) is empty since if not, in the process to be seen shortly, we can simply replace the variable with the constant \( \rho[A] \) for each attribute \( A \) in \( X \setminus \{A_1, \ldots, A_m\} \).

The algorithm will ensure that for each \( i \in [1, n] \), the instance \( I_i \) contains at most one tuple, which guarantees that the algorithm uses only linear space.

(c) Starting with \( p_1 := 1 \) for all pointers \( p_i \) in \( P \) (\( i \in [1, n-1] \)), the algorithm processes CINDs in \( \Sigma_{p, \psi} \) one by one, as follows.

We set \( i = 1 \), and let \( \phi \) be the \( p_i \)-th CIND \( (R_i[U; U_p] \subseteq R_j[V; V_p], t_{p_j}) \) in \( \Sigma_i \).

- Guess an instantiation \( \rho_j \) for the list \( V \) of all finite-domain attributes in \( \text{attr}(R_j) \setminus (V \cup V_p) \), in which for each \( C \) in \( V \), \( \rho_j(C) \) is a value drawn from the finite \( \text{dom}(C) \).
- If there is a tuple \( t_i \) in \( I_i \), but there exists no tuple \( t_j \) in \( I_j \) such that \( t_j[V] = t_i[U] \), \( t_j[V_p] = t_{p_j}[V_p] \), it first creates a new tuple \( t'_j \) such that \( t'_j[V] = t_i[U] \), \( t'_j[V_p] = t_{p_j}[V_p] \), and \( t'_j[V_f] = \rho_j(V_f) \). Then updates the instance \( I_j := t'_j \) to contain this new tuple.
- If \( i < (n-1) \), it increases the variable \( i \) by 1, and repeats the process above.
- Otherwise, it checks whether there exists a tuple \( t_n \) in \( I_n \) such that \( t_n[B_1, \ldots, B_m] = t_1[A_1, \ldots, A_m] = (v_1, \ldots, v_m) \) and \( t_n[Y_p] = t_{p_j}[Y_p] \). Observe that the current database instance \( D_p \models \Sigma_{p, \psi} \).

(d) If there exists a pointer \( p_j \) (\( 1 \leq j \leq n-1 \)) such that \( p_j \neq n_j \), the algorithm then adjusts the list \( P = [p_1, \ldots, p_{n-1}] \) of pointers by the following pseudo-codes.

\[
\begin{align*}
&\text{let } j := n - 1, \text{ and increase } p_n-1 \text{ by 1;} \\
&\text{while } (j > 1) \text{ do} \\
&\quad \text{if } p_j = (n_j + 1) \text{ then} \\
&\quad \quad \text{let } p_j := 1, \text{ and increase } p_{j-1} \text{ by 1;} \\
&\quad \text{decrease the variable } j \text{ by 1.}
\end{align*}
\]

(e) If there exist no such pointers in \( P \), i.e., \( p_j = n_j \) for all \( j \in \{1, n-1\} \), the algorithm stops and returns ‘yes’ if the tuple \( t_n \) is not found in all the cases of the list \( P \) of pointers, i.e., from \([1, \ldots, n-1] \) to \([n_1, \ldots, n_{n-1}] \). Otherwise, the algorithm starts again from step (a).

(1.3) We next show that the algorithm is in PSPACE and that it is correct.

Observe that at any step of the process, the database instance \( D \) contains at most \( n \) tuples, where \( n \) is the number of relation schemas in \( \mathcal{R} \). Hence, it is obvious that the non-deterministic algorithm runs in linear space.

To show the correctness of the algorithm, first observe the following.

(a) The algorithm examines each combination of those CINDs in all \( \Sigma_i \)’s (\( i \in [1, n-1] \)), represented by the pointer list \( P = [p_1, \ldots, p_{n-1}] \) of pointers. Recall that for each \( i \in [1, n-1] \), the pointer \( p_i \) denotes the \( p_i \)-th CIND in \( \Sigma_i \).

For acyclic CINDs \( \Sigma \), this suffices to determine whether \( \Sigma \models \psi \).

(b) The algorithm examines various instantiation of variables for finite-domain attributes, by backtracking. More specifically, it changes the list \( P = [p_1, \ldots, p_{n-1}] \) of pointers starting from the last pointer \( p_n-1 \), and does not change \( p_i \) until \( p_j = n_j \) for each \( j > i \) (recall that for each \( i \in [1, n-1] \), \( n_i \) is the number of CINDs in \( \Sigma_i \)). This allows us to avoid random valuations of finite-domain attributes and moreover, to use the same space in the entire process.

Having seen these, we finally show the correctness of the algorithm, i.e., it returns ‘yes’ if and only if \( \Sigma \models \psi \).

First assume that \( \Sigma \models \psi \). Then there exists a database instance \( D = (I_1, \ldots, I_n) \) of \( \mathcal{R} \) such that \( D \models \Sigma \) but \( D \not\models \psi \). By the definition of \( \psi \), there must exist a tuple \( t_1 \) in the instance \( I_1 \) of schema \( R_1 \) such that \( t_1[X_p] = t_{p_1}[X_p] \), but there exists no tuple \( t_n \) in the instance \( I_n \) of schema \( R_n \) such that \( t_n[Y] = t_1[X] \) and \( t_n[Y_p] = t_{p_1}[Y_p] \). If we choose the instantiation \( \rho_1(X_f) = t_1[X_f] \) at step (a), and the instantiations \( \rho_j(V_f) \) at step (c) based on the instance \( D \), then for each combination of the list \( P \) of pointers, there exists no tuple \( t_n \) in the instance \( I_n \) of schema \( R_n \) such that \( t_n[Y] = t_1[X] \) and \( t_n[Y_p] = t_{p_1}[Y_p] \). Thus the algorithm must stop and return ‘yes’.

Conversely, assume that the algorithm returns ‘yes’. We construct a nonempty database instance \( D \) of \( \mathcal{R} \) such that \( D \models \Sigma \) but \( D \not\models \psi \). Let \( D \) be the union of all database instances \( D_p \) at step (c), where \( D_p \models \Sigma_{p, \psi} \). Then as observed earlier, \( D \models \Sigma \) and \( D \not\models \psi \).

(2) We next show that the problem is PSPACE-hard by reduction from the Q3SAT problem, which is PSPACE-complete [cf. [29]].

An instance of Q3SAT is a first-order logic sentence \( \theta = \forall x_1 \exists x_2 \exists x_3 \cdots \exists x_m x_m \phi \), where \( Q_m \) is \( \forall \) if \( m \) is odd and it is \( \exists \) if \( m \) is even; \( \phi = C_1 \land \cdots \land C_n \) is an instance of the 3SAT problem in which all the variables are \( x_1, \ldots, x_m \), and for each \( j \in [1, n] \), the clause \( C_j \) is \( y_{i_1} \lor y_{i_2} \lor y_{i_3} \) such that for \( i \in [1, 3] \), \( y_{i} \) is either \( x_{p_{i}} \) or \( \overline{x_{p_{i}}} \) for \( p_i \in [1, m] \). Here we use \( x_{p_i} \) to denote the occurrence of a variable in the literal \( i \) of clause \( C_j \). The Q3SAT problem is to decide whether \( \theta \) is true.

Given an instance \( \theta \) of Q3SAT, we construct an instance of the implication problem for acyclic CINDs, which consists of a database schema \( \mathcal{R} \) with finite-domain attributes, a set \( \Sigma \) of acyclic CINDs defined on \( \mathcal{R} \) and another CIND \( \psi \) on \( \mathcal{R} \). We show that \( \Sigma \not\models \psi \) if and only if \( \theta \) is true. This suffices. For it holds, then it is also PSPACE-hard to decide whether \( \Sigma \models \psi \).
The database schema $\mathcal{R}$ consists of $m + 2$ relation schemas $R_0(A_1, A_2), \ldots, R_m(A_1, \ldots, A_m)$, and $S(B)$. All the attributes in $\mathcal{R}$ have a finite domain $\{0, 1\}$. Intuitively, each $R_i$ is to encode a quantifier in $\theta$, which is either $\forall$ or $\exists$, for each $i \in [1, m]$. In an instance $I_m$ of schema $R_m$, each tuple $t[A_1, \ldots, S_i]$ is to carry a truth assignment of the variables $\{x_1, \ldots, x_m\}$ in $\theta$. In addition, we shall use an instance of $S$ to indicate whether $\theta$ is satisfied.

(B) The set $\Sigma$ contains the following ACINDs.

- For each odd number $1 \leq i \leq m$, we define two CINDs $\psi_{i,0}$ and $\psi_{i,1}$ from $R_{i-1}$ to $R_i$ as follows:

$$\psi_{i,0} = \left(R_{i-1}[A_1, \ldots, A_{i-1}; nil] \subseteq R_i[A_1, \ldots, A_{i-1}; \{nil \mid 0\}], \{nil \mid 1\} \right).$$

$$\psi_{i,1} = \left(R_{i-1}[A_1, \ldots, A_{i-1}; nil] \subseteq R_i[A_1, \ldots, A_{i-1}; \{nil \mid 1\}] \right).$$

When $i = 1$, $R_{i-1}[A_1, \ldots, A_{i-1}]$ is $R_0[\text{nil}]$.

- For each tuple $t$ in an $R_i$ relation ($i \in [1, 3, \ldots, m-1]$), there exist two tuples $t_0$ and $t_1$ in the $R_i$ relation such that $t_0[i] = 0$ and $t_1[i] = 1$.

Intuitively, we encode a universal quantifier $\forall$ by using these CINDs.

- For each even number $1 < i \leq m$, we define a CIND $\psi_i$ from $R_{i-1}$ to $R_i$:

$$\psi_i = \left(R_{i-1}[A_1, \ldots, A_{i-1}; nil] \subseteq R_i[A_1, \ldots, A_{i-1}; \{nil \mid nil\}] \right).$$

This CIND ensures that for each tuple $t$ in an $R_i$ relation ($i \in [2, 4, \ldots, m-2]$), there exists a tuples $t'$ in the $R_i$ relation such that $t'[A_1, \ldots, A_{i-1}] = t[A_1, \ldots, A_{i-1}]$ while $t'[A_i] = 0$ or $1$.

Intuitively, we encode an existential quantifier $\exists$ by using these CINDS.

- For each clause $C_j = y_{j_1} \lor y_{j_2} \lor y_{j_3}$ in the 3SAT instance $\phi$, we define CIND $\psi_{S,j}$ from $R_m$ to $S$:

$$\psi_{S,j} = \left(\{nil; A_{p_1,j}, A_{p_2,j}, A_{p_3,j}\} \subseteq S[nil, B]; \{t_{p_{m+1},j}[B] = 0, \text{for each } i \in [1, 3], t_{p_{S,j}}[A_{p_i}] = \xi_j(x_{p_i})\}. \right)$$

where $t_{p_{m+1},j}[B] = 0$, and for each $i \in [1, 3]$, $t_{p_{S,j}}[A_{p_i}] = \xi_j(x_{p_i})$. Here $\xi_j$ is the unique truth assignment of the 3SAT instance $\phi$ that makes clause $C_j$ false, and $\xi_j(x_{p_i})$ is the truth value of variable $x_{p_i}$ by treating true as 1 and false as 0. Intuitively, these CINDs assert that for each tuple $t$ in an $R_m$ relation, $t[A_1, \ldots, A_m]$ denotes a truth assignment $\xi_j$ for the 3SAT instance $\phi$, such that for each $i \in [1, m]$, $\xi_j(x_i) = true$ if $t[A_i] = 1$, and $\xi_j(x_i) = false$ if $t[A_i] = 0$. If the truth assignment $\xi_j$ makes $\phi$ false, then there exists a tuple $t'$ in relation $S$ such that $t'[B] = 0$.

The set $\Sigma$ has no more than $2m + n$ of CINDs in total. Note that $\Sigma$ is acyclic.

For example, consider the following instance of the 3SAT problem: $\theta = \forall x_1 \exists x_2 \exists x_3 \exists x_4 (C_1 \land C_2)$, where $C_1 = x_1 \lor x_2 \lor x_3$, and $C_2 = x_2 \lor x_3 \lor x_4$. Then the set $\Sigma$ for $\theta$ consists of 7 CINDs, as shown in Figure 7.

(C) We define CIND $\psi = (R_0[\text{nil}; nil] \subseteq S[\{nil; B\}, \{nil \mid 0\}].$ It ensures that if the $R_0$ relation is not empty, then there exists a tuple $t$ in relation $S$ such that $t[B] = 0$.

We next show that the 3SAT instance $\theta$ is true if and only if $\text{ACINDs} \not\models \psi$.

First, assume that $\theta$ is satisfiable. We define an instance $D$ of $\mathcal{R}$ as follows. For $i \in [1, m]$, the instance $I_i$ of $R_i$ in $D$ consists of all truth assignments for $x_1, \ldots, x_m$ that satisfy the 3SAT instance $\phi$. The instance $I_S$ of $S$ in $D$ consists of a single tuple $s$ with $s[B] = 0$. One can readily verify that $D \models \Sigma$ but $D \not\models \psi$. Hence $\text{ACINDs} \not\models \psi$.

Conversely, assume that $\Sigma \not\models \psi$. Then there exists an instance $D$ of $\mathcal{R}$ such that $D \models \Sigma$ but $D \not\models \psi$. By $D \not\models \psi$, the instance of $I_0$ of schema $R_0$ in $D$ is nonempty, and hence so is the instance $I_1$ of $R_1$ in $D$ for all $i \in [1, m]$, by $D \models \Sigma$. Observe that the instance $I_m$ of $R_m$ encodes truth assignments for $x_1, \ldots, x_m$. By $D \models \Sigma$, we know that $I_m$ includes all the truth assignments required by the quantifiers in $\theta$. By $D \not\models \psi$ again, the instance $I_S$ of $S$ in $D$ does not have any tuple $s$ with $s[B] = 0$. Hence by the definition of the CINDs $\psi_{m,j}$, none of those truth assignments in $I_m$ violates the 3SAT instance $\phi$. Therefore, $\theta$ is true. \qed
A.6. Proof of Theorem 13

Theorem 13. The implication problem for UCINDs is in polynomial time in the absence of finite-domain attributes.

Proof. It suffices to give a PTIME algorithm for checking whether \( \Sigma \models \psi \) or not. Similar to the upper bound proof of Theorem 7, the algorithm converts the problem to the graph reachability problem, i.e., checking whether there exists a path from a node to another in a graph. Recall that the graph reachability problem is solvable in quadratic time [29].

The proof consists of three parts. We first present the algorithm. We then show that the algorithm is correct. Finally, we show that the algorithm is in PTIME.

Consider a set \( \Sigma \cup \{ \psi \} \) of UCINDs over a database schema \( \mathcal{R} = (R_1, \ldots, R_n) \), where the UCIND \( \psi = (R_d[A; X_p] \subseteq R_b[B; Y_p], t_p) \).

1. The algorithm simulates the chase procedure given in the proof of Theorem 2, fine-tuned to leverage unary CINDs.

2. The algorithm checks whether there exists a node \( \psi \) in the graph \( G \) of \( (R_i[\mathcal{C}; U_p], t(U_p)) \) (1 ≤ \( i \) ≤ \( n \)) such that:

   (i) \( R_i \) is a schema in \( \mathcal{R} \).
   (ii) \( C \) is a single attribute in \( \text{attr}(R_i) \).
   (iii) \( U_p \) is a list of attributes in \( \text{attr}(R_i) \), and
   (iv) \( t(U_p) \) is a partial tuple of \( R_i \) defined on \( U_p \).

The set \( V \) of nodes in \( G \) includes the following: (a) a single node \( u_b = (R_b[A; X_p], t_{p_b}[X_p]) \), which corresponds to the LHS of the UCIND \( \psi \); and (b) for each UCIND \( \psi' = (R_i[C; U_p] \subseteq R_j[F; V_p], t_p) \) in \( \Sigma \), a node \( u = (R_j[F; V_p], t_p[V_p]) \), which denotes the RHS of \( \psi' \).

The set \( E \) of edges contains a directed edge \((u_1, u_2)\) for all nodes \( u_1 = (R_i[C; U_p], t_1[U_p]) \) and \( u_2 = (R_j[F; V_p], t_j[V_p]) \) in \( \psi \) if there exists a UCIND \( (R_i[C; U_p'] \subseteq R_j[F; V_p'], t_p') \) in \( \Sigma \) such that \( U_p' \subseteq Z_p \), \( V_p' \subseteq V_p \), \( t_p[U_p'] = t_i[U_p], \) and \( t_p[V_p] = t_j[V_p] \).

(b) The algorithm then checks whether \( \Sigma \models \psi \), based on graph \( G \).

Let \( S_b \) be the set of nodes that are of the form \( u_b = (R_b[B; U_p], t_b(U_p)) \) in \( G \) such that \( Y_p \subseteq Z_p \) and \( t_b[Y_p] = t_{p_b}[Y_p] \). Recall that \( Y_p \) and \( t_{p_b} \) are from the UCIND \( \psi \).

The algorithm checks whether there exists a node \( u_b \) in the node set \( S_b \) such that there is a path from \( u_b \) to \( u_b \) in graph \( G \) (recall that \( u_b \) denotes the LHS of \( \psi \)). If there exists such \( u_b \), it returns ‘yes’, and it returns ‘no’ otherwise.

We now verify the correctness of the algorithm that returns ‘yes’ iff \( \Sigma \models \psi \).

First assume that the algorithm returns ‘yes’. Then there must exist a path from the node \( u_b \) to a node \( u_b = (R_b[B; U_p], t_b(U_p)) \) in the graph \( G \), where \( Y_p \subseteq Z_p \) and \( t_b[Y_p] = t_{p_b}[Y_p] \). Along the same lines as the proof of Theorem 2, one can construct a proof to show that \( \Sigma \models \exists \leftarrow (1 \rightarrow 6) \psi \) based on IR1–IR6 in the inference system \( \mathcal{I} \) (see Section 3.2.1). By Theorem 2, IR1–IR6 are sound and complete for the implication analysis of CINDs in the absence of finite-domain attributes, by which we have \( \Sigma \models \psi \).

Conversely, assume that the algorithm returns ‘no’. We show that \( \Sigma \not\models \psi \) by constructing a database instance \( D \) of \( \mathcal{R} \) such that \( \mathcal{D} \models \Sigma \), but \( \mathcal{D} \not\models \psi \). The instance \( D \) is constructed step by step as follows.

(a) Initialize \( D := \{I_1, \ldots, I_l\} \) such that \( I_1 = \cdots = I_n = \emptyset \).

(b) Create a tuple \( t_a \) such that \( t_a[A] = v, t_a[X_p] = t_{p_a}[X_p] \), and \( t_a[A'] = v_0 \) for all the other attributes \( A' \) in \( \text{attr}(R_a) \), and let \( I_a = I_a \cup \{t_a\} \). Here \( v \) and \( v_0 \) are two distinct variables.

(c) For each node \( u = (R_i[C; U_p], t_i[U_p]) \) in \( G \) such that there exists a path from nodes \( u_a \) to \( u \), construct a tuple \( t \) such that \( t[C] = v, t[U_p] = t_{i[U_p]} \), and \( t[C'] = v_0 \) for all the other attributes \( C' \) in \( \text{attr}(R_i) \), and let \( I_i = I_i \cup \{t\} \).

(d) Extend the instance \( D \) by using the chase procedure given in the proof of Theorem 2 until \( D \) reaches a fixpoint.

Then as argued in the proof of Theorem 2 about the chase procedure, one can verify that \( \mathcal{D} \models \Sigma \), but \( \mathcal{D} \not\models \psi \). Therefore, \( \Sigma \not\models \psi \).

(3) We next show that the algorithm is in polynomial time.

It is obvious that the graph \( G \) can be built in polynomial time. Observe that the size of the set \( S_b \) is bounded by the number of nodes in the graph \( G \), of which the size is bounded by a polynomial in the size of \( \Sigma \cup \{ \psi \} \). Based on these, it is easy to verify that the algorithm is indeed in PTIME.

Putting these together, we conclude that the implication problem for UCINDs is in PTIME, in the absence of finite-domain attributes.

A.7. Proof of Theorem 15

Theorem 15. The implication problem for UCINDs is coNP-complete in the general setting.

Proof. (1) We first show that the problem is in co NP. Consider a set \( \Sigma \cup \{ \psi \} \) of UCINDs defined on a database schema \( \mathcal{R} = (R_1, \ldots, R_n) \).


To show that the problem is in coNP, it suffices to give NP algorithms for checking whether \( \Sigma \not< \psi \). We first show that UCINDs can be transformed into two normal forms. We then present two NP algorithms based on the forms of \( \psi \), and show that the algorithms are indeed correct.

(A) We show that UCINDs can be expressed in certain “normal” forms. Consider a UCIND \( \psi = (R_1[C; U_p] \subseteq R_1[F; V_p], \tau_{p_0}) \) such that the attribute \( C \) has a finite domain \( \text{dom}(C) = \{c_1, \ldots, c_k\} \). Let \( \Sigma_\psi = \{\psi_1, \ldots, \psi_h\} \), where for each \( i \in [1, k] \), \( \psi_i = (R_i[\text{nil}; C, U_{p_i}] \subseteq R_i[\text{nil}; F, V_{p_i}], \tau_{p_{0,i}}) \) such that \( \tau_{p_{0,i}}[U_p \parallel V_p] = \tau_{p_i}[U_p \parallel V_p] \) and \( \tau_{p_{0,i}}[C] = \tau_{p_i}[F] = \lceil c_i \rceil \). It is easy to verify that \( \Sigma_\psi \equiv (\psi) \), by IR4 and IR8 in the inference system \( \mathcal{I} \) for CINDs (see Section 3.2.1).

As a result, given a set \( \Sigma \) of UCINDs, we can derive an equivalent set \( \Sigma' \) of CINDs by using the transformations above.

Note that the number of CINDs in \( \Sigma' \) is bounded by a polynomial of the size of \( \mathcal{R} \) and the number of UCINDS in \( \Sigma \). In light of this, we can assume w.l.o.g. that all the UCINDS in \( \Sigma \cup \{\psi\} \) are of one of the following forms:

- \( (R_i[\text{nil}; U_{p_i}] \subseteq R_i[\text{nil}; V_p], \tau_{p_i}); \)  
- \( (R_i[C; U_p] \subseteq R_i[F; V_p], \tau_{p_i}); \) where attributes \( C \) and \( F \) have an infinite domain.

Given a set \( \Sigma \cup \{\psi\} \) of CINDS in these two forms, we develop two NP algorithms for checking whether \( \Sigma \not< \psi \), depending on the form of the CIND \( \psi \). We use \( \Sigma_1 \) and \( \Sigma_2 \) to denote those CINDS in \( \Sigma \) in the first form and those in the second one, respectively, where \( \Sigma = \Sigma_1 \cup \Sigma_2 \) and \( \Sigma_1 \cap \Sigma_2 = \emptyset \).

(B) We now provide an NP algorithm for the first case, where the CIND \( \psi \) is of the form \( (R_i[\text{nil}; U_{p_i}] \subseteq R_i[\text{nil}; V_p], \tau_{p_i}) \). To treat \( \Sigma_1 \) and \( \Sigma_2 \) uniformly, we further transform the CINDS of \( \Sigma_2 \) into CINDS in the first form. More specifically, for each CIND \( \psi = (R_i[C; U_p] \subseteq R_i[F; V_p], \tau_{p_i}) \) in \( \Sigma_2 \), we define a set \( \Sigma_{\psi} \) of CINDS in the first form as follows.

- Let \( \text{dom} := \{a_1, \ldots, a_q\} \) be the set of constants appearing in either \( \Sigma \cup \{\psi\} \) or in the finite domains of \( \mathcal{R} \).
- Let \( \psi_0 \) to be the set \( \{\psi_0, \psi_1, \ldots, \psi_h\} \), where (a) \( \psi_0 = (R_i[\text{nil}; U_{p_i}] \subseteq R_i[\text{nil}; V_p], \tau_{p_i}) \), and (b) for each \( i \in [1, h] \), \( \psi_i = (R_i[C; U_p] \subseteq R_i[F; V_p], \tau_{p_i}) \) such that \( \tau_{p_i}[U_p \parallel V_p] = \tau_{p_i}[U_p \parallel V_p] \) and \( \tau_{p_i}[C] = \tau_{p_i}[F] = \lceil a_i \rceil \). Here CIND \( \psi_0 \) is derived from \( \psi \) by IR2, and the other CINDS in \( \Sigma_{\psi} \) are derived from \( \psi \) by IR4 in the inference system \( \mathcal{I} \). Observe that the total number of CINDS in \( \Sigma_\psi \) is bounded by a polynomial in the size of \( \mathcal{R} \), \( \Sigma \), and \( \psi \).

Let \( \Sigma_2 \) be the union of \( \Sigma_{\psi} \)’s when \( \psi \) ranges over all CINDS in \( \Sigma_2 \). We show that it suffices to consider CINDS in \( \Sigma_1 \cup \Sigma_2 \). Indeed, a close examination of the chase procedure given in the proof of Theorem 3 tells us that the chase process for \( \Sigma_1 \cup \Sigma_2 \cup \{\psi\} \) is equivalent to the one for \( \Sigma_1 \cup \Sigma_2 \) since whenever a CIND in \( \Sigma_1 \cup \Sigma_2 \) is applied, we can use one in \( \Sigma_1 \cup \Sigma_2 \) instead to reach the same result, and vice versa. Recall that the chase procedure determines whether \( \Sigma \vdash \psi \) for CINDS in the general setting.

We now give the details of the NP algorithm, which is a non-deterministic extension of the PTIME algorithm given in the proof of Theorem 13, to handle finite-domain attributes. Given the set \( \Sigma_1 \cup \Sigma_2 \cup \{\psi\} \) of CINDS, it extends the PTIME algorithm as follows.

- It adds an extra step after generating the node set \( V \) but before generating the edge set \( E \). For each node \( u = (R_i[\text{nil}; U_p], \tau(U_p)) \in V \), it guesses an instantiation \( \rho_u \) for the list \( U_p \) of all finite-domain attributes in \( \text{attr}(R_i) \setminus \{U_p\} \) such that for each attribute \( C' \in U_f \), \( \rho_u(C') \) is a data value drawn from the finite domain \( \text{dom}(C') \).

Note that the format of nodes is a little different from those used in the PTIME algorithm. However, this has no impact on the algorithm itself.

- The NP algorithm returns an answer opposite to that of the PTIME algorithm.

That is, if there exists no node \( v_b \) in the set \( S_p \) such that there exists a path from the node \( u_b \) to \( v_b \) in the graph \( G \), the algorithm returns ‘yes’, and it returns ‘no’ otherwise. Recall that \( u_b \) and \( v_b \) denote the LHS and the RHS of \( \psi \), respectively. This is because the NP algorithm checks whether \( \Sigma \not< \psi \), while the PTIME algorithm checks whether \( \Sigma \vdash \psi \).

It is easy to see that the algorithm is in NP.

We next show that algorithm returns ‘yes’ if and only if \( \Sigma \not< \psi \). Assume first that the NP algorithm returns ‘yes’. Along the same lines as the argument for the ‘no’ answer of the PTIME algorithm given in the proof of Theorem 13, we can construct a nonempty instance \( \mathcal{D} \) of \( \mathcal{R} \) such that \( \mathcal{D} \vdash \Sigma \) but \( \mathcal{D} \not< \psi \), i.e., \( \Sigma \not< \psi \). Here when populating \( \mathcal{D} \), it suffices to randomly guess an instantiation for the finite-domain attributes of the new tuple to be inserted into \( \mathcal{D} \) (the one shown in the NP algorithm will do).

Conversely, assume that \( \Sigma \not< \psi \). We show that there exists a set of instantiations for the node set \( V \) such that the algorithm returns ‘yes’. Since \( \Sigma \not< \psi \), there exists a database instance \( \mathcal{D} = (I_1, \ldots, I_n) \) such that \( \mathcal{D} \vdash \Sigma \) and \( \mathcal{D} \not< \psi \). That is, there exists a tuple \( t \in I_k \) such that \( t[a_i] = \rho_u[a_i] \), but there exists no tuple \( t \in I_k \) such that \( t[a_i] = \rho_u[a_i] \). The instantiations are defined as follows. For the node \( u = (R_i[\text{nil}; X_p], \tau_{p_i}[Y_p]) \in V \), define an instantiation \( \rho_u \) such that \( \rho_u[X_p] = \tau(U_p) \) for the list \( X_p \) of all finite-domain attributes in \( \text{attr}(R_i) \setminus X_p \). For all the other nodes \( u = (R_i[\text{nil}; U_p], \tau(U_p)) \in V \), if there exists a tuple \( t \in I_k \) such that \( t[U_p] = \tau(U_p) \), define an instantiation \( \rho_u \) such that \( \rho_u[U_f] = \tau(U_f) \) for the list \( U_f \) of all finite-domain attributes in \( \text{attr}(R_u) \setminus U_p \). Otherwise, let \( \rho_u[U_f] \) be defined in terms of arbitrary values.
in the domains. The algorithm must return 'yes' for this specific graph $G$ since there exist no tuples $t_b$ in $I_b$ such that $t_b[Y_p] = t_p[Y_p]$.

(C) We next present an NP algorithm for the second case where the CIND $\psi$ is of the form $(R_a[A] \subseteq R_b[B], t_p)$.

This case is simpler. Indeed, the chase procedure given in the proof of Theorem 3 tells us that those CINDs in $\Sigma_1$ can be simply left out. In this case, the NP algorithm is the same as the one for the first case, except that only those CINDs in $\Sigma_2$ are considered when generating the graph $G$. Here the nodes have the same format as those used in the PTIME algorithm given in the proof of Theorem 13.

An argument similar to the one for the first case can verify that this NP algorithm returns 'yes' if and only if $\Sigma \upharpoonright \psi$. (2) We next show that the problem is coNP-hard by reduction from the 3SAT problem to the complement of the problem (i.e., to decide whether $\Sigma \not\models \psi$). It is known that 3SAT is NP-complete (cf. [30]).

Consider an instance $\phi = C_1 \land \cdots \land C_n$ of 3SAT, where $x_1, \ldots, x_m$ are all the variables in $\phi$, and for each $j \in [1, n]$, $C_j$ is of the form $y_j \lor y_j \lor y_j$ such that for $i \in [1, 3], y_i$ is either $x_{p,j}$ or $\overline{x}_{p,j}$ for $p_j \in [1, m]$. Here $x_{p,j}$ denotes the occurrence of a variable in the literal $I_i$ of clause $C_j$. The 3SAT problem is to decide whether $\phi$ is satisfiable.

Given an instance $\phi$ of 3SAT, we construct an instance of the implication problem for unary CINDs, which consists of a database schema $R$ with finite-domain attributes, and a set $\sigma \cup \{\psi\}$ of CINDs defined on $R$. We show that $\Sigma \not\models \psi$ if and only if $\phi$ is satisfiable.

(A) The database schema $R$ consists of two relation schemas $(R, A_1, \ldots, A_m)$ and $(S, B, C)$, where all attributes have a finite domain $[0, 1]$.

Intuitively, a tuple $t(A_1, \ldots, A_m)$ in an instance $I_R$ of schema $R$ denotes a truth assignment $\xi(t)$ of the 3SAT instance $\phi$, such that for each $i \in [1, m], \xi(t_i) = 1$ if $t[A_i] = 1$, and $\xi(t_i) = 0$ if $t[A_i] = 0$. As will be seen shortly, an instance $I_S$ of schema $S$ is used to indicate whether $\phi$ is satisfiable.

(B) The set $\sigma$ consists of $n$ CINDs given as follows. For each $j \in [1, n]$, let $\xi_j$ be the unique truth assignment of the 3SAT instance $\phi$ that makes the clause $C_j = y_j \lor y_j \lor y_j$ false. Then we define a CIND $\psi_j = (R[B], A_{p,j}, A_{q,j}, A_{r,j}) \subseteq S[B \cup C], t_p)$, where $t_{p,j}[C] = 0$, and for each $i \in [1, 3], t_{p,j}[A_{i,j}] = 1$ if $\xi_j(x_{p,j})$, and $t_{p,j}[A_{i,j}] = 0$ otherwise.

These CINDs assert that for a tuple $t$ in an $R$ relation, if its carries a truth assignment $\xi(t)$ that makes $\phi$ false, then there must exist a tuple $t'$ in the $S$ relation such that $t'[B] = t[B]$ and $t'[C] = 0$. Observe that if there exists a database instance $D = (I_R, I_S)$ such that $I_R$ is nonempty, $I_S$ is empty, and $D \models \sigma$, then the 3SAT instance $\phi$ must be satisfiable.

(C) The UCIND $\psi = (R[B], nil) \subseteq S[B \cup C], (nil \parallel 0)$, it enforces that for each tuple $t$ in relation $I_R$, there exists a tuple $t'$ in relation $I_S$ such that $t'[B] = t[B]$ and $t'[C] = 0$.

As an example, consider an instance $\phi = C_1 \land C_2 \land C_3$ of the 3SAT problem, where $C_1 = x_1 \lor x_2 \lor x_3$, $C_2 = x_2 \lor x_3 \lor x_4$ and $C_3 = x_3 \lor x_4 \lor x_5$. The reduction for $\phi$ is shown in Fig. 8.

The reduction above is obviously in polynomial time.

We next show that $\Sigma \not\models \psi$ if and only if the 3SAT instance $\phi$ is satisfiable.

We first assert that $\Sigma \not\models \psi$. It suffices to construct a database $D$ such that $D \models \sigma$ but $D \not\models \psi$. Since there exists a truth assignment $\xi$ that satisfies $\phi$. Based on $\xi$, we define a tuple $t$ on $R$ such that for each $i \in [1, m], t[A_i] = 1$ if $\xi(t_i) = 1$, and $t[A_i] = 0$ if $\xi(t_i) = 0$. Then the instance $D = (I_R, I_S)$ such that $I_R = \{t\}$ and $I_S = \emptyset$, and $D \models \sigma$ but $D \not\models \psi$.

Conversely, we assert that $\Sigma \not\models \psi$, and show that $\phi$ is satisfiable. It suffices to find a truth assignment $\xi$ that satisfies $\phi$. Since $\Sigma \not\models \psi$, there exists a database instance $D = (I_R, I_S)$ such that $D \models \sigma$ but $D \not\models \psi$. By $D \not\models \psi$, there is a tuple $t$ in $I_R$ but there exists no tuple $t'$ in $I_S$ with $t'[B] = t[B]$ and $t'[C] = 0$. Define a truth assignment $\xi$ for $\Phi$ such that for each $i \in [1, m], \xi(t_i) = 1$ if $t[A_i] = 1$, and $\xi(t_i) = 0$ if $t[A_i] = 0$. Then $\xi$ satisfies $\phi$. Since otherwise, there must exist a tuple $t'$ in $I_S$ with $t'[B] = t[B]$ and $t'[C] = 0$ by the definition of those UCINDs in $\Sigma$. Hence $\phi$ is satisfiable. □

References


