A Paraconsistent Logic Programming Approach for Querying Inconsistent Databases

Sandra de Amo\textsuperscript{a} and Mônica Sakuray Pais \textsuperscript{b,1}

\textsuperscript{a}Faculdade de Computação, Universidade Federal de Uberlândia, Brazil
\textsuperscript{b}Departamento de Informática, CEFET-Urataí, Brazil

Abstract

When integrating data coming from multiple different sources we are faced with the possibility of inconsistency in databases. A paraconsistent approach for knowledge base integration allows keeping inconsistent information and reasoning in its presence. In this paper, we use a paraconsistent logic (LFI1) as the underlying logic for the specification of P-Datalog, a deductive query language for databases containing inconsistent information. We present a declarative semantics which captures the desired meaning of a recursive query executed over a database containing inconsistent facts and whose rules allow inferring information from inconsistent premises. We also present a bottom-up evaluation method for P-Datalog programs based on an alternating fixpoint operator.

Key words: Inconsistent Information, Logic Programming, Paraconsistent Logic, Deductive Databases, Query Languages.

1 Introduction

The treatment of inconsistencies arising from the integration of multiple sources has been a topic increasingly studied in the past years and has become an important field of research in databases. Roughly, there are two approaches to handle the inconsistency problem in knowledge bases: belief revision [20,27] and paraconsistent logic [9]. The goal of the first approach is to make an inconsistent theory consistent, either by revising it or by representing it by a

\textsuperscript{1} Supported by CAPES-Brazil
consistent semantics. The main concern of this approach is to avoid contradictions. On the other hand, the paraconsistent approach allows reasoning in the presence of inconsistency, and contradictory information can be derived or introduced without trivialization. Besides these two approaches, there are “hybrid” approaches based on formalisms which associates degrees of belief, reliability or uncertainty to the source knowledge bases [12,7]. In this paper, we introduce P-Datalog, a logic programming language for querying databases containing inconsistencies. Our approach is paraconsistent: inconsistencies are not rejected. Following the arguments presented in [18], our choice is motivated by the assumption that, in most situations, inconsistent information can be useful, unavoidable and even desirable. Thus, in most situations, the goal of retaining all available information is quite legitimate since discarding inconsistent information would imply losing information.

P-Datalog is a language which allows inferring facts from a database $\mathcal{K}$ obtained by integrating local consistent sources which may be globally inconsistent: they may be contradictory with respect to each other. In this paper, we focus our attention in the integrated database, which possibly contains inconsistent information. We are not interested in the integration process, that is, which information is kept in the source databases or how the integrated database is built. These issues have been treated in a previous paper [15], where we proposed a system based on a deductive proof mechanism to integrate information coming from multiple databases. The intuitive idea of the semantics of facts stored in the integrated database $\mathcal{K}$ is the following: a fact $A \in \mathcal{K}$ is true (resp. false) if it is true (resp. false) in all source databases. It is inconsistent if it is true in some local source and false in some other. The language P-Datalog allows to infer new facts from the facts stored in the integrated database $\mathcal{K}$. These new facts inferred are related to the facts which would be inferred in each individual consistent source. If an inferred fact $A$ is true in the global database (the integrated one), then it would be locally inferred as true in all individual sources: it is a safe information and it is surely true. If it is globally inconsistent then it would be locally inferred as true in some individual sources and as false in others: there is some evidence of $A$ in $\mathcal{K}$. If it is globally false, then it would be locally inferred as false in all individual sources.

The syntax of P-Datalog slightly differs from Datalog\neg syntax [1]. As in Datalog\neg, P-Datalog programs are normal logic programs [22] (the same as the general logic programs of [29]): a set of rules where negation may appear in the body but not in the head of rules. In fact, the main difference between P-Datalog and Datalog\neg concerns their semantics. The answers to a Datalog\neg query constitute a set of facts where each fact has an associated truth-value $t$ (true), $f$ (false) or $u$ (unknown). In our approach, the rules of a P-Datalog

\footnote{This paper is a full version of the conference paper [16].}
program are Horn clauses like in Datalog\(^\neg\), but their semantics is related to the paraconsistent logic \textsc{LFI1}, which was originally introduced in [14,15] as a logical framework to model database integration\(^3\). An answer to a P-Datalog query is a set of facts, where each fact has an associated truth-value which can be \(t\) (true), \(f\) (false), \(u\) (unknown) or \(i\) (inconsistent).

In order to define the 4-valued well-founded semantics of a P-Datalog query, we take advantage of the natural 3-valued semantics of the paraconsistent logic \textsc{LFI1} (where the truth-values are \(t\), \(f\) and \(i\)). We adapt the ideas of [25,28] originally presented in the context of Datalog\(^\neg\) programs. In this classical setting, 2-valued first-order logic models are augmented with a third truth-value \(u\) (unknown). In our setting, 3-valued \textsc{LFI1} models are augmented with the truth-value \(u\) (unknown) as well. Thus, the 4-valued semantics for P-Datalog programs we propose is a natural extension of the 3-valued well-founded semantics of Datalog\(^\neg\) programs.

The following example illustrates our approach.

\textbf{Example 1 (Motivation)} Suppose we have the following rule in a dishonest public contest for hiring civil servants: “if there is \emph{some evidence} that the candidate is supported by an influential person which is not a civil servant himself and if the candidate has no debts towards the income tax services, then there is \emph{some evidence} that this candidate will get the job.” The intuitive meaning behind the expression \emph{there is some evidence} is that this information is supported by \emph{at least} one source, even though some sources may affirm the contrary. We can translate the story above in the following P-Datalog program \(P_{\text{job}}\):

\[
\text{job}(x) \leftarrow \neg \text{owe}(x), \text{supportedby}(x,y), \neg \text{job}(y)
\]

Notice that the unique rule of \(P_{\text{job}}\) follows the syntax of Datalog\(^\neg\). However, its semantics is evaluated according to the semantic laws of the paraconsistent logic \textsc{LFI1}. In this logic, an atomic formula \(R(\vec{x})\) is \emph{verified} if its truth-value is \(t\) or \(i\) (in a paraconsistent approach, inconsistencies are not rejected). Thus in the \(P_{\text{job}}\) program, literals \(\text{supportedby}(x,y)\) (in the body) and \(\text{job}(x)\) (in the head) represent information that are true or inconsistent. On the other hand, the literals \(\sim \text{owe}(x)\) and \(\sim \text{job}(y)\) represent negative information. Intuitively, this means that all information sources affirm the fact that \(x\) has no records in the income tax services files concerning debts and that \(y\) is not a civil servant. Let us suppose that we have the following facts stored in the integrated database:

\[\text{owe}(x), \text{job}(y)\]

\(^3\) An extensive and comprehensive presentation of the paraconsistent logic \textsc{LFI1}, can be found in [14]. In this paper, several surprising properties concerning \textsc{LFI1} formulas are discussed.
\[ K = \{ \circ \text{supportedby(charles,joseph)}, \circ \text{supportedby(joseph,charles)}, \]
\[ \circ \text{supportedby(paul,james)}, \bullet \text{supportedby(john,kevin)}, \]
\[ \circ \text{supportedby(james,kevin)}, \bullet \text{owe(james)} \} \]

The symbols \( \circ \) and \( \bullet \) attached to each fact in the database mean that the fact is \textit{sure} and \textit{controversial}, respectively. We notice that the facts stored in the database must be explicitly declared as sure or controversial (by attaching these symbols \( \circ \) and \( \bullet \)). Following the closed-world assumption, facts that are not in the database are considered false. In order to better understand the semantics of each fact stored in the database \( K \), we can view it as the integration of several local consistent sources \( K_1, \ldots, K_n \). In each \( K_i \), each fact \textit{supportedby}(a,b) and \textit{owe}(c) has truth-value \( t \) or \( f \). The fact \textit{supportedby(charles,joseph)} is \textit{true} in every \( K_i \) (\( i = 1, \ldots, n \)) and \textit{owe(john)} is \textit{false} in every \( K_i \) (\( i = 1, \ldots, n \)). Whereas \textit{owe(james)} is \textit{true} in some \( K_j \) (\( j = 1, \ldots, n \)) and \textit{false} in some \( K_p \) (\( p = 1, \ldots, n \)).

We now show a 4-valued model \( J \) of \( P_{job} \) which includes the facts of the database \( K \), i.e., \( J \) agrees with \( K \) on the values of \textit{owe} and \textit{supportedby} atoms. This 4-valued model \( J \) contains the facts \textit{job}(x) which correspond to the answer to the query “For which people is there some evidence that they will get the job?” As we will show later (see Example 5), this model \( J \) is the P-Datalog well-founded semantics of \( P_{job} \) on input \( K \). The values of the \textit{job} atoms in the derived database \( J \) are the following:

- \textbf{true (t)}: \textit{job(paul)}
- \textbf{false (f)}: \textit{job(kevin), job(james)}
- \textbf{inconsistent (i)}: \textit{job(john)}
- \textbf{unknown (u)}: \textit{job(charles), job(joseph)}

This model asserts that James surely does not get the job: there is some evidence that he owes to the taxation office. From this we can infer that Paul surely get the job. Indeed, Paul does not owe any tax return and he is supported by James who is not a civil servant. It also can be deduced that Kevin does not succeed in getting the job because nobody supports him. In John’s case, he does not owe the taxation office but it is controversial that he is supported by Kevin, who is not a public servant himself. Thus it is controversial that John gets the job. On the other hand, it is unknown that Charles and Joseph succeed in the public contest. They fulfill almost all the requirements: they do not have debts, they have the support of an influential person but they depend on each other: Charles supports Joseph and Joseph supports Charles. The only chance for Charles getting the job is if Joseph (his only support) does not get it. And vice-versa, the only chance for Joseph getting the job is if Charles (his only support) does not get it. Therefore it is
not possible to infer which one will get the job: either Charles or Joseph. This means that this information is unknown: we are not able to infer the existence or nonexistence of any source supporting it.

The answer to our query: “For which people is there some evidence that they will get the job?” is Paul and John. Besides, we know that Paul surely gets the job, but in John’s case, we only can affirm that it is controversial that he gets the job. This is derived from a inconsistent fact (\(\bullet\text{supportedby}(\text{john}, \text{kevin})\)) in the integrated database \(\mathcal{K}\). That means: (1) From the point of view of local sources \(\mathcal{K}_i\) affirming that John is supported by Kevin, it is inferred that John gets the job, (2) From the point of view of sources in \(\mathcal{K}_j\) affirming that John is not supported by Kevin, it is inferred that John does not get the job. Therefore, in the integrated database \(\mathcal{K}\), as expected, the information \(\text{job}(\text{john})\) is derived as inconsistent.

Differently from some approaches treating paraconsistent query languages [24,26,9,27], our well-founded semantics is a natural extension of the well-founded semantics for Datalog\(^-\) programs proposed by [25]. In this paper, we also present a bottom-up evaluation procedure for computing the well-founded semantics based on the alternating fixpoint computation introduced in [28].

The paper is organized as follows: In Section 2 we briefly describe the basic notions of the logic LFI1. In Section 3, we introduce P-Datalog programs and generalize the notion of database instance to allow the storage of inconsistent information in databases. In Section 4 we describe the well-founded semantics of a P-Datalog program. In Section 5, we present a bottom-up method for evaluating P-Datalog programs based on an alternating fixpoint operator and briefly discuss its implementation. In Section 6, we discuss some related work. Finally in Section 7, we briefly discuss the implementation of the method we proposed for evaluating P-Datalog queries and also present our perspectives for further research.

2 LFI1: A 3-valued Paraconsistent Logic

In this section we briefly describe the syntax and semantics of LFI1 (Logic of Formal Inconsistency). A detailed presentation can be found in [14]. The semantics of a P-Datalog program is based on the semantics of LFI1. Even if P-Datalog programs constitute a small fragment of the set of LFI1 formulas (we only consider Prolog-like Horn clauses), inference in P-Datalog is based on the paraconsistent framework of LFI1.

\(^4\) Other approaches for associating a meaning to a program by means of a special Herbrand model have been proposed in the literature [29,8].
Let $R$ be a finite signature without functional symbols and $\text{Var}$ a set of variables symbols. We assume that formulas of LFI1 are defined in the usual way, as in the classical first-order logic setting, with the addition of a new symbol $\bullet$ (read “it is inconsistent”). A formula of LFI1 is defined inductively by the following statements (and only by them):

1. If $R$ is a predicate symbol of arity $k$ and $x_1, ..., x_k$ are constants or variables, then $R(x_1, ..., x_k)$ and $x_1 = x_2$ are atomic formulas or atoms. The former is called a relational atom and the later an equality atom.

2. If $F, G$ are formulas and $x$ is a variable then $F \lor G$, $\neg F$, $\forall x F$, $\exists x F$ and $\bullet F$ are formulas.

A sentence is a formula without free variables. A fact is a relational atom without free variables. We denote by $F$ the set of facts.

In the remainder, we often will use the symbols $\bullet$ and $\circ$ in two different contexts: a syntactic one and a semantic one. In LFI1, the symbols $\bullet$ and $\circ$ are new modalities which allow extending the syntax of first-order formulas. So, in LFI1, these symbols are used in a syntactic context. In the Example 1, these symbols have been used in a semantic context, in order to identify the truth-value of facts stored in the database. We think that these two forms of using the symbols $\circ$ and $\bullet$ will not cause confusion, because the P-Datalog syntax do not include them in its specification. So, in this paper, the use of $\circ$ and $\bullet$ will be restricted to the semantic context.

Definition 1 Let $R$ be a finite signature and $F$ the set of facts defined over $R$. An interpretation over $R$ is an application $\delta : F \rightarrow \{ f \text{ (false)}, t \text{ (true)}, i \text{ (inconsistent)} \}$.

An interpretation of facts can be extended to the propositional sentences in a natural way by using the connective matrices described in Figure 1. The connective $\land$ is derived from $\lor, \neg$: $A \land B = \neg (\neg A \lor \neg B)$ and the connective $\rightarrow$ is derived from $\neg, \lor, \bullet$: $A \rightarrow B = \neg (A \lor \bullet A) \lor B$. In the next section, we will use a derived connective $\sim$, called negation by default, defined as $\sim A = \neg A \land \neg \bullet A$.

The extension of $\delta$ to the quantified sentences is not treated here. The reason is that the formulas we will deal with in the next sections are Horn clauses, which are interpreted over a finite Herbrand Universe. So, the universal quantifiers
appearing in the clauses can be viewed as a bounded conjunction. The details of the semantics of quantifiers can be found in [14].

We denote by $\text{Dom}$ the Herbrand Universe of $\mathbf{R}$ (the constant symbols of $\mathbf{R}$). In fact, we are supposing that the universe domain of any interpretation $\delta$ is $\text{Dom}$ (thus, $\delta$ interprets the constant symbols by themselves). A valuation is an application $v : \text{Var} \rightarrow \text{Dom}$.

**Definition 2** Let $F(x_1, \ldots, x_n)$ be a formula of $\text{LFI1}$ with free variables $x_1, \ldots, x_n$, $v$ a valuation and $\delta$ an interpretation. We say that $(\delta, v)$ satisfies $F(x_1, \ldots, x_n)$ (denoted by $(\delta, v) \models F(x_1, \ldots, x_n)$) iff $\delta(F[v(x_1), \ldots, v(x_n)/x_1, \ldots, x_n])$ is $\text{t}$ or $\text{i}$. If $(\delta, v) \models F$ for each valuation $v$, we say that $\delta$ is a model of $F$ (denoted $\delta \models F$). We also say that $F$ is verified or satisfied by $\delta$.

**Remark.** We notice that the notion of equivalence between formulas in $\text{LFI1}$ is different from the one we are used in classical logic. We say that two formulas $F$ and $G$ are equivalent if for all interpretations $\delta$, $\delta$ satisfies $F$ if and only if $\delta$ satisfies $G$. In fact, this is the same definition used in classical logic, except that in the latter, being equivalent is the same as having the same truth-value. This is not the case for $\text{LFI1}$. For instance, the formulas $\sim \sim A$ and $A$ are equivalent in $\text{LFI1}$, but they don’t have the same truth-value under some interpretations. If $\delta$ is an interpretation such that $\delta(A) = \text{i}$, we have $\delta(\sim \sim A) = \text{t}$. So, the two formulas have not the same truth-values, but both are satisfied by $\delta$. Notice that $\neg A$ is equivalent to $\sim A \lor \bullet A$ in $\text{LFI1}$.

### 3 The Query Language P-Datalog

In this section we use the logical formalism $\text{LFI1}$ to generalize the notion of database instance to allow the storage of inconsistent information in our databases. We also introduce the query language P-Datalog which is designed to query databases containing inconsistent information. We assume that the reader is familiar with traditional database terminology [1]. In what follows, we denote by $R(\bar{a})$ the formula $R(u_1, \ldots, u_k)$, where $u_1, \ldots, u_k$ are variables and we denote by $R(\bar{a})$ the ground atom (or fact) $R(a_1, \ldots, a_k)$, where $a_1, \ldots, a_k$ are constants ($k$ is called the arity of $R$).

**Definition 3** (Paraconsistent Databases) Let $\mathbf{R}$ be a database schema (or signature), i.e., a set of relation names (or predicate names) and a set $\text{Dom}$ of constants (the Herbrand Universe of its instances). A 3-valued instance over $\mathbf{R}$ (or a paraconsistent database) is an interpretation $\mathbf{I}$ such that for each $R \in \mathbf{R}$ the set $\mathbf{I}_R = \{\bar{a} : \mathbf{I}(R(\bar{a})) = \text{t} \text{ or } \mathbf{I}(R(\bar{a})) = \text{i}\}$ is finite. So, an instance over $\mathbf{R}$ can be viewed as a finite set of facts over $\mathbf{R}$, having truth-values $\text{t}$ or $\text{i}$. The facts which are not in the instance $\mathbf{I}$ have truth-value $\text{f}$. A fact $R(\bar{a})$ such that $\mathbf{I}(R(\bar{a})) = \text{i}$ is intended to be controversial. On the other hand, if $\mathbf{I}(R(\bar{a})) =$
\( t, R(\vec{a}) \) is intended to be a safe information.

**Notation.** In what follows, we will denote by \( \circ R(\vec{a}) \) the fact that \( I(R(\vec{a})) = t \) and by \( \bullet R(\vec{a}) \) the fact that \( I(R(\vec{a})) = i \). Here, we use the symbols \( \circ \) and \( \bullet \) in a semantic context.

P-Datalog is an extension of Datalog\(^\neg\) [1]. This well-known deductive query language uses the classical first order logic as its underlying logic, and a Datalog\(^\neg\) query applies over a classical database instance, i.e. a finite first-order interpretation. Rather than classical first-order logic, P-Datalog uses the paraconsistent logic \( LFI_{1} \) as its underlying logic, and P-Datalog queries apply over paraconsistent databases. P-Datalog programs are first-order Horn clauses as in Datalog\(^\neg\) programs, i.e. first-order clauses with positive and negative literals in their bodies. Negation in P-Datalog (as well as in Datalog\(^\neg\)) is understood as the *negation by default (or strong negation) \( \sim \).* The negation \( \sim \) used in \( LFI_{1} \) is called *weak negation*.

The intuitive meaning behind default and weak negations is the following: (1) the ground formula \( \sim R(\vec{a}) \) is verified by a paraconsistent database \( I \) if the fact \( R(\vec{a}) \) is not in \( I \); (2) the ground formula \( \neg R(\vec{a}) \) is verified by \( I \) if the fact \( R(\vec{a}) \) is in \( I \) as controversial or if it is not in \( I \).

**Definition 4 (P-Datalog Programs)** A P-Datalog *program* is a finite set of rules \( A \leftarrow L_{1}, ..., L_{n}, \) where \( A \) is an atom of the form \( R(\vec{u}) \), and \( L_{i} \) are literals of the form: \( R(\vec{u}) \) or \( \sim R(\vec{u}) \). \( R \) is a relation name and \( \vec{u} \) is a free tuple of appropriate arity. The atom \( A \) is called the *head* of the rule. The literals \( L_{1}, ..., L_{n} \) constitute the *body* of the rule. One requires also that each variable occurring in the head of the rule must occur in at least one of the free tuples in the body.

We denote by \( sch(P) \) the set of relations (predicates) appearing in \( P \), by \( adom(P) \) the set of constants appearing in \( P \) and by \( B(P) \) all facts of the form \( R(\vec{a}) \) where \( R \in sch(P) \) and \( \vec{a} \) is a tuple of constants in \( adom(P) \) (the Herbrand Base of \( P \)). The set of relations which appear in the head of rules are called the *intensional relations* and is denoted by \( idb(P) \). The set of those appearing only in the body of rules are called *extensional relations* and is denoted by \( edb(P) \).

**Definition 5 (P-Datalog Query)** A P-Datalog *query* is a pair \( (P, Q(u_{1}, ..., u_{n})) \) where \( P \) is a P-Datalog program, \( Q \in idb(P) \) and \( u_{1}, ..., u_{n} \) are variables or constants in \( adom(P) \) (\( n \) is the *arity* of the relation \( Q \)).

**Example 2 (Running Example)** Let us consider the same situation presented in Example 1. The rule \( P_{job} \) and the 3-valued instance \( I \) described in that example constitute a P-Datalog program \( P \) where \( sch(P) = \{ \text{supportedby, job, owe} \} \), \( adom(P) = \{ \text{charles, john, james, joseph, paul, kevin} \} \) and \( B(P) = \)
\{owe(charles), owe(joseph), supportedby(charles,joseph), supportedby(joseph, charles), \ldots\}. The intensional and extensional schemas are \(edb(P) = \{supportedby, owe\}, idb(P) = \{job\}. The pair \((P_{job}, job(x))\) is a P-Datalog query (“For which people is there some evidence that they will get the job?”). The pair \((P_{job}, job(Kevin))\) corresponds to the boolean query “Is there some evidence that Kevin may get the job?”

4 Answering P-Datalog Queries

In this section we introduce the well-founded semantics for P-Datalog programs. The well-founded semantics of a P-Datalog program \(P\) is designed to capture the natural semantics of queries \((P, Q(u_1, \ldots, u_n))\) where \(Q \in idb(P)\), that is, what their answers are expected to be. Our approach is a natural extension of the well-founded semantics for Datalog \(^{25}\). Our definition of a P-Datalog query makes use of \(4\)-valued instances, in which facts may assume one of the four truth-values in the set \(Val = \{true(t), false(f), inconsistent(i), unknown(u)\}\}. We assume that the reader is familiar with the notions of lattices, lattice operators, monotonicity and continuity, fixpoints, etc. For details, see \cite{22}.

4.1 \(4\)-valued Models

Let us consider the complete lattice \((Val, \leq)\), where \(f \leq u \leq i \leq t\).

**Definition 6** Let \(P\) be a P-Datalog program. A \(4\)-valued instance \(I\) over \(sch(P)\) is an application \(I : B(P) \longrightarrow \{t, f, u, i\}\). If \(I(A) \neq u\) for all \(A \in B(P)\), we say that \(I\) is total.

The answer of a P-Datalog program \(P\) is a special \(4\)-valued instance which corresponds to the well-founded semantics of \(P\). The main goal of this section is to define this particular instance.

There is a natural ordering \(\preceq\) among \(4\)-valued instances over \(sch(P)\), defined by: \(I \preceq J\) iff for each \(A \in B(P)\), \(I(A) \leq J(A)\).

The set of \(4\)-valued instances of a P-Datalog program \(P\) is denoted by \(4\-Inst_P\). It is easy to verify that \((4\-Inst_P, \preceq)\) constitutes a complete lattice. We denote by \(\top\) the maximal \(4\)-valued instance (where all facts have truth-value \(t\)) and by \(\bot\) the minimal \(4\)-valued instance (where all facts have truth-value \(f\)). We also represent a \(4\)-valued instance by listing the positive, inconsistent and negative facts, and omitting the unknown ones.

**Example 3 (4-valued instance)** Let \(J\) be a \(4\)-valued instance, where \(J(p) = t\),
We extend the 3-valued connective matrices of \( \text{LFI1} \) (see Figure 1), to the 4-valued matrices showed in Figure 2.

If \( F \) is the body of a P-Datalog rule and \( J \) is a 4-valued instance, we denote by \( J(F) \) the truth-value associated to \( F \) according to the matrices for the \( \land \) and \( \sim \) connectives given in Figure 2.

We notice that the matrix corresponding to the \( \rightarrow \) connective in Figure 2 is not a straight extension of its counterpart in Figure 1: in P-Datalog, the truth-value of \( t \rightarrow i \) is \( f \) and in \( \text{LFI1} \) this truth-value is \( i \). In fact, in the P-Datalog context an inconsistent fact cannot be derived from a set of consistent true facts. This kind of inference is accepted in \( \text{LFI1} \) (the truth-value of the implication being \( i \), and so, accepted in the paraconsistent logic).

Let \( I \) be a 4-valued instance and \( A \leftarrow G \) be a P-Datalog rule (here, \( G \) denotes a conjunction of literals of the form \( B \) or \( \sim B \)). Then

\[
I(A \leftarrow G) = \begin{cases} 
  i & \text{if } I(A) = I(G) = i, \\
  t & \text{if } I(A) > I(G) \text{ or } I(A) = I(G) \neq i \\
  f & \text{otherwise}
\end{cases}
\]

**Definition 7** Let \( P \) be a P-Datalog program. An *instantiated rule* of \( P \) is a rule where all variables are replaced by constants in \( \text{adom}(P) \). We denote by \( \text{ground}(P) \) the set of instantiated rules of \( P \). A 4-valued instance \( J \) over \( \text{sch}(P) \) satisfies a boolean combination \( \alpha \) of atoms in \( \text{B}(P) \) iff \( J(\alpha) \in \{t, i\} \).

A *4-valued model* of \( P \) is a 4-valued instance \( J \) over \( \text{sch}(P) \) satisfying each rule in \( \text{ground}(P) \), i.e., the truth-value of each rule in \( \text{ground}(P) \) is \( t \) or \( i \). A 4-valued model \( M \) is *minimal* iff for all \( M' \subset M \), \( M' \) is not a model.

As we claimed in the introduction, \( \text{LFI1} \) is the underlying logic of P-Datalog. This is justified by the following proposition:

**Proposition 1** Let \( J \) be a paraconsistent instance. Then:
(1) If \( J \) satisfies a ground P-Datalog rule \( A \leftarrow L_1, \ldots, L_n \) in the P-Datalog semantics (using the matrices of Figure 2) then \( J \) satisfies the rule in the \textbf{LFI1} semantics (using the matrices of Figure 1).

(2) If \( J \) satisfies a P-Datalog program \( P \) then \( J \) satisfies the \textbf{LFI1} formula corresponding to the finite conjunction of all instantiated rules of \( P \).

\textbf{Proof.}

(1) Suppose the instance \( J \) satisfies the rule \( A \leftarrow G \) in P-Datalog, where \( A \) is an atom and \( G \) is a conjunction of literals of the form \( B \) or \( \sim B \). By definition 7, \( J(A \leftarrow G) \in \{i,t\} \). By the definition of the semantics of the \( \leftarrow \) connective, \( J(A) \geq J(G) \):

(a) \( J(A)=t \) and \( J(G) \in \{t,i,f\} \) ⇒ \( J \) satisfies \( A \leftarrow G \) in \textbf{LFI1}, because \( J(A \leftarrow G) = t \) in \textbf{LFI1}.

(b) \( J(A)=i \) and \( J(G) \in \{i, f\} \) ⇒ \( J \) satisfies \( A \leftarrow G \) in \textbf{LFI1}, because \( J(A \leftarrow G) \in \{t,i\} \) in \textbf{LFI1}.

(c) \( J(A)=f \) and \( J(G)=f \) ⇒ \( J \) satisfies \( A \leftarrow G \) in \textbf{LFI1}, because \( J(A \leftarrow G) = t \) in \textbf{LFI1}.

(2) Follows from the fact that a P-Datalog program is a conjunction of rules and each rule is the conjunction of its ground instances. \( \square \)

4.2 Extended P-Datalog programs

The well-founded semantics of P-Datalog programs is based on the notion of stable models [19]. Stable models are usually defined as fixpoint of an immediate consequence operator. Following the same idea underlying the definition of 3-stable models in [25], we introduce the notion of \textit{extended P-Datalog programs}. We will see that for such programs we can define an immediate consequence operator which is monotonic and has a unique least fixpoint.

\textbf{Definition 8} An \textit{extended P-Datalog program} is a P-Datalog program where (1) negative facts \( \sim A \) do not appear in the body of rules; (2) truth-values \( t, f, u \) and \( i \) may occur as literals in the body of rules.

Next we define the \textit{immediate consequence operator} \( 4-T_P \) associated to an extended program \( P \).

\textbf{Definition 9} Let \( P \) be an extended P-Datalog program. The \textit{immediate consequence operator} \( 4-T_P \) associated to \( P \) is a mapping \( 4-T_P : 4-\text{Inst}_P \rightarrow 4-\text{Inst}_P \) defined as follows. Let \( J \) be a 4-valued instance and \( A \in \text{B}(P) \), then

---

5 A rule is a closed universally quantified sentence in \textbf{LFI1}. Because the Herbrand Universe is finite, the rule reduces to the conjunction of its ground instances.
Let the 4-valued instance $I_0$, $I_1$, $I_2$, ... of 4-valued instances as follows: $I_0 = \bot$, $I_1 = 4-T_P(I_0)$, $I_2 = 4-T_P(I_1)$, ...

The following lemma says that the immediate consequence operator for an extended program $P$ has a least fixpoint which coincides with the unique minimal 4-valued model of $P$.

**Lemma 1** Let $P$ be an extended P-Datalog program.

1. The operator $4-T_P$ is monotonic;
2. The 4-valued instance $M$ is a 4-valued model of $P$ iff $4-T_P(M) \triangleq M$;
3. The sequence $\{4-T_P(\bot)\}_{i \geq 0}$ is increasing and converges to the least fixpoint of $4-T_P$;
4. $P$ has a unique 4-valued minimal model (denoted by $P(\bot)$) that equals the least fixpoint of $4-T_P$.

**Proof.** (1) Suppose that $A \in \textbf{B}(P)$, $I(A) \leq J(A)$. We have to show that $4-T_P(I)(A) \leq 4-T_P(J)(A)$. We have the following possibilities: (a) There are rules of the form $A \leftarrow F$ in $\text{ground}(P)$, where $F$ corresponds to a conjunction of atoms with truth-values $t$, $i$, $u$, $f$. Then $4-T_P(I)(A) = \max\{I(F)\}$ and $4-T_P(J)(A) = \max\{J(F)\}$. By the fact that $I(A) \leq J(A)$, $\max\{I(F)\} \leq \max\{J(F)\}$. So $4-T_P(I)(A) \leq 4-T_P(J)(A)$. (b) There are no rules in $\text{ground}(P)$ with $A$ in the head. In this case, $4-T_P(I)(A) = 4-T_P(J)(A) = f$.

(2 $\rightarrow$): Let $M$ a 4-valued model of $P$ and let $A \in \textbf{B}(P)$. If there are no rules in $\text{ground}(P)$ with $A$ in the head, then $4-T_P(M)(A) = f \triangleq M(A)$. Otherwise, for all rules $r_k$ of the form $A \leftarrow F_k$ in $\text{ground}(P)$, $M(A \leftarrow F_k) \in \{t$, $i\}$, and so, $\max\{M(F_k)\} \leq M(A)$. Then, $4-T_P(M) \triangleq M$. (2 $\leftarrow$): Let $4-T_P(M) \triangleq M$. We show that $M$ is a 4-model of $P$. Let $r_k, 0 \leq k \leq l$ the rules in $\text{ground}(P)$ of the form $A \leftarrow F_k$. Then, $\max\{M(F_k) : 0 \leq k \leq l\} = 4-T_P(M)(A) \leq M(A)$. So, $M(F_k) \leq M(A)$, for all $k$. Thus, $M$ satisfies the rules $r_k$.

(3) Because $4-$Inst$_P$ is a finite set, the sequence $\{4-T_P(\bot)\}_{i \geq 0}$ reaches a fixpoint after a finite number $N$ of steps. Because $4-T_P$ is monotonic, $I_0 \triangleq I_1 \triangleq I_2 \ldots \triangleq I_N$. So the sequence $\{4-T_P(\bot)\}_{i \geq 0}$ is increasing and converges to the least fixpoint of $4-T_P(I_N)$.

(4) $M$ is a minimal 4-valued model of $P$ $\iff$ $M = \inf\{M| M$ is a 4-valued model of $P\} \iff M = \inf\{M| 4-T_P(M) \triangleq M \}$, by (2) $\iff M = I_N = \text{least}$
fixpoint of $4-T_P$ (by monotonicity of $4-T_P$, and by the fact that $I_0 \preceq M$). □

4.3 4-stable Models

According to [25], the semantics of a Datalog$^-$ program $P$ is an appropriate 3-valued model of $P$. We extend this idea to P-Datalog programs and introduce the 4-stable models, a class of special models. The semantics of a P-Datalog query will be the intersection of all 4-stable models.

Let $P$ be a P-Datalog program and $I$ a paraconsistent database instance (a 3-valued instance). We denote by $P_I$ the program obtained from $P$ by adding to $P$ unit clauses $A \leftarrow$ for each $A$ such that $I(A) = t$, and clauses $A \leftarrow i$ for each $A$ such that $I(A) = i$. From now on, we suppose that our programs include these clauses corresponding to the facts of $I$.

Let $I$ be a 4-valued instance over $\text{sch}(P)$. The positivised ground version of $P$ according to $I$ (denoted $\text{pg}(P, I)$), is the P-Datalog program obtained from $\text{ground}(P)$ by replacing each negative literal $\neg A$ by $I(\neg A)$ (i.e., by its respective truth value: t, f, u, i). So, $\text{pg}(P, I)$ is an extended P-Datalog program, i.e., a program without negation. By lemma 1, the least fixpoint $\text{pg}(P, I)(\perp)$ of its immediate consequence operator exists. It contains all facts that are inferred from $P$ and $I$, by assuming the values for the negative premises as given by $I$. We denote $\text{pg}(P, I)(\perp)$ by $\text{conseq}_P(I)$, i.e. $\text{conseq}_P(I)$ is the least fixpoint of the extended P-Datalog program $\text{pg}(P, I)$.

Definition 10 Let $P$ be a P-Datalog program. A 4-valued instance $I$ over $\text{sch}(P)$ is a 4-stable model of $P$ iff $\text{conseq}_P(I) = I$.

The following example illustrates the notion of 4-stable model:

Example 4 (4-stable model) Consider the P-Datalog program $P_{\text{job}}$ and the input instance $J$ given in the example 1. Let us check that $J$ is a 4-stable model of $P_{\text{job}}$. For this, we have to compute $\text{conseq}(J)$ and show that $\text{conseq}(J) = J$. The program $P' = \text{pg}(P, J)$ is:

$$
\text{job(charles)} \leftarrow t, \text{supportedby(charles,joseph)}, u
$$

$$
\text{job(joseph)} \leftarrow t, \text{supportedby(joseph, charles)}, u
$$

$$
\ldots
$$

$$
\text{supportedby(paul,james)} \leftarrow
$$

$$
\text{supportedby(charles,joseph)} \leftarrow
$$

$$
\text{supportedby(john,kevin)} \leftarrow i
$$

$$
\ldots
$$
The minimal 4-valued model of $P'$ is obtained by iterating $4-T_P(\bot)$ up to a fixpoint. The first execution of $4-T_P$ yields $4-T_P^1(\bot) = \{\sim \text{job(charles)}, \sim \text{job(joseph)}, \sim \text{job(paul)}, \sim \text{job(john)}, \sim \text{job(james)}, \sim \text{job(kevin)}\}$. We can verify that $4-T_P^2(\bot) = 4-T_P^3(\bot) = \{\circ \text{job(paul)}, \bullet \text{job(john)}, \sim \text{job(james)}, \sim \text{job(kevin)}\}$. Thus $\text{conseq}_P(J) = J$ and $J$ is a 4-stable model of $P$. The instance $J$ coincides with $I$ (given in the example 1) for the atoms supportedby and owe.

4.4 Well-founded Semantics

P-Datalog programs generally may have several 4-stable models, and each P-Datalog program has at least one 4-stable model (see theorem 3). Then it is reasonable to say that the desired answer to a P-Datalog query consists of the positive, inconsistent and negative facts belonging to all 4-stable models of the program.

Definition 11 Let $P$ be a P-Datalog program. The well-founded semantics of $P$ is a 4-valued instance consisting of the positive, inconsistent and negative facts belonging to all 4-stable models of $P$. This semantics is denoted by $P^{4wf}$.

5 Bottom-up Evaluation of P-Datalog Queries

The previous description of the well-founded semantics, although effective, is inefficient. It involves checking all possible 4-valued instances of a program, determining which are 4-stable models, and then taking their intersection.

A much simpler method is based on an alternating fixpoint computation [28], that converges to the well-founded semantics. The idea of the method is as follows. We define an alternating sequence $\{I_i\}_{i \geq 0}$ of 4-valued instances that are underestimates and overestimates of the facts known in every 4-stable model of $P$. The alternating sequence $\{I_i\}_{i \geq 0}$ is defined as follows: $I_0 = \bot$ and $I_{i+1} = \text{conseq}_P(I_i)$, for $i \geq 0$.

We notice that each $I_i$ is constructed starting from the total instance $\bot$ by repeated applications of $\text{conseq}_P$. Thus each $I_i$ is a total instance, i.e., has no undefined truth-values.

Theorem 1 The operator $\text{conseq}_P$ is antimonotonic. That is, if $I \preceq J$ then $\text{conseq}_P(J) \preceq \text{conseq}_P(I)$.

Proof. Let us suppose that $I \preceq J$ and let $A \in B(P)$. We will show that $\text{conseq}_P(J)(A) \preceq \text{conseq}_P(I)(A)$. We remind that $\text{conseq}_P(I)$ is the least fixpoint of the extended P-Datalog program $pg(P, I)$, denoted by $pg(P, I)(\bot)$. We have the following cases:
(1) \text{conseq}_P(J)(A) = \text{t}.

We affirm that for all \( n \geq 1 \), if \( n \) is such that \( 4-T_{pg(P,J)}^n(\bot)(A) = \text{t} \), and \( \forall j, j \geq n, 4-T_{pg(P,J)}^j(\bot)(A) = \text{t} \), then \text{conseq}_P(I)(A) = \text{t}. \) This assertion is proved by induction on \( n \).

- Induction Base: \( n=1 \). By our hypothesis, in \( \text{ground}(P) \) there is a rule of the form: \( A \leftarrow B_1, \ldots, B_n \), where \( B_k \) are atoms, \( J(B_k) = \text{f}, \forall k, 1 \leq k \leq n \).

By \( I \models J \), \( I(B_k) = \text{f} \forall k, 1 \leq k \leq n \).

Thus \text{conseq}_P(I)(A) = \text{t} and the induction base is proved.

- Induction step: by our hypothesis, in \( \text{ground}(P) \) there exists a rule of the form: \( A \leftarrow B_1, \ldots, B_n, D_1, \ldots, D_m \), where \( B_k, D_g \) are atoms, \( J(B_k) = \text{f} \forall k, 1 \leq k \leq n \) and \( 4-T_{pg(P,J)}^j(\bot)(D_g) = \text{t}, \forall g, 1 \leq g \leq m \).

By \( I \models J \), \( I(B_k) = \text{f} \forall k, 1 \leq k \leq n \).

By the induction step hypothesis, \text{conseq}_P(I)(D_g) = \text{t}, \forall g, 1 \leq g \leq m.

Thus \text{conseq}_P(I)(A) = \text{t}.

The induction is proved and \text{conseq}_P(I)(A) = \text{t}.

(2) \text{conseq}_P(J)(A) = \text{i}.

We affirm that for all \( n \geq 1 \), if \( n \) is such that \( 4-T_{pg(P,J)}^n(\bot)(A) = \text{i} \), and \( \forall j, j \geq n, 4-T_{pg(P,J)}^j(\bot)(A) = \text{i} \), then \text{conseq}_P(I)(A) = \text{i}. \) This assertion is proved by induction on \( n \).

- Induction base: \( n=1 \). By our hypothesis, in \( \text{ground}(P) \) there is a rule of the form: \( A \leftarrow B_1, \ldots, B_n, c_1, \ldots, c_p \), where \( B_k \) are atoms and \( c_g \) are truth-values, \( J(B_k) = \text{f} \forall k, 1 \leq k \leq n, c_g = \text{i} \forall g, 1 \leq g \leq p \), there is a \( w, c_w = \text{i}, 1 \leq w \leq p \).

By \( I \models J \), \( I(B_k) = \text{f} \forall k, 1 \leq k \leq n \).

Thus \text{conseq}_P(I)(A) = \text{i} and the induction base is proved.

- Induction step: By our hypothesis, in \( \text{ground}(P) \) there is a rule of the form: \( A \leftarrow B_1, \ldots, B_n, D_1, \ldots, D_m \), where \( B_k, D_g \) are atoms, \( J(B_k) = \text{f} \forall k, 1 \leq k \leq n \) and \( 4-T_{pg(P,J)}^n(\bot)(D_g) = \text{i}, \forall g, 1 \leq g \leq m \).

By \( I \models J \), \( I(B_k) = \text{f} \forall k, 1 \leq k \leq n \).

By the induction step hypothesis, \text{conseq}_P(I)(D_g) = \text{i}, \forall g, 1 \leq g \leq m.

Thus \text{conseq}_P(I)(A) = \text{i}.

The induction is proved and we can conclude that \text{conseq}_P(I)(A) = \text{i}.

(3) \text{conseq}_P(J)(A) = \text{u}.

We affirm that for all \( n \geq 1 \), if \( n \) is such that \( 4-T_{pg(P,J)}^n(\bot)(A) = \text{u} \), and \( \forall j, j \geq n, 4-T_{pg(P,J)}^j(\bot)(A) = \text{u} \), then \text{conseq}_P(I)(A) = \text{u}. \) This assertion is proved by induction on \( n \).

- Induction base: \( n=1 \). By our hypothesis, in \( \text{ground}(P) \) there is a rule of the form: \( A \leftarrow B_1, \ldots, B_n, c_1, \ldots, c_p \), where \( B_k \) are atoms and \( c_g \) are truth-values, \( J(B_k) \in \{ \text{f}, \text{u} \} \forall k, 1 \leq k \leq n, c_g = \text{u} \forall g, 1 \leq g \leq p \), and there is a \( w, 1 \leq w \leq p, c_w = \text{u} \) or there is a \( y, 1 \leq y \leq n, J(B_y) = \text{u} \).

By \( I \models J \), \( I(B_k) \in \{ \text{f}, \text{u} \} \forall k, 1 \leq k \leq n \).

Thus \text{conseq}_P(I)(A) = \text{u} and the induction base is proved.
• Induction step: by our hypothesis, in \( \text{ground}(P) \) there is a rule of the form: 
\( A \leftarrow \sim B_1, \ldots, \sim B_n, D_1, \ldots, D_m \), where \( B_k, \ D_g \) are atoms, \( J(B_k) \in \{f,u\} \ \forall k, \ 1 \leq k \leq n \) and \( 4-T^m_{pg}(P,J)(\bot)(D_g) \geq u, \ \forall g, \ 1 \leq g \leq m \).

By \( I \not\subseteq J, I(B_k) \in \{f,u\} \ \forall k, \ 1 \leq k \leq n \).

By the induction step hypothesis, \( \text{conseq}_P(I)(D_g) \geq u, \ \forall g, \ 1 \leq g \leq m \).

Thus \( \text{conseq}_P(I)(A) \geq u \).

The induction is proved and we can conclude that \( \text{conseq}_P(I)(A) \geq u \).

(4) \( \text{conseq}_P(J)(A)=f \).

As \( \text{conseq}_P(I)(A) \geq f \), we conclude that \( \text{conseq}_P(I)(A) \geq \text{conseq}_P(J)(A) \). □

From theorem 1, we can easily see that in the alternating sequence \( \{I_i\}_{i \geq 0} \) we have:

\[
I_0 \preceq I_2 \preceq \cdots \preceq I_{2i} \preceq I_{2i+1} \preceq \cdots \preceq I_{2i-1} \preceq I_{2i+2} \preceq \cdots \preceq I_3 \preceq I_1 \quad (1)
\]

Thus the even subsequence is increasing and the odd one is decreasing. Because there are finitely many 4-valued instances relatively to a given program \( P \), each of these sequences becomes constant at some point: \( I_{2k_0} = I_{2k_0+2} = \cdots, I_{2k_0+4} = \cdots \) and \( I_{2j_0+1} = I_{2j_0+3} = I_{2j_0+5} = \cdots \), for some \( k_0 \geq 0 \) and some \( j_0 \geq 0 \).

Let \( I_* \) be the least upper bound of the increasing sequence: \( I_* = \text{lub}\{I_{2i}\}_{i \geq 0} \),

and let \( I^* \) be the greatest lower bound of the decreasing sequence:

\( I^* = \text{glb}\{I_{2i+1}\}_{i \geq 0} \). From (1), it follows that \( I_* \preceq I^* \).

**Lemma 2** Let \( I \) be a 4-valued instance of a P-Datalog program. Then \( \text{conseq}_P(I_*) = I^* \) and \( \text{conseq}_P(I^*) = I_* \).

**Proof:** Let \( \{I_i\}_{i \geq 0} \) be an alternating sequence of 4-valued instances, where \( I_* = I_{2k} = I_{2k+2} \) e \( I^* = I_{2k+1} = I_{2k+3}, \ k \geq 0 \). We need to show that \( \text{conseq}_P(I^*) = I_* \) and \( \text{conseq}_P(I_*) = I^* \). In \( \{I_i\}_{i \geq 0} \) we have \( I_{i+1} = \text{conseq}_P(I_i) \):

\[
I_{2k} = \text{conseq}_P(I_{2k-1}), \ I_{2k+1} = \text{conseq}_P(I_{2k}), \ I_{2k+2} = \text{conseq}_P(I_{2k+1}), \ I_{2k+3} = \text{conseq}_P(I_{2k+2}).
\]

Thus, \( I_* = \text{conseq}_P(I^*) \) and \( I^* = \text{conseq}_P(I_*) \). □

From the 4-valued instances \( I_* \) and \( I^* \) we can define the 4-valued instance \( I^*_* \) which coincides with the well-founded semantics of a P-Datalog program, as we will see in Theorem 3.

**Definition 12** Let \( I^*_* \) be a 4-valued instance of a P-Datalog program \( P \), consisting of the facts known in both \( I_* \) and \( I^* \), that is:

\[
I^*_*(A) = \begin{cases} 
  t & \text{if } I_*(A) = I^*(A) = t \\
  i & \text{if } I_*(A) = I^*(A) = i \\
  f & \text{if } I_*(A) = I^*(A) = f \\
  u & \text{otherwise} 
\end{cases}
\]
Theorem 2 Let \( \mathbf{I} \) be a 4-valued instance of a P-Datalog program \( P \). Then \( \mathbf{I}_s \preceq \mathbf{I}_s^* \preceq \mathbf{I}^* \).

Proof. By the definition of \( \mathbf{I}_s^* \), \( \mathbf{I}_s \preceq \mathbf{I}_s^* \preceq \mathbf{I}^* \) is verified for all cases except when \( \mathbf{I}_s(A) = \mathbf{i} \) and \( \mathbf{I}^*(A) = \mathbf{t} \). We show that this cannot happen. Let us suppose that \( \mathbf{I}_s(A) = \mathbf{i} \) and \( \mathbf{I}^*(A) = \mathbf{t} \). By the fact that \( \text{conseq}_P(\mathbf{I}_s) = \mathbf{I}_s^* \), \( \text{conseq}_P(\mathbf{I}^*) = \mathbf{I}_s \), it follows that \( \text{conseq}_P(\mathbf{I}_s^*)(A) = \mathbf{i} \) and \( \text{conseq}_P(\mathbf{I}_s)(A) = \mathbf{t} \). We can prove by induction on \( n \) that: for all \( n \geq 1 \), if \( n \) is such that \( 4 - T_{pg(P, I_s)}^n(\bot)(A) = \mathbf{t} \), and \( \forall j, j \geq n, 4 - T_{pg(P, I_s)}^j(\bot)(A) = \mathbf{t} \), then \( \text{conseq}_P(\mathbf{I}_s^*)(A) \in \{ \mathbf{f}, \mathbf{t} \} \). So, we can conclude that \( \text{conseq}_P(\mathbf{I}_s^*)(A) \in \{ \mathbf{f}, \mathbf{t} \} \). Contradiction.

The fixpoint construction yields the well-founded semantics for P-Datalog programs. The following theorem is the main result of this paper. It shows that each P-Datalog program has at least one 4-stable model \( (\mathbf{I}_s^*) \) and that the well-founded semantics coincides with \( \mathbf{I}_s^* \).

Theorem 3 For each P-Datalog program \( P \), (a) \( \mathbf{I}_s^* \) is a 4-stable model of \( P \) and (b) \( \mathbf{I}_s^* \) equals the well-founded semantics of \( P \) (\( P^{4wf} \)).

Proof. For statement (a), we need to show that \( \text{conseq}_P(\mathbf{I}_s^*) = \mathbf{I}_s^* \).
From Theorems 1 and 2, we can easily conclude that:

\[
\mathbf{I}_s \preceq \text{conseq}_P(\mathbf{I}_s^*) \preceq \mathbf{I}^*
\]

We affirm that \( \text{conseq}_P(\mathbf{I}_s^*)(A) = \mathbf{I}_s^*(A) \), for all \( A \in \mathbf{B}(P) \). Indeed: (1) If \( \mathbf{I}_s^*(A) \in \{ \mathbf{f}, \mathbf{t}, \mathbf{i} \} \), by the definition of \( \mathbf{I}_s^* \) and from (2), we can conclude that \( \text{conseq}_P(\mathbf{I}_s^*)(A) = \mathbf{I}_s^*(A) \).

(2) Let \( \mathbf{I}_s^*(A) = \mathbf{u} \). By the definition of \( \mathbf{I}_s^* \) we have the following possibilities:
(2a) \( \mathbf{I}_s(A) = \mathbf{i} \) and \( \mathbf{I}^*(A) = \mathbf{t} \). By Theorem 2, this case does not happen.
(2b) \( \mathbf{I}_s(A) = \mathbf{f} \) and \( \mathbf{I}^*(A) \in \{ \mathbf{i}, \mathbf{t} \} \). This can be written as \( \text{conseq}_P(\mathbf{I}_s^*)(A) = \mathbf{f} \) and \( \text{conseq}_P(\mathbf{I}_s)(A) \geq \mathbf{i} \). By definition, \( \text{conseq}_P(\mathbf{I}_s) \) is the least fixpoint of \( pg(P, I_s) \) \( \bot \). We affirm that for all \( n \geq 1 \), if \( n \) is such that \( 4 - T_{pg(P, I_s)}^n(\bot)(A) \geq \mathbf{i} \), and \( \forall j, j \geq n, 4 - T_{pg(P, I_s)}^j(\bot)(A) = 4 - T_{pg(P, I_s)}^n(\bot)(A) \), then \( \text{conseq}_P(\mathbf{I}_s^*)(A) = \mathbf{u} \).
This assertion can be proved by induction on \( n \). So, we can conclude that \( \text{conseq}_P(\mathbf{I}_s^*)(A) = \mathbf{I}_s(A) = \mathbf{u} \).

For statement (b), we will show that \( P^{4wf} = \mathbf{I}_s^* \). Let \( A \in \mathbf{B}(P) \).
• Let \( P^{4wf}(A) = \mathbf{t}(\mathbf{i}, \mathbf{f}) \). By definition, \( P^{4wf} \) is the 4-valued instance consisting of facts that are true (resp. false, inconsistent) in all 4-stable models of \( P \).
From statement (1), we can affirm that \( \mathbf{I}_s^*(A) \) is a 4-stable model. So \( \mathbf{I}_s^*(A) = P^{4wf}(A) \).
• Let \( P^{4wf}(A) = \mathbf{u} \). So, there exist 4-stable models \( M_1 \) and \( M_2 \) such that
$M_1(A) \neq M_2(A)$. Let us suppose that $I^*_*(A) = \text{t}$ (i, f). We affirm that for all 4-stable model $M$ of program $P$, and for all $i \geq 0$, we have:

$$I_{2i} \preceq M \preceq I_{2i+1}. \quad (3)$$

This assertion can be proved by induction on $i$ and its proof does not present any difficulty.

By (3), $I_{2j} \preceq M \preceq I^*$, for $j = 1, 2$.

• If $I^*_*(A) = \text{t}$, then $I_*^*(A) = t$, and by $I_*^*(A) \preceq M_j(A)$, we conclude that $M_1(A) = M_2(A) = t$.

• If $I^*_*(A) = i$, then $I_*^*(A) = I^*(A) = i$, and by $I_*^*(A) \preceq M_j(A) \preceq I^*(A)$, we conclude that $M_1(A) = M_2(A) = i$.

• If $I^*_*(A) = f$, then $I_*^*(A) = f$, and by $I_*^*(A) \preceq M_j(A)$, we conclude that $M_1(A) = M_2(A) = f$.

Contradiction. So, $I_*^*(A) = u$. \hfill \square

We illustrate this computation in our running example:

**Example 5 (I^*_* computation)** Consider again the program $P_{job}$ and the database instance $I$ of the running example 1. Note that for $I_0$ the value of all facts is $f$, and for each $j \geq 1$, $I_j$ agrees with the input $I$ on the predicates supportedby and owe. Therefore we only show the inferred $job$-facts:

$I_0 = \{\sim job(charles), \sim job(james), \sim job(john), \sim job(joseph),$

$\sim job(kevin), \sim job(paul)\}.$

$I_1 = \{\circ job(charles), \circ job(james), \bullet job(john), \circ job(joseph), \sim job(kevin),$

$\circ job(paul)\}.$

$I_2 = \{\sim job(charles), \sim job(james), \bullet job(john), \sim job(joseph), \sim job(kevin),$

$\sim job(paul)\}.$

$I_3 = \{\circ job(charles), \sim job(james), \bullet job(john), \circ job(joseph), \sim job(kevin),$

$\circ job(paul)\}.$

$I_4 = \{\sim job(charles), \sim job(james), \bullet job(john), \sim job(joseph), \sim job(kevin),$

$\circ job(paul)\}.$

$I_5 = I_3$ and $I_6 = I_4$. So, $I_* = I_6$ and $I^* = I_5$. Thus $I^*_* = \{\sim job(james), \bullet job(john), \sim job(kevin), \circ job(paul)\}$. This is exactly the natural answer for $P_{job}$ we have informally discussed in example 1.
6 Related Work

The problem of inconsistent information management arising from the integration of heterogeneous sources has been widely studied in recent research. We can distinguish three main approaches to handle the inconsistency problem in knowledge bases: a consistency-based approach (where inconsistency is eliminated), a paraconsistent logic approach (where inconsistency is not rejected and inference methods can draw plausible conclusions from it) and a hybrid approach (formalisms which do not reject any information but instead associate degrees of belief, reliability or uncertainty to each source knowledge base).

In what follows, we discuss some recent papers which deal with the subject of integrating multi-source information and querying the resulting integrated database which maybe contain inconsistencies. The methods are grouped following the three main approaches mentioned above. We begin the discussion by summarizing the main issues treated in this paper.

Our approach. The present paper follows the paraconsistent logic approach. We assume that data integration has already been achieved following the method we proposed in [15]. This method follows a paraconsistent approach for data integration and is based on a tableau proof system of \( LFI_1 \) [11]. It supposes the existence of \( n \) source databases \( K_1, K_2, ..., K_n \), where each database \( K_i \) satisfies a set of integrity constraints \( C_i \) (properties expressed by first-order formulas). The method builds an integrated database \( K \) which satisfies the set \( \{C_1, C_2, ..., C_n\} \) of integrity constraints, where the notion of satisfiability is taken from the paraconsistent logic \( LFI_1 \). The integrated database \( K \) is a paraconsistent database: each fact stored in it can be either sure (positive or negative) or inconsistent. In the present paper, we propose a deductive language P-Datalog which allows to query this integrated database containing inconsistencies. The querying process semantics follows the paraconsistent semantics of the logic \( LFI_1 \).

Consistent-based approaches. In [4,3] a logic framework based on annotated predicate logic [20] is proposed for obtaining consistent answers when querying a database with inconsistent information with respect to a set of integrity constraints. A consistent answer to a given query \( q \) is the one that every database repair would give in response to \( q \). A procedure for building consistent answers without actually repairing the inconsistent database is described. In [2], a method based on the classical tableau system proof for first-order logic is introduced for producing consistent answers in a relational database that may violate given integrity constraints. Roughly speaking, these answers are those obtained in all repaired versions of the database (a repaired version is obtained by inserting or deleting information from the original database in
order to make it consistent with the integrity constraints). In [10], a method for specifying the database repairs of a mediated integration system under the paradigm “Local-as-View“ for data integration has been proposed. The repairs are specified by means of a disjunctive logic program. The consistent answers to queries posed to such a system are computed by running a disjunctive logic program together with the specification of database repairs. Following this same line, in [17], a generic logic programming framework for computing consistent answers to queries posed to a data integration system (in which inconsistency possibly arises) is proposed. This approach is also based on the specification of repairs and queries by means of a logic program. The main objective is optimizing the evaluation of queries expressed as logic programs.

In [6], one proposes a method to specify database repairs using simple classical normal disjunctive logic programs with a stable model semantics. The database predicates in these programs contain annotations as extra arguments. The annotations come from the classical annotated predicated logic, contrarily to annotated or paraconsistent logic programs [9,27]. Even though their method produces all the possible minimal sets of changes required to restore the consistency of a relational database, the main goal of this method is to obtain the consistent answers to a first order query. Our language P-Datalog is designed to query paraconsistent databases, i.e., where inconsistent information is kept inside the database with a special truth-value which distinguish it from safe information. When using P-Datalog to query a paraconsistent database, one can specify the type of answers we are interested in: consistent or inconsistent, or both. P-Datalog classifies each fact in a query answer as true, false, inconsistent or unknown. In [5], the Belnap 4-valued logic is used to represent different degrees of contradiction and partial information. The lattice of truth-values they use is different from ours. The intuition behind their semantics is that contradictory data corresponds to inadequate information about the real world, and therefore it should be minimized. In [23], an abductive method for coherent integration of independent datasources is proposed. The method is based on SLDNF-Resolution. The general idea consists in computing a list of data-facts that should be inserted to the concatenated database or deleted from it in order to restore its consistency.

**Hybrid approaches.** In [13] a majority merging operator is used in the definition of a logic that allows us to reason with data coming from different sources. The possible inconsistencies that would come from the integration process are solved according to that majority operator. In our approach, we have assumed that our data have already been integrated and in this integrated databases, inconsistent information has been detected and marked. In [7], several methods allowing inference of information coming from different sources are studied. The methods proposed can be grouped in two categories: those following a coherent-based approach (where inconsistencies are discarded) and those where inconsistent information is kept. The later ones are closer to our
approach since they do not rely on restoring consistency in the integrated database. The authors suppose the knowledge bases are sets of formulas which are stratified, that is, there exists a total ordering between the sources based on a notion of reliability: source $\mathcal{K}_i$ is more reliable than source $\mathcal{K}_j$ if $j > i$. In our approach, the information kept in the integrated database is not stratified, since in most situations, there is no means of attaching a degree of reliability to information coming from different sources (for instance, simple information like client addresses may have different values in two different databases, and there is no means of saying that the first database is more reliable than the second one). Besides, the problem we proposed is different: our integrated database is a set of (possibly inconsistent) facts and we propose a language to query this set of facts, producing a new set of facts, where each element is clearly identified as positive sure information, negative sure information or inconsistent information.

Paraconsistent Approach. The approach proposed in [26] is close to the approach we introduced in this paper. Among all the methods we mentioned so far for querying databases containing inconsistent information, this is the only one based on a paraconsistent logic programming framework, like ours. This work proposes a deductive query language and a declarative and fixpoint-based well-founded semantics for this language. The underlying logic is a paraconsistent logic with 7 truth-values: true, inconsistent, false, indifferent, true by default, false by default, and unknown. An interpretation satisfies a formula if the truth-value it associates to the formula is true or inconsistent. The main differences between this approach and ours are the following: (1) The underlying logic is far more complex than LFI1 since it involves 7 truth-values (2) The body of the clauses contains explicit negation and default negation, and the head may contain a negative atom. Thus the language is not a simple extension of Datalog$\neg$. P-Datalog has been designed to extend to a paraconsistent context the classical Datalog$\neg$ query language used for querying consistent databases. This has been achieved by augmenting the three-value semantics of Datalog$\neg$ with only one new value (inconsistent).

7 Conclusion and Further Research

In this paper we have introduced the deductive query language P-Datalog for querying databases resulting for the concatenation of several sources which together can contain inconsistent information. We have provided a declarative well-founded semantics which can be evaluated by means of an alternating fixpoint. We have decided to implement the P-Datalog prover as a separate system and further to integrate it in a relational database system. For now, the P-Datalog prover is a helpful tool for validating the well-founded semantics we have proposed for P-Datalog programs. It has been implemented in Objec-
tive Caml [21]. The OCaml compiler generates code whose executing time is comparable to a C/C++ code, and it includes libraries for several platforms. Those characteristics and also the functional programming qualities allowed us to focus on the difficulties of our application and to develop a preliminary succinct solution.

Several work remains to be done. We will extend the syntax of P-Datalog in order to allow literals of the form $\circ A$, $\bullet A$ and $\neg A$ in the body of the rules. The $\circ$, $\bullet$ and $\neg$ operators are non-monotonic and that will imply in modifications on the definition of extended P-Datalog program. Another direction for further research consists in relating the well-founded semantics of Datalog$^\neg$ queries posed to the source databases with the well-founded semantics of P-Datalog we have proposed in this paper. How the answers of Datalog$^\neg$ queries submitted to the (consistent) source databases $D_1, D_2, ..., D_n$ are related to the answer of a P-Datalog query submitted to the integrated database $D = D_1 \cup ... \cup D_n$. More precisely, given a P-Datalog query $Q$ over $D$, there exists Datalog$^\neg$ queries $Q_1, Q_2, ..., Q_n$ over $D_1, ..., D_n$ respectively, such that the integration of their answers (using the integration method introduced in [15]) coincides with the answer of the P-Datalog $Q$? A complete investigation of how Datalog$^\neg$ queries over the consistent sources relate to P-Datalog queries over the concatenated database would formalize the intuitive idea we gave of P-Datalog query answering in Example 1.

Finally, we intend to present a more efficient implementation of P-Datalog, based on a proof system for the LFI1 paraconsistent logic. In [11], a sound and complete tableau proof system for LFI1 has been developed and in [15], a database integration method based on this proof system has been introduced. We intend to use this proof system or an equivalent resolution proof system for evaluating the well-founded semantics of P-Datalog queries.

Acknowledgments: We would like to thank the anonymous referees whose really careful reading, useful comments and corrections have lead to significant improvement of our work.

References


