This paper introduces a new approach for conflict resolution: given a set of tuples pertaining to the same entity, it is to identify a single tuple in which each attribute has the latest and consistent value in the set. This problem is important in data integration, data cleaning and query answering. It is, however, challenging since in practice, reliable timestamps are often absent, among other things. We propose a model for conflict resolution, by specifying data currency in terms of partial currency orders and currency constraints, and by enforcing data consistency with constant conditional functional dependencies. We show that identifying data currency orders helps us repair inconsistent data, and vice versa. We investigate a number of fundamental problems associated with conflict resolution, and establish their complexity. In addition, we introduce a framework and develop algorithms for conflict resolution, by integrating data currency and consistency inferences into a single process, and by interacting with users. We experimentally verify the accuracy and efficiency of our methods using real-life and synthetic data.

Categories and Subject Descriptors: H.2 [Database Management]: General—integrity
General Terms: Theory, Algorithms, Experimentation
Additional Key Words and Phrases: Conditional functional dependency, Data cleaning

1. INTRODUCTION
Conflict resolution is the process that, given a set \( I \) of tuples pertaining to the same entity, fuses the tuples into a single tuple and resolves conflicts among the tuples of \( I \) [Dong and Naumann 2009]. Traditional work resolves conflicts typically by taking, e.g., the \( \text{max} \), \( \text{min} \), \( \text{avg} \), any of attribute values. We refer to [Bleiholder and Naumann 2008] for a survey on conflict resolution.

We study a new approach for conflict resolution, by highlighting both data currency and data consistency. Given \( I \), we want to identify a single tuple in which each attribute has consistent and the most current value taken from \( I \), referred to as the true values of the entity relative to \( I \). The need for studying this problem is evident in data integration, where conflicts often emerge from values that refer to the same entity and come from different sources. It is also common to find multiple values of the same entity residing in a database. While these values were once correct, i.e., they were the true values of the entity at some time, some of them may have become stale and thus inconsistent. Indeed, it is estimated that in a customer database, about 50\% of the records may become obsolete within two years [Eckerson 2002]. With these comes the need for resolving conflicts for, e.g., data fusion [Bleiholder and Naumann 2008; Dong and Naumann 2009], data cleaning [Arenas et al. 1999] and query answering with current values [Fan et al. 2012].

No matter how important, the problem is rather challenging. Indeed, it is already highly nontrivial to find consistent values for an entity [Arenas et al. 1999; Cong et al. 2007]. Moreover, it is hard to identify the most current entity values [Fan et al. 2012] since in the real world, reliable timestamps are often absent [Zhang et al. 2010; Goldring 1995]. Add to this the complication that when resolving conflicts one has to find the entity values that are both consistent and most current.

Example 1: The photo in Fig. 1 is known as “V-J Day in Times Square”. The nurse and sailor in the photo have been identified as Edith Shain and George Mendonça,
currency constraints and constant CFDs

respectively, and their information is collected in sets $E_1$ and $E_2$ of tuples, respectively, shown in Fig. 2.

We want to find the true values of these entities, i.e., a tuple $t_1$ for Edith (resp. a tuple $t_2$ for George) such that the tuple has the most current and consistent attribute values for her (resp. his) status, job, the number of kids, city, AC (area code), zip and county in $E_1$ (resp. $E_2$). However, the values in $E_1$ ($E_2$) have conflicts, and worse still, they do not carry timestamps. They do not tell us, for instance, whether Edith still lives in NY, or even whether she is still alive.

The situation is bad, but not hopeless. We can often deduce certain currency orders from the semantics of the data. In addition, dependencies such as conditional functional dependencies (CFDs) [Fan et al. 2008] have proven useful in improving the consistency of the data. Better still, data currency and consistency interact with each other. When they are taken together, we can often infer some true values from inconsistent tuples, even in the absence of timestamps, as illustrated below.

Example 2: From the semantics of the data, we can deduce the currency constraints and CFDs shown in Fig. 3.

(1) Currency constraints. We know that for each person, status only changes from working to retired and from retired to deceased, but not from deceased to working or retired. These can be expressed as $\varphi_1$ and $\varphi_2$ given in Fig. 3, referred to as currency constraints.
Here \( t_1 \prec_{\text{status}} t_2 \) denotes a partial currency order defined on the attribute status, indicating that \( t_2 \) is more current than \( t_1 \) in attribute status. Similarly, we know that job can only change from \textit{sailor} to \textit{veteran} but not the other way around. We can express this as currency constraint \( \varphi_3 \), shown in Fig. 3. Moreover, the number of kids typically increases monotonically. We can express this as \( \varphi_4 \), assuring that \( t_2 \) is more current than \( t_1 \) in attribute kids if \( t_1[{\text{kids}}] < t_2[{\text{kids}}] \).

In addition, we know that for each person, if tuple \( t_2 \) is more current than \( t_1 \) in attribute status, then \( t_2 \) is also more current than \( t_1 \) in job, AC and zip. Furthermore, if \( t_2 \) is more current than \( t_1 \) in attributes city and zip, it also has a more current county than \( t_1 \). These can be expressed as currency constraints \( \varphi_5 \sim \varphi_8 \).

(2) \textit{Constant CFDs.} In the US, if the AC is 213 (resp. 212), then the city must be LA (resp. NY). These are expressed as conditional functional dependencies \( \psi_1 \) and \( \psi_2 \) shown in Fig. 3.

We can apply these constraints to the set \( E_1 \) of tuples given in Fig. 2, to improve the currency and consistency of the data. By interleaving inferences of data currency and data consistency, we can actually identify the true values of entity Edith, as follows:

(a) from the currency constraints \( \varphi_1 \) and \( \varphi_2 \), we can conclude that her latest status is deceased;
(b) similarly, by \( \varphi_4 \), we find that her true kids value is 3 (assuming null \( < k \) for any number \( k \));
(c) from (a) above and \( \varphi_5 \sim \varphi_7 \), we know that her latest job, AC and zip are n/a, 213 and 90058, respectively;
(d) after currency inferences (a) and (c), we can apply the CFD \( \psi_1 \) and find her latest job as LA; and
(e) after the consistency inference (d), from (c) and (d) we get her latest county as Vermont, by applying the currency constraint \( \varphi_8 \).

Now we have identified a single tuple
\[
t_1 = \text{(Edith Shain, deceased, n/a, 3, LA, 213, 90085, Vermont)}
\]
as the true values of Edith Shain relative to the set \( E_1 \) of tuples (the address is for her cemetery). \( \Box \)

This example suggests the following. (1) Data currency and consistency should be interleaved when resolving conflicts. Indeed, not only deducing currency orders helps us improve the consistency (e.g., from steps (a), (c) to (d)), but data consistency inferences also help us identify the most current values (e.g., step (e) is doable only after (d)). (2) Both data currency and data consistency can be specified with constraints, and hence, can be processed in a uniform logical framework.

While the need for deducing the consistent and most current values has been advocated for conflict resolution [Dong and Naumann 2009; Motro and Anokhin 2006], prior work typically assumes the availability of timestamps. Previous work on data quality focuses on either data consistency (e.g., [Arenas et al. 1999; Fan et al. 2008; Cong et al. 2007; Yakout et al. 2010]) or data currency (e.g., [Fan et al. 2012]). However, no models or algorithms are yet in place to combine data consistency and currency for conflict resolution.

**Contributions.** We study conflict resolution by inferring both data currency and data consistency.

(1) We propose a model for conflict resolution (Section 2). We specify data currency in terms of (a) partial currency orders denoting available (yet possibly incomplete) temporal information on the data, and (b) simple currency constraints, to express currency
relationships derived from the semantics of the data. Data consistency is specified in terms of constant CFDs [Fan et al. 2008] on the latest values of the data. Given such a specification \( S_e \) on a set \( E \) of tuples pertaining to the same entity \( e \), we aim to derive the true values of \( e \) from \( S_e \).

(2) We introduce a framework for conflict resolution (Section 3). One may find some true values of an entity from its specification, but not all, as illustrated below.

**Example 3:** Consider the set \( E_2 \) of tuples for entity George Mendonça (Fig. 2). Along the same lines as Example 2, we find that its true \((\text{name}, \text{kids})\) values are \((\text{George Mendonça}, 2)\). However, we do not have sufficient information to infer the true values of the other attributes of this entity.

In light of this, our framework **automatically derives** as many true values as possible from a given specification \( S_e \) of an entity \( e \), identifies attributes for which the true values of \( e \) are not derivable from \( S_e \), and interacts with users to solicit additional input for those attributes, so that all the true values of all the attributes of \( e \) can be derived from \( S_e \) and users’ input.

(3) We study problems fundamental to conflict resolution (Section 4). Given a specification \( S_e \), we determine whether partial currency orders, currency constraints and CFDs in \( S_e \) have conflicts among themselves? Whether some other currency orders are implied by \( S_e \)? Whether true values of an entity can be derived from \( S_e \)? If not, what additional minimum currency information has to be provided so that the true values are derivable? We establish their complexity bounds, ranging from \( NP \)-complete and co\( NP \)-complete to \( \Sigma^p_2 \)-complete. These results reveal the complexity inherent to conflict resolution.

(4) We develop several practical algorithms (Section 5). We propose methods for finding (a) whether a specification \( S_e \) has conflicts, (b) what true values can be derived from \( S_e \), and (c) a minimum set of attributes that require users’ input to find their true values. All these problems are **intractable**; in particular, the last problem is \( \Sigma^p_2 \)-complete. Nevertheless, we provide efficient heuristic algorithms, by integrating inferences of data consistency and currency into a single process.

(5) We evaluate the accuracy and efficiency of our method using real-life and synthetic data (Section 6). We find that unifying currency and consistency substantially improves the accuracy of traditional methods, by 201% (f-measure), even with only a small number of constraints. It is also more effective than taking consistency and currency separately. Furthermore, our algorithms are efficient, and scale well with the number of tuples pertaining to an entity and with the number of constraints; for example, it takes an average of 7 seconds to resolve conflicts in sets of 8k-10k tuples representing an entity, with 1983 constraints.

We contend that this work provides fundamental results for conflict resolution, and propose a practical solution via inferences of data currency and data consistency in the absence of timestamps.

**Related work.** This work extends [Fan et al. 2013] by including (1) a comprehensive analysis of the fundamental problems in connection with conflict resolution in terms of currency constraints and constant CFDs (Section 4); and (2) a detailed discussion of algorithms and procedures needed for finding the most current values (Section 5). None of the proofs of (1) was presented in [Fan et al. 2013]. These proofs are interesting in their own right. Most procedures of (2) were not given in [Fan et al. 2013].

Conflict resolution has been studied for decades, started from [Dayal 1983]. It aims to combine data from different sources into a single representation (see [Bleiholder...}
Conflict Resolution with Data Currency and Consistency

and Naumann 2008; Dong and Naumann 2009] for surveys). In that context, inconsistencies are typically resolved by selecting the \( \max \), \( \min \), \( \text{avg} \), or any value [Bleiholder and Naumann 2008]. While the need for data currency was also observed there (e.g., [Dong and Naumann 2009; Motro and Anokhin 2006]), previous work identifies current values only by using timestamps. This work differs from the traditional work in the following. (1) We revise the conflict resolution problem to identify values of entities that are both consistent and most current. (2) We do not assume the availability of timestamps, which are often missing in practice [Zhang et al. 2010; Goldring 1995]. (3) We resolve conflicts by using currency constraints and CFDs [Arenas et al. 1999; Fan et al. 2008; Cong et al. 2007], instead of picking \( \max \), \( \min \), \( \text{avg} \) or any value. (4) We employ automated reasoning to identify true values by unifying the inferences of data currency and consistency.

There has been work on truth discovery from data sources [Dong et al. 2009; Galland et al. 2010; Yin et al. 2008]. Their approaches include (1) vote counting and probabilistic computation based on the trustworthiness of data sources [Galland et al. 2010; Yin et al. 2008]; (2) source dependencies to find copy relationships and reliable sources [Dong et al. 2009]; and (3) employing lineage information and probabilities [Widom 2005]. In contrast, we assume no information about the accuracy of data sources, but derive true values based on data currency and consistency. In addition, we adopt a logical approach via automated reasoning about constraints, as opposed to probabilistic computation. This work is complementary to the previous work and can be combined with the prior approaches.

This work extends [Fan et al. 2012; Fan et al. 2008]. A data currency model was presented in [Fan et al. 2012] with partial currency orders and denial constraints [Arenas et al. 1999]. CFDs were studied for specifying data consistency [Fan et al. 2008]. This work differs from [Fan et al. 2012; Fan et al. 2008] in the following. (1) We propose a conflict resolution model that combines data currency and consistency. In contrast, [Fan et al. 2012] only studies data currency, while [Fan et al. 2008] only considers data consistency. (2) We interleave inferences of data currency and consistency, which is far more intriguing than handling currency and consistency separately, and requires new techniques to capture the interaction between the two. (3) We use currency constraints, which are simpler than denial constraints, to strike a balance between the complexity of inferring true values and the expressivity needed for specifying currency (Section 4). (4) No practical algorithms were given in [Fan et al. 2012] for deriving current values.

Previous work on data consistency [Arenas et al. 1999; Fan et al. 2008; Cong et al. 2007; Yakout et al. 2010; Greco et al. 2003] has been focusing on consistent query answering and data repairing [Bertossi 2011], topics different from conflict resolution. The study of preferred repairs [Greco et al. 2003] also advocates partial orders. It differs from the currency orders we study here in that they use PTIME functions to rank different repairs over the entire database, whereas we derive the currency orders by automated reasoning about both available partial temporal information and currency constraints. Preferred repairs are implemented by [Cong et al. 2007] via a cost metric, and by [Yakout et al. 2010] based on a decision theory, which can be incorporated into our framework.

There has been a large body of work on temporal databases (see [Chomicki and Toman 2005] for a survey). In contrast to that line of work, we do not assume the availability of timestamps. It has also recently been shown that temporal information helps record linkage identify records that refer to the same entity [Li et al. 2011]. Here we show that data currency helps conflict resolution as well, a different process that takes place after record linkage has identified tuples pertaining to the same entity. While [Li et al. 2011] is based on timestamps, we do not assume it here.
2. A CONFLICT RESOLUTION MODEL

In this section we show how to capture conflicts in terms of data currency and data consistency. We start with data currency (Section 2.1) and consistency (Section 2.2) specifications. We then present our model for characterizing conflicts commonly found in the real world by means of the specifications (Section 2.3).

2.1. Data Currency

We specify the currency of data by means of (a) partial currency orders, and (b) currency constraints.

**Data with partial currency orders.** Consider a relation \( R = (A_1, \ldots, A_n) \), where each attribute \( A_i \) has a domain \( \text{dom}(A_i) \). In this work we focus on *entity instances* \( I_e \) of \( R \), which are sets of tuples of \( R \) all pertaining to the same real-world entity \( e \), and are typically much smaller than a database instance. Such entity instances can be identified by e.g., record linkage techniques (see [Elmagarmid et al. 2007] for a survey).

For an attribute \( A_i \in R \) and an entity instance \( I_e \) of \( R \), we denote by \( \text{adom}(I_e, A_i) \) the set of \( A_i \)-attribute values that occur in \( I_e \), referred to as the *active domain* of \( A_i \) in \( I_e \).

We have seen two entity instances given in Fig. 2: \( E_1 = \{r_1, r_2, r_3\} \) for entity “Edith”, and \( E_2 = \{r_4, r_5, r_6\} \) for “George”. Here \( \text{adom}(E_1.\text{city}) = \{\text{NY}, \text{SFC}, \text{LA}\} \); similarly for other attributes.

A temporal instance \( I_t \) of \( I_e \) is given as \( (I_e, \preceq_{A_1}, \ldots, \preceq_{A_n}) \), where each \( \preceq_{A_i} \) is a partial order on \( I_e \), referred to as the *currency order for attribute* \( A_i \) for the entity \( e \) represented by \( I_e \). For all \( t_1, t_2 \in I_e \), \( t_1 \preceq_{A_i} t_2 \) iff either \( t_1 \) and \( t_2 \) share the same \( A_i \)-attribute value (i.e., \( t_1[A_i] = t_2[A_i] \)), or \( t_2[A_i] \) is more current than \( t_1[A_i] \) (denoted by \( t_1 \prec_{A_i} t_2 \)).

Intuitively, currency orders represent available temporal information about the data. Observe that \( \preceq_{A_i} \) is a *partial order*, possibly empty. For example, for \( E_1 \) above, we only know that \( r_3 \preceq_{\text{kids}} r_1 \) and \( r_3 \preceq_{\text{kids}} r_2 \) since \( r_3[\text{kids}] \) is null, which are in the currency order \( \preceq_{\text{kids}} \), while the currency orders for other attributes are empty, excluding the case when tuples carry the same attribute value. Similarly for \( E_2 \). In particular, \( t_1 \preceq_{A_i} t_2 \) if \( t_1[A_i] \) is null, i.e., an attribute with value missing is ranked the lowest in the currency order.

**Current instances.** Currency orders are often incomplete. Hence we consider possible completions of currency orders. More specifically, a *completion* \( I_t^* \) of \( I_t \) is a temporal instance \( I_t^* = (I_e, \preceq_{A_1}^*, \ldots, \preceq_{A_n}^*) \), such that for each \( i \in [1, n] \), (1) \( \preceq_{A_i} \subseteq \preceq_{A_i}^* \), and (2) for all tuples \( t_1, t_2 \in I_e \), either \( t_1 \preceq_{A_i}^* t_2 \) or \( t_2 \preceq_{A_i}^* t_1 \). That is, \( \preceq_{A_i}^* \) induces a *total order* on tuples in \( I_e \). Intuitively, \( I_t^* \) totally sorts the attribute values in \( I_e \) such that the most current value of each attribute is the last in the order.

We define the *most current* \( A_i \)-attribute value of \( I_t^* \) to be \( t[A_i] \) that comes last in the total order \( \preceq_{A_i}^* \). The *current tuple* of \( I_t^* \), denoted by \( \text{LST}(I_t^*) \) (i.e., last), is defined to be the tuple \( t_1 \) such that for each attribute \( A_i \), \( t_1[A_i] \) is the most current \( A_i \)-value of \( I_t^* \), i.e., \( t_1 \) contains the most current values from \( I_t^* \).
Currency constraints. In addition to partial currency orders, we can derive additional currency information from the semantics of the data modeled as currency constraints. A currency constraint $\varphi$ is of the form
$$\forall t_1, t_2 (\omega \rightarrow t_1 \prec_A t_2),$$
where $\omega$ is a conjunction of predicates of the form:

1. $t_1 \prec_A t_2$, i.e., $t_2$ is more current than $t_1$ in attribute $A_i$;
2. $t_1[A_i] \omega t_2[A_i]$, where $\omega$ is $=, \neq, >, <, \leq$ or $\geq$; and
3. $t_i[A_i] \omega c$ for $i \in \{1, 2\}$, where $c$ is a constant.

In contrast to denial constraints adopted in the model of [Fan et al. 2012], currency constraints help us find the most current values of some attributes, but not necessarily that currency constraints help us find the most current values of some attributes, but not necessarily.

Example 4: Recall entity instances $E_1$ and $E_2$ from Fig. 2. Currency constraints on $E_1$ and $E_2$ include $\varphi_1 \land \varphi_2$, as specified in Fig. 3 and interpreted in Example 2.

It is readily verified that for any completion $E'_C$ of $E_1$, if it satisfies these constraints, it yields $\text{LST}(E'_C)$ of the form $(\text{Edith}, \text{deceased}, \text{n/a}, 3, x_{\text{city}}, 213, 90058, x_{\text{county}})$ for Edith, in which the most current values for attributes name, status, job, kids, AC and zip are deduced from the constraints and remain unchanged, while $x_{\text{city}}$ and $x_{\text{county}}$ are values determined by the total currency order given in $E'_C$. Observe that the values of the current tuple are taken from different tuples in $E_1$, e.g., kids = 3 from $r_2$ and AC = 213 from $r_3$.

Similarly, for any completion of $E_2$, its current tuple has the form $(\text{George}, x_{\text{status}}, x_{\text{job}}, 2, x_{\text{city}}, x_{\text{AC}}, x_{\text{zip}}, x_{\text{county}})$, if they satisfy all the constraints. From these we can see that currency constraints help us find the most current values of some attributes, but not necessarily all. □

2.2. Data Consistency

To specify the consistency of data in our conflict model, we use a simple class of conditional functional dependencies (CFDs) [Fan et al. 2008], known as constant CFDs. A constant CFD [Fan et al. 2008] $\psi$ is of the form
$$\forall t(A_r) (\nu \rightarrow t[A_r] = c_r),$$
where $c_r$ is a constant from $\text{dom}(A_r)$ and $\nu$ is a conjunction of predicates of the form $t[A_i] = c_i$, for $c_i \in \text{dom}(A_i)$. For example, $\psi_1$ and $\psi_2$ in Fig. 3 are constant CFDs, as interpreted in Example 2.

Such CFDs are defined on the current tuple of a completion. Consider a completion $I'_C$ of $I_t$, for which the current tuple is $t = \text{LST}(I'_C)$. We say that the completion $I'_C$ satisfies a $\psi$, denoted by $I'_C \models \psi$, if whenever $t_i$ satisfies the predicate $\nu$, following the standard semantics of first-order logic, then $t_i[A_r] = c_r$.

Intuitively, this assures that if $t_i$ agrees with the constants specified in $\nu$ and if these are most current, then $t_i[A_r]$ should take value $c_r$, and $t_i[A_r]$ is the most current value in attribute $A_r$.

We say that $I'_C$ satisfies a set $\Gamma$ of constant CFDs, denoted by $I'_C \models \Gamma$, iff $I'_C \models \psi$ for each $\psi \in \Gamma$.

Observe that a constant CFD is defined on a single tuple $\text{LST}(I'_C)$. In light of this, we do not need general CFDs of [Fan et al. 2008], which are typically defined on two tuples.
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_e$</td>
<td>an entity instance of relation schema $R$</td>
</tr>
<tr>
<td>$I_t$</td>
<td>a temporal instance of $I_e$ (partial currency orders)</td>
</tr>
<tr>
<td>$I'_t$</td>
<td>a completion of partial currency orders in $I_t$</td>
</tr>
<tr>
<td>$S_e$</td>
<td>$(I_t, \text{currency constraints } \Sigma, \text{constant CFDs } \Gamma)$</td>
</tr>
<tr>
<td>$\text{LST}(I'_t)$</td>
<td>the current tuple of a completion $I'_t$</td>
</tr>
<tr>
<td>$O_t$</td>
<td>a partial temporal order</td>
</tr>
<tr>
<td>$T(S_e)$</td>
<td>the true values of the entity specified by $S_e$</td>
</tr>
</tbody>
</table>

Fig. 4. A summary of notations

**Example 5:** Recall the current tuples for $E_1$ described in Example 4. Then all completions of $E_1$ that satisfy $\psi_1$ in Fig. 3 have the form $(\text{Edith}, \text{deceased}, \text{n/a}, 3, \text{LA}, 213, 90058, \text{Vermont})$, in which $x_{\text{city}}$ is instantiated as $\text{LA}$ to be $\psi_1$, and as a result, $x_{\text{county}}$ then becomes $\text{Vermont}$ by the currency constraint $\varphi_8$.

2.3. Conflict Resolution

We are ready to specify entities.

**Specifications.** A specification $S_e = (I_t, \Sigma, \Gamma)$ of an entity consists of (1) a temporal instance $I_t = (I_e, \preceq_{A_1}, \ldots, \preceq_{A_n})$; (2) a set $\Sigma$ of currency constraints; and (3) a set $\Gamma$ of constant CFDs. A completion $I'_t = (I_e, \preceq'_{A_1}, \ldots, \preceq'_{A_n})$ of $I_t$ is called a valid completion of specification $S_e$ if $I'_t$ satisfies both $\Sigma$ and $\Gamma$.

We say that $S_e$ is valid if there exists a valid completion $I'_t$ of $S_e$. For instance, the specification of $E_1$ (or $E_2$) with the constraints given in Fig. 3 is valid.

**True values.** There may be many valid completions $I'_t$, each leading to a possibly different current tuple $\text{LST}(I'_t)$. When two current tuples differ in some attribute, there is a conflict. We aim to resolve such conflicts. If all such current tuples agree on all attributes, then the specification is conflict-free, and a unique current tuple exists for the entity $e$ specified by $S_e$. In this case, we say that this tuple is the true value of $e$.

The true value of $S_e$, denoted by $T(S_e)$, is the single tuple $t_e$ such that for all valid completions $I'_t$ of $S_e$, $t_e = \text{LST}(S_e)$, if it exists. For each attribute $A_i$ of $R$, we call $t_e[A_i]$ the true value of $A_i$ in specification $S_e$.

**The conflict resolution problem.** Consider a specification $S_e = (I_t, \Sigma, \Gamma)$ of an entity $e$, where $I_t = (I_e, \preceq'_{A_1}, \ldots, \preceq'_{A_n})$. Given $S_e$, conflict resolution is to find the minimum amount of additional currency information such that the true value exists.

The additional currency information is specified in terms of a partial temporal order $O_t = (I_e, \preceq'_{A_1}, \ldots, \preceq'_{A_n})$. We use $S_e \oplus O_t$ to denote the extension $S'_e = (I_t', \Sigma, \Gamma)$ of $S_e$ by enriching $I_t$ with $O_t$, where $I'_t = (I_e \cup I, \preceq_{A_1} \cup \preceq'_{A_1}, \ldots, \preceq_{A_n} \cup \preceq'_{A_n})$. In the sequel we only consider partial temporal orders $O_t$ such that $\preceq_{A_i} \cup \preceq'_{A_i}$ is a partial order for all $i \in [1, n]$.

We use $|O_t|$ to denote $\Sigma_{i \in [1, n]}|\preceq'_{A_i}|$, i.e., the sum of the sizes of all the partial orders in $O_t$ (note that each $\preceq'_{A_i}$ is a binary relation, and $|\preceq'_{A_i}|$ is its cardinality).

Given a valid specification $S_e = (I_t, \Sigma, \Gamma)$ of an entity $e$, the conflict resolution problem is to find a partial temporal order $O_t$ such that (a) the true value $T(S_e \oplus O_t)$ of $e$ exists and (b) $|O_t|$ is minimum.

**Example 6:** Recall from Example 4 the current tuples for George. Except for name and kids, we do not have a unique current value for the other attributes. Nonetheless, if a partial temporal order $O_t$ with, e.g., $r_6 \preceq_{\text{status}} r_5$ is provided by the users (i.e., status changes from unemployed to retired), then the true value of George in $E_2$ can...
be derived as (George, retired, veteran, 2, NY, 212, 12404, Accord) from the currency constraints and the constant CFDs of Fig. 3.

We summarize the notations in Fig. 4.

3. A CONFLICT RESOLUTION FRAMEWORK

We propose a framework for conflict resolution. As depicted in Fig. 5, given a specification $S_e = (I_t, \Sigma, \Gamma)$ of an entity $e$, the framework is to find the true value $T(S_e)$ of $e$ by reasoning about data currency and consistency, and by interacting with the users to solicit additional data currency information.

The framework provides the users with suggestions. A suggestion is a minimum set $A$ of attributes of $e$ such that if the true values of these attributes are provided by the users, $T(S_e)$ is automatically deduced from the users’ input, $\Sigma$, $\Gamma$ and $I_t$. The true values for $A$ are represented as a temporal order $O_t$. More specifically, the framework deduces $T(S_e)$ as follows.

1. **Validity checking.** It first inspects whether $S_e \oplus O_t$ is valid, via automated reasoning, where $O_t$ is a partial temporal order provided by the users, initially empty (see step (4) below for details about $O_t$). If valid, it follows the “Yes” branch. Otherwise the users need to revise $O_t$ by following the “No” branch.

2. **True value deducing.** After $S_e \oplus O_t$ is validated, it derives as many true values for the attributes of $e$ as possible, via automated reasoning.

3. **Finding the true value.** If $T(S_e \oplus O_t)$ exists, it computes and returns it following the “Yes” branch. Otherwise, it follows the “No” branch and goes to step (4).

4. **Generating suggestions.** It computes a suggestion $A$ along with its candidate values taken from the active domain of $S_e$, such that if the users pick and validate the true values for $A$, then $T(S_e \oplus O_t)$ is warranted to be found. The users are expected to provide $V$, the true values of some attributes in $A$, represented as a partial temporal order $O_t$. Given $O_t$, $S_e \oplus O_t$ is constructed and the process goes back to step (1).

The process proceeds until $T(S_e \oplus O_t)$ is found, or when the users opt to settle with true values for a subset of attributes of $e$. That is, if users do not have sufficient knowledge about the entity, they may let the system derive true values for as many attributes as possible, and revert to the traditional methods to pick the max, min, avg, any values for the rest of the attributes.
Remarks. (1) To specify users’ input, let \( I_e \) in \( S_e \) be \( (I_e, \preceq_{A_1}, \ldots, \preceq_{A_n}) \) and \( A \cup A' \cup B = \{A_1, \ldots, A_n\} \), where (i) \( A \) is the set of attributes identified in step (4) for which the true values are unknown; (ii) for \( B \), their true values \( V_B \) have been deduced (step (2)); and (iii) \( A' \) is the set of attributes whose true values can be deduced from \( V_B \) and the suggestion for \( A \). Given a suggestion, the user is expected to provide a set \( V \) of true values for (a subset of) \( A \) that they are confident of. Here \( V \) consists of either the candidate values taken from the suggestion, or some new values not in the active domains of \( S_e \) that users opt to choose. The users do not have to enter values for all attributes in \( A \).

From the input \( V \), a partial temporal order \( O_t \) is automatically derived, by treating \( V \) as the most current values of those attributes involved. Indeed, \( O_t \) has the form \( (I_e \cup \{t_o\}, \preceq_{A_1}', \ldots, \preceq_{A_n}') \), where \( t_o \) is a new tuple such that for all attributes \( A \), \( t_o[A] = V(A) \) if \( V \) has a value \( V(A) \) for \( A \), and \( t_o[A] = \text{null} \) otherwise, while \( t_o[B] = V_B \) remains unchanged. Moreover, \( \preceq_{A}' \) extends \( \preceq_{A} \) by including \( t[A] \preceq_{A}' t_o[A] \) if \( t_o[A] \neq \text{null} \), for all tuples \( t \in I_e \). From this, \( S_e \oplus O_t \) can be readily defined.

(2) There have been efficient methods for discovering constant CFDs, e.g., [Fan et al. 2011]. Along the same lines as CFD discovery [Chiang and Miller 2008; Fan et al. 2011], automated methods can be developed for discovering currency constraints from (possibly dirty) data. With certain quality metrics in place [Chiang and Miller 2008], the constraints discovered can be as accurate as those manually designed (such as those given in Fig. 3), and can be used by the framework as input.

(3) To simplify the discussion we do not allow users to change constraints in \( S_e \). We defer this to Section 7.

(4) We assume that the values in entity instances were once correct. When a temporal instance contains errors, one may inspect different samples and only take those currency orders that are consistent among the samples, or have sufficient support (e.g., frequency).

4. FUNDAMENTAL PROBLEMS

In this section, we identify fundamental problems associated with conflict resolution based on both data currency and consistency, and establish their complexity. These results are not only of theoretical interest, but also tell us where the complexity arises, and hence guide us to develop effective (heuristic) algorithms.

Satisfiability. The first one is the satisfiability problem for entity specifications. It is to decide, given a specification \( S_e = (I_e, \Sigma, \Gamma) \) of an entity, whether \( S_e \) is valid, i.e., whether there exists a valid completion of \( S_e \).

Intuitively, it is to check whether \( S_e \) makes sense, i.e., whether the currency constraints, constant CFDs and partial orders in \( S_e \), when put together, have conflicts themselves. The analysis is needed in step (1) of the framework of Fig. 5. In practice, this analysis tells us whether we have to revise constraints in \( S_e \), or ask users to validate its partial orders.

The problem is obviously important, but is NP-complete. One might think that the absence of currency constraints or CFDs would simplify the analysis. Unfortunately, its intractability is rather robust.

**Theorem 4.1.** The satisfiability problem for entity specifications is NP-complete. It remains NP-hard for valid specifications \( S_e = (I_e, \Sigma, \Gamma) \) of an entity when (1) both \( \Sigma \) and \( \Gamma \) are fixed; (2) \( \Gamma = \emptyset \), i.e., when only currency constraints are present; or (3) \( \Sigma = \emptyset \), i.e., when only constant CFDs are present.
Proof: For the upper bound it suffices to observe that the following NP algorithm correctly decides whether a given specification has a valid completion. Given a specification \( S_c = (I_t, \Sigma, \Gamma) \), the algorithm simply guesses a completion \( I'_t \) of \( I_t \) and then checks whether (i) \( I'_t \models \Sigma \); and (ii) \( I'_t \models \Gamma \). If the guessed completion passes these checks, then the algorithm returns “yes”. Otherwise, the guessed completion is rejected. Note that a “guess” simply completes the partial orders on the values of \( I_t \), and there are finitely many guesses in total. The algorithm is in NP since checking can be done in PTIME.

The NP-lower bound is established by reduction from the 3-satisfiability problem. An instance of the 3-satisfiability problem is formula \( \varphi = C_1 \land \cdots \land C_r \) with \( C_j = \ell^1_j \lor \ell^2_j \lor \ell^3_j \), where for \( k \in \{1, 2, 3\} \) and \( j \in [1, r] \), \( \ell^k_j \) is either a variable or a negation of a variable from a set \( X = \{x_1, \ldots, x_n\} \) of variables. It is to determine whether \( \varphi \) is satisfiable, i.e., whether there exists a truth assignment of variables in \( X \) that satisfies \( \varphi \). This problem is known to be NP-complete (cf. [Papadimitriou 1994]).

Given \( \varphi \), we define a specification \( S_c = (I_t, \Sigma, \Gamma) \) such that there exists a valid completion of \( S_c \) iff \( \varphi \) is satisfiable. The specification \( S_c \) consists of a temporal instance \( I_t \) of schema \( R(D, C, P, U, V, W) \) and a fixed set of currency constraints \( \Sigma \). No constant CFDs are defined in \( S_c \). Intuitively, \( D \) is to distinguish between tuples that encode truth assignments and tuples that correspond to clauses in \( \varphi \); \( C \) is to identify variables (by \( x_i \)) and clauses (by \( j \in [1, r] \); \( P \) is used to enforce the validity of clauses and finally, \( U, V \) and \( W \) represent the positions \( (1, 2 \text{ and } 3, \text{ resp}) \) of variables in each clause.

We first explain how the temporal instance \( I_t \) of \( R \) together with the currency constraints in \( \Sigma \) is to encode truth assignments for \( X \) and clauses in \( \varphi \). More specifically, for each variable \( x_i \in X \), we use two constants \( a_i \) and \( b_i \) such that \( a_i <_A b_i \) encodes that \( x_i \) is set to true, whereas \( b_i <_A a_i \) encodes that \( x_i \) is set to false. Here \( A \) ranges over attributes \( U, V \) and \( W \). More specifically, for each variable \( x_i \in X \) we include two tuples in \( I_t \):

\[
(0, x_i, 0, a_i, a_i) \quad \text{and} \quad (0, x_i, 0, b_i, b_i).
\]

These encode truth assignments of \( X \). To ensure that the choice of truth value for variables is consistent, we include the following currency constraints in \( \Sigma \):

\[
\forall t_1, t_2 \ (t_1[D] = 0 \land t_2[D] = 0 \land t_1[C] = t_2[C] \land t_1[A] \prec t_2[A] \rightarrow t_1[B] \prec t_2[B]),
\]

where \( A \) and \( B \) range over distinct pairs taken from \( \{U, V, W\} \). These currency constraints enforce that variables \( x_i \) are set to true (resp. false) independent of the position at which they appear in clauses (i.e., in attribute \( U, V \) or \( W \)).

We next consider the clauses in \( \varphi \). Let \( C_j = \ell^1_j \lor \ell^2_j \lor \ell^3_j \). observe that this can be equivalently written as \( \ell^1_j \land \ell^2_j \rightarrow \ell^3_j \). For instance, consider a clause \( C = x_1 \lor \bar{x}_2 \lor \bar{x}_3 \). This is equivalent to \( \bar{x}_1 \land x_2 \rightarrow \bar{x}_3 \). Given this, we include two tuples in \( I_t \) for each clause \( C_j \):

\[
(1, j, 1, v_1, v_2, v_3) \quad \text{and} \quad (1, j, 2, v_1', v_2', v_3'),
\]

where \( v_i = a_k \) and \( v'_i = b_k \) if \( \ell^k_j = x_k \), and \( v_i = b_k \) and \( v'_i = a_k \) if \( \ell^k_j = \bar{x}_k \), for \( i = 1, 2 \), and conversely for \( i = 3 \). The example clause \( C \) is thus encoded by \( (1, _, 1, b_1, a_2, b_3) \) and \( (1, _, 2, a_1, b_2, a_3) \). The connection between truth assignments selected by completions and the validity of clauses is established by means of the following currency constraint:

\[
\forall t_1, t_2 \ (t_1[D] = 1 \land t_2[D] = 1 \land t_1[C] = t_2[C] \land t_1[P] = 1 \land t_2[P] = 2 \\
\land t_1[U] \prec t_2[U] \land t_1[V] \prec t_2[V] \rightarrow t_1[W] \prec t_2[W]).
\]

This constraint tells us that whenever the truth assignment (represented by a completion) makes \( \ell^1_j \land \ell^2_j \) true, then it must also make \( \ell^3_j \) true.

We next show the correctness of the reduction. Suppose that \( \varphi \) is true and let \( \mu_X \) be a satisfying truth assignment. We define a valid completion of \( S_c \) as follows: For
attributes \( D, C \) and \( P \) we order the tuples in \( I_t \) arbitrarily. For attributes \( U \) (and consequently also for \( V \) and \( W \) by the currency constraints) we set \( a_i \preceq_U b_i \) if \( \mu_X(x_i) \) is true, and \( b_i \prec_U a_i \) otherwise. We need to verify that the second currency constraint is satisfied. This follows immediately from the fact that each clause is satisfied by \( \mu_X \).

Conversely, suppose that we have a valid completion of \( S_\varphi \). From this, we define \( \mu_X \) by simply setting \( \mu_X(x_i) = 1 \) if \( a_i \preceq_U b_i \) and \( \mu_X(x_i) = 0 \) otherwise. Similarly as above, it is readily verified that \( \mu_X \) satisfies all the clauses. Indeed, this follows from the second currency constraint given above.

It remains to show that the satisfiability problem is NP-complete when (a) \( \Sigma \) and \( \Gamma \) are fixed; (b) \( \Gamma = \emptyset \); or (c) \( \Sigma = \emptyset \). Since we have shown that the satisfiability problem is in NP, for general \( \Sigma \) and \( \Gamma \), it suffices to show the lower bounds. Furthermore, observe the proof above uses (i) a fixed set of currency constraints, i.e., the currency constraints \( \Sigma \) are independent of the input instance \( \varphi \), and (ii) it does not use any constant CFDs. In other words, (a) and (b) follow directly from the lower bound proof given above. It remains to show (c), i.e., the satisfiability problem is NP-hard even when only constant CFDs are present.

We establish this lower bound by reduction from the complement of the tautology problem, which is known to be coNP-complete (cf. Papadimitriou 1994)). An instance of the tautology problem is a formula \( \varphi = C_1 \lor \cdots \lor C_r \), where \( C_j = \ell_j^1 \lor \ell_j^2 \land \ell_j^3 \) and each \( \ell_j^i \) is either a variable or a complement of a variable from \( X = \{x_1, \ldots, x_n\} \). It is to determine whether \( \varphi \) is true for all truth assignments of \( X \). We define a specification \( S_e = (I_e, \Sigma = \emptyset, \Gamma) \) such that \( S_e \) has a valid completion iff \( \varphi \) is not a tautology.

The temporal instance \( I_t \) of \( S_e \) is an instance of schema \( R'(X_1, \ldots, X_n, C) \); it consists of two tuples \((0, 0, \ldots, 0)\) and \((1, 1, \ldots, 1)\). We impose no currency order or currency constraints on \( I_t \). Note that each completion \( I_t^e \) yields a current tuple \( \text{LST}(I_t^e) \) that encodes a truth assignment \( \mu_X \) of \( X \) in its first \( n \) attributes.

The set \( \Gamma \) of constant CFDs is given as follows. For each clause \( C_j \), we define \( \psi_j \):

\[
\forall t \left( t[L_1] = c_1 \land t[L_2] = c_2 \land t[L_3] = c_3 \rightarrow t[C] = 1 \right),
\]

where \( L_1 = X_k \) if \( \ell_j^1 \) or \( \ell_j^2 \) is \( x_k \) and \( c_i = 1 \) if \( \ell_j^i = x_k \) and \( c_i = 0 \) if \( \ell_j^i = \bar{x}_k \), for \( i = 1, 2, 3 \). Clearly, a completion \( I_t^e \models \psi_j \) if the truth assignment \( \mu_X \) encoded by the current tuple \( \text{LST}(I_t^e) \) makes \( C_j \) true. We further add the CFD \( \psi_C = \forall t \left( t[C] = 1 \rightarrow t[C] = 0 \right) \) to \( \Gamma \), which intuitively prevents any clause to be satisfied. Indeed, a completion \( I_t^e \) such that \( I_t^e \models \psi_C \) must set the \( C \)-attribute of its current tuple to 0. Contrast this with the requirement on the \( C \)-attribute of current tuples imposed by the \( \psi_j \)’s.

We next show the correctness of the reduction. If \( \varphi \) is a tautology then every truth assignment \( \mu_X \) makes at least one clause \( C_j \) true. That is, any valid \( I_t^e \) must set the \( C \)-attribute of its current tuple to 1 (by \( \psi_j \)) and at the same time it must set the \( C \)-attribute of its current tuple to 0 (by \( \psi_C \)). Hence, no valid completion can exists. Conversely, if there exists a valid completion \( I_t^e \) of \( S_e \) such that \( I_t^e \models \Gamma \), then its current tuple must have its \( C \)-attribute set to 0. In other words, none of the left-hand sides of the \( \psi_j \)’s can be true, and hence \( \mu_X \) must make all clauses false. In other words, \( \mu_X \) is a counterexample to the validity of \( \varphi \) and hence \( \varphi \) is not a tautology.

**Implication.** The second problem aims to deduce partial temporal orders that are logical consequences of the given currency order and currency constraints. Consider a valid specification \( S_e = (I_e, \Sigma, \Gamma) \) of an entity \( e \) and a partial temporal order \( O_t = (I_e, \preceq_{A_1}, \ldots, \preceq_{A_n}) \). We say that \( O_t \) is implied by \( S_e \), denoted by \( S_e \models O_t \), iff for all valid completions \( I_t^e \) of \( S_e \), \( O_t \subseteq I_t^e \). Here \( O_t \subseteq I_t^e \) if \( \preceq_{A_i} \subseteq \preceq_{A_i}^e \) for all \( i \in [1, n] \), where \( I_t^e = (I_e, \preceq_{A_1}^e, \ldots, \preceq_{A_n}^e) \).
The implication problem for conflict resolution is to decide, given a valid specification $S_c$ and a partial temporal order $O_i$, whether $S_c \models O_i$.

That is, no matter how we complete the temporal instance $I_i$ of $S_c$, as long as the completion is valid, the completion includes $O_i$ in its currency orders. The implication analysis is conducted at step (2) of the framework of Fig. 5, for deducing true values of attributes.

Unfortunately, this problem is also intractable.

**Theorem 4.2.** The implication problem for conflict resolution is coNP-complete. It remains coNP-hard for valid specifications $S_c = (I_i, \Sigma, \Gamma)$ of an entity when (1) both $\Sigma$ and $\Gamma$ are fixed; (2) $\Gamma = \emptyset$; or (3) $\Sigma = \emptyset$.

**Proof:** The coNP upper bound is verified by providing an NP algorithm for the complement problem. In a nutshell, given a specification $S_c = (I_i, \Sigma, \Gamma)$ and a partial temporal order $O_i$, the algorithm simply guesses a completion $I_i'$ of $I_i$ and then verifies whether (i) $I_i' \models \Sigma$; (ii) $I_i' \models \Gamma$; and (iii) $O_i \not\subseteq I_i'$. If $I_i'$ satisfies all these checks successfully, then the algorithm returns “yes” since $S_c \not\models O_i$. Otherwise, the current guess is rejected. This is clearly an NP algorithm for the complement problem and hence the implication problem is in coNP.

For the lower bounds, we show that the implication problem is coNP-hard when (1) both $\Sigma$ and $\Gamma$ are fixed; (2) $\Gamma = \emptyset$; or (3) $\Sigma = \emptyset$. The lower bounds for (1) and (2) are established by a revision of the proof of Theorem 4.1. More specifically, we revise the reduction used there as follows. First, the relation schema used in that proof is extended with an additional attribute $A$ and a tuple $t$ with its $A$-attribute set to a constant $a$, i.e., $t^a[A] = a$, and a tuple $t^b$ with its $A$-attribute set to a constant $b$. Finally, the premise of each currency constraint used in that proof carries an additional condition “$t^a[A] = a \land t^b[A] = b \land t_1[A] \approx A t_2[A]$”. These conditions enforce the constraints to have an effect only on completions in which $b$ is more current than $a$ in attribute $A$.

Denote by $S_c' = (I_i', \Sigma', \Gamma = \emptyset)$ the specification obtained from $S_c$ in the proof of Theorem 4.1 after such revisions. Let $O_i$ be the partial temporal order $(I_i', \{t^a \approx_A t^b\}, \emptyset, \ldots, \emptyset)$, where $t^a$ and $t^b$ are the two copies of an arbitrary tuple $t$ in $I_i$. We claim the following: (i) $S_c' \models I_i'$ is valid; and (ii) $S_c' \models O_i$ iff the formula $\varphi$ is not satisfiable. For (i) it suffices to observe that for any completion $(I_i')^c$, as long as it puts $t^b \approx_A t^a$ in its currency order for $A$ and arbitrarily completes currency orders for all the other attributes, it makes a valid completion. Indeed, this is simply because the conditions added to the premise of constraints used in the proof of Theorem 4.1 are false, and hence the currency constraint vacuously hold. Hence, $S_c'$ is valid.

For (ii), assume first that there exists a truth assignment $\mu_X$ that makes $\varphi$ true. We define a completion $(I_i')^c$ of $I_i'$ by setting $t^a \approx_A t^b$, where $t$ is the tuple used to define $O_i$, and by completing the currency orders for the attributes based on $\mu_X$ as in the proof of Theorem 4.1. As a result, $O_i \not\subseteq (I_i')^c$ and $S_c' \not\models O_i$. Conversely, suppose that $S_c' \not\models O_i$. This implies the existence of a valid completion $(I_i')^c$ of $I_i'$ that includes $t^a \approx_A t^b$ and satisfies all currency constraints in $\Sigma'$. Similar to the proof of Theorem 4.1 it is readily verified that a truth assignment $\mu_X$ can be constructed from $(I_i')^c$ that makes $\varphi$ true. Hence, $S_c' \models O_i$ iff $\varphi$ is not satisfiable. Observe that the proof only uses a fixed set of CPU constraints and does not require any constant CFDs.

Similarly, the coNP-lower bound for (3) is established by a similar modification of the specification for its counterpart given in the proof of Theorem 4.2, by reduction from the tautology problem. More specifically, given an instance $\varphi$ of the tautology problem as stated in the proof of Theorem 4.1, we extend the schema $R'$ given there with an additional attribute $A$. Its temporal instance $I_i'$ now consists of two tuples
It suffices to add two tuples $\Gamma'$ consistent with the constant CFDs $\psi_j$ and $\psi_C$ in the proof of Theorem 4.1 with the extra condition $"[[A] = \alpha\"$. That is, these constant CFDs only have an effect when the current tuple has $\alpha$ as its $A$-attribute value. Denote by $S'_e = (I'_e, \Sigma = \emptyset, \Gamma')$ the specification obtained in this way. Clearly, $S'_e$ is consistent since we just need to enforce $t_0 \preceq_A t_1$ in a completion to assure that the corresponding current tuple vacuously satisfies the CFDs in $\Gamma'$. Consider $O_e = (I'_e, \{ t_0 \preceq_A t_1 \}, \emptyset, \emptyset)$. Then, similar to the argument given above, one can readily verify that $S'_e \models O_e$ iff $\varphi$ is a tautology. \hfill $\Box$

**True value deduction.** The third problem is the true value problem for conflict resolution. It is to decide, given a valid specification $S_e$ for an entity $e$, whether $T(S_e)$ exists. That is, there exists a tuple $t_e$ such that for all valid completions $I'_e$ of $S_e$, LST($I'_e$) = $t_e$.

This analysis is needed by step (3) of the framework (Fig. 5) to decide whether $S_e$ has enough information to deduce $T(S_e)$, i.e., whether additional temporal information is needed to determine the true value of $e$.

No matter how important this problem is, it is also nontrivial: it is coNP-complete, and remains intractable in several practical special cases.

**Theorem 4.3.** The true value problem for conflict resolution is coNP-complete. It remains coNP-hard for valid specifications $S_e = (I, \Sigma, \Gamma)$ for an entity even when (1) both $\Sigma$ and $\Gamma$ are fixed; (2) $\Gamma = \emptyset$; or (3) $\Sigma = \emptyset$.

**Proof:** The upper bound is verified by providing an NP algorithm for the complement problem. Given a specification $S_e = (I, \Sigma, \Gamma)$, the algorithm simply guesses two completions $I'_1$ and $I'_2$ of $I$, and then checks whether both completions are valid and generate different current tuples. If so, the algorithm returns "yes" and concludes that no true value of $S_e$ can be determined. Otherwise, the current guesses are rejected. This is clearly an NP algorithm for the complement problem, and hence the true value problem is in coNP.

For the lower bounds, we need to show that the true value problem is coNP-hard when (1) both $\Sigma$ and $\Gamma$ are fixed; (2) $\Gamma = \emptyset$; or (3) $\Sigma = \emptyset$. The lower bounds for (1) and (2) are verified by a modification of the proof of its counterpart for Theorem 4.2. Indeed, it suffices to add two tuples $t_{\#}^a = (a, \#, \ldots, \#)$ and $t_{\#}^b = (b, \#, \ldots, \#)$ to the temporal instance given there, together with additional currency constraints that enforce $\#$ to come after any other constant in the currency orders for all attributes of the schema except for $A$ (which does not carry $\#$). Denote by $S''_{\#} = (I''_{\#}, \Sigma''_{\#}, \Gamma = \emptyset)$ the specification obtained in this way from $S'_e$ given in the proof of Theorem 4.2. As a consequence, any completion of $S''_{\#}$ can only yield current tuples $t_{\#}^a$ or $t_{\#}^b$.

As argued there, $S''_{\#}$ is valid since one only has to consider a completion that includes $t_{\#}^b \preceq_A t_{\#}^a$. Furthermore, we next show that a true value exists iff $\varphi$ is not satisfiable. Indeed, suppose that $\varphi$ is not satisfiable. Then for any valid completion $(I''_{\#})^\varphi$ of $I''_{\#}$, if $(I''_{\#})^\varphi \models \Sigma''$, then it has to set $t_{\#}^b \preceq_A t_{\#}^a$. Indeed, otherwise the currency constraints will be triggered and the completion would generate a satisfying truth assignment for $\varphi$, which by assumption does not exist. Hence, the true value will be the tuple $t_{\#}^a$. Conversely, suppose that no true value exists. This implies that there exist two completions of $I''_{\#}$, such that one leads to current tuple $t_{\#}^b$, and the other one leads to current tuple $t_{\#}^a$. In the second case, $t_{\#}^b \preceq_A t_{\#}^a$ and hence, as argued in the proof of Theorem 4.2, one can construct a satisfying truth assignment for $\varphi$ from the completions. Hence, if no true value exists, then $\varphi$ must be satisfiable.

The coNP-lower bound for (3) is established by a modification of the specification given in the proof for the case of constant CFDs in Theorem 4.2, by reduction from
the tautology problem. The modification is as follows. Given an instance \( \varphi \) of the tautology problem as stated in the proof of Theorem 4.2, we introduce a third tuple \( t_b = (b, b, \ldots, b) \) to the temporal instance given there, and extend the set \( \Gamma' \) of constant CFDs by including \( \psi_{a-b}^{\ell} = \forall t \left( t[A] = a \land t[X_i] = b \rightarrow t[A] = b \right), \) for \( i \in [1, n] \). These constant CFDs prevent the current tuple \( t \) in completions to have \( t[A] = a \) and \( t[X_i] = b \) for all \( i \in [1, n] \). In addition, we add \( \psi_{b-b}^{\ell} = \forall t \left( t[A] = b \rightarrow t[C'] = c \right) \) and \( \psi_{b-b}^{\ell} = \forall t \left( t[A] = b \rightarrow t[C] = c \right) \). These assure that for all current tuples \( t_i \), if \( t[A] = b \), then \( t \) has the constant \( b \) in all of its attributes.

Denote by \( S''_e = (I''_e, \Sigma = \emptyset, \Gamma'') \) the specification obtained this way from \( S' \) given in the proof of Theorem 4.2. A completion that results in current tuple \( t_b \) is clearly a valid completion, and hence \( S''_e \) is valid itself. Moreover, it is readily verified that a true value exists iff \( \varphi \) is a tautology. Indeed, observe first that completions either result in the current tuple \( t_b \) or a tuple of the form \( (a, \mu_X, 0) \), where \( \mu_X \) is a truth assignment for \( X \). While \( t_b \) can always be witnessed by a valid completion of \( S''_e \) (as mentioned above), \( (a, \mu_X, 0) \) can only be witnessed provided that \( \mu_X \) makes \( \varphi \) false (using the argument given in the proof of Theorem 4.2). Hence \( t_b \) is the true value iff \( \varphi \) is a tautology.

Coverage analysis. Finally, the minimum coverage problem is to determine, given a valid specification \( S_e = (I_e, \Sigma, \Gamma) \) of an entity and a positive integer \( k \), whether there exists a partial temporal order \( O \) such that (1) \( T(S_e \oplus O) \) exists, and (2) \(|O| \leq k \).

Intuitively, this is to check whether one can add a partial temporal order \( O \), of a bounded size to a specification such that the enriched specification has sufficient information to deduce all the true values of an entity. The ability to solve this problem helps us identify what minimum additional temporal information is needed to deduce the true value. The analysis of minimum \( O \), is required by step (4) of the framework of Fig. 5. This problem is \( \Sigma^p_2 \)-complete (NP\textsuperscript{coNP} or NP\textsuperscript{NP}), unfortunately. Worse still, it remains \( \Sigma^p_2 \)-hard even in several practical special cases, as stated below.

**Theorem 4.4.** The minimum coverage problem is \( \Sigma^p_2 \)-complete. It remains \( \Sigma^p_2 \)-hard for valid specifications \( S_e = (I_e, \Sigma, \Gamma) \) for an entity even when (1) both \( \Sigma \) and \( \Gamma \) are fixed; (2) \( \Gamma = \emptyset \); or (3) \( \Sigma = \emptyset \).

**Proof:** For the \( \Sigma^p_2 \) upper bound it suffices to observe that the following NP\textsuperscript{NP} algorithm correctly decides whether there exists a partial temporal order \( O \), of size \(|O| \leq k \) such that \( T(S_e \oplus O) \) exists. Given a valid specification \( S_e = (I_e, \Sigma, \Gamma) \), the algorithm first guesses a partial temporal order \( O \), and then checks whether \(|O| \leq k \) and whether \( T(S_e \oplus O) \) exists. The latter can be done in \textit{coNP} (see Theorem 4.3). If the guessed partial temporal order passes these checks, then the algorithm returns “yes”. Otherwise, the guessed order is rejected. The algorithm is in \( \Sigma^p_2 \) since it is a non-deterministic PTIME algorithm by calling a \textit{coNP} oracle (see, e.g., [Papadimitriou 1994] for detailed discussion about \( \Sigma^p_2 \)).

We now show that the problem is \( \Sigma^p_2 \)-hard when (1) \( \Sigma \) and \( \Gamma \) are fixed; (2) \( \Gamma = \emptyset \); or (3) \( \Sigma = \emptyset \).

For (1) and (2) we establish the \( \Sigma^p_2 \)-lower bound by reduction from the \( \exists \forall \forall \text{DNF} \) problem, which is known to be \( \Sigma^p_2 \)-complete [Stockmeyer 1976]. An instance of the \( \exists \forall \forall \text{DNF} \) problem is a formula of the form \( \varphi = \exists X \forall Y \psi \), where \( X = \{x_1, \ldots, x_n\} \) and \( Y = \{y_1, \ldots, y_m\} \), \( \psi = C_1 \lor \cdots \lor C_r \) for \( j \in [1, r] \), \( C_j = \ell_1^j \land \ell_2^j \land \ell_3^j \), and for \( k = 1, 2, 3 \), the literal \( \ell_1^j \) is either a variable or the complement of a variable in \( X \cup Y \). It is to determine whether \( \varphi \) is true.

Given an instance \( \varphi \) of the \( \exists \forall \forall \text{DNF} \) problem, we define a specification \( S_e = (I_e, \Sigma, \Gamma) \) and a constant \( k \) such that the minimal coverage problem for \( S_e \) and \( k \) has a solution
iff $\varphi$ is true. In particular, in $S_e$ we have a fixed set of currency constraints and no constant CFs. Hence, the reduction shows (1) and (2).

Recall the relation schema $R(A, D, C, P, U, V, W)$ used in the proof of Theorem 4.2. We populate its temporal instance $I_t = (I, \preceq_A, \preceq_D, \preceq_C, \preceq_P, \preceq_U, \preceq_V, \preceq_W)$ as follows: We assume the presence of $2(n + m)$ distinct constants $a_i$ and $b_i$ for $i \in [1, n]$ and $c_i$ and $d_i$ for $i \in [1, m]$. As in the proof of Theorem 4.2, truth values for variables in $X$ are encoded by means of two tuples:

$$(a, 0, x_i, 0, a_i, a_i) \quad \text{and} \quad (a, 0, x_i, 0, b_i, b_i),$$

with their $A$-attribute set to $a$, and two tuples

$$(b, 0, x_i, 0, a_i, a_i) \quad \text{and} \quad (b, 0, x_i, 0, b_i, b_i),$$

with their $A$-attribute set to $b$. Similarly, truth values for variables in $Y$ are encoded by the following tuples:

$$(a, 0, y_i, 0, c_i, c_i) \quad \text{and} \quad (a, 0, y_i, 0, d_i, d_i)$$

$$(b, 0, y_i, 0, c_i, c_i) \quad \text{and} \quad (b, 0, y_i, 0, d_i, d_i).$$

Moreover, we add a currency constraint to $\Sigma$ for every pair of attributes $(L, L')$ taken from $\{U, V, W\}$:

$$\forall t_1, t_2 \ (t_1[D] = 0 \land t_2[D] = 0 \land t_1[C] = t_2[C] \land t_1[A] = t_2[A] = b \land t_1[A] \prec_A t_2[A] \land t_1[L] \prec t_2[L] \rightarrow t_1[L'] \prec t_2[L'].$$  

These constraints ensure that whenever a $\prec_A b$, the order between $a_i$ and $b_i$ (resp. $c_i$ and $d_i$) is consistent for all attributes $U$, $V$, and $W$. As before, $a_i \prec_U b_i$ indicates that $x_i$ is set to true, whereas $b_i \prec_U a_i$ indicates that $x_i$ is false; similarly for variables in $Y$ but using the constants $c_i$ and $d_i$ instead. In other words, with every completion of $I_t$ in which $a \prec_A b$, we can associate truth assignments $\mu_X$ and $\mu_Y$ of $X$ and $Y$, respectively.

We next encode the clauses in $\varphi$ in a similar way the one given in the proof of Theorem 4.2. More specifically, given $C_1 \land \cdots \land C_r$, we encode its negation $\bar{C}_1 \land \cdots \land \bar{C}_r$ with $\bar{C}_i = \bar{C}_i^1 \lor \bar{C}_i^2 \lor \bar{C}_i^3$. Such clauses can be equivalently written as $\bar{C}_i^1 \land \bar{C}_i^2 \rightarrow \bar{C}_i^3$ by means of the tuples

$$(a, 1, j, 1, v_1, v_2, v_3) \quad \text{and} \quad (a, 1, j, 2, v'_1, v'_2, v'_3),$$

and their $b$-variants

$$(b, 1, j, 1, v_1, v_2, v_3) \quad \text{and} \quad (b, 1, j, 2, v'_1, v'_2, v'_3).$$

Here $v_i = a_k$ and $v'_i = b_k$ if $\bar{C}_i^j = x_k$, and $v_i = b_k$ and $v'_i = a_k$ if $\bar{C}_i^j = x_k$, for $i = 1, 2$; we define $v_i$ and $v'_i$ the other way around for $i = 3$; similarly for variables in $Y$ but then using constants $c_i$ and $d_i$ instead. For example, consider the clause $C = x_1 \land \bar{x}_2 \land \bar{y}_3$ whose complement is $\bar{C} = \bar{x}_1 \lor \bar{x}_2 \lor \bar{y}_3$. Equivalently, we write $\bar{C}$ as $x_1 \land \bar{x}_2 \land y_3$. Hence, we encode $\bar{C}$ by $(a, 1, \ldots, 1, a_1, b_2, c_3)$ and $(a, 1, \ldots, 2, b_1, a_2, d_3)$, together with their $b$-counterparts $(b, 1, \ldots, 1, a_1, b_2, c_3)$ and $(b, 1, \ldots, 2, b_1, a_2, d_3)$.

The link between truth assignments selected by completions and the validity of (complemented) clauses is established by the following currency constraint:

$$\forall t_1, t_2 \ (t_1[D] = 1 \land t_2[D] = 1 \land t_1[C] = t_2[C] \land t_1[P] = 1 \land t_2[P] = 2 \land t_1[A] = a \land t_2[A] = b \land t_1[A] \prec_A t_2[A] \land t_1[U] \prec t_2[U] \land t_1[V] \prec t_2[V] \rightarrow t_1[W] \prec t_2[W].$$

This constraint tells that whenever the truth assignment (represented by a completion) makes $\bar{C}_i^1 \land \bar{C}_i^2$ true, then it must also make $\bar{C}_i^3$ true, provided that $a \prec_A b$.

We also include two tuples $t^{sa}_# = (a, \ldots, \#)$ and $t^{sb}_# = (b, \ldots, \#)$ in $I_t$, which serves as potential true values of the entity represented by $S_e$. We enforce the symbol # to come after any other constant in currency orders by means of currency constraints.
(one for each attribute in \( R \)), as defined in the proof of Theorem 4.3. Clearly, in valid completions, if \( a \prec_A b \) then \( t^b_\# \) is the current tuple; when \( b \prec_A a \), \( t^a_\# \) is the current tuple.

Finally, we ensure that the partial temporal order \( O_t \) can only add currency information related to the values \( a_i \) and \( b_i \) in the instance, so that \( O_t \) can only affect the choice of truth values for variables in \( X \). To achieve this, observe that given instance \( I_t \) constructed so far, \( |O_t| \) can be no larger than \( 7|I|^2 \), where \( 7 \) is simply the number of attributes in \( R \). Let \( k = 7|I|^2 \). Next, for each constant \( v \) different from the \( a_i \)'s and \( b_i \)'s we add \( p > k \) tuples of the form \((v_{id}, v, \ldots, v)\), where \( v_{id} \) is a unique identifier for each of these tuples. Let \( I^t \) denote the temporal instance obtained in this way and let \( S_\# = (I^t, \Sigma, \Gamma = \emptyset) \). Clearly, for any \( O_t \) that relates tuples in an attribute with values different from \( a_i \) and \( b_i \), \( O_t \) will cause the addition of more than \( k \) tuples. Indeed, let \( B \) be an arbitrary attribute. Then the addition of \( t \prec_B t' \) implies that \( s \prec_B s' \) for all tuples \( s \) and \( s' \) that share the same \( B \)-attribute value with \( t \) and \( t' \), respectively. By the choice of \( k \) and the addition of \( p > k \) tuples for each constant, any \( O_t \) of size \( \leq k \) can only relate tuples that contain \( a_i \) or \( b_i \) values in one of its attributes.

Observe that the specification \( S_\# \) defined above is valid. Indeed, any completion that makes \( a \) more current than \( b \) in the \( A \)-attribute vacuously satisfies the currency constraints in \( \Sigma \). As a consequence \( t^b_\# \) will always be one of the possible current tuples.

We next show that the minimum coverage problem has a solution iff \( \varphi \) is true. Suppose first that \( \varphi \) is false. In other words, for every \( \mu_X \) of \( X \), there exists a truth assignment \( \mu_Y \) of \( Y \) that makes \( C_1 \lor \cdots \lor C_r \) false. Consider a partial temporal order \( O_t \) with \( |O_t| \leq k \). By the construction, \( O_t \) can only add temporal information between tuples that concern variables in \( X \). In other words, the impact of \( O_t \) is that it restricts the set of truth assignments of \( X \) that can be obtained by means of valid completions. However, since \( \varphi \) is false, even for each \( \mu_X \) in this restricted set, there exists a \( \mu_Y \) that makes \( C_1 \lor \cdots \lor C_r \) false. This in turn implies that \( t^b_\# \) can be a current tuple in a completion that sets \( a \prec_A b \). Indeed, simply consider the completion that (i) sets \( a \prec_A b \); (ii) selects a \( \mu_X \) that belongs to the restricted set; (iii) selects \( \mu_Y \) such that the clauses are false; and (iv) arbitrarily complete partial currency orders for the other attributes. It is easily verified that this completion indeed satisfies all currency constraints since it satisfies the constraints related to truth assignments and all constraints corresponding to the negated clauses (recall that \( \mu_X \) and \( \mu_Y \) make all \( C_j \) true). Hence, when \( \varphi \) is false, both \( t^a_\# \) and \( t^b_\# \) are current tuples and no true value can exist, no matter what \( O_t \) is.

Conversely, suppose that \( \varphi \) is true. That is, there exists a truth assignment \( \mu_X \) of \( X \) such that for all \( \mu_Y \) of \( Y \), \( C_1 \lor \cdots \lor C_r \) is true. We let \( O_t \) be the partial temporal order that restricts the choices of truth assignments for \( X \) to be \( \mu_X \). By the construction, this can be done by using \( \leq k \) added pairs. Then it is impossible that \( t^b_\# \) becomes a current tuple. Indeed, for this to happen we need a completion that sets \( a \prec_A b \) and in addition satisfies all constraints in \( \Sigma \). This, however, would imply the existence of a truth assignment \( \mu_Y \) of \( Y \), which, together with \( \mu_X \), makes \( C_1 \lor \cdots \lor C_r \) false. This in contrast to the assumption that \( \varphi \) holds for \( \mu_X \). As a consequence, \( T(S_\# \oplus O_t) \) exists and is equal to \( t^b_\# \).

ACM Journal of Data and Information Quality, Vol. V, No. N, Article A, Publication date: January YYYY.
Finally, we show that the problem is $\Sigma^p_2$-hard for case (3), when $\Sigma = \emptyset$. This is verified again by reduction from the $3\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lor\lo
denial constraints: they reduce the complexity of inferring data currency [Fan et al. 2012], from $\Sigma^\ast_1$-complete, $\Pi^\ast_2$-complete (coNP$^{coNP}$) and $\Pi^\ast_2$-complete down to NP-complete, coNP-complete and coNP-complete, respectively.

When it comes to data consistency, it is known that the satisfiability and implication problems for general CFDs are NP-complete and coNP-complete, respectively [Fan et al. 2008]. Theorems 4.1 and 4.2 give a stronger result: these lower bounds already hold for constant CFDs.

5. ALGORITHMS FOR CONFLICT RESOLUTION

We next provide algorithms underlying the framework depicted in Fig. 5. We first present an algorithm for checking whether a specification is valid (step (1) of the framework; Section 5.1). We then study how to deduce true attribute values from a valid specification (step (2); Section 5.2). Since not all true attribute values can be deduced automatically, we further discuss algorithms to generate suggestions such that the users may solicit true values of some attributes (step (4); Section 5.3), which can in turn help the deduction procedure.

5.1. Validity Checking

We start with algorithm IsValid that, given a specification $S_e = (I_e, \Sigma, \Gamma)$, returns true if $S_e$ is valid, and false otherwise. As depicted in Fig. 5, IsValid is invoked for an initial specification $S$ and its extensions $S_e \oplus O_t$ with the input $O_t$ from the users.

Theorem 4.1 tells us that it is NP-complete to determine whether $S_e$ is valid. In other words, IsValid is necessarily heuristic if it is to be efficient. Instead of designing an efficient algorithm from scratch, we approach this by reducing the problem to SAT, one of the most studied NP-complete problems, which is to decide whether a Boolean formula is satisfiable (see e.g., [Biere et al. 2009]). Several high-performance tools for SAT (SAT-solvers) are already in place [Biere et al. 2009], which have proved effective in e.g., software verification, AI and operations research. For instance, MiniSAT [Giunchiglia and Tacchella 2004] can effectively solve a formula with 4,500 variables and 100K clauses in 1 second.

Algorithm IsValid leverages existing SAT-solvers. We convert a given specification $S_e$ to a propositional formula in the conjunctive normal form (CNF), and then employ a SAT-solver to decide the satisfiability of $S_e$.

Algorithm. More specifically, given a specification $S_e$ of an entity $e$, IsValid works in three steps as follows.

(i) Instantiation: First, the specification $S_e$ is expressed as a set of (propositional) predicate formulas.

(ii) ConvertToCNF: Then, the predicate formulas from (i) are converted into a CNF such that the given specification is valid iff the CNF is satisfiable.

(iii) Finally, a SAT-solver is applied to the CNF resulting from (ii) to determine the satisfiability of the CNF.

We next present procedures Instantiation and ConvertToCNF. Consider $S_e = (I_e, \Sigma, \Gamma)$, where $I_e = (I_e, \leq_{A_1}, \ldots, \leq_{A_n})$ is a temporal instance of schema $R$. We denote also by $R$ the set $\{A_i \mid i \in [1, n]\}$ of attributes in $R$. We define the extended active domain of $A_i$, denoted by $\text{adom}(I_e, A_i)$, to be the set including all the values in $\text{adom}(I_e, A_i)$ and all the constants that appear in attribute $A_i$ of some constant CFD in $\Gamma$. To check whether $S_e$ is satisfiable, it suffices to consider the values from the extended active domains only.

Instantiation. To uniformly treat partial currency orders, currency constraints and constant CFDs as predicate formulas, we introduce a notion of instance constraints. The set of instance constraints of $S_e$, denoted as $\Omega(S_e)$, is defined in terms of values in
the extended active domains, and a strict partial order $\prec_{\mathcal{A}_i}$ on $\text{dom}^\nu(I_{\mathcal{A}}, A_i)$. These constraints are derived from $S_e$ as follows.

1. **Currency orders.** To encode the partial currency orders in $I_t$, for each $A_i \in R$, we include the following instance constraints in $\Omega(S_e)$.
   (a) Partial orders in $I_t$: ($true \rightarrow t[A_i] \prec_{\mathcal{A}_i} t[A_i]$) for each $t \prec_{\mathcal{A}_i} t_2$ in $I_t$, as long as $t_1[A_i] \neq t_2[A_i]$.
   (b) Transitivity of $\prec_{\mathcal{A}_i}$: $(a \prec_{\mathcal{A}_i} b \land b \prec_{\mathcal{A}_i} c \rightarrow a \prec_{\mathcal{A}_i} c)$ for all distinct values $a, b, c$ in $\text{dom}^\nu(I_{\mathcal{A}}, A_i)$.
   (c) Asymmetry: $(a \prec_{\mathcal{A}_i} b) \rightarrow \neg (b \prec_{\mathcal{A}_i} a)$ for all distinct values $a, b \in \text{dom}^\nu(I_{\mathcal{A}}, A_i)$.

   Intuitively, these assure that each $\prec_{\mathcal{A}_i}$ is a strict partial order (via both (b) and (c)), and express available temporal information in $I_t$ as predicate formulas (via (a)).

2. **Currency constraints.** For each currency constraint $\varphi = \forall t_1, t_2 (\omega \rightarrow t_1 \prec_{\mathcal{A}_i} t_2)$ in $\Sigma$ and for all distinct tuples $s_1, s_2 \in I_{\mathcal{A}_i}$, we include the following in $\Omega(S_e)$:

   \[
   \text{ins}(\omega, s_1, s_2) \rightarrow s_1[A_i] \prec_{\mathcal{A}_i} s_2[A_i],
   \]

   where $\text{ins}(\omega, s_1, s_2)$ is obtained from $\omega$ by (a) substituting $s_1[A_i]$ for $t_1$ and $s_2[A_i]$ for $t_2$ in each predicate $t \prec_{\mathcal{A}_i} t_2$, for $i \in [1, 2]$; and (b) evaluating each conjunct of $\omega$ with a comparison operator to its truth value $\omega$ r.t. $s_1$ and $s_2$. Intuitively, $\text{ins}(\omega, s_1, s_2)$ "instantiates" $\omega$ with values in $s_1$ and $s_2$.

**Example 7:** For currency constraint $\varphi_1$ in Fig. 3, and tuples $r_1$ and $r_2$ in Fig. 2 for Edith, its instance constraint is derived to be ($true \rightarrow working \prec_{\text{status}} \text{retired}$). Observe that the precondition of $\varphi_1$ is evaluated to be true on these two particular tuples, by instantiating variables of $\varphi_1$ with values in $r_1$ and $r_2$.

Similarly, from currency constraint $\varphi_6$ and tuples $r_1$ and $r_2$, we derive instance constraint ($\text{working} \prec_{\text{status}} \text{retired} \rightarrow 212 \prec_{\mathcal{A}_c} 415$), by replacing $\prec_{\text{status}}$ with $\prec^\nu_{\mathcal{A}_c}$ and by replacing variables in $\varphi_6$ with the corresponding attribute values from $r_1$ and $r_2$.

3. **Constant CFDs.** For each CFD $\forall t (\nu \rightarrow t[A_i] = c_r)$ in $\Gamma$ and each $c \in \text{dom}^\nu(I_{\mathcal{A}_i}, A_i) \setminus \{c_r\}$, $\Omega(S_e)$ includes

   \[
   \psi = (\nu' \rightarrow c \prec_{\mathcal{A}_i} c_r),
   \]

   where $\nu'$ is obtained from $\nu$ by replacing each conjunct $t[A_i] = c_j$ in $\nu$ with the conjunction $\land_{\mathcal{A}_i} c \prec_{\mathcal{A}_i} c_j$, where $c$ ranges over all the values in $\text{dom}^\nu(I_{\mathcal{A}_i}, A_i) \setminus \{c_r\}$.

   Intuitively, constraint $\psi$ asserts that if $\nu$ describes true values in the attributes in $\nu$, then $c_r$ is the true value of attribute $A_i$. Indeed, the CFD is defined on LST($I'_t$) for a completion $I'_t$ of $I_t$ (see Section 2.2), and $\psi$ assures that this semantics is enforced.

**Example 8:** Recall constant CFD $\psi_1$ from Fig. 3. For the entity instance $E_1$ of Edith, the CFD is encoded by two instance constraints given below, included in $\Omega(E_1)$:

   \[
   212 \prec_{\mathcal{A}_c} 213 \land 415 \prec_{\mathcal{A}_c} 213 \rightarrow \text{NY} \prec_{\text{city}}^{\nu} \text{LA},
   212 \prec_{\mathcal{A}_c} 213 \land 415 \prec_{\mathcal{A}_c} 213 \rightarrow \text{SFC} \prec_{\text{city}}^{\nu} \text{LA},
   \]

   i.e., $\text{LA}$ is her true city value if her true AC is 213.

**ConvertToCNF.** After we derive $\Omega(S_e)$, we convert the instance constraints of $\Omega(S_e)$ into a CNF $\Phi(S_e)$ as follows. We first substitute a Boolean variable $x_{A_i a_2}$ for each predicate $a_1 \prec_{\mathcal{A}_i} a_2$ in $\Omega(S_e)$. We then rewrite each formula of the form $(x_1 \land \cdots \land x_k \rightarrow x_{k+1})$ into the equivalent form $(\neg x_1 \lor \cdots \lor \neg x_k \lor x_{k+1})$. Finally, $\Phi(S_e)$ is defined to be the conjunction of all such formulas obtained from $\Omega(S_e)$, which is obviously in CNF.

One can readily verify the following (by contradiction), which justifies the reduction from the satisfiability of the specification $S_e$ to the SAT instance $\Phi(S_e)$.
5.2. Deducing True Values

Given a valid specification \( S_e = (I_e, \Sigma, \Gamma) \) of an entity \( e \), we want to deduce a maximum partial order \( O_d \) such that \( S_e \models O_d \). As an immediate corollary of Theorem 4.2, one can show that the decision problem of this problem is also coNP-complete, even when either \( \Sigma \) or \( \Gamma \) is fixed or absent. Thus we give a heuristics to strike a balance between its complexity and accuracy. The algorithm is based on the following lemma, which is easy to verify.

**Lemma 5.2.** For the CNF \( \Phi(S_e) \) converted from a valid specification \( S_e \), and for all tuples \( t_1, t_2 \) in \( S_e \) such that \( t_1[A_i] = a_1 \) and \( t_2[A_i] = a_2 \), \( S_e \models t_1 \prec_{A_i} t_2 \) iff \( \Phi(S_e) \rightarrow x^A_{a_1,a_2} \) is a tautology, where \( x^A_{a_1,a_2} \) is the variable denoting \( a_1 \prec_{A_i} a_2 \) in \( \Phi(S_e) \).

**Algorithm DeducOrder**

*Input:* A valid specification \( S_e = (I_e, \Sigma, \Gamma) \) of an entity.

*Output:* A partial temporal order \( O_d \) such that \( S_e \models O_d \).

1. \( \Omega(S_e) := \text{Instantiation}(S_e) \);
2. \( \Phi(S_e) := \text{ConvertToCNF}(\Omega(S_e)) \);
3. \( O_d := (I_e, \emptyset, \ldots, \emptyset) \);
4. while there exists a one-literal clause \( C \) in \( \Phi(S_e) \) do
   
   /* \( x^A_{a_1,a_2} \) in \( C \) is the variable denoting \( a_1 \prec_{A_i} a_2 \) */
   
   5. if \( C \) is a one-literal clause \( (x^A_{a_1,a_2}) \) then
5.1. add \( a_1 \prec_{A_i} a_2 \) to \( O_d \);
5.2. \( C_\gamma := \neg x^A_{a_1,a_2} \);
6. if \( C \) is a one-literal clause \( (\neg x^A_{a_1,a_2}) \) then
6.1. add \( a_2 \prec_{A_i} a_1 \) to \( O_d \);
6.2. \( C_\gamma := x^A_{a_1,a_2} \);
7. for each \( C' \in \Phi(S_e) \) do
7.1. if \( C' \) contains \( C_\gamma \) then
7.1.1. \( C' := C' \setminus C_\gamma \);
7.1.2. if \( C' \) contains \( C \) then
7.1.2.1. Remove \( C' \) from \( \Phi(S_e) \);
16. return \( O_d \).

Fig. 7. Algorithm DeducOrder

**Lemma 5.1.** Specification \( S_e \) is valid iff its converted CNF \( \Phi(S_e) \) is satisfiable.

**Complexity:** Observe the following. (a) The size \( |\Omega(S_e)| \) of \( \Omega(S_e) \) is bounded by \( O(|\Sigma| + |\Gamma||I_i|^2 + |I_i|^3) \), since encoding currency orders, currency constraints and constant CPDs is in time \( O(|I_i|^3) \), \( O(|\Sigma||I_i|^2) \) and \( O(|\Gamma||I_i|^2) \), respectively. (b) It takes \( O(|\Omega(S_e)|) \) time to convert \( \Omega(S_e) \) into \( \Phi(S_e) \). Hence the size of the CNF \( \Phi(S_e) \) is bounded by \( O(|\Sigma| + |\Gamma||I_i|^2 + |I_i|^3) \). In practice, an entity instance \( I_e \) is typically much smaller than a database, and the sets \( \Sigma \) and \( \Gamma \) of constraints are also small. As will be seen in Section 6, SAT-solvers can efficiently process CNFs of this size.

5.2. Deducing True Values

We next present an algorithm that, given a valid specification \( S_e = (I_e, \Sigma, \Gamma) \) of an entity \( e \), deduces true values for as many attributes of \( e \) as possible.

Intuitively, it is to find a maximum partial order \( O_d \) such that \( S_e \models O_d \), i.e., (a) for all valid completions \( I_i \) of \( S_e \), \( O_d \subseteq I_i \) (Section 4), and (b) for tuples \( t_1,t_2 \in I_e \) and \( A_i \in R \), if \( S_e \models t_1 \prec_{A_i} t_2 \) then \( t_1 \prec_{A_i} t_2 \) is in \( O_d \).

5.2.1. Partial Order Deduction. To deduce true values, below we first present a heuristic approach, and then discuss an exact algorithm.

A heuristic approach. Given \( S_e \), we want to deduce a maximum partial order \( O_d \) such that \( S_e \models O_d \). As an immediate corollary of Theorem 4.2, one can show that the decision problem of this problem is also coNP-complete, even when either \( \Sigma \) or \( \Gamma \) is fixed or absent. Thus we give a heuristics to strike a balance between its complexity and accuracy. The algorithm is based on the following lemma, which is easy to verify.

**Lemma 5.2.** For the CNF \( \Phi(S_e) \) converted from a valid specification \( S_e \), and for all tuples \( t_1, t_2 \) in \( S_e \) such that \( t_1[A_i] = a_1 \) and \( t_2[A_i] = a_2 \), \( S_e \models t_1 \prec_{A_i} t_2 \) iff \( \Phi(S_e) \rightarrow x^A_{a_1,a_2} \) is a tautology, where \( x^A_{a_1,a_2} \) is the variable denoting \( a_1 \prec_{A_i} a_2 \) in \( \Phi(S_e) \).
Algorithm NaiveDeduce

Input: A valid specification \( S_c = (I_c, \Sigma, \Gamma) \) of an entity.
Output: A partial temporal order \( O_d \) such that \( S_c \models O_d \).

1. \( \Omega(S_c) := \) Instantiation\((S_c)\); 
2. \( \Phi(S_c) := \) ConvertToCNF\((\Omega(S_c))\); 
3. \( O_d := (I_d, \emptyset, \ldots, \emptyset) \); 
4. for each \( A \in R \) do 
   5. for each \( a_1, a_2 \in \text{dom}^v(A) \) do 
      6. if \( \text{Sat}(\Phi(S_c) \cup \{ \neg a_1^A \}) \) then 
         7. add \( a_1 \prec_a^v a_2 \) to \( O_d^i \); 
   8. return \( O_d^i \).

Observe that the condition \( \Phi(S_c) \rightarrow x_{a_1a_2}^A \) indicates that for any truth assignment \( \mu \), if \( \mu \) satisfies \( \Phi(S_c) \), then \( \mu(x_{a_1a_2}^A) \) is true. That is, the one-literal clause \( x_{a_1a_2}^A \) is implied by \( \Phi(S_c) \), which in turn encodes \( S_c \). Based on this observation, our algorithm checks one-literal clauses in \( \Phi(S_c) \) one by one, and enriches the known partial order accordingly.

Algorithm. The algorithm for deducing true values, referred to as DeduceOrder, is given in Fig. 7. It first converts a specification \( S_c \) to the CNF \( \Phi(S_c) \) (lines 1-2; see Section 5.1). For each literal \( C \) of the form \( x_{a_1a_2}^A \) or \( \neg x_{a_1a_2}^A \), it checks whether \( C \) is a clause in \( \Phi(S_c) \) (line 4); and if so, it will enrich the partial order \( O_d \) (lines 5-10). It then reduces \( \Phi(S_c) \) by using \( C \) and its negation \( C' \) (lines 11-15). That is, for each clause \( C' \) that contains \( C \), the entire \( C' \) is removed since \( C' \) is true if \( C \) has to be satisfied (lines 12-13). Similarly, for each clause \( C'' \) that contains \( C', C' \) is removed from \( C'' \), as \( C' \) has to be false (lines 14-15). The deduced partial order \( O_d \) is then returned (line 16).

Example 9: Given the entity instance \( E_2 \) of Fig. 2 and the constraints of Fig. 3, DeduceOrder finds \( O_d \) including: (1) \( x_{\text{job}, \text{sailor}}^2 \) by \( \varphi_4 \), (2) working \( \prec \text{status} \) retired by \( \varphi_1 \), (3) sailor \( \prec \text{veteran} \) veteran, 401 \( \prec \text{AC} \) 212 and \( 02840 \prec \text{zip} \) 12404, by (2) and \( \varphi_5 \) and \( \varphi_6 \) respectively. A current tuple of George is then of the form \( \langle \text{George}, x_{\text{status}}, x_{\text{job}}, 2, x_{\text{city}}, x_{\text{AC}}, x_{\text{zip}}, x_{\text{county}} \rangle \), with variables.

Assume that the users assure that the true value of the attribute status is \( \text{retired} \). Then the algorithm can deduce the following from the extended specification:
(a) \( x_{\text{job}} \), \( x_{\text{AC}} \) and \( x_{\text{zip}} \) as n/a, 212 and 12404, from tuple \( r_5 \) via currency constraints \( \varphi_5 \), \( \varphi_6 \) and \( \varphi_7 \), respectively;
(b) \( x_{\text{city}} = \text{NY} \), from the true value of \( \text{AC} \) (i.e., 212 deduced in step (a) above) and the constant CFD \( \psi_2 \);
(c) \( x_{\text{county}} \) as \( \text{Accord} \), from \( \varphi_8 \) and the true values of city and zip deduced in steps (b) and (a), respectively.

The automated deduction tells us that the true value for George is \( t_2 = (\text{George}, \text{retired}, \text{n/a}, 2, \text{NY}, \text{212}, 12404, \text{Accord}) \). This shows that inferences of currency constraints help consistency inference (from step (a) to (b)), and vice versa (e.g., from (b) to (c)).

Complexity. (1) It takes \( O(|\Sigma| + |\Gamma|)|I_c|^2 + |I_c|^3 \) time to convert \( S_c \) into \( \Phi(S_c) \) (lines 1-2; see Section 5.1). (2) The total time taken by the while loop (lines 4-15) is in \( O(|\Sigma| + \ldots) \).
enumerates values pair $a$ for each variable together, the algorithm works in $S^Φ(|Γ|)$. Indeed, we maintain a hash-based index for literals $C$, in which the key is $C$ and its value is the list of clauses in $Φ(S_e)$ that contain $C$ or $C$... In the process, $Φ(S_e)$ decreases monotonically. Hence in total it takes at most $O(|Φ(S_e)|)$ time to reduce $Φ(S_e)$ for all literals, where $|Φ(S_e)|$ is bounded by $O((|Σ| + |Γ|)|I_1|^2 + |I_1|^3)$. Putting these together, the algorithm works in $O((|Σ| + |Γ|)|I_1|^2 + |I_1|^3)$ time.

An exact approach. By Lemma 5.2, one might want to compute a temporal order $O_d'$ consisting of all such variables $x_{a_1,a_2}$, that $Φ(S_e) ∧ ¬a_{a_1,a_2}$ is not satisfiable. That is, for each variable $x_{a_1,a_2}$, we inspect the satisfiability of $Φ(S_e) ∧ ¬a_{a_1,a_2}$ by invoking an SAT-solver.

This approach, referred to as NaiveDeduce, is given in Fig. 8. It first converts specification $S_e$ to the CNF $Φ(S_e)$ (lines 1-2). For each attribute $A$ (line 4), it then enumerates values pair $a_1, a_2$ in $adom^v(A)$ (line 5). It examines whether $Φ(S_e) → x_{a_1,a_2}$ is a tautology by invoking the SAT-solver to check whether $Φ(S_e) ∧ ¬x_{a_1,a_2}$ is not satisfiable (line 6). If $Φ(S_e) → x_{a_1,a_2}$ is a tautology, it adds $a_1 ∧ ¬a_2$ to $O_d'$ (line 7). The procedure returns $O_d'$ when all the possible partial orders are examined (line 8).

NaiveDeduce is an exact algorithm for deducing $O_d'$ provided the SAT-solver it invokes is an exact algorithm. However, NaiveDeduce calls the SAT-solver $|I_1|^2$ times. As will be seen in Section 6, DeduceOrder finds an $O_d$ with accuracy comparable to $O_d'$, without incurring the cost of repeatedly calling an SAT-solver.

5.2.2. True Value Deduction. Using the partial temporal order $O_d$ found by DeduceOrder or NaiveDeduce, one can readily deduce true attributes values as follows: a value $a_1$ is the true value of attribute $A_1$ if for all values $a_2 ∈ adom^v(I_e.A_1) \ {a_1}$, the currency order $a_2 ≺^v_a a_1$ is in $O_d$. Let $B ⊆ R$ be the set of attributes for which true values $V_B$ can be deduced.

5.3. Generating Suggestions

To identify the true value of the entity $e$ specified by $S_e = (I_e, Σ, Γ)$, instead of asking the users for input on all those attributes whose true values remain unknown, i.e., for attributes in $R \ B$, we compute a suggestion for a set of attributes $A ⊆ R$, disjoint from $B$, such that if the true values for $A$ are validated, the true value of the entire $e$ can be determined, even for attributes in $R \ (B ∪ A)$ (see Fig. 5). Below we first formally define suggestions and a notion of derivation rules (Section 5.3.1). We then provide an algorithm for computing suggestions (Section 5.3.2).

5.3.1. Suggestions and Derivation Rules. For each attribute $A_1 ∈ R \ B$, we denote by $V(A_1)$ the set of candidate true values for $A_1$, i.e., for any candidate $a_1 ∈ V(A_1)$, there exists no $a_2 ∈ adom^v(I_e.A_1) \ {a_1}$ such that $a_1 ≺^v_a a_2$ is in $O_d$. For a set $X$ of attributes, we write $V(X) = \{V(A_1) | A_1 ∈ X\}$.

Suggestion. A suggestion for $S_e$ is a pair $(A, V(A))$, where $A = (A_1, ..., A_m)$ is a set of attributes of $R$ such that $A ∩ B = ∅$ and moreover, (1) there exist values $(a_1, ..., a_m)$ such that if $(a_1, ..., a_m)$ are validated as the true values of $A$, then the true value $T(S_e)$ of $S_e$ exists; and (2) for all possible values $(a_1', ..., a_m')$ that satisfy condition (1), $a_i'$ is in $V(A_i)$ for $i ∈ [1, m]$.

Intuitively, condition (1) says that when the true values of $A$ are validated, so is $T(S_e)$. That is, the true values of attributes in $A = R \ (B ∪ A)$ can be automatically deduced from $V_B$ and the true values of $A$. Condition (2) says that $V(A)$ gives “complete” candidates for the true values of $A$ from their active domains.

One naturally wants a suggestion to be as “small” as possible, so that it takes the users minimal efforts to validate the true values of $A$. This motivates us to study the minimum suggestion problem, which is to find a suggestion $(A, V(A))$ with the
minimum number |A| of attributes. Unfortunately, this problem is \( \Sigma_p^p \)-complete (NP-complete), which can be verified by reduction from the minimum coverage problem (Theorem 4.4).

**Corollary 5.3.** The (decision version of) minimum suggestion problem for conflict resolution is \( \Sigma_p^p \)-complete.

**Proof:** It suffices to observe that a solution of the minimal suggestion problem of size \( \ell \) relates to a solution of the minimal coverage problem of size \( k = \ell/|I|^2 \). Conversely, one can show that a solution of the minimal coverage problem of size \( k \) relates to a solution of the minimal suggestion problem of size \( k/|I|^2 \).

In light of the high complexity, we develop an effective heuristic algorithm to compute suggestions. To do this, we examine how true values are inferred by using currency constraints and constant CFDs in a specification \( S_e \), by expressing them as a uniform set of rules.

**Derivation rules.** A true-value derivation rule for \( S_e \) has the form \((X, P[X]) \rightarrow (B, b)\), where (1) \( X \) is a set of attributes, \( B \) is a single attribute, and (2) \( b \) is a value that is either in adm\( ^V(I_e, B) \); and (3) for each \( A_i \in X \), \( P[A_i] \) is drawn from adm\( ^V(I_e, A_i) \). It assures if \( P[X] \) is the true value of \( X \), then \( b \) is the true value of \( B \).

Derivation rules are computed from instance constraints \( \Omega(S_e) \) of \( S_e \), which is illustrated below and will be elaborated in Section 5.3.2.

**Example 10:** Sample rules for George in Fig. 2 include:

\[
\begin{align*}
n_1 : (\{\text{status}\}, \{\text{retired}\}) & \rightarrow (\text{job, veteran}) \\
n_2 : (\{\text{status}\}, \{\text{retired}\}) & \rightarrow (\text{AC, 212}) \\
n_3 : (\{\text{status}\}, \{\text{retired}\}) & \rightarrow (\text{zip, 12404}) \\
n_4 : (\{\text{city, zip}\}, \{\text{NY, 12404}\}) & \rightarrow (\text{county, Accord}) \\
n_5 : (\{\text{AC, 212}\}) & \rightarrow (\text{city, NY}) \\
n_6 : (\{\text{status}\}, \{\text{unemployed}\}) & \rightarrow (\text{job, n/a}) \\
n_7 : (\{\text{status}\}, \{\text{unemployed}\}) & \rightarrow (\text{AC, 312}) \\
n_8 : (\{\text{status}\}, \{\text{unemployed}\}) & \rightarrow (\text{zip, 60653}) \\
n_9 : (\{\text{city, zip}\}, \{\text{Chicago, 60653}\}) & \rightarrow (\text{county, Bronzeville})
\end{align*}
\]

Here rule \( n_5 \) is derived from CFD \( \phi_2 \), which states that if his true AC is 212, then his true city must be NY. Rule \( n_1 \) is from tuple \( r_3 \) and constraint \( \phi_5 \) (Fig. 3), which states that if his true status is retired, then his true job is veteran. Note that in \( n_1 \), status is instantiated with retired. Similarly, \( n_6 \) is derived from \( r_6 \) and \( \phi_7 \); \( n_2 \) and \( n_3 \) (resp. \( n_7 \) and \( n_8 \)) are derived from tuple \( r_5 \) (resp. \( r_6 \)) and constraints \( \phi_6 \) and \( \phi_7 \), respectively; and \( n_4 \) (resp. \( n_9 \)) is derived from \( r_5 \) (resp. \( r_6 \)) and \( \phi_8 \).

To find a suggestion, we want to find a set \( A \) of attributes so that a maximum number of derivation rules can be applied to them at the same time. As a consequence, the true
values of as many other attributes as possible can be derived from these rules. To capture this idea, we introduce the following notion.

**Compatibility graphs.** Consider a set \( \Pi \) of derivation rules. The compatibility graph \( G(N, E) \) of \( \Pi \) is an undirected graph, where (1) each node \( x \) in \( N \) is a rule \( (X_r, P_z[X_x]) \rightarrow (B_x, b_x) \) in \( \Pi \), and (2) an edge \( (x, y) \) is in \( E \) iff \( B_x \neq B_y \) and \( P_z[X_{xy}] = P_y[X_{xy}] \), where \( X_{xy} = (X_x \cup B_x) \cap (X_y \cup B_y) \).

Intuitively, two nodes are connected (i.e., compatible) in \( G \) if their associated derivation rules derive different attributes (i.e., \( B_x \neq B_y \)), and they agree on the values of their common attributes (i.e., \( P_z[X_{xy}] = P_y[X_{xy}] \)). Hence these rules have no conflict with each other and can be applied at the same time.

**Example 11:** The compatibility graph of the rules given in Example 10 is shown in Fig. 9. There is an edge \( (n_1, n_2) \) in the graph since their common attribute status has the same value retired; similarly for the other edges. In contrast, there is no edge between \( n_5 \) and \( n_7 \) since the values of their common attribute AC are different: 212 for \( n_5 \) and 312 for \( n_7 \).

Observe that each clique \( C \) in the compatibility graph indicates a set of derivation rules that can be applied together. Let \( A' \) be the set of attributes whose true values can be derived from the rules in \( C \), if \( C \) and \( S_e \) have no conflicts (will be discussed shortly).

To find a suggestion, we compute a maximum clique \( C' \) from the graph, and derive a suggestion as \( (A, V(A)) \) from \( C \), where \( A \) consists of attributes in \( R \setminus (A' \cup B) \), and \( V(A) \) is the set of candidate true values for \( A \).

**Example 12:** Example 6 shows that for George with entity instance \( E_2 \), only the true values of name and kids are known, i.e., \( B = \{ \text{name}, \text{kids} \} \) and \( V_B = (\text{George}, 2) \). To find a suggestion for George, we identify a clique \( C_1 \) with five nodes \( n_1 \sim n_5 \) in the compatibility graph of Fig. 9. Observe the following: (a) The values of job, AC and zip depend on the value of status by rules \( n_1 \sim n_3 \) and \( n_5 \) respectively. (b) The value of AC in turn decides city by \( n_5 \). (c) From city and zip one can derive county by \( n_4 \). Hence, the set of attributes that can be derived from clique \( C_1 \) is \( A' = \{ \text{job}, \text{AC}, \text{zip}, \text{city}, \text{county} \} \). This yields a suggestion \( (A, V(\text{status})) \), where \( A = R \setminus (A' \cup B) = \{ \text{status} \}, \) and \( V(\text{status}) = \{ \text{retired}, \text{unemployed} \} \). As long as users identify the true value of status, the true value of George exists, and can be automatically deduced as described in Example 9.

However, a clique \( C \) and \( S_e \) may still have conflicts and as a result, \( C \) may not yield a valid completion of \( S_e \), as illustrated by the example below.

**Example 13:** Consider the clique \( C_2 \) shown in Fig. 9 with three nodes \( n_{5}, n_{6}, \) and \( n_{8} \). Observe the following: (a) \( n_5 \) indicates that 312 \( \preceq_{AC} \) 212, since 212 is assumed the latest AC value; whereas (b) \( n_6, n_8 \) and constraint \( \psi_6 \) in Fig. 3 state that 312 is the latest AC value, i.e., 212 \( \preceq_{AC} \) 312. Therefore, the values embedded in clique \( C_2 \) may not lead to a valid completion for entity instance \( E_2 \), i.e., \( C_2 \) and \( S_e \) have conflicts.

To handle conflicts between \( C \) and \( S_e \), we use MaxSat to find a maximum subgraph \( C' \) of \( C \) that has no conflicts with \( S_e \) (MaxSat is to find a maximum set of satisfiable clauses in a Boolean formula; see e.g., [Selman and Kautz 2004]). For instance, for clique \( C_2 \) of Example 13, we use a MaxSat-solver [Selman and Kautz 2004] to identify clique \( C'_2 \) with nodes \( n_6 \) and \( n_8 \), which has no conflicts with the specification for George. We then derive \( A' = \{ \text{job}, \text{zip} \} \) from \( C'_2 \). Since \( B = \{ \text{name}, \text{kids} \} \) (Example 12), we find \( A = R \setminus (A' \cup B) = \{ \text{status}, \text{city}, \text{AC}, \text{county} \} \) for a suggestion.

5.3.2. Computing Suggestions. We are now ready to present the algorithm for computing suggestions, referred to as Suggest, which is shown in Fig. 10. It takes as input a specification \( S_e \) of \( e \), partial orders \( O_d \) deduced from \( S_e \) (\( S_e \models O_d \), by Algo-

ACM Journal of Data and Information Quality, Vol. V, No. N, Article A, Publication date: January YYYY.
Algorithm Suggest

Input: A specification $S_e = (I_e, \Sigma, \Gamma)$ of an entity, order $O_d(S_e) \subseteq O_B$, and $V_B$.


1. $V(R) := \text{DeriveVR}(I_e, O_d)$;
2. $\Omega(S_e) := \text{Instantiation}(S_e)$;
3. $\Pi := \text{TrueDer}(\Omega(S_e), V(R))$;
4. $G := \text{CompGraph}(\Pi, S_e)$;
5. $C := \text{MaxClique}(G)$;
6. $A := \text{GetSug}(S_e, C, \Omega(S_e), V_B)$;
7. return $(A, V(A))$;

We next present the details of the procedures used in algorithm Suggest one by one.

DeriveVR: For each attribute $A$ whose true value is yet unknown, it computes a set $V(A)$ of candidate true values for $A$. Observe that given an attribute $A$, for any value $a_1$ in its active domain $\text{dom}^\nu(A)$, if there exists another value $a_2$ also from $\text{dom}^\nu(A)$, such that $a_1 \prec^\nu_a a_2$, then $a_1$ is not the true value for $A$. In other words, $a_1$ is known not to be the most current value.

Based on this observation, DeriveVR works as follows (not shown). Initially, $V(A)$ takes the active domain $\text{dom}^\nu(I_e.A)$. It then removes all values $a_1$ in $\text{dom}^\nu(I_e.A)$ from $V(A)$ for which there exists a value $a_2$ in $\text{dom}^\nu(I_e.A) \setminus \{a_1\}$ such that $a_1 \prec^\nu_a a_2$ is in the deduced partial order $O_d$, as $a_2$ is more current than $a_1$ in $A$. DeriveVR takes $O(|I_e|^2)$ time with an index, since it checks at most $|O_d|$ partial orders, and $|O_d| \leq |I_e|^2$.

TrueDer: Given a set $\Omega(S_e)$ of instance constraints, procedure TrueDer deduces a set $\Pi$ of derivation rules following the same way as shown in Example 10.

(1) From constant CFDs, the derivation rules could be deduced directly as long as they do not have conflicts with the candidate true values derived by DeriveVR.

(2) From those instance constraints that represent currency constraints and currency orders, it deduces derivation rules of the form $(X, P(X)) \rightarrow (B, b)$, for each attribute $B$ whose true value is unknown and for each $b \in V(B)$, if such a rule exists. While it is prohibitively expensive to enumerate all these rules, we use a heuristics to find a set of derivation rules. For each candidate true value $b$, it first identifies instance constraints that could complement the missing partial orders when assuming $b$ as true value. Then it maintains and populates pattern $(X, P(X))$ from unknown attributes and candidate true values that could satisfy the premise of each of those instance constraints.

For example, the rule $r_1$ in Example 10 could be deduced as follows. Observe that with $\varphi_5$ (Fig. 3) and $r_5, r_6$ (Fig. 2), the instance constraint

$$(\text{unemployed} \prec^\nu_a \text{status\ retired} \rightarrow \text{n/a} \prec^\nu_a \text{veteran})$$

could be derived. Here $V(\text{job})$ consists of 2 values “n/a” and “veteran”. The users may inspect the two values and choose one from the two as the true value of $V(\text{job})$. If one
Deduce rules from those instance constraints in $\Omega(\xi)$, a set of instance constraints $\Omega(\xi)$, and the candidate true values $V(R)$.

Output: A set $\Pi$ of derivation rules.

1. $\Pi := \emptyset$;
2. for each $(\nu \rightarrow t[A_v] = c_v) \in R$, $\nu = \bigwedge_{i=1}^k t[A_i] = c_i$, we check the following (line 2). If for each conjunct $t[A_v] = c_v$ in $\nu$, with $A_v \in B$, we have that $c_v \in V(A)$, i.e., when the values of the CFD have no conflict with those candidate true values (line 3), then we add $(\nu, (A_v, \xi_v), \xi_v) \rightarrow (A_v, \xi_v)$ as a new derivation rule (line 4).

(2) Deduce rules from those instance constraints in $\Omega(\xi)$ that represent currency constraints and partial currency orders in $\xi$, as follows:

(i) for each attribute $B$ whose true value is unknown (line 5) and each value $b$ in $V(B)$ that can possibly be its true value (line 6), let $U(B, b) = \{ b_i \prec_B b \mid b_i \in V(B) \setminus \{b\} \}$, which is the set consisting of all the missing partial orders when $b$ is assumed to be the true value of $B$ (line 7);
partition instance constraints based on \(U(B,b)\): for each value \(b\) in \(U(B,b)\), let 
\(\Omega(B,b)\) consist of all instance constraints \(\phi \in \Omega(S_c)\) such that \(\phi\) is of the form 
\(\omega \rightarrow b_i \prec b\) (line 8); note that each \(\phi\) appears in at most one of the partitions;

(iii) for each \(b_i \in U(B,b)\) (line 10), we pick a rule \(\phi = \omega \rightarrow b_i \prec b\) from \(\Omega(B,b)\) (line 11); 
we then expand \(X\) and the pattern \(P(X)\) so that the premise \(\omega\) can be satisfied 
(line 13), until either \((X, P(X))\) can no longer satisfy \(\omega\) (success = false) (line 14), 
or each \(b_i \prec b\) in \(U(B,b)\) is covered by such a rule \(\phi\) (success = true); in the latter 
case, we add the rule \((X, P(X)) \rightarrow (B,b)\) to \(\Pi\) (line 15). Note that \(|X| \leq |R|\).

The procedure is in \(O((|\Sigma|+|\Gamma|)|I_1|^2 + |I_1|^3)\) time. Indeed, the cost of step (1) is bounded 
by \(O(|\Gamma|)\); and for step (2), since \(U(B,b)\)'s are disjoint, \(\Omega(B,b)\)'s partition \(\Omega(S_c)\), and 
moreover, each \(\phi\) in \(\Omega(S_c)\) is used at most once, it takes at most \(O((|\Sigma|+|\Gamma|)|I_1|^2 + |I_1|^3)\) time.

CompGraph:. Given a set of derivation rules, procedure CompGraph generates their 
compatibility graph \(G(N,E)\) (see Example 11 for a running example).

More specifically, CompGraph is presented in Fig. 12. It takes a set \(\Pi\) of derivation 
rules as input. It constructs and returns a compatibility graph \(G\). The procedure 
works as follows. It first initializes a compatibility graph (line 1). It then generates a 
node for each derivation rule (line 2). The edges are then added (lines 3-6). For any 
two distinct nodes, if their associated rules are compatible (line 5, see the definition 
of compatibility graphs given earlier), an edge is added to connect these two nodes 
(line 6). It terminates and returns a compatibility graph \(G\) (line 7).

It is readily to verify that the procedure takes at most \(O(|\Pi|^2)\) time, where \(|\Pi|\) is no 
larger than \(|R||I_1|\).

MaxClique:. Given a compatibility graph \(G(N,E)\), this procedure computes a 
maximum clique \(C\) of \(G(N,E)\). While it is intractable to find a maximum clique, several 
tools have been developed for computing maximum cliques, with a good approximation 
bound (e.g., [Feige 2005]). We employ one of these tools as MaxClique.

GetSug:. Given a specification \(S_{\omega}\) of an entity, a set \(\Omega(S_{\omega})\) of instance constraints, a 
clique \(C\), and \(V_B\), the procedure computes a suggestion as output. As shown in Examples 
12 and 13, the clique returned by MaxClique represents a suggestion, but the 
suggestion may contain conflicts. This procedure is to convert the clique to a suggestion, 
and revise it in the presence of conflicts by invoking a weighted MaxSat-solver.

More specifically, GetSug is given in Fig. 13. It works as follows. It first identifies 
the required attributes \(A'\) and pattern \(P'[A]\) by applying derivation rules in \(\Pi\) (lines 1-4). It then converts \(P'\) to a 
CNF \(\Phi'\), along the same line as procedures Instantiation and ConvertToCNF given earlier (line 5). Since \(\Phi'\) may have conflicts with the \(\Phi(S_{\omega})\), it invokes a weighted MAXSAT-solver to 
milliminer revise \(\Phi'\) such that \(\Phi' \cup \Phi(S_{\omega})\) is satisfiable (lines 6-7). It then finds the subset of \(C\) corresponding to the revised \(\Phi'\), 
which has no conflicts with \(S_{\omega}\) (lines 8-12). It also derives a set \(A\) of attributes from the 
subset of \(C\) (line 12; see Example 12). Finally, it returns \(A\) (line 13). Recall that \(V(A)\) is 
computed by procedure DeriveVR given earlier. Note that the input to the MaxSat-solver 
is no larger than \(|R|^2|I_1|^2\). Moreover, there are efficient MaxSat-solvers available, with 
a reasonable approximation bound [Selman and Kautz 2004].

**Correctness.** Algorithm Suggest guarantees to generate a suggestion \((A,V(A))\). Indeed, (1) 
the clique \(C'\) revised by MaxSat has no conflicts with \(S_{\omega}\), and thus \(C'\) and \(S_{\omega}\) 
warrant to have a valid completion \(I'_1\). Let \(t_e = \text{LST}(I'_1)\). If \(V(A)\) are validated for \(A\), 
then \(t_e\) must be the true value \(V(S_{\omega})\) of \(S_{\omega}\), since \(t_e[B] = V_B\) remains unchanged for 
all valid completions of \(S_{\omega}\), and \(t_e[A']\) is uniquely determined by \(t_e[A]\) and \(V_B\) by the
6. EXPERIMENTAL STUDY

We conducted experiments using both real-life and synthetic data. We evaluated the accuracy and scalability of (1) isValid for validating a specification, (2) DeduceOrder for deducing true values, (3) Suggest for computing suggestions, and (4) the overall performance of conflict resolution supporting (1-3) above. Note that isValid and DeduceOrder are useful in their own right, since users may want to check their specifications and infer true values outside the interaction framework.

Experimental data. We used two real-life datasets (NBA and CAREER) and synthetic data (Person). Constraints were discovered using profiling algorithms [Chiang and Miller 2008; Fan et al. 2011], and examined manually. Timestamps for the datasets were either missing (for CAREER and Person) or incomplete (NBA). We assumed empty currency orders in all the experiments even when partial timestamps were given. The available (incomplete) timestamps were used for designing currency constraints and for evaluating the quality of the derived true values.

NBA player statistics. This dataset was retrieved from the following sites: (1) http://databasebasketball.com/, (2) http://www.infochimps.com/marketplace, and (3) http://en.wikipedia.org/wiki/List_of_National_Basketball_Association_arenas. It consists of three tables: (a) Player (from sources 1 and 3) contains information about players, identified by player id (pid). (b) Stat (from 1) includes the statistics of these players from the 2005/2006 season to the 2010/2011 season. (c) Arenas (from 3) records the historical team names and arenas of each team. We created a table, referred to as NBA, by first joining Player and Stat via equi-join on the pid attribute, and then joining Arenas via equi-join on the team attribute. The NBA table consists of 19573 tuples for 760 entities (i.e., players). Its schema is (pid, name, true name, team, league, tname, points, poss, allpoints, min, arena, opened, capacity, city). When producing the NBA table we took care of the attributes containing multiple values for a player, e.g., multiple teams for the same player, and multiple teams for one arena. We ensure that only one attribute value (e.g., team) appears in any tuple. Only data from (1) and (3) carries (partial)
timestamps. Therefore, the true values of entities in the NBA table cannot be directly derived when putting (1), (2) and (3) together.

The number of tuples pertaining to an entity ranges from 2 to 136, about 27 in average. We consider entity instances, i.e., tuples referring to the same entity, which are much smaller than a database.

We found 54 currency constraints: 15 for team names (t[name]) as shown by \( \varphi_1 \) below; 32 for arena, similar to \( \varphi_2 \); and 4 (resp. 3) for attribute allpoints that were scored since 2005 (resp. arena), similar to \( \varphi_3 \) (resp. \( \varphi_4 \)), where \( B \) ranges over points, pos, min and tname (resp. opened, capacity and years). We deduced 58 constant CFDs, e.g., \( \psi_1 \) below. Note that some rules are derived automatically, while the others are designed manually based on the semantics of the data.

\[
\varphi_1: \forall t_1, t_2 \ (t_1[t\text{name}] = \text{“New Orleans Jazz”} \land t_2[t\text{name}] = \text{“Utah Jazz”} \rightarrow t_1 \prec_{t\text{name}} t_2);
\]

\[
\varphi_2: \forall t_1, t_2 \ (t_1[t\text{arena}] = \text{“Long Beach Arena”} \land t_2[arena] = \text{“Staples Center”} \rightarrow t_1 \prec_{t\text{arena}} t_2);
\]

\[
\varphi_3: \forall t_1, t_2 \ (t_1[t\text{allpoints}] \prec t_2[t\text{allpoints}] \land t_1[B] \neq t_2[B] \rightarrow t_1 \prec_B t_2);
\]

\[
\varphi_4: \forall t_1, t_2 \ (t_1 \prec_{t\text{city}} t_2 \land t_1[B] \neq t_2[B] \rightarrow t_1 \prec_B t_2);
\]

\[
\psi_1: \forall t \ (t[arena] = \text{“United Center”} \rightarrow t[city] = \text{“Chicago, Illinois”})
\]

(2) CAREER. This data set was retrieved as is from the following source: http://www.cs.purdue.edu/commugrate/data/citeseer. Its schema is (first name, last name, affiliation, city, country). We chose 65 persons from the dataset, and for each one, we collected all of his/her publications, one tuple for each. No reliable timestamps were available for this dataset. The number of tuples pertaining to an entity ranges from 2 to 175, about 32 in average.

We derived 503 currency constraints: if two papers \( A \) and \( B \) are by the same person and \( A \) cites \( B \), then the affiliation and address (city and country) used in paper \( A \) are more current than those used in paper \( B \). We also found a single CFD of the form: \( \forall t \ (t[affiliation] = c_1 \rightarrow t[city] = c_2 \land t[country] = c_3) \), with 347 patterns with different constants \( (c_1, c_2, c_3) \).

The constraints for each dataset (NBA and CAREER) have essentially the same form, and only differ in their constants. Indeed, we find that the number of constraints with different forms is rather small in practice.

(3) Person data. The synthetic data adheres to the schema given in Table 2. We found 983 currency constraints (of the same form but with distinct constant values for status, job and kid) and a single CFD \( \forall t \ (t[AC] = c_1 \rightarrow t[city] = c_2) \) with 1000 patterns \( (c_1, c_2) \) (counted as distinct constant CFDs), similar to those in Table 3. The data generator used two parameters: \( n \) denotes the number of entities, and \( s \) is the size of entity instances (the number of tuples pertaining to an entity). For each entity, it first generated a true value \( t_e \), and then produced a set \( E \) of tuples that have conflicts but do not violate the currency constraints; we treated \( E \setminus \{t_e\} \) as the entity instance. We generated \( n = 10k \) entities, with \( s \) from 1 to 10k. We used empty currency orders.

Algorithms. We implemented the following algorithms in C++: (a) IsValid (Section 5.1); it calls MiniSat [Giunchiglia and Tacchella 2004] as the SAT-solver; (b) DeduceOrder and NaiveDeduce, where NaiveDeduce repeatedly invokes MiniSat [Giunchiglia and Tacchella 2004], as described in Section 5.2; and (c) Suggest: it uses MaxClique [Feige 2005] to find a maximal clique, and MaxSat-solver [Selman and Kautz 2004] to derive a suggestion (Section 5.3). We simulated user interactions by providing true values for suggested attributes, some with new values, i.e., values not in the active domain. We also implemented (d) Pick, a traditional method that randomly takes a value [Bleiholder and Naumann 2008]; to favor Pick, we picked a value from those that are not less current than any other values, based on currency constraints \( \forall t_1, t_2 (\omega \rightarrow t_1 \prec_A t_2) \) in which \( \omega \) is a conjunction of comparison predicates only, e.g., \( \varphi_1 \sim \varphi_3 \) above.
Accuracy. To measure the quality of suggestions, we used the notion of F-measure (http://en.wikipedia.org/wiki/F-measure):

\[
F\text{-measure} = \frac{2 \cdot (\text{recall} \cdot \text{precision})}{(\text{recall} + \text{precision})}
\]

Here precision is the ratio of the number of values correctly deduced to the total number of values deduced; and recall is the ratio of the number of values correctly deduced to the total number of attributes with conflicts or stale values.

All experiments were conducted on a Linux machine with a 3.0GHz Intel CPU and 4GB of Memory. Each experiment was repeated 5 times, and the average is reported here.

Experimental results. We next present our findings. Due to the small size of the CAREER data for each entity, experiments conducted on it took typically less than 10 milliseconds (ms). Hence we do not report its result in the efficiency study.

ACM Journal of Data and Information Quality, Vol. V, No. N, Article A, Publication date: January YYYY.
Exp-1: Validity checking. We first evaluated the scalability of IsValid. The average time taken by entity instances of various sizes is reported in Fig. 14(a), where the lower x-axis shows the sizes of NBA, and the upper x-axis is for Person data. The results show that IsValid suffices to validate specifications of a reasonably large size. For example, it took 220 ms for NBA entity instances of 109-135 tuples and 112 constraints, with 14 attributes in each tuple. For Person, it took an average of 4.7 seconds on entities of 8k-10k tuples and 1983 constraints. We also find IsValid accurate: specifications reported (in)valid are indeed (in)valid.

Exp-2: Deducing true values. We next evaluated the performance of algorithms DeduceOrder and NaiveDeduce. The results on both NBA and Person data are reported in Fig. 14(b), which tell us the following: (a) DeduceOrder scales well with the size of entity instances, and (b) DeduceOrder substantially outperforms NaiveDeduce on both datasets, for reasons given in Section 5.2. Indeed, DeduceOrder took 51 ms on NBA entity instances with 109-135 tuples, and 914 ms on Person entities of 8k-10k tuples; in contrast, NaiveDeduce spent 13585 ms and over 20 minutes (hence not shown in Fig. 14(b)) on the same datasets, respectively.

We also find that DeduceOrder derived as many true values as NaiveDeduce on both datasets (not shown). This tells us that DeduceOrder can efficiently deduce true values on large entity instances without compromising the accuracy of the true values found.

Exp-3: Suggestions for user interactions. We evaluated the accuracy of suggestions generated from currency constraints $\Sigma$ and CFDs $\Gamma$ put together. The results on NBA, CAREER and Person are given in Figures 14(e), 14(i) and 14(m), respectively, where the x-axis indicates the percentage of true attribute values deduced.

These results tell us the following. (a) Few rounds of interactions are needed to find all the true attribute values for an entity: at most 2, 2 and 3 rounds for NBA, CAREER and Person data, respectively. (b) A large part of true values can be automatically deduced by means of currency and consistency inferences: 35%, 78% and 22% of true values are identified from $\Sigma + \Gamma$ without user interaction, as indicated by the 0-interaction in Figures 14(e), 14(i) and 14(m), respectively.

Impact of $|\Sigma|$ and $|\Gamma|$. To be more precise when evaluating the accuracy, we use F-measure, which combines precision and recall, and take the cases of using $|\Gamma|$ only or $|\Sigma|$ only into consideration. Figures 14(f)–14(h), 14(j)–14(l) and 14(n)–14(p) show the results for NBA, CAREER and Person, respectively, when varying both $|\Sigma|$ and $|\Gamma|$, $|\Sigma|$ only, and varying $|\Gamma|$ alone, respectively. The x-axis shows the percentage of $\Sigma$ or $\Gamma$ used, and the y-axis shows the corresponding F-measure values.

These results tell us the following. (a) As shown in Figures 14(f), 14(j) and 14(n), our method substantially outperforms the traditional method Pick, by 201% in average on all datasets, even when we favor Pick by allowing it to capitalize on currency orders. This verifies that data currency and consistency can significantly improve the accuracy of conflict resolution. (b) When $\Sigma$ and $\Gamma$ are taken together, the F-measure value is up to 0.930 for NBA (Fig. 14(f), the top right point), 0.958 for CAREER (Fig. 14(j)), and 0.903 for Person (Fig. 14(n)), in contrast to 0.830 in Fig. 14(g), 0.907 in Fig. 14(k), and 0.826 in Fig. 14(o), respectively, when $\Sigma$ is used alone, and as opposed to 0.210 in Fig. 14(h), 0.741 in Fig. 14(l), and 0.234 in Fig. 14(p), respectively, with $\Gamma$ only. These further verify that the inferences of data currency and consistency should be unified instead of being taken separately. (c) The more currency constraints and/or CFDs are available, the higher the F-measure is, as expected. (d) The two curves for the 2- and 1-interaction overlap in Figures 14(f)–14(h) for NBA, 2- and 1-interaction in Figures 14(j)–14(l) for CAREER, and 3- and 2-interaction in Figures 14(n)–14(p) for Person. These indicate that
the user interactions are needed to provide true values for those attributes that we do not have enough information to deduce their true values.

**Exp-4: Efficiency.** The overall performance for resolving conflicts in the NBA (resp. Person) data is reported in Fig. 14(c) (resp. Fig. 14(d)). Each bar is divided into the elapsed time taken by (a) validity checking, (b) true value deducing, and (c) suggestion generating, including computing maximal cliques and running MaxSat. The result shows that conflict resolution can be conducted efficiently in practice, e.g., each round of interactions for NBA took 380 ms. Here validating specifications takes most time, dominated by the cost of SAT-solver, while deducing true values takes the least time.

**Summary.** From the experimental results we find the following. (a) Conflict resolution by reasoning about data currency and consistency substantially outperforms the traditional method Pick, by 201%. (b) It is more effective to unify the inferences of data currency and consistency than treating them independently. Indeed, when \( \Sigma \) and \( \Gamma \) are taken together, the F-measure improves over \( \Sigma \) only and \( \Gamma \) only by 11% and 236%, respectively. (c) Our conflict resolution method is efficient: it takes less than 0.5 second on the real-life datasets even with interactions. (d) Our method scales well with the size of entities and the number of constraints. Indeed, it takes an average of 7 seconds to resolve conflicts in Person entity instances of 8k-10k tuples, with 1983 constraints. (e) At most 2-3 rounds of interactions are needed for all datasets.

7. CONCLUSION

We have proposed a model for resolving conflicts in entity instances, based on both data currency and data consistency. We have also identified several problems fundamental to conflict resolution, and established their complexity. Despite the inherent complexity of these problems, we have introduced a framework for conflict resolution, along with practical algorithms supporting the framework. Our experimental study has verified that our methods are effective and efficient using real-life and synthetic data. We contend that these yield a promising approach to resolving conflicts in practice.

We are now exploring efficient algorithms with better performance guarantees for generating suggestions, and testing them with data in various domains. Another topic concerns the discovery of data quality rules. Prior work on discovery of such rules [Chiang and Miller 2008] shows that a large number of high-quality rules can be identified from possibly dirty data. A third topic is to repair data by using currency constraints and partial temporal orders. This is more challenging than conflict resolution, since a database to be repaired is typically much larger than entity instances. Finally, a challenging topic is to extend our framework by allowing users to edit constraints, and by soliciting other information (such as semantic dependencies specifying how attributes are correlated) as users’ feedback when the users do not have sufficient currency knowledge about their data.

REFERENCES

Marcelo Arenas, Leopoldo E. Bertossi, and Jan Chomicki. 1999. Consistent Query Answers in Inconsistent Databases. In PODS.


ACM Journal of Data and Information Quality, Vol. V, No. N, Article A, Publication date: January YYYY.
A.34 W. Fan, F. Geerts, N. Tang & W. Yu


Xin Dong and Felix Naumann. 2009. Data fusion - Resolving Data Conflicts for Integration. In VLDB.


Wenfei Fan, Floris Geerts, Ma Shuai, Nan Tang, and Wenyuan Yu. 2013. Inferring Data Currency and Consistency for Conflict Resolution. In ICDE.


Xiaoxin Yin, Jiawei Han, and Philip S. Yu. 2008. Truth Discovery with Multiple Conflicting Information Providers on the Web. TKDE 20, 6 (2008).

Haopeng Zhang, Yanlei Diao, and Neil Immerman. 2010. Recognizing Patterns in Streams with Imprecise Timestamps. In VLDB.