CDMTCS
Research
Report
Series

A Fourth Normal Form for Possibilistic Data

Ziheng Wei
The University of Auckland

Sebastian Link
The University of Auckland

CDMTCS-533
March 2019

Centre for Discrete Mathematics and
Theoretical Computer Science

A Fourth Normal Form for Possibilistic Data

Ziheng Wei
The University of Auckland, New Zealand
z.wei@auckland.ac.nz

Sebastian Link
The University of Auckland, New Zealand
s.link@auckland.ac.nz

March 18, 2019

Abstract

Relational database design addresses applications for data that is certain. Modern applications require the handling of uncertain data. Indeed, one dimension of big data is veracity. Ideally, the design of databases helps users quantify their trust in the data. For that purpose, we need to establish a design framework that handles responsibly any knowledge of an organization about the uncertainty in their data. Naturally, such knowledge helps us find database designs that process data more efficiently. In this paper, we apply possibility theory to introduce the class of possibilistic multivalued dependencies that are a significant source of data redundancy. Redundant data may occur with different degrees, derived from the different degrees of uncertainty in the data. We propose a family of fourth normal forms for uncertain data. We justify our proposal showing that its members characterize schemata that are free from any redundant data occurrences in any of their instances at the targeted level of uncertainty in the data. We show how to automatically transform any schema into one that satisfies our proposal, without loss of any information. Our results are founded on axiomatic and algorithmic solutions to the implication problem of possibilistic functional and multivalued dependencies which we also establish.

Keywords: Database design; Functional dependency; Multivalued dependency; Normal Form; Redundancy; Uncertainty

1 Introduction

Big data has given us big promises. However, their realization comes with big responsibilities. One dimension of big data is veracity, or the uncertainty in data. According to an IBM study, one in three managers distrust the data that they use to make decisions[1]. Our community needs to establish design frameworks that produce responsible systems, that is, systems which establish trust in the data that they manage. A first step is to provide a capability for data stewards to quantify the degrees of uncertainty in data that they deem appropriate. It is natural to think that such information helps tailor database designs according to application requirements on the certainty in data. With such a framework data would meet the requirements of applications by design. This would provide a solid foundation for trust in data and data-driven decision making.

In [20, 21] the authors presented a design framework of relational databases for uncertain data. Based on possibility theory [5], records are assigned a discrete degree of possibility (p-degree) with which they occur in a relation. Intuitively, the p-degree quantifies the level of confidence an organization is prepared to assign to a record. The assignment of p-degrees can be based on many factors, specific to applications and irrelevant for developing the framework. In addition, an integrity constraint is assigned a degree of certainty (c-degree) that quantifies to which records it applies. Intuitively, the higher the c-degree of a constraint the lower the minimum p-degree of records to which the constraint applies. For example, a constraint with the highest c-degree applies to all records, and one with the lowest c-degree only applies to records with the highest p-degree. The design framework of [21] was developed for possibilistic functional dependencies (pFDs) [20]. Generalizations of Boyce-Codd and Third Normal Forms were established that characterized schemata that eliminated, minimized across all dependency-preserving decompositions respectively, from every instance, every redundant data value occurrence from every record whose p-degree meets a given target. Consequently, the normalization effort is tailored to the application requirements for the minimum p-degrees required from data.

Our main goal is to establish a design framework for the combined class of pFDs and possibilistic multivalued dependencies (pMVDs). That is, tailor relational schema design up to Fagin's Fourth Normal Form (4NF) [6] to uncertain data. One key motivation follows the relational framework since MVDs cause data redundancy that is not covered by FDs. However, modern applications provide stronger motivation as they need to integrate data from various sources, in which companies have varying degrees of trust. In fact, the integration process often relies on joins which are a frequent source of data redundancy that is caused by MVDs. In fact, a relation exhibits an MVD if and only if the relation is the lossless join over two of its projections [6]. As modern information systems have the responsibility to process uncertain data efficiently, we require a design framework for uncertain data that can eliminate sources of data redundancy as required by applications. Our main goal poses new challenges. Indeed, the framework in [20,21] is founded on the *downward closure property* of FDs, which says that an FD that is satisfied by a relation is also satisfied by every subset of records in the relation. Unfortunately, MVDs are not closed downwards and it is unclear whether the achievements of [21] for pFDs can be extended to pMVDs.

Contributions. 1) We introduce the class of pMVDs, generalizing both MVDs and pFDs, as a rich source of redundancy in uncertain data. 2) We establish axiomatic

[1] http://www-01.ibm.com/software/data/bigdata/

and algorithmic characterizations for the implication problem of pFDs and pMVDs. 3) For pFDs and pMVDs we characterize schemata that do not permit any instances with any redundant data value occurrence of a targeted minimum p-degree. This is achieved by a fourth normal form proposal on conditions of pMVDs that apply to data in which redundancy of the targeted p-degree can occur. 4) We show how to transform any schema into one that meets the fourth normal form condition, without loss of information. Our contributions subsume Fagin's 4NF as the special case with one p-degree.

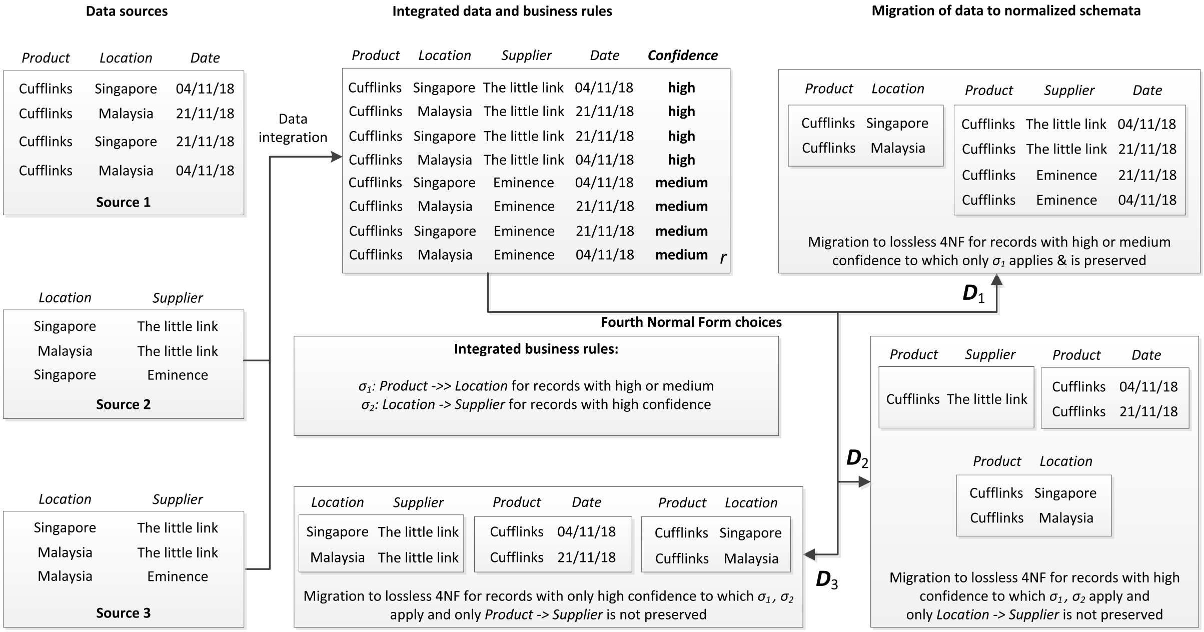
Organization. Section 2 introduces our running example. We discuss related work in Section 3. Preliminaries are provided in Section 4. A design theory is established in Section 5. Our normal form proposal and semantic justification is developed in Section 6. Section 7 explains how our normal form can always be achieved. Section 8 concludes and comments briefly on future work. Proofs can be found in the appendix.

2 Running Application Scenario

We introduce a running example chosen small enough to illustrate our concepts and findings throughout the paper. The application scenario integrates data about the supply of products that record which *p)roducts* are supplied by which *s)upplier* from which *l)ocation* on which *d)ate*. The data originates from different sources. Data about products, suppliers, and date originate from source 1, while data about suppliers and their location originate from source 2 and source 3. The integrated relation is the result of joining source 1 with the union of sources 2 and 3. Importantly, there is more confidence in records that occur in both source 2 and 3 than in records that occur in either source 2 or source 3. As a means to distinguish between these levels of confidence, records that occur in both source 2 and 3 are assigned the confidence label *high*, while the remaining ones have label *medium*.

A specific instance of this application scenario is depicted in Figure 1, in which relation r originates from the given instances of the three sources. We can observe that the subrelation r_1 of r, that consists of the four records with confidence label *high*, exhibits the FD σ_2 : *location* $\rightarrow$ *supplier* while r does not satisfy σ_2. Indeed, the location of supplier *Eminence* is not certain. Recall that an FD $X \rightarrow Y$ with attribute subsets X, Y is satisfied by a relation whenever two records with matching values on all the attributes in X have also matching values on all the attributes in Y. On the other hand, the entire relation r does exhibit the MVD σ_1 : *product* $\twoheadrightarrow$ *location*. Indeed, an MVD $X \twoheadrightarrow Y$ with attribute subsets $X, Y \subseteq R$ is satisfied by a relation whenever for every two records with matching values on all the attributes in X there is a record that has matching values with the first record on all the attributes in $X \cup Y$ and matching values with the second record on all the attributes in $X \cup (R - Y)$. In other words, r satisfies $X \twoheadrightarrow Y$ if and only if r is the join of its projections $r(XY)$ and $r(X(R - Y))$.

Following Vincent's notion of data redundancy [25], σ_2 causes all occurrences of supplier *The little link* to be redundant: every change in value for one of these occurrences causes a violation of σ_2. These redundant values only occur in records of *high* confidence, since σ_2 only applies to those records. However, every value occurrence in columns *location*, *supplier*, and *date* is also redundant. These redundancies are caused by the MVD


Data sources
Integrated data and business rules
Migration of data to normalized schemata
Data integration
Source 1
Source 2
Source 3
r
Fourth Normal Form choices
D_1
D_2
D_3

Product	Location	Date
Cufflinks	Singapore	04/11/18
Cufflinks	Malaysia	21/11/18
Cufflinks	Singapore	21/11/18
Cufflinks	Malaysia	04/11/18

Location	Supplier
Singapore	The little link
Malaysia	The little link
Singapore	Eminence

Location	Supplier
Singapore	The little link
Malaysia	The little link
Malaysia	Eminence

Product	Location	Supplier	Date	Confidence
Cufflinks	Singapore	The little link	04/11/18	high
Cufflinks	Malaysia	The little link	21/11/18	high
Cufflinks	Singapore	The little link	21/11/18	high
Cufflinks	Malaysia	The little link	04/11/18	high
Cufflinks	Singapore	Eminence	04/11/18	medium
Cufflinks	Malaysia	Eminence	21/11/18	medium
Cufflinks	Singapore	Eminence	21/11/18	medium
Cufflinks	Malaysia	Eminence	04/11/18	medium

Integrated business rules:

σ_1: *Product ->> Location* for records with high or medium
σ_2: *Location -> Supplier* for records with high confidence

Product	Location
Cufflinks	Singapore
Cufflinks	Malaysia

Product	Supplier	Date
Cufflinks	The little link	04/11/18
Cufflinks	The little link	21/11/18
Cufflinks	Eminence	21/11/18
Cufflinks	Eminence	04/11/18

Migration to lossless 4NF for records with high or medium confidence to which only σ_1 applies & is preserved

Product	Supplier
Cufflinks	The little link

Product	Date
Cufflinks	04/11/18
Cufflinks	21/11/18

Product	Location
Cufflinks	Singapore
Cufflinks	Malaysia

Migration to lossless 4NF for records with high confidence to which σ_1, σ_2 apply and only *Location -> Supplier* is not preserved

Location	Supplier
Singapore	The little link
Malaysia	The little link

Product	Date
Cufflinks	04/11/18
Cufflinks	21/11/18

Product	Location
Cufflinks	Singapore
Cufflinks	Malaysia

Migration to lossless 4NF for records with only high confidence to which σ_1, σ_2 apply and only *Product -> Supplier* is not preserved

Figure 1: 4NF normalization strategies for data with different confidence levels

σ_1 which applies to all records. This illustrates that MVDs are a major source of data redundancy that cannot be captured by FDs, but also that the uncertainty in data causes different degrees of data redundancy. The latter point is further highlighted in Figure 1. In applications with records of high or medium confidence, we only need the MVD σ_2 for normalization, because the FD σ_1 does not apply. The 4NF decomposition $\mathbf{D}_1$ illustrates this case in Figure 1. In applications with records of high confidence only, we require both the FD σ_1 and the MVD σ_2 for normalization. Here, $\mathbf{D}_2$ and $\mathbf{D}_3$ are lossless decompositions into 4NF, respectively. The scenario also illustrates that redundancy in records with higher p-degrees require a higher normalization effort than redundancy in records with lower p-degrees, for the simple reason that more dependencies apply to fewer records. Arguably, records with lower p-degree are more susceptible to updates. In this case, the normalization effort is smaller as fewer dependencies cause redundancy in records with lower p-degrees.

3 Related Work

While there is a plethora of interest in uncertain data, research on query languages and probabilistic approaches dominate. Instead, we are interested in schema design using a possibilistic approach. Hence, our discussion of related work centers on schema design.

The article is a natural continuation of our work on schema design [21], based on pFDs introduced in [20]. As illustrated before, pFDs cannot capture many instances of redundancy in uncertain data for which pMVDs are responsible. This is particularly relevant in modern applications where data is joined from various sources of information. Our article is the first to introduce pMVDs, and therefore also the first with a 4NF proposal targeted at uncertain data. As we capture the combined class of pFDs and pMVDs, our results subsume previous findings on schema design based on pFDs alone [21].

Our framework subsumes Fagin's well-known MVDs and 4NF from relational databases as the special case with only one p-degree [6]. This includes results on the implication problem for the combined class of FDs and MVDs [7, 17], as well as the semantic justification of 4NF by the absence of data redundancy caused by MVDs [25].

Few other papers address schema design for uncertain data [4, 24]. In [24] an "FD theory for data models whose basic construct for uncertainty is *alternatives*" is developed. That work is thus fundamentally different from the current approach. In particular, p-relations cannot always be expressed by the uncertain relations of [24]. For example, the simple two-tuple p-relation $\{(t_1, \alpha_1), (t_2, \alpha_2)\}$ with the possible worlds $w_1 = \{t_1\} \subseteq \{t_1, t_2\} = w_2$ cannot be expressed: The world w_1 says there is at most one record in which t_1 is an alternative, while the world w_2 says that there must be at least two records, namely one record in which t_1 is an alternative and one record in which t_2 is an alternative. Indeed, t_1 and t_2 cannot be alternatives of the same record since possible worlds in [24] result from choosing one alternative from each record. The article [4] models fuzziness in an Entity-Relationship model, so addresses schema design by a conceptual approach and not by a logical approach as we do. [4] derives the uncertainty of their fuzzy FDs from fuzzy similarity relations between attribute values, as proposed in [22]. That means that classical normalization is applied to data value redundancy with weaker notions of value equality, but always to the same set of records and the same set of FDs. Instead, the certainty of pFDs in our approach is derived from the largest possible world of records to which they apply. That means classical normalization is still applied to data value redundancy based on value equality, but optimized in terms of the number of FDs that apply to a possible world. Both approaches are therefore incomparable, even if only FDs are considered.

4 Possibilistic Relations

We summarize our model of uncertain data, introduced in [20] and further used in [21].

A relation schema, usually denoted by R, is a finite non-empty set of *attributes*. Each attribute $A \in R$ has a *domain* $dom(A)$ of values. A *tuple* t over R is an element of the Cartesian product $\prod_{A\in R} dom(A)$ of the attributes' domains. For $X \subseteq R$ we denote by $t(X)$ the *projection* of t on X. A *relation* over R is a finite set r of tuples over R. As a running example we use the relation schema SUPPLY with attributes *Product*, *Location*, *Supplier*, and *Date*, as introduced before.

We define possibilistic relations as relations where each tuple is associated with some confidence. The confidence of a tuple expresses up to which degree of possibility a tuple occurs in a relation. Formally, we model the confidence as a *scale of possibility*, that is, a finite, strictly linear order $\mathcal{S} = (S, <)$ with $k + 1$ elements where k is some positive integer, which we denote by $\alpha_1 > \cdots > \alpha_k > \alpha_{k+1}$, and whose elements $\alpha_i \in S$ we call *possibility degrees* (p-degrees). The top p-degree α_1 is reserved for tuples that are 'fully possible' to occur in a relation, while the bottom p-degree α_{k+1} is reserved for tuples that are ‘not possible at all’, that is ‘impossible’, to occur in a relation currently. The use of the bottom p-degree α_{k+1} in our model is the counterpart of the classical closed world assumption. Humans like to use simple scales in everyday life, for instance to communicate, compare, or rank. Simple usually means to classify items qualitatively, rather than quantitatively by putting a precise value on it. Note that classical relations use a scale with two elements, that is, where $k = 1$.

In our running example, source one says which products are supplied on which days from which location, while sources two and three say which suppliers are found at these locations. While both sources confirm “The little link” as supplier at “Singapore” and “Malaysia”, only source two mentions supplier “Eminence” at “Singapore” and only source three mentions supplier “Eminence” at “Malaysia”. Accordingly, (The little link, Singapore) and (The little link, Malaysia) have p-degree α_1, while (Eminence, Singapore) and (Eminence, Malaysia) only have p-degree α_2. This is a natural technique to assign p-degrees to records based on the number k of sources in which they appear: they obtain the highest p-degree α_1 when they appear in all sources, the second highest p-degree α_2 when they appear in all but one source, and so on until records that appear in only one source obtain the lowest p-degree α_k and the bottom p-degree α_{k+1} is reserved for records that do not appear in any source.

Formally, a *possibilistic relation schema* (p-schema) $(R, \mathcal{S})$ consists of a relation schema R and a possibility scale $\mathcal{S}$. A *possibilistic relation* (p-relation) over $(R, \mathcal{S})$ consists of a relation r over R, together with a function $Poss_r$ that maps each tuple $t \in r$ to a p-degree $Poss_r(t)$ in the possibility scale $\mathcal{S}$. Sometimes, we simply refer to a p-relation $(r, Poss_r)$ by r, assuming that $Poss_r$ has been fixed. For example, Table 1 shows our p-relation $(r, Poss_r)$ over $(\text{SUPPLY}, \mathcal{S} = \{\alpha_1, \alpha_2, \alpha_3\})$.

Table 1: Possibilistic relation $(r, Poss_r)$

Product	*Location*	*Supplier*	*Date*	$Poss_r$
Cufflinks	Singapore	The little link	04/11/2018	α_1
Cufflinks	Malaysia	The little link	21/11/2018	α_1
Cufflinks	Singapore	The little link	21/11/2018	α_1
Cufflinks	Malaysia	The little link	04/11/2018	α_1
Cufflinks	Singapore	Eminence	04/11/2018	α_2
Cufflinks	Malaysia	Eminence	21/11/2018	α_2
Cufflinks	Singapore	Eminence	21/11/2018	α_2
Cufflinks	Malaysia	Eminence	04/11/2018	α_2

P-relations enjoy a well-founded semantics in terms of possible worlds. In fact, a p-relation gives rise to a possibility distribution over possible worlds of relations. For $i = 1, \ldots, k$ let r_i denote the relation that consists of all tuples in r that have a p-degree of at least α_i, that is, $r_i = \{t \in r \mid Poss_r(t) \geq \alpha_i\}$. The linear order of the p-degrees results in a linear order of possible worlds of relations. Indeed, we have $r_1 \subseteq r_2 \subseteq \cdots \subseteq r_k$. The possibility distribution π_r for this linear chain of possible worlds is defined by $\pi_r(r_i) = \alpha_i$. Note that r_{k+1} is not considered to be a possible world, since its possibility $\pi(r_{k+1}) = \alpha_{k+1}$ means ‘not possible at all’. Vice versa, the possibility $Poss_r(t)$ of a tuple $t \in r$ is the possibility of the smallest possible world in which t occurs, that is, the maximum possibility $\max\{\alpha_i \mid t \in r_i\}$ of a world to which t belongs. If $t \notin r_k$, then $Poss_r(t) = \alpha_{k+1}$. The top p-degree α_1 takes on a distinguished role: every tuple that is ‘fully possible’ occurs in every possible world - and is thus - ‘fully certain’. This confirms our intuition that p-relations subsume relations (of fully certain tuples) as a special case.

4.1 Functional and Multivalued Dependencies

Recall that an FD $X \rightarrow Y$ is satisfied by a relation r whenever every pair of tuples in r that have matching values on all the attributes in X have also matching values on all the attributes in Y [6]. For example, the FD *Location* → *Supplier* is not satisfied by r_2 but by r_1. An MVD $X \twoheadrightarrow Y$ is satisfied by a relation r over relation schema R whenever for every pair of tuples in r that have matching values on all the attributes in X there is some tuple in r that has matching values on all the attributes in XY with the first tuple and matching values on all the attributes in $X(R-Y)$ with the second tuple [6]. For example, the MVD $\mathit{Product} \twoheadrightarrow \mathit{Location}$ is satisfies by relations r_1 and r_2.

For a given FD or MVD σ, the marginal certainty with which a σ holds in a p-relation corresponds to the p-degree of the smallest possible world in which σ is violated. Therefore, dually to a scale $\mathcal{S}$ of p-degrees for tuples we use a scale $\mathcal{S}^T$ of certainty degrees (c-degrees) for FDs and MVDs. We commonly use subscripted versions of the Greek letter β to denote c-degrees associated with FDs and MVDs. Formally, the duality between p-degrees in $\mathcal{S}$ and the c-degrees in $\mathcal{S}^T$ can be defined by the mapping $\alpha_i \mapsto \beta_{k+2-i}$, for $i = 1, \ldots, k+1$. Assuming that the world r_{k+1} cannot satisfy any FD or MVD, the marginal certainty $C_{(r,Poss_r)}(\sigma)$ with which the FD or MVD σ holds on the p-relation $(r, Poss_r)$ is the c-degree β_{k+2-i} that corresponds to the smallest world r_i in which σ is violated, that is,

$$C_{(r,Poss_r)}(\sigma) = \min\{\beta_{k+2-i} \mid \not\models_{r_i} \sigma\}.$$

In particular, if r_k satisfies σ, then $C_{(r,Poss_r)}(\sigma) = \beta_1$. We can now define the syntax and semantics of pFDs and pMVDs.

Definition 1 *A* possibilistic FD *(pFD) over a p-schema* $(R, \mathcal{S})$ *is an expression* $(X \rightarrow Y, \beta)$ *where* $X, Y \subseteq R$ *and* $\beta \in \mathcal{S}^T$*. A* possibilistic MVD *(pMVD) over a p-schema* $(R, \mathcal{S})$ *is an expression* $(X \twoheadrightarrow Y, \beta)$ *where* $X, Y \subseteq R$ *and* $\beta \in \mathcal{S}^T$*. A p-relation* $(r, Poss_r)$ *over* $(R, \mathcal{S})$ satisfies *the pFD or pMVD* (σ, β) *if and only if* $C_{(r,Poss_r)}(\sigma) \geq \beta$*.*

The following comment illustrates a difference between FDs and MVDs that is important for Definition 1 of their possibilistic variants. FDs are downwards-closed: whenever a relation r satisfies an FD, then every sub-relation $s \subseteq r$ also satisfies the FD. This is not the case for MVDs: For example, the relation r_1 satisfies the MVD $\mathit{Product} \twoheadrightarrow \mathit{Supplier}$, while every sub-relation with either 2 or 3 records violates this MVD. The downward-closure property of FDs provides a very natural definition of its possibilistic counterpart: If the FD is satisfied by some world, then it is satisfied by every smaller world as well. This means that every FD that holds with c-degree β also holds with every smaller c-degree β'. However, even though MVDs are not downward-closed, they still permit the same natural definition for their possibilistic counterpart: whenever there are two tuples $t, t' \in r$ with p-degrees α_i and α_j, respectively, then an MVD that is satisfied by $r_{\max\{i,j\}}$ also generates a tuple $t'' \in r_{\max\{i,j\}}$, that is, t'' has minimum p-degree $\min\{\alpha_i, \alpha_j\}$. In particular, if $t, t' \in r_i$, then $t'' \in r_i$, too. In other words, if an MVD is satisfied by some world, then it is also satisfied by every smaller world. This means that every MVD that holds with c-degree β also holds with every smaller c-degree β'. For example, the world r_2 satisfies the MVD $\mathit{Product} \twoheadrightarrow \mathit{Supplier}$, and also the world r_1. In summary, our definition of the marginal certainty applies to both FDs and MVDs, but is motivated by different properties: namely by the downward-closure property of FDs and the tuple-generating property of MVDs.

5 Possibilistic Design Theory

Relational normalization for 4NF [6] is founded on the theory of functional and multivalued dependencies, in particular the axiomatic and algorithmic solutions to their implication problem [6,7]. Consequently, we now establish a design theory for pFDs and pMVDs as a foundation for a possibilistic 4NF normal form, its semantic justification, and normalization of p-schemata. First, we establish a strong link between the implication problem of pFDs and pMVDs and the implication problem of FDs and MVDs, which is a consequence of the downward closure property of FDs and the tuple-generating property of MVDs. Based on this link, we then establish axiomatic and algorithmic solutions to the implication problem of pFDs and pMVDs. Our design theory for pFDs and pMVDs subsumes the design theory for classical FDs and MVDs as the special case where $k = 1$, and also subsumes the design theory for pFDs as the special case with no pMVDs.

5.1 β-Cuts

We establish a precise correspondence between instances of the implication problem for pFDs and pMVDs and instances of the implication problem for relational FDs and MVDs. Let $\Sigma \cup \{\varphi\}$ denote a set of pFDs and pMVDs over a p-schema $(R, \mathcal{S})$. We say that Σ *implies* φ, denoted by $\Sigma \models \varphi$, if every p-relation (r, Poss_r) over $(R, \mathcal{S})$ that satisfies every pFD and pMVD in Σ also satisfies φ.

Example 1 *Let Σ consist of the pMVD $(\mathit{Product} \twoheadrightarrow \mathit{Location}, \beta_1)$ but also the pFD $(\mathit{Location} \rightarrow \mathit{Supplier}, \beta_2)$ over $(\textsc{Supply}, \{\alpha_1, \alpha_2, \alpha_3\})$. Further, let φ denote the pFD $(\mathit{Product} \rightarrow \mathit{Supplier}, \beta_1)$. Then Σ does not imply φ as the following p-relation witnesses.*

Product	*Location*	*Supplier*	*Date*	*P-degree*
Cufflinks	*Singapore*	*The little link*	*04/11/2018*	α_1
Cufflinks	*Singapore*	*Eminence*	*04/11/2018*	α_2

For a set Σ of pFDs and pMVDs on some p-schema $(R, \mathcal{S})$ and c-degree $\beta \in \mathcal{S}^T$ where $\beta > \beta_{k+1}$, let

$$\Sigma_\beta = \{X \rightarrow Y \mid (X \rightarrow Y, \beta') \in \Sigma \text{ and } \beta' \geq \beta\}$$

be the *β-cut* of Σ. The major strength of our framework is engraved in the following result. It says that a pFD or pMVD (σ, β) with c-degree β is implied by a set Σ of pFDs and pMVDs if and only if the FD or MVD σ is implied by the β-cut Σ_β of Σ.

Theorem 1 *Let $\Sigma \cup \{(\sigma, \beta)\}$ be a set of pFDs and pMVDs over a p-schema $(R, \mathcal{S})$ where $\beta > \beta_{k+1}$. Then $\Sigma \models (\sigma, \beta)$ if and only if $\Sigma_\beta \models \sigma$.*

The following example illustrates Theorem 1.

Table 2: Axiomatization $\mathfrak{P}$ of pFDs and pMVDs

$$\frac{}{(XY \to X, \beta)}$$
(FD reflexivity, $\mathcal{R}$)

$$\frac{(X \to Y, \beta)}{(X \to XY, \beta)}$$
(FD extension, $\mathcal{E}$)

$$\frac{(X \to Y, \beta) \quad (Y \to Z, \beta)}{(X \to Z, \beta)}$$
(FD transitivity, $\mathcal{T}$)

$$\frac{}{(\emptyset \twoheadrightarrow R, \beta)}$$
(MVD complement, $\mathcal{C}$)

$$\frac{(X \twoheadrightarrow Y, \beta) \quad (X \twoheadrightarrow Z, \beta)}{(X \to YZ, \beta)}$$
(MVD union, $\mathcal{U}$)

$$\frac{(X \twoheadrightarrow Y, \beta) \quad (Y \twoheadrightarrow Z, \beta)}{(X \twoheadrightarrow Z - Y, \beta)}$$
(MVD pseudo-transitivity, $\mathcal{P}_M$)

$$\frac{}{(\sigma, \beta_{k+1})}$$
(bottom rule, $\mathcal{B}$)

$$\frac{(X \to Y, \beta)}{(X \twoheadrightarrow Y, \beta)}$$
(FD-MVD implication, $\mathcal{I}$)

$$\frac{(X \twoheadrightarrow Y, \beta) \quad (Y \to Z, \beta)}{(X \to Z - Y, \beta)}$$
(FD-MVD pseudo-transitivity, $\mathcal{P}_{FM}$)

$$\frac{(\sigma, \beta)}{(\sigma, \beta')}\beta \geq \beta'$$
(weakening, $\mathcal{W}$)

Example 2 *Let Σ and φ be as in Example 1, in particular Σ does not imply φ. Theorem 1 reduces the implication problem of pFDs and pMVDs to that of FDs and MVDs, namely Σ_{β_1} does not imply Product $\to$ Supplier. Indeed, the possible world r_2 of the p-relation from Example 1*

Product	*Location*	*Supplier*	*Date*
Cufflinks	*Singapore*	*The little link*	*04/11/2018*
Cufflinks	*Singapore*	*Eminence*	*04/11/2018*

satisfies the MVD Product $\twoheadrightarrow$ Location that forms Σ_{β_1}, and violates the FD Product $\to$ Supplier.

5.2 Axiomatic Characterization

The *semantic closure* $\Sigma^* = \{\varphi \mid \Sigma \models \varphi\}$ contains all pFDs and pMVDs implied by Σ. We compute Σ^* by applying *inference rules* of the form $\frac{\text{premise}}{\text{conclusion}}$ condition, where rules without premise are *axioms*. For a set $\mathfrak{R}$ of inference rules let $\Sigma \vdash_{\mathfrak{R}} \varphi$ denote that there is an *inference* of φ from Σ by $\mathfrak{R}$. That is, there is some sequence $\sigma_1, \ldots, \sigma_n$ such that $\sigma_n = \varphi$ and every σ_i is in Σ or the result of applying a rule in $\mathfrak{R}$ to some premises in $\{\sigma_1, \ldots, \sigma_{i-1}\}$. Let $\Sigma^+_{\mathfrak{R}} = \{\varphi \mid \Sigma \vdash_{\mathfrak{R}} \varphi\}$ denote the *syntactic closure* of Σ under inferences by $\mathfrak{R}$. $\mathfrak{R}$ is *sound* (*complete*) if for every p-schema $(R, \mathcal{S})$ and for every set Σ we have $\Sigma^+_{\mathfrak{R}} \subseteq \Sigma^*$ ($\Sigma^* \subseteq \Sigma^+_{\mathfrak{R}}$). The (finite) set $\mathfrak{R}$ is a (finite) *axiomatization* if $\mathfrak{R}$ is both sound and complete.

Table 2 shows the axiomatization $\mathfrak{P}$ of pFDs and pMVDs. For a rule in $\mathcal{R} \in \mathfrak{P} - \{\mathcal{B}, \mathcal{W}\}$ let $\mathcal{R}'$ denote the rule that is obtained from $\mathcal{R}$ by omitting the c-degree β and the parenthesis, and let $\mathfrak{P}' = \{\mathcal{R}' \mid \mathcal{R} \in \mathfrak{P} - \{\mathcal{B}, \mathcal{W}\}\}$. It is known that $\mathfrak{P}'$ forms an axiomatization for FDs and MVDs [9]. In fact, $\mathfrak{P}$ subsumes the axiomatization for FDs and MVDs as the special case where the scale $\mathcal{S}^T$ consists of just two c-degrees. In the rules all attribute sets X, Y, Z are subsets of the given relation schema R, the c-degrees β

and β' belong to the given certainty scale $\mathcal{S}^T$, and β_{k+1} denotes the bottom c-degree. For a completeness proof of $\mathfrak{P}$ we could use the classical strategy to write down a two-tuple p-relation that violates a given pFD or pMVD (σ, β) that cannot be inferred from a given pFD and pMVD set Σ using $\mathfrak{P}$. Instead, we establish the completeness of $\mathfrak{P}$ directly by showing that a given pFD or pMVD (σ, β) that is implied by Σ can be also be inferred from Σ using the rules in $\mathfrak{P}$. This direct proof shows how the bottom axiom $\mathcal{B}$ and weakening rule $\mathcal{W}$ can be applied to reduce the inference of (σ, β) from Σ to an inference of σ from Σ_β. However, the completeness of $\mathfrak{P}'$ guarantees immediately that σ can be inferred from Σ_β, due to Theorem 1 and the assumption that (σ, β) is implied by Σ.

Theorem 2 *$\mathfrak{P}$ forms a finite axiomatization for the implication of pFDs and pMVDs.*

Example 3 *Let $\Sigma = \{(\textit{Product} \twoheadrightarrow \textit{Location}, \beta_1), (\textit{Location} \rightarrow \textit{Supplier}, \beta_2)\}$ and $\varphi = (\textit{Product} \rightarrow \textit{Supplier}, \beta_2)$ over* Supply, *and $\varphi' = (\textit{Product} \rightarrow \textit{Supplier}, \beta_1)$. We show an inference of φ from Σ by $\mathfrak{P}$.*

$$\dfrac{\mathcal{W}: \dfrac{(\textit{Product} \twoheadrightarrow \textit{Location}, \beta_1)}{(\textit{Product} \twoheadrightarrow \textit{Location}, \beta_2)} \quad (\textit{Location} \rightarrow \textit{Supplier}, \beta_2)}{\mathcal{P}_{FM}: \quad (\textit{Product} \rightarrow \textit{Supplier}, \beta_2)}$$

Of course, φ' cannot be inferred from Σ by $\mathfrak{P}$, as Example 1 shows.

5.3 Algorithmic Characterization

In practice it is often unnecessary to compute the closure Σ^* from a given set Σ. Instead, rather frequently occurs the problem of deciding whether a given Σ implies a given φ.

PROBLEM:	IMPLICATION
INPUT:	Relation schema R, Scale $\mathcal{S}$ with $k+1$ possibility degrees, Set $\Sigma \cup \{\varphi\}$ of pFDs and pMVDs over $(R, \mathcal{S})$
OUTPUT:	Yes, if $\Sigma \models \varphi$, and No, otherwise

One may compute Σ^* and check if $\varphi \in \Sigma^*$, but this is inefficient and does not make effective use of the additional input φ. For pFDs and pMVDs, we can exploit Theorem 1 to derive an almost linear time algorithm that decides the implication problem. Given a pFD and pMVD set $\Sigma \cup \{(\sigma, \beta)\}$ we return *true* if $\beta = \beta_{k+1}$ (since this is the trivial case where β_{k+1} is the bottom c-degree), otherwise it is sufficient to check if $\Sigma_\beta \models \sigma$. The latter test can be done in almost linear time [7].

Theorem 3 *The implication problem $\Sigma \models \varphi$ of pFDs and pMVDs can be decided in time $\mathcal{O}(||\Sigma|| \cdot \log ||\Sigma||)$.*

We illustrate the algorithm on our running example.

Example 4 *Let $\Sigma = \{(\textit{Product} \twoheadrightarrow \textit{Location}, \beta_1), (\textit{Location} \rightarrow \textit{Supplier}, \beta_2)\}$ and $\varphi = (\textit{Product} \rightarrow \textit{Supplier}, \beta_2)$ over* Supply, *and $\varphi' = (\textit{Product} \rightarrow \textit{Supplier}, \beta_1)$. Indeed, Product $\rightarrow$ Supplier is not implied by Σ_{β_1} while Product $\rightarrow$ Supplier is implied by Σ_{β_2}. That is, Σ implies φ but Σ does not imply φ'.*

6 Possibilistic Fourth Normal Form Design

In p-relations different tuples may have different p-degrees, and different pFDs and pMVDs may apply to them. Hence, data value redundancy is caused by pFDs or pMVDs with different c-degrees and occurs in tuples of different p-degrees. Indeed, the smaller the p-degree for which data value redundancy is to be eliminated, the smaller the number of pFDs and pMVDs that can cause this redundancy. Consequently, the smaller the normalization effort will be, too. We will now exploit this observation to tailor relational schema design for applications with different requirements for the uncertainty of their data. For this purpose, we will introduce notions of data value redundancy that target the p-degree of tuples in which they occur. This results in a family of semantic normal forms by which data value redundancy of varying p-degrees are eliminated. We characterize each of the semantic normal forms by a corresponding syntactic normal form, and establish strong correspondences with Fagin's 4NF in relational databases [6, 25].

6.1 Redundancy-Free Normal Form

Motivated by our running example we propose different degrees of data value redundancy that are tailored towards the different p-degrees of tuples in a p-relation. For this, we exploit the classical proposal by Vincent [25]. Let R denote a relation schema, A an attribute of R, t a tuple over R, and Σ a set of constraints over R. A *replacement* of $t(A)$ is a tuple $\bar{t}$ over R such that: i) for all $\bar{A} \in R - \{A\}$ we have $\bar{t}(\bar{A}) = t(\bar{A})$, and ii) $\bar{t}(A) \neq t(A)$. For a relation r over R that satisfies Σ and $t \in r$, the data value occurrence $t(A)$ in r is *redundant* for Σ if and only if for every replacement $\bar{t}$ of $t(A)$, $\bar{r} := (r - \{t\}) \cup \{\bar{t}\}$ violates some constraint in Σ. A relation schema R is in *Redundancy-Free normal form* (RFNF) for a set Σ of constraints if and only if there are no relation r over R that satisfies Σ, tuple $t \in r$, and attribute $A \in R$ such that the data value occurrence $t(A)$ is redundant for Σ [25].

Definition 2 *Let $(R, \mathcal{S})$ denote a p-schema, Σ a set of pFDs and pMVDs over $(R, \mathcal{S})$, $A \in R$ an attribute, $(r, Poss_r)$ a p-relation over $(R, \mathcal{S})$ that satisfies Σ, and t a tuple in r_i. The data value occurrence $t(A)$ is* α_i-redundant *if and only if $t(A)$ is redundant for $\Sigma_{\alpha_i} = \{X \to Y \mid (X \to Y, \beta) \in \Sigma$ and $\beta \geq \beta_{k+1-i}\}$.*

This definition meets the intuition of data value redundancy in our running example. For instance, each occurrence of *The little link* is α_1-redundant, and each occurrences of *Eminence* is α_2-redundant. Importantly, α_i-redundant data value occurrences can only be caused by pFDs or pMVDs (σ, β) that apply to the world of the occurrence, that is, where $\beta \geq \beta_{k+1-i}$. Hence, α_1-redundancy can be caused by pFDs or pMVDs with any c-degree $\beta_1, \ldots, \beta_k$, while α_k-redundancy can only be caused by pFDs or pMVDs with c-degree β_1. This motivates the following definition.

Definition 3 *A p-schema $(R, \mathcal{S})$ is in* α_i-Redundancy-Free Normal Form *(α_i-RFNF) for a set Σ of pFDs and pMVDs over $(R, \mathcal{S})$ if and only if there do not exist a p-relation $(r, Poss_r)$ over $(R, \mathcal{S})$ that satisfies Σ, an attribute $A \in R$, and a tuple $t \in r_i$ such that $t(A)$ is α_i-redundant.*

For example, (Supply, $\mathcal{S}$) is neither in α_1-RFNF nor in α_2-RFNF for Σ. The negative results follow directly from the redundant occurrences in Figure 1. Indeed, α_i-RFNF characterizes p-schemata that permit only p-relations whose possible world r_i is free from data redundancy caused by the classical FDs and MVDs that apply to it.

Theorem 4 *$(R, \mathcal{S})$ is in α_i-RFNF for Σ if and only if R is in RFNF for Σ_{α_i}.*

6.2 Possibilistic Fourth Normal Form

Our goal is now to characterize α-RFNF, which is a semantic normal form, purely syntactically. Therefore, we propose qualitative variants of the classical 4NF condition [6]. Recall that a relation schema R is in Fourth normal form (4NF) for a set Σ of FDs and MVDs over R if and only if for all $X \twoheadrightarrow Y \in \Sigma^+_{\mathfrak{P}'}$ where $Y \not\subseteq X$ and $Y \neq R$, we have $X \to R \in \Sigma^+_{\mathfrak{P}'}$. Here, $\Sigma^+_{\mathfrak{P}'}$ denotes the syntactic closure of Σ with respect to the axiomatization $\mathfrak{P}'$ of FDs and MVDs [9]. While α-RFNF is defined semantically using the p-degree α of a possible world, qualitative variants of 4NF are defined syntactically using the c-degrees of the given pFDs and pMVDs.

Definition 4 *A p-schema $(R, \mathcal{S})$ is in β-*4NF *for a set Σ of pFDs and pMVDs over $(R, \mathcal{S})$ if and only if for every pMVD $(X \twoheadrightarrow Y, \beta) \in \Sigma^+_{\mathfrak{P}}$ where $Y \not\subseteq X$ and $Y \neq R$, we have $(X \to R, \beta) \in \Sigma^+_{\mathfrak{P}}$.*

Recall that sets Σ and Θ are *covers* of one another if $\Sigma^* = \Theta^*$ holds. The property of being in β-4NF for Σ is independent of the representation of Σ. That is, for every cover Σ' of Σ, $(R, \mathcal{S})$ is in β-4NF for Σ if and only if $(R, \mathcal{S})$ is in β-4NF for Σ'. The β-4NF condition for a pFD and pMVD set Σ can be characterized by the 4NF condition for the FD and MVD set Σ_β.

Theorem 5 *$(R, \mathcal{S})$ is in β-4NF for Σ if and only if R is in 4NF for Σ_β.*

We can now characterize the semantic α_i-RFNF by the syntactic β_{k+1-i}-4NF.

Theorem 6 *For all $i = 1, \ldots, k$, $(R, \mathcal{S})$ with $|\mathcal{S}| = k+1$ is in α_i-RFNF with respect to Σ if and only if $(R, \mathcal{S})$ is in β_{k+1-i}-4NF with respect to Σ.*

$$\begin{tikzcd}
\Sigma_{\beta_2}\text{-}BCNF \arrow[d, Leftrightarrow] & \Sigma_{\beta_1}\text{-}BCNF \arrow[d, Leftrightarrow] \\
\beta_2\text{-}BCNF \arrow[d, Leftrightarrow] & \beta_1\text{-}BCNF \arrow[d, Leftrightarrow] \\
\alpha_1\text{-}RFNF \arrow[d, Leftrightarrow] & \alpha_2\text{-}RFNF \arrow[d, Leftrightarrow] \\
\Sigma_{\alpha_1}\text{-}RFNF & \Sigma_{\alpha_2}\text{-}RFNF
\end{tikzcd}$$

Figure 2: Normal Forms

Figure 2 shows the correspondences between the syntactic and semantic normal forms, and their relationships to classical normal forms.

Due to the cover-insensitivity of the β-4NF condition, one may wonder about the efficiency of checking whether a given p-schema $(R, \mathcal{S})$ is in β-4NF with respect to a set Σ. Indeed, as in the classical case it suffices to check some pFDs and pMVDs in Σ instead of checking all pMVDs in $\Sigma^+_{\mathfrak{P}}$.

Theorem 7 *A p-schema $(R, \mathcal{S})$ is in β-4NF for a set Σ of pFDs and pMVDs over $(R, \mathcal{S})$ if and only if for every pFD $(X \to Y, \beta') \in \Sigma$ where $\beta' \geq \beta$ and $Y \not\subseteq X$ and for every pMVD $(X \twoheadrightarrow Y, \beta') \in \Sigma$ where $\beta' \geq \beta$ and $Y \not\subseteq X$ and $Y \neq R$, we have $(X \to R, \beta) \in \Sigma^+_{\mathfrak{P}}$.*

Example 5 *Let* (Supply, $\mathcal{S}$) *and* Σ *. Using Theorem 7 we can observe that the schema is neither in β_1- nor β_2-4NF for Σ. By Theorem 6 we conclude that the schema is neither in α_2- nor α_1-RFNF for Σ. By Theorem 5 it follows that* Supply *is neither in 4NF for Σ_{β_1}, nor Σ_{β_2}. Finally, by Theorem 4, it follows that* Supply *is neither in RFNF for Σ_{α_2} or Σ_{α_1}.*

7 Qualitative Normalization

We now establish algorithmic means to design relational database schemata for applications with uncertain data. For that purpose, we normalize a given p-schema $(R, \mathcal{S})$ for the given set Σ of pFDs and pMVDs. Our strategy is to fix some c-degree $\beta \in \mathcal{S}^T$ that determines which possible world we normalize for which FDs and MVDs. For each choice of a c-degree, we pursue 4NF normalizations to obtain lossless decompositions free from any data value redundancy but potentially not dependency-preserving (that is, some FDs or MVDs may require validation on the join of some relations). Applying our strategy to different c-degrees provides organizations with a variety of normalized database schemata, each targeted at different levels of data integrity, data losslessness, and the efficiency of different updates and queries. In this sense, our c-degrees are parameters that allow stakeholders to control trade-offs between data integrity and data losslessness, as well as between query efficiency and update efficiency, as illustrated in Figure 3.

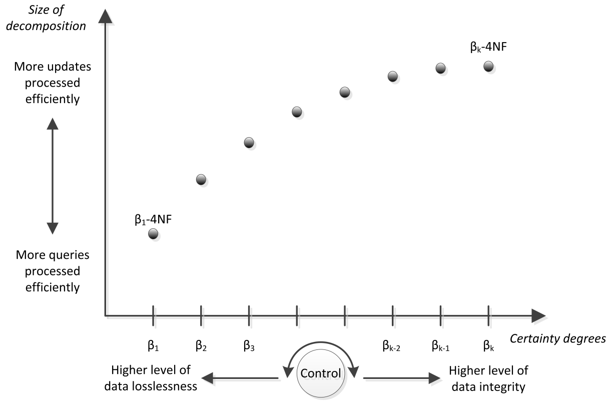

Size of decomposition
βk-4NF
More updates processed efficiently
β1-4NF
More queries processed efficiently
β1
β2
β3
βk-2
βk-1
βk
Certainty degrees
Higher level of data losslessness
Control
Higher level of data integrity

Figure 3: C-degrees to control design trade-offs

7.1 4NF Decomposition

We recall basic terminology from relational databases. A *decomposition* of relation schema R is a set $\mathcal{D} = \{R_1, \ldots, R_n\}$ of relation schemata such that $R_1 \cup \cdots \cup R_n = R$. For $R_j \subseteq R$ and an FD and MVD set Σ over R, $\Sigma[R_j] = \{X \to Y \mid X \to Y \in \Sigma^+_{\mathfrak{A}}$ and $X, Y \subseteq R_j\}$ denotes the *projection* of Σ onto R_j. A decomposition $\mathcal{D}$ of a relation schema R with FD and MVD set Σ is *lossless* if and only if every relation r over R that satisfies Σ is the join of its projections on the elements of $\mathcal{D}$, that is, $r = \bowtie_{R_j \in \mathcal{D}} r[R_j]$. Here, $r[R_j] = \{t(R_j) \mid t \in r\}$. A *4NF* decomposition of a relation schema R with FD and

MVD set Σ is a decomposition $\mathcal{D}$ of R where every $R_j \in \mathcal{D}$ is in 4NF for $\Sigma[R_j]$. Theorem 5 motivates the following definition of a 4NF decomposition that is lossless for a given p-degree.

Definition 5 *Given a p-schema $(R, \{\alpha_1, \ldots, \alpha_{k+1}\})$, an α_{k+1-i}-lossless 4NF decomposition for the pFD and pMVD set Σ is a lossless 4NF decomposition of R for Σ_{β_i}.*

Instrumental to Definition 5 is the following decomposition theorem, which follows directly from Theorem 1. It covers the classical decomposition theorem [6] as the special case of having just one possible world.

Theorem 8 *Let $(X \to Y, \beta_i)$ be a pFD and $(X \twoheadrightarrow Y, \beta_i)$ be a pMVD with $1 \leq i < k+1$ that satisfies the p-relation $(r, Poss_r)$ over the p-schema $(R, \mathcal{S})$. Then $r_{k+1-i} = r_{k+1-i}[XY] \bowtie r_{k+1-i}[X(R-Y)]$, that is, the possible world r_{k+1-i} of r is the lossless join of its projections on XY and $X(R-Y)$.*

Therefore, an α_{k+1-i}-lossless 4NF decomposition for a pFD and pMVD set Σ can simply be obtained by performing a classical lossless 4NF decomposition for the β_i-cut Σ_{β_i} of Σ. This suggests a simple lossless 4NF decomposition strategy.

PROBLEM:	Qualitative 4NF Decomposition
INPUT:	Possibilistic Relation Schema $(R, \mathcal{S})$ Set Σ of pFDs and pMVDs over $(R, \mathcal{S})$ Certainty degree $\beta_i \in \mathcal{S}^T - \{\beta_{k+1}\}$
OUTPUT:	α_{k+1-i}-lossless 4NF decomposition of $(R, \mathcal{S})$ with respect to Σ
METHOD:	Perform a lossless 4NF decomposition of R with respect to Σ_{β_i}

We illustrate the decomposition on our running example.

Example 6 *Let* $(\textsc{Supply}, \mathcal{S})$ *and Σ be as in Example 1. As* $(\textsc{Supply}, \mathcal{S})$ *is not in β_1-4NF for Σ, we perform an α_2-lossless 4NF decomposition for Σ_{β_1}. The result consists of $R_1 = \{Product, Location\}$ and $R_2 = \{Product, Supplier, Date\}$. Note that the MVD in Σ_{β_1} is satisfied by the join of any relation over R_1 with any relation over R_2. That means our 4NF decomposition is β_1-dependency-preserving.*

The last example is rather special, since one cannot expect to preserve all FDs and MVDs in the 4NF decomposition process. Recall that a decomposition $\mathcal{D}$ of relation schema R with FD and MVD set Σ is *dependency-preserving* if and only if the join $\bowtie_{r \text{ relation over } D \in \mathcal{D}} r$ of every relation r over every $D \in \mathcal{D}$ satisfies Σ. Note here that we consider only relations r over D of the join that satisfy all the FDs and MVDs in the projection of Σ onto D.

This is illustrated with another example.

Example 7 *Let* $(\textsc{Supply}, \mathcal{S})$ *and Σ be as in Example 1. As* $(\textsc{Supply}, \mathcal{S})$ *is not in β_2-4NF for Σ, we perform an α_1-lossless 4NF decomposition for Σ_{β_2}. The result is $R_1 = \{Product, Location\}$ and $R_2 = \{Product, Supplier\}$ with projected FD/MVD set*

$$\Sigma_{\beta_2}[R_2] = \{Product \to Supplier\}$$

and $R_3 = \{Product, Date\}$. Note that the FD Location → Supplier is not preserved by this decomposition. Since the pFDs and pMVDs in Σ apply only to world r_1 from Figure 1, this decomposition may be applied to r_1 to obtain the instance $\mathcal{D}_2$ shown in Figure 1.

Definition 6 *A β-*dependency-preserving *decomposition of a p-schema $(R, \mathcal{S})$ for the pFD and pMVD set Σ is a dependency-preserving decomposition of R for Σ_β.*

The α_2-lossless 4NF decomposition from Example 6 is β_1-dependency-preserving, but the α_1-lossless 4NF decomposition from Example 7 is not β_2-dependency-preserving. In practice, lost dependencies can only be validated by joining relations after inserts or modification. For example, to validate the FD *Location* → *Supplier* after an update, one would have to join R_1 and R_2 from Example 7. This can be prohibitively expensive. While 3NF synthesis algorithms exist that can transform any relational database schema into one that is dependency-preserving, such normal form is unknown for MVDs.

8 Conclusion and Future Work

We have extended Fagin's 4NF from certain [6] to uncertain data. Our results show how traditional database design can make the most of information about the uncertainty in data, provided that the information about the data is given in the form of possibility degrees. Future work includes extensions to partial [10, 12, 14, 16, 18] and nested data [8,11] over fixed and undetermined schemata [1,2,19], and to referential integrity [13,15]. Investigations in the context of probabilistic databases must take different routes as any classes of probabilistic data dependencies that require a non-unary interaction are not finitely axiomatizable [3, 23].

References

[1] J. Biskup. Inferences of multivalued dependencies in fixed and undetermined universes. *Theor. Comput. Sci.*, 10:93–105, 1980.

[2] J. Biskup and S. Link. Appropriate inferences of data dependencies in relational databases. *Ann. Math. Artif. Intell.*, 63(3-4):213–255, 2011.

[3] P. Brown and S. Link. Probabilistic keys. *IEEE Trans. Knowl. Data Eng.*, 29(3):670–682, 2017.

[4] N. A. Chaudhry, J. R. Moyne, and E. A. Rundensteiner. An extended database design methodology for uncertain data management. *Inf. Sci.*, 121(1-2):83–112, 1999.

[5] D. Dubois and H. Prade. Possibility theory. In R. A. Meyers, editor, *Computational Complexity: Theory, Techniques, and Applications*, pages 2240–2252. Springer New York, 2012.

[6] R. Fagin. Multivalued dependencies and a new normal form for relational databases. *ACM Trans. Database Syst.*, 2(3):262–278, 1977.

[7] Z. Galil. An almost linear-time algorithm for computing a dependency basis in a relational database. *J. ACM*, 29(1):96–102, 1982.

[8] S. Hartmann and S. Link. Multi-valued dependencies in the presence of lists. In C. Beeri and A. Deutsch, editors, *Proceedings of the Twenty-third ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 14-16, 2004, Paris, France*, pages 330–341. ACM, 2004.

[9] S. Hartmann and S. Link. On a problem of Fagin concerning multivalued dependencies in relational databases. *Theor. Comput. Sci.*, 353(1-3):53–62, 2006.

[10] S. Hartmann and S. Link. The implication problem of data dependencies over SQL table definitions: Axiomatic, algorithmic and logical characterizations. *ACM Trans. Database Syst.*, 37(2):13:1–13:40, 2012.

[11] S. Hartmann, S. Link, and K. Schewe. Functional and multivalued dependencies in nested databases generated by record and list constructor. *Ann. Math. Artif. Intell.*, 46(1-2):114–164, 2006.

[12] H. Köhler and S. Link. SQL schema design: Foundations, normal forms, and normalization. In F. Özcan, G. Koutrika, and S. Madden, editors, *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 267–279. ACM, 2016.

[13] H. Köhler and S. Link. Inclusion dependencies and their interaction with functional dependencies in SQL. *J. Comput. Syst. Sci.*, 85:104–131, 2017.

[14] H. Köhler and S. Link. SQL schema design: foundations, normal forms, and normalization. *Inf. Syst.*, 76:88–113, 2018.

[15] M. Levene and M. W. Vincent. Justification for inclusion dependency normal form. *IEEE Trans. Knowl. Data Eng.*, 12(2):281–291, 2000.

[16] Y. E. Lien. On the equivalence of database models. *J. ACM*, 29(2):333–362, 1982.

[17] S. Link. Charting the completeness frontier of inference systems for multivalued dependencies. *Acta Inf.*, 45(7-8):565–591, 2008.

[18] S. Link. On the implication of multivalued dependencies in partial database relations. *Int. J. Found. Comput. Sci.*, 19(3):691–715, 2008.

[19] S. Link. Characterisations of multivalued dependency implication over undetermined universes. *J. Comput. Syst. Sci.*, 78(4):1026–1044, 2012.

[20] S. Link and H. Prade. Possibilistic functional dependencies and their relationship to possibility theory. *IEEE Trans. Fuzzy Systems*, 24(3):757–763, 2016.

[21] S. Link and H. Prade. Relational database schema design for uncertain data. In S. Mukhopadhyay, C. Zhai, E. Bertino, F. Crestani, J. Mostafa, J. Tang, L. Si, X. Zhou, Y. Chang, Y. Li, and P. Sondhi, editors, *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 1211–1220. ACM, 2016.

[22] K. Raju and A. K. Majumdar. Fuzzy functional dependencies and lossless join decomposition of fuzzy relational database systems. *ACM Trans. Database Syst.*, 13(2):129–166, 1988.

[23] T. Roblot, M. Hannula, and S. Link. Probabilistic cardinality constraints - validation, reasoning, and semantic summaries. *VLDB J.*, 27(6):771–795, 2018.

[24] A. D. Sarma, J. D. Ullman, and J. Widom. Schema design for uncertain databases. In M. Arenas and L. E. Bertossi, editors, *Proceedings of the 3rd Alberto Mendelzon International Workshop on Foundations of Data Management, Arequipa, Peru, May 12-15, 2009.*, volume 450 of *CEUR Workshop Proceedings*, 2009.

[25] M. W. Vincent. Semantic foundations of 4NF in relational database design. *Acta Inf.*, 36(3):173–213, 1999.

A Proofs

Theorem 9 (Theorem 1 restated) *Let $\Sigma \cup \{(\sigma, \beta)\}$ be a set of pFDs and pMVDs over a p-schema $(R, \mathcal{S})$ where $\beta > \beta_{k+1}$. Then $\Sigma \models (\sigma, \beta)$ if and only if $\Sigma_\beta \models \sigma$.*

Proof Suppose $(r, Poss_r)$ is some p-relation over $(R, \mathcal{S})$ that satisfies Σ, but violates (σ, β). In particular, $C_{(r,Poss_r)}(\sigma) < \beta$ implies that there is some relation r_i that violates σ and where

$$\beta_{k+2-i} < \beta. \tag{1}$$

Let $\sigma' \in \Sigma_\beta$, where $(\sigma', \beta') \in \Sigma$. Since r satisfies $(\sigma', \beta') \in \Sigma$ we have

$$C_{(r,Poss_r)}(\sigma') \geq \beta' \geq \beta. \tag{2}$$

If r_i violated σ', then

$$\begin{array}{rlr} \beta & > \beta_{k+2-i} & \text{by (1)} \\ & \geq C_{(r,Poss_r)}(\sigma') & \text{by Definition of } C_{(r,Poss_r)} \\ & \geq \beta & \text{by (2)} \end{array}$$

a contradiction. Hence, the relation r_i satisfies Σ_β and violates σ'.

Let r' denote some relation that satisfies Σ_β and violates σ. Without loss of generality, we assume that $r' = \{t, t'\}$ consists of only two tuples. If that is not the case, then it is well-known that there is a sub-relation of r' with two tuples that satisfies Σ_β and violates σ. Let r be the p-relation over $(R, \mathcal{S})$ that consists of the relation r' and where $Poss_{r'}(t) = \alpha_1$ and $Poss_{r'}(t') = \alpha_i$, such that $\beta_{k+1-i} = \beta$. Then r violates (σ, β) since $C_{(r,Poss_r)}(\sigma) =$

β_{k+2-i}, as $r_i = r'$ is the smallest relation that violates σ, and $\beta_{k+2-i} < \beta_{k+1-i} = \beta$. For $(\sigma', \beta') \in \Sigma$ we distinguish two cases. If r_i satisfies σ', then $C_{(r,Poss_r)}(\sigma') = \beta_1 \geq \beta$. If r_i violates σ', then $\sigma' \notin \Sigma_\beta$, i.e., $\beta' < \beta = \beta_{k+1-i}$. Therefore, $\beta' \leq \beta_{k+2-i} = C_{(r,Poss_r)}(\sigma')$ as $r_i = r'$ is the smallest relation that violates σ'. We conclude that $C_{(r,Poss_r)}(\sigma') \geq \beta'$. Consequently, $(r, Poss_r)$ is a p-relation that satisfies Σ and violates (σ, β).

Theorem 10 (Theorem 2 restated) *The set $\mathfrak{P}$ forms a finite axiomatization for the implication of pFDs and pMVDs.*

Proof The proofs of soundness are straightforward, keeping in mind the soundness of inference rules for FDs and MVDs, as well as Theorem 1. For the completeness proof, we take full advantage of Theorem 1 and the fact that the inference rules are sound and complete for the implication of FDs and MVDs. Let $(R, \mathcal{S})$ be an arbitrary possibilistisc relation schema with $|\mathcal{S}| = k+1$, and $\Sigma \cup \{(\sigma, \beta)\}$ an arbitrary set of pFDs and pMVDs over the schema, such that $\Sigma \models (\sigma, \beta)$ holds. We need to show that $\Sigma \vdash_{\mathfrak{P}} (\sigma, \beta)$ holds, too.

We distinguish two cases. If $\beta = \beta_{k+1}$, then $\Sigma \models (\sigma, \beta)$ means that $\Sigma \vdash_{\mathfrak{P}} (\sigma, \beta)$ holds by a single application of the bottom rule $\mathcal{B}$. For the remainder of the proof we therefore assume that $\beta < \beta_{k+1}$. From $\Sigma \models (\sigma, \beta)$ we conclude $\Sigma_\beta \models \sigma$ by Theorem 1. Since the inference system $\mathfrak{A}$ is complete for the implication of FDs and MVDs, we conclude that $\Sigma_\beta \vdash_{\mathfrak{A}} \sigma$ holds. Now, for the FD and MVD set Σ_β we define $\Sigma_\beta^\beta = \{(\sigma, \beta) \mid \sigma \in \Sigma_\beta\}$. Therefore, the inference of σ from Σ_β using the inference system can be easily turned into an inference of (σ, β) from Σ_β^β by $\mathfrak{P}$, simply by adding the certainty degree β to each FD and MVD that occurs in the inference. Therefore, whenever an inference rule $\mathcal{R}'$ is applied, one can now apply the corresponding rule $\mathcal{R}$, respectively. It follows that $\Sigma_\beta^\beta \vdash_{\mathfrak{P}} (\sigma, \beta)$ holds. Finally, the definition of Σ_β^β ensures that every pFD and pMVD in Σ_β^β can be inferred from a pFD or pMVD in Σ by a single application of the weakening rule $\mathcal{W}$. Hence, $\Sigma_\beta^\beta \vdash_{\mathfrak{P}} (\sigma, \beta)$ means, in particular, $\Sigma \vdash_{\mathfrak{P}} (\sigma, \beta)$. This completes the proof.

Theorem 11 (Theorem 4 restated) *$(R, \mathcal{S})$ is in α_i-RFNF with respect to Σ if and only if R is in RFNF with respect to Σ_{α_i}.*

Proof We show first the following: if $(R, \mathcal{S})$ is not in α_i-RFNF with respect to Σ, then R is in not in RFNF with respect to Σ_{α_i}. According to our hypothesis, there is some possibilistic relation $(r, Poss_r)$ over $(R, \mathcal{S})$ that satisfies Σ, an attribute $A \in R$, and a tuple $t \in r_i$ such that $t(A)$ is redundant with respect to Σ_{α_i}. In particular, it follows that r_i satisfies Σ_{α_i} since r satisfies Σ. Hence, R is not in RFNF with respect to Σ_{α_i}.

We now show: if R is in not in RFNF with respect to Σ_{α_i}, then $(R, \mathcal{S})$ is not in α_i-RFNF with respect to Σ. According to our hypothesis, there is some relation r_i over R that satisfies Σ_{α_i}, and some $t \in r_i$ and $A \in R$ such that $t(A)$ is redundant with respect to Σ_{α_i}. In particular, r_i must contain some tuple $t_1 \neq t$. We now extend the relation r_i to a possibilistic relation $(r, Poss_r)$ by defining $Poss_r(t_1) = \alpha_1$ and $Poss_r(t') = \alpha_i$ for all $t' \in r_i - \{t_1\}$. We show that r satisfies Σ. If $(\sigma, \beta) \in \Sigma$ is in Σ_{α_i}, then $C_r(\sigma) = \beta_1 \geq \beta$. If $(\sigma, \beta) \notin \Sigma_{\alpha_i}$, then $\beta < \beta_{k+1-i}$. If r_i satisfies σ, then $C_{(r,Poss_r)}(\sigma) = \beta_1 \geq \beta$. If r_i violates σ, then $C_{(r,Poss_r)}(\sigma) = \beta_{k+2-i} \geq \beta$. As $t(A)$ is α_i-redundant we have shown that $(R, \mathcal{S})$ is not in α_i-RFNF with respect to Σ.

Theorem 12 (Theorem 4 restated) *$(R, \mathcal{S})$ is in α_i-RFNF for Σ if and only if R is in RFNF for Σ_{α_i}.*

Proof We show first the following: if $(R, \mathcal{S})$ is not in α_i-RFNF for Σ, then R is in not in RFNF for Σ_{α_i}. According to our hypothesis, there is some p-relation $(r, Poss_r)$ over $(R, \mathcal{S})$ that satisfies Σ, an attribute $A \in R$, and a tuple $t \in r_i$ such that $t(A)$ is redundant for Σ_{α_i}. In particular, it follows that r_i satisfies Σ_{α_i} since r satisfies Σ. Hence, R is not in RFNF for Σ_{α_i}.

We now show: if R is in not in RFNF for Σ_{α_i}, then $(R, \mathcal{S})$ is not in α_i-RFNF for Σ. According to our hypothesis, there is some relation r_i over R that satisfies Σ_{α_i}, and some $t \in r_i$ and $A \in R$ such that $t(A)$ is redundant for Σ_{α_i}. In particular, r_i must contain some tuple $t_1 \neq t$. We now extend the relation r_i to a p-relation $(r, Poss_r)$ by defining $Poss_r(t_1) = \alpha_1$ and $Poss_r(t') = \alpha_i$ for all $t' \in r_i - \{t_1\}$. We show that r satisfies Σ. If $(\sigma, \beta) \in \Sigma$ is in Σ_{α_i}, then $C_r(\sigma) = \beta_1 \geq \beta$. If $(\sigma, \beta) \notin \Sigma_{\alpha_i}$, then $\beta < \beta_{k+1-i}$. If r_i satisfies σ, then $C_{(r,Poss_r)}(\sigma) = \beta_1 \geq \beta$. If r_i violates σ, then $C_{(r,Poss_r)}(\sigma) = \beta_{k+2-i} \geq \beta$. As $t(A)$ is α_i-redundant we have shown that $(R, \mathcal{S})$ is not in α_i-RFNF for Σ.

Theorem 13 (Theorem 5 restated) *$(R, \mathcal{S})$ is in β-4NF for Σ if and only if R is in 4NF for Σ_β.*

Proof By definition, $(R, \mathcal{S})$ is in β-4NF for Σ if and only if for every pMVD $(X \twoheadrightarrow Y, \beta) \in \Sigma^+_{\mathfrak{P}}$ and $Y \not\subseteq X$ and $Y \neq R$ we have $(X \to R, \beta) \in \Sigma^+_{\mathfrak{P}}$. Due to Theorem 2 the latter condition is equivalent to saying that for every pMVD $(X \twoheadrightarrow Y, \beta)$ that is implied by Σ and $Y \not\subseteq X$ and $Y \neq R$, the pFD $(X \to R, \beta)$ is implied by Σ, too. According to Theorem 1, this statement is equivalent to saying that for every MVD $X \to Y$ that is implied by Σ_β and $Y \not\subseteq X$ and $Y \neq R$, the FD $X \to R$ is implied by Σ_β, too. Due to the completeness of the inference rules for the implication of FDs and MVDs, this statement is equivalent to saying that for every MVD $X \twoheadrightarrow Y \in (\Sigma_\beta)^+_{\mathfrak{P}'}$ where $Y \not\subseteq X$ and $Y \neq R$ we have $X \to R \in (\Sigma_\beta)^+_{\mathfrak{P}'}$, too. The latter statement, however, is equivalent to saying that R is in 4NF for Σ_β.

Theorem 14 (Theorem 6 restated) *For all $i = 1, \ldots, k$, $(R, \mathcal{S})$ with $|\mathcal{S}| = k + 1$ is in α_i-RFNF for Σ if and only if $(R, \mathcal{S})$ is in β_{k+1-i}-4NF for Σ.*

Proof By Theorem 4, $(R, \mathcal{S})$ is in α_i-RFNF for Σ if and only if R is in RFNF for Σ_{α_i}. Since $\Sigma_{\alpha_i} = \Sigma_{\beta_{k+1-i}}$, and since RFNF and 4NF coincide in relational databases, the latter statement is equivalent to saying that R is in 4NF for $\Sigma_{\beta_{k+1-i}}$. However, the last statement is equivalent to saying that $(R, \mathcal{S})$ is in β_{k+1-i}-4NF for Σ, by Theorem 5.

Theorem 15 (Theorem 7 restated) *A p-schema $(R, \mathcal{S})$ is in β-4NF for a set Σ of pFDs and pMVDs over $(R, \mathcal{S})$ if and only if for every pFD $(X \to Y, \beta') \in \Sigma$ where $\beta' \geq \beta$ and $Y \not\subseteq X$ and for every pMVD $(X \twoheadrightarrow Y, \beta') \in \Sigma$ where $\beta' \geq \beta$ and $Y \not\subseteq X$ and $Y \neq R$, we have $(X \to R, \beta) \in \Sigma^+_{\mathfrak{P}}$.*

Proof By Theorem 5, $(R, \mathcal{S})$ is in *β-4NF* for Σ if and only if R is in 4NF for Σ_β. However, the latter condition is well-known to be equivalent to checking for every FD

$X \to Y \in \Sigma_\beta$ where $Y \not\subseteq X$ and every MVD $X \twoheadrightarrow Y \in \Sigma_\beta$ where $Y \not\subseteq X$ and $Y \neq R$, that $X \to R \in (\Sigma_\beta)^+_{\mathfrak{P}'}$ [25]. The condition that $\sigma \in \Sigma_\beta$ is equivalent to saying that $(\sigma, \beta') \in \Sigma$ for $\beta' \geq \beta$. Lastly, the condition $X \to R \in (\Sigma_\beta)^+_{\mathfrak{A}}$ is equivalent to saying that $(X \to R, \beta') \in \Sigma^+_{\mathfrak{P}}$ for some $\beta' \geq \beta$, according to Theorem 2 and Theorem 1.