Completeness in Databases with Maybe-Tuples

Fabian Panse and Norbert Ritter

University of Hamburg, Vogt-Kölln Straße 33, 22527 Hamburg, Germany
{panse,ritter}@informatik.uni-hamburg.de
http://vsis-www.informatik.uni-hamburg.de/

Abstract. Some data models use so-called *maybe tuples* to express the uncertainty, whether or not a tuple belongs to a relation. In order to assess this relation's quality the corresponding vagueness needs to be taken into account. Current metrics of quality dimensions are not designed to deal with this uncertainty and therefore need to be adapted. One major quality dimension is data completeness. In general, there are two basic ways to distinguish *maybe tuples* from *definite tuples*. First, an attribute serving as a maybe indicator (values YES or NO) can be used. Second, tuple probabilities can be specified. In this paper, the notion of data completeness is redefined w.r.t. both concepts. Thus, a more precise estimating of data quality in databases with *maybe tuples* (e.g. probabilistic databases) is enabled.

Keywords: data completeness, maybe tuple, probabilistic database.

1 Introduction

Since in databases using the three-valued logic uncertain query results can appear (e.g. resulting from operations on null values), in some cases, it is not exactly known whether a tuple belongs to a query result set or not. For indicating *possible* result tuples several data models ([1], [2] et al.) use the concept of *maybe tuples*. Additionally, as a consequence of a poor information elicitation, sometimes it is not clear, whether a tuple belongs to a database relation or not. For modeling these cases *maybe tuples* can be used, too. Besides a simple indication of *maybe tuples* a more exact specification by individual tuple probabilities as it is known from probabilistic databases (e.g. a tuple belongs to a relation with a certainty of 70 percent) is possible ([9], [3] et al.). Altogether, both types of models enable the indication of tuples which may belong to a relation with less confidence.

For estimating a database's quality, for example in order to compare different databases containing information on the same issue, in the last years various data quality dimensions have been defined. Current metrics of these dimensions do not consider the uncertainty represented by *maybe tuples*. Thus, for corresponding databases some of these metrics are insufficient. Since, data completeness is one of the relevant quality dimensions, in this paper new completeness metrics with respect to the *maybe tuple* concept are defined.

Generally, we consider completeness from a theoretical point of view and try to define it as precise and exact as possible. In reality, often some required

C.A. Heuser and G. Pernul (Eds.): ER 2009 Workshops, LNCS 5833, pp. 202–211, 2009.

information are not available and more approximate and hence more imprecise methods have to be used. Since, such a practical point of view is out of the scope of this paper, it will be considered in future work.

The paper is structured as follows: In Section 2 related work is examined. Furthermore, we discuss and correct deficiencies of current data completeness metrics w.r.t. relations without *maybe tuples*. After presenting relations with *maybe tuples* in more detail (Section 3), we introduce three approaches for extending the corrected metrics to relations with *maybe tuples* (for simple maybe indications as well as for individual tuple probabilities) in Section 4. A final comparison relates these metrics to each other and points out the most suitable one. Section 5 summarizes the paper and gives an outlook to future work.

2 Related Work

Metrics of data completeness are considered in different works (Scannapieco ([7]), Naumann ([6]), Motro ([5]) et al.), but none of them regards the uncertainty resulting from *maybe tuples*. In [6], data completeness is composed by the two measures data coverage and data density[1]. Data coverage represents the completeness of the extension and is the ratio of all stored to all actually existing entities of the modeled world. Therefore w.r.t. a single relation $\mathcal{R}$, the coverage $c(\mathcal{R})$ is the ratio of all tuples of this relation to the number of entities of the corresponding entity type $\mathcal{E}$ (equation 1). Data density represents the completeness of the stored entities (intension) and can be considered at different levels of granularity (e.g. attribute value, tuple, relation). The density of an attribute value measures the information content of this value with respect to its maximal potential information content. In existing approaches (e.g. [6]) this density is either 1 if the value is specified, or 0 if it is a null value, but another value density is possible if partial information is respected ($\Rightarrow d(t) \in [0,1]$). The density $d(t)$ of a tuple t is the average of its values' densities and the density $d(\mathcal{R})$ of a relation $\mathcal{R}$ which, in turn, is the average of its tuples' densities (equation 2).

$$c(\mathcal{R}) = \frac{|\mathcal{R}|}{|\mathcal{E}|} \qquad (1) \qquad\qquad d(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}} d(t)}{|\mathcal{R}|} \qquad (2)$$

Using these two measures, the data completeness of $\mathcal{R}$ results in:

$$comp(\mathcal{R}) = c(\mathcal{R}) \cdot d(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}} d(t)}{|\mathcal{E}|} \qquad (3)$$

2.1 Metric Deficiencies

As we show by the following example, the metrics given above (equations 1-3) are deficient for relations containing tuples which do not represent an entity of the corresponding entity type: A company is assumed to have 10 employees currently.

[1] Since this decomposition increases the interpretability of completeness, we adapt the metrics defined by Naumann in the following.

Thus, a relation *employee* contains one tuple for each of them. Additionally, the relation contains a tuple for an employee who was fired last month. Resulting from a failure of the responsible secretary, the tuple has not been deleted by now. Calculating the coverage of *employee* by equation 1, $c(emplyoee) = 11/10 = 1.1$ results. Usually, quality metrics are normalized and hence a quality value has always to be within the range $[0, 1]$. Since normalization is one of the most important requirements for an adequate quality metric ([4]), this is a deficiency which must not be underrated.

Furthermore, if completeness is used to compare two or more data sources an unsound source[2] can mistakenly be regarded as the best source. For avoiding such errors only the tuples which correctly belong to the relation have to be considered (see [5]). Given $\mathcal{R}$ is the regarded relation, and $\mathcal{E}$ is the entity type which is represented by this relation, and $m : \mathcal{E} \rightarrow \mathcal{R}$ is the mapping of the entities (extension) of $\mathcal{E}$ on tuples of $\mathcal{R}$, the relation $\mathcal{R}_C(\mathcal{E})$ (short $\mathcal{R}_\mathcal{C}$) contains all tuples which correctly belong to $\mathcal{R}$ w.r.t. the entity type $\mathcal{E}$.

$$\mathcal{R}_C(\mathcal{E}) = \{t \mid t \in \mathcal{R} \wedge (\exists e \in \mathcal{E}) : m(e) = t\}$$

Considering this 'tuple cleaning', the metrics of data coverage $c(\mathcal{R})$ and data density $d(\mathcal{R})$ have to be adapted to:

$$c(\mathcal{R}) = \frac{|\mathcal{R}_C|}{|\mathcal{E}|} \qquad (4) \qquad\qquad d(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}_C} d(t)}{|\mathcal{R}_C|} \qquad (5)$$

The metric of data completeness (equation 3) has to be adapted accordingly.

3 Relations with Maybe-Tuples

In contrast to *definite tuples*, as the name already says, *maybe tuples* are tuples for which it is undefined whether they belong to the associated relation or not. *Maybe tuples* can appear in database relations as well as in (intermediate) query result sets. The appearance in database relations can be traced back to a poor information elicitation. Sometimes from the available information it cannot be certainly concluded whether an entity is part of the extension of an entity type or not. As a consequence, for representing this uncertainty, the associated tuple can neither be exlcuded from nor included into the corresponding database relation. Thus, these tuples have to be indicated as 'maybe' (see attribute M of relation $\mathcal{R}_2$ in Figure 1). In addition, if a database contains null values or values which represent partial information (e.g. interval values), during query evaluation some tuples cannot be evaluated to TRUE or FALSE. In such cases, it cannot be determined, whether or not the query condition is satisfied. Thus, these tuples are *possible* query results and have to be indicated as *maybe tuples*, too.

A relation $\mathcal{R}$ with *maybe tuples* (in the following denoted as *maybe relation*) can be lossless divided into two subrelations ($\mathcal{R} = \mathcal{R}^\mathcal{D} \cup \mathcal{R}^\mathcal{M}$): Relation $\mathcal{R}^\mathcal{D}$

[2] A source containing many tuples which do not correctly belong to the corresponding source's relation.

contains all tuples which definitely belong to $\mathcal{R}$ and relation $\mathcal{R}^{\mathcal{M}}$ contains all tuples which may be belong to $\mathcal{R}$. If $\mathcal{R}$ does not contain duplicates (e.g. if $\mathcal{R}$ is a database relation), the two subsets have to be disjunct ($\mathcal{R}^{\mathcal{D}} \cap \mathcal{R}^{\mathcal{M}} = \emptyset$).

The individual tuple probability $p(t)_{\mathcal{R}}$ for a tuple t of the relation $\mathcal{R}$ is defined as the probability that this tuple belongs to the associated relation. Since all tuples of the subrelation $\mathcal{R}^{\mathcal{D}}$ are definitely in $\mathcal{R}$, the individual tuple probabilities of these tuples always have to be 1. Since every *maybe tuple* only possibly belongs to the relation, its individual tuple probability has to be lower than 1. However, because these tuples can certainly not be excluded from this relation, the individual tuple probability has to be within the range $]0, 1[$.

In the following, $\mathcal{S}(\mathcal{R})$ represents the set of all possible instances and $\mathcal{R}'$ represents the real instance of the relation $\mathcal{R}$ under a closed world assumption[3]. Since all tuples of $\mathcal{R}^{\mathcal{D}}$ definitely belong to $\mathcal{R}$, each possible instance of $\mathcal{R}$ contains these tuples. In general, for every possible combination of the *maybe tuples* (the power set ($\mathcal{P}(\mathcal{R}^{\mathcal{M}})$)) one possible instance of $\mathcal{R}$ results:

$$\mathcal{S}(\mathcal{R}) = \{\mathcal{R}^{\mathcal{D}} \cup M \mid M \in \mathcal{P}(\mathcal{R}^{\mathcal{M}})\} \tag{6}$$

If $\mathcal{R}$ does not contain *maybe tuples*, all tuples of $\mathcal{R}$ are known and the real set of tuples belonging to $\mathcal{R}$ is completely described by $\mathcal{R}$ itself. As a consequence, $\mathcal{S}(\mathcal{R})$ contains just one element and the relations $\mathcal{R}$ and $\mathcal{R}'$ are equal. If, in contrast, $\mathcal{R}$ contains *maybe tuples*, the set of tuples which really belong to $\mathcal{R}$ and hence the relation $\mathcal{R}'$ are not completely known. This uncertainty can be represented by a discrete probability distribution of $\mathcal{R}'$ on the set $\mathcal{S}(\mathcal{R})$. For example, we assume a relation $\mathcal{R}$ containing one *definite tuple* t_1 and one *maybe tuple* t_2 ($p(t_2)_{\mathcal{R}} = 0.6$). The set of all possible instances is $\mathcal{S}(\mathcal{R}) = \{S_0 = \{t_1\}, S_1 = \{t_1, t_2\}\}$ and the real instance $\mathcal{R}'$ is distributed over $\mathcal{S}(\mathcal{R})$ with the probability distribution $P(\mathcal{R}' = S_0) = 0.4$ and $P(\mathcal{R}' = S_1) = 0.6$.

4 Data Completeness Regarding Maybe-Tuples

Since a *maybe tuple* only possibly belongs to a relation, for measuring data completeness this imprecision has to be taken into account. In order to demonstrate this necessity, we consider the three relations $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$ as illustrated in Figure 1. $\mathcal{R}_1$ and $\mathcal{R}_3$ are relations without *maybe tuples* containing 2 or 3 tuples respectively. Relation $\mathcal{R}_2$ contains two *definite* (the same tuples as $\mathcal{R}_1$) and one *maybe tuple*. It is obvious that the completeness of $\mathcal{R}_2$ has to be greater than the completeness of $\mathcal{R}_1$. The uncertain membership of t_3 to $\mathcal{R}_2$ is also a kind of incomplete information. Since this incompleteness can influence the output of a quality driven query answering, it is also comprehensible that the completeness of $\mathcal{R}_2$ has to be smaller than the completeness of $\mathcal{R}_3$. As a consequence, the completeness of $\mathcal{R}_2$ can be limited to $comp(\mathcal{R}_1) < comp(\mathcal{R}_2) < comp(\mathcal{R}_3)$.

[3] Totally missing tuples are ignored and uncertain memberships of *maybe tuples* are the only incomplete information. Thus, w.r.t. the calculation of all possible instances only the tuples of $\mathcal{R}^{\mathcal{D}}$ and $\mathcal{R}^{\mathcal{M}}$ are considered.

	firstname	surname
t_1	Georg	Washington
t_2	Abraham	Lincoln

$\mathcal{R}_1$

	firstname	surname	M
t_1	Georg	Washington	NO
t_2	Abraham	Lincoln	NO
t_3	Theodor	Roosevelt	YES

$\mathcal{R}_2$

	firstname	surname
t_1	Georg	Washington
t_2	Abraham	Lincoln
t_3	Theodor	Roosevelt

$\mathcal{R}_3$

Fig. 1. Completeness classification of *maybe relations*

In order to calculate an exact value for the completeness of a *maybe relation*, we introduce three different but each intuitive approaches. The first one uses the average completeness of the subrelations which can result from a so-called α-selection, the second one is based on the expectation value of the completeness of the relation's real instance, and the last one considers the uncertainty of *maybe tuples* as a lower priority. Partially, we trace our new metrics to the current ones. For distinction, the newly defined metrics of completeness, coverage and density with respect to a relation $\mathcal{R}$ and an approach Ai are denoted as $comp'_{Ai}(\mathcal{R})$, $c'_{Ai}(\mathcal{R})$ and $d'_{Ai}(\mathcal{R})$.

4.1 Approach 1 (α-Selection)

The first approach is based on the α-selection introduced by Tseng ([9]). An α-selection ($\hat{\sigma}^{\alpha}(\mathcal{R})$) selects each tuple $t \in \mathcal{R}$ which belongs to $\mathcal{R}$ with a probability $p(t)_{\mathcal{R}}$ greater or equal than $\alpha \in [0, 1]$:

$$\hat{\sigma}^{\alpha}(\mathcal{R}) = \{t \mid t \in \mathcal{R} \wedge p(t)_{\mathcal{R}} \geq \alpha\} \tag{7}$$

If an α-selection is used for a probability based tuple filtering, the completeness of the resulting subrelation depends on the value α. Since the higher α the more tuples are filtered, the completeness $comp(\hat{\sigma}^{\alpha}(\mathcal{R}))$ is monotonically decreasing (see Figure 2). Additionally, the completeness of a filtered relation $\hat{\sigma}^{\alpha}(\mathcal{R})$ is always greater or equal than the completeness of $\mathcal{R}^{\mathcal{D}}$ and always smaller or equal than the completeness of $\mathcal{R}$ if maybe indications are ignored ($\alpha = 0$).

One intuitive possibility is to esteem the completeness of a *maybe relation* $\mathcal{R}$ as the average completeness of the subrelations resulting from all possible α-selections on $\mathcal{R}$.

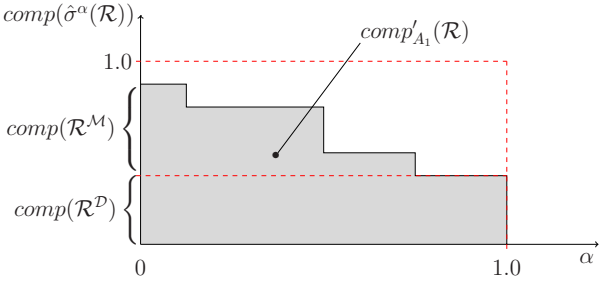

Fig. 2. Completeness of a *maybe relation* $\mathcal{R}$ w.r.t. all possible α-selections

Individual Tuple Probability: If individual tuple probabilities are given, for each α another subrelation can result from applying an α-selection. Thus, α has to be considered within the continuous range $[0, 1]$ and the completeness $comp'_{A1}(\mathcal{R})$ can be defined as the integral of $comp(\hat{\sigma}^{\alpha}(\mathcal{R}))$ over α (see gray area in Figure 2):

$$comp'_{A1}(\mathcal{R}) = \int_0^1 comp(\hat{\sigma}^{\alpha}(\mathcal{R}))d\alpha \tag{8}$$

Since the coverage and the density of each subrelation are not independent of each other, by using this approach a decomposition into these two measures is not possible:

$$\int_0^1 c(\hat{\sigma}^{\alpha}(\mathcal{R})) \cdot d(\hat{\sigma}^{\alpha}(\mathcal{R}))d\alpha \neq \int_0^1 c(\hat{\sigma}^{\alpha}(\mathcal{R}))d\alpha \cdot \int_0^1 d(\hat{\sigma}^{\alpha}(\mathcal{R}))d\alpha \tag{9}$$

Simple Maybe Indication: Intuitively, in the simple case, the tuple probability of each *maybe tuple* is assumed to be 0.5. Therefore, from applying α-selections only two subrelations can result: the subrelation $\mathcal{R}^{\mathcal{D}}$, if α is within the range $]0.5, 1]$, and the whole relation $\mathcal{R} = \{\mathcal{R}^{\mathcal{D}} \cup \mathcal{R}^{\mathcal{M}}\}$ otherwise. Consequently, the completeness $comp'_{A1}(\mathcal{R})$ defined in equation 8 can be simplified to:

$$\begin{aligned} comp'_{A1}(\mathcal{R}) &= \int_0^{0.5} comp(\{\mathcal{R}^{\mathcal{D}} \cup \mathcal{R}^{\mathcal{M}}\})d\alpha + \int_{0.5}^1 comp(\mathcal{R}^{\mathcal{D}})d\alpha \\ &= comp(\mathcal{R}^{\mathcal{D}}) + \frac{1}{2} comp(\mathcal{R}^{\mathcal{M}}) \end{aligned} \tag{10}$$

4.2 Approach 2 (Expectation Value)

Another illustrative way is to calculate the completeness of $\mathcal{R}$ by using the expectation value of the completeness of $\mathcal{R}'$. As for approach 1, a decomposition of completeness into coverage and density is not possible:

$$comp'_{A3}(\mathcal{R}) = E(comp(\mathcal{R}')) = E(c(\mathcal{R}') \cdot d(\mathcal{R}')) \neq E(c(\mathcal{R}')) \cdot E(d(\mathcal{R}'))$$

Individual Tuple Probability: Defining the completeness of $\mathcal{R}$ as the expectation value of $comp(\mathcal{R}')$, the completeness[4] and probability for every possible instance of $\mathcal{R}$ have to be known.

$$\begin{aligned} E(comp(\mathcal{R}')) &= \sum_{S_i \in \mathcal{S}(\mathcal{R}_{\mathcal{C}})} P(\mathcal{R}'_{\mathcal{C}} = S_i)\, comp(S_i) \\ &= \frac{1}{|\mathcal{E}|} \sum_{S_i \in \mathcal{S}(\mathcal{R}_{\mathcal{C}})} P(\mathcal{R}'_{\mathcal{C}} = S_i) \sum_{t \in S_i} d(t) \end{aligned} \tag{11}$$

[4] Since every possible instance S_i has to be handled as a relation without *maybe tuples*, for calculating completeness the metric $comp(S_i)$ can be used.

The probability of a possible instance $S_i \in \mathcal{S}(\mathcal{R}_\mathcal{C})$ results from the product of the tuple probabilities of all tuples in S_i and the inverse probabilities of all tuples of $\mathcal{R}_\mathcal{C}$ not in S_i.

$$P(\mathcal{R}'_\mathcal{C} = S_i) = \prod_{t \in S_i} p(t)_\mathcal{R} \prod_{t \in \{\mathcal{R}_\mathcal{C} \setminus S_i\}} (1 - p(t)_\mathcal{R})$$

Simple Maybe Indication: In the simple case, the possible instances are uniformly distributed. Thus, there exist $|\mathcal{S}(\mathcal{R}_\mathcal{C})| = |\mathcal{P}(\mathcal{R}^\mathcal{M})| = 2^{|\mathcal{R}^\mathcal{M}_\mathcal{C}|}$ possible instances, and the expectation value $E(comp(\mathcal{R}'))$ and hence the completeness $comp'_{A3}(\mathcal{R})$ defined in equation 11 can be simplified to:

$$comp'_{A3}(\mathcal{R}) = E(comp(\mathcal{R}')) = \frac{1}{2^{|\mathcal{R}^\mathcal{M}_\mathcal{C}|}} \frac{1}{|\mathcal{E}|} \sum_{S_i \in \mathcal{S}(\mathcal{R}_\mathcal{C})} \sum_{t \in S_i} d(t) \tag{12}$$

4.3 Approach 3 (Tuple Priorities)

In the third approach, the uncertainty resulting from *maybe tuples* is expressed by specifying lower priorities for subrelation $\mathcal{R}^\mathcal{M}$ (simple case) or each individual *maybe tuple* (exact case) respectively. In contrast to the first two approaches, completeness here can be decomposed into coverage and density. In the case of a simple maybe indication, for $\mathcal{R}$ the new metrics $comp'_{A2}(\mathcal{R})$, $c'_{A2}(\mathcal{R})$ and $d'_{A2}(\mathcal{R})$ can be traced back to $comp(\mathcal{R}^\mathcal{D})$ and $comp(\mathcal{R}^\mathcal{M})$, $c(\mathcal{R}^\mathcal{D})$ and $c(\mathcal{R}^\mathcal{M})$ or $d(\mathcal{R}^\mathcal{D})$ and $d(\mathcal{R}^\mathcal{M})$, respectively. In the exact case such a derivation is not possible. Thus, the metrics $comp'_{A2}(\mathcal{R})$, $c'_{A2}(\mathcal{R})$ and $d'_{A2}(\mathcal{R})$ have to be newly defined by regarding the individual tuple probabilities.

Simple Maybe Indication: Both, the subrelation $\mathcal{R}^D$ as well as the subrelation $\mathcal{R}^M$ cover parts of the corresponding entity type's extension. As a consequence, the coverage $c'_{A2}(\mathcal{R})$ of a relation $\mathcal{R}$ can be calculated from the coverages of these two subrelations. Assuming the probability that a *maybe tuple* belongs to a relation is equal to the probability that the *maybe tuple* does not belong to this relation, the coverage of the subrelation $\mathcal{R}^M$ is taken into account with a priority which is half as high as the priority of the coverage $c(\mathcal{R}^\mathcal{D})$:

$$c'_{A2}(\mathcal{R}) = c(\mathcal{R}^\mathcal{D}) + \frac{1}{2} c(\mathcal{R}^\mathcal{M}) = \frac{|\mathcal{R}^\mathcal{D}_\mathcal{C}| + \frac{1}{2}|\mathcal{R}^\mathcal{M}_\mathcal{C}|}{|\mathcal{E}|} \tag{13}$$

The densities of $\mathcal{R}^\mathcal{D}$ and $\mathcal{R}^\mathcal{M}$ are the averages of their tuples' densities (equation 5). As with the coverage, the effect of the density $d(\mathcal{R}^\mathcal{M})$ on $d'_{A2}(\mathcal{R})$ is only half as high as the effect of the *definite tuples*' densities. Since the two densities $d(\mathcal{R}^\mathcal{D})$ and $d(\mathcal{R}^\mathcal{M})$ are only relative, for the total density both have to be correlated by taking into account the associated relation's size:

$$d'_{A2}(\mathcal{R}) = \frac{|\mathcal{R}^\mathcal{D}_\mathcal{C}| d(\mathcal{R}^\mathcal{D}) + \frac{1}{2}|\mathcal{R}^\mathcal{M}_\mathcal{C}| d(\mathcal{R}^\mathcal{M})}{|\mathcal{R}^\mathcal{D}_\mathcal{C}| + \frac{1}{2}|\mathcal{R}^\mathcal{M}_\mathcal{C}|} = \frac{\sum_{t \in \mathcal{R}^\mathcal{D}_\mathcal{C}} d(t) + \frac{1}{2} \sum_{t \in \mathcal{R}^\mathcal{M}_\mathcal{C}} d(t)}{|\mathcal{R}^\mathcal{D}_\mathcal{C}| + \frac{1}{2}|\mathcal{R}^\mathcal{M}_\mathcal{C}|} \tag{14}$$

As for approach 1, the completeness $comp'_{A2}(\mathcal{R}) = c'_{A2}(\mathcal{R}) \cdot d'_{A2}(\mathcal{R})$ results in:

$$comp'_{A2}(\mathcal{R}) = comp(\mathcal{R}^{\mathcal{D}}) + \frac{1}{2} comp(\mathcal{R}^{\mathcal{M}}) \tag{15}$$

Individual Tuple Probability: If the data model supports individual tuple probabilities instead of one global maybe priority, each tuple has a different impact on the coverage and the density of $\mathcal{R}$. Considering the individual tuple probability as the degree of this impact, coverage and density are defined as:

$$c'_{A2}(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}_C} p(t)_{\mathcal{R}}}{|\mathcal{E}|} \tag{16} \qquad d'_{A2}(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}_C} p(t)_{\mathcal{R}} \cdot d(t)}{\sum_{t \in \mathcal{R}_C} p(t)_{\mathcal{R}}} \tag{17}$$

Thus, the completeness $comp'_{A2}(\mathcal{R}) = c'_{A2}(\mathcal{R}) \cdot d'_{A2}(\mathcal{R})$ results in:

$$comp'_{A2}(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}_C} p(t)_{\mathcal{R}} \cdot d(t)}{|\mathcal{E}|} \tag{18}$$

4.4 Correlated Tuples

As in most works on *maybe relations*, dependencies between tuples have not been addressed so far. Since in reality data is often correlated, a complete independence among tuples is a simplistic assumption which distorts the representation of the modeled world. Therefore, in some newer proposals ([8] et al.) probabilistic data models are extended by representing such dependencies. Since tuple dependencies restrict the set of all possible instances of a relation $\mathcal{R}$, these dependencies are completely represented by the set $\mathcal{S}(\mathcal{R})$. For example, relation $\mathcal{R}$ contains one *definite* tuple t_1 and two *maybe tuples* t_2 and t_3. A tuple dependency defines that either both *maybe tuples* belong to $\mathcal{R}$ or none of them. As a consequence, instead of four possible instances $S(\mathcal{R}) = \{\{t_1\}, \{t_1, t_2\}, \{t_1, t_3\}, \{t_1, t_2, t_3\}\}$ only two possible instances $S(\mathcal{R}) = \{\{t_1\}, \{t_1, t_2, t_3\}\}$ exist. Thus, it is obvious, that our completeness metrics which are based on the expectation value of $\mathcal{R}'$ can be used in models with tuple dependencies without any adaption.

In general, the total probability of each tuple t is (independent of correlations) always $p(t)_{\mathcal{R}}$. Thus, it does not matter in which way this probability is distributed on the possible instances. As a consequence, the completeness of a *maybe relation* is generally independent from tuple correlations and the metrics of the other two approaches do not need to be adapted to such cases, too.

4.5 Comparison of Proposed Approaches

In the approaches outlined above, we defined metrics for calculating completeness of *maybe relations*. The next step is to compare these metrics to each other and try to determine which of them is most suitable. In general, all these completeness metrics supply the same results whether tuple correlations exist or not.

This fact enhances the certainty that the resulting value is actually an adequate representation of the completeness of the considered *maybe relation*.

Regarding the requirements proposed by Heinrich ([4]) the metrics of all approaches satisfy the requirements of normalization, interval scale and adaptivity. Furthermore, the input parameters and hence the feasability of all approaches are equal. Thus, the most severe differences w.r.t. these requirements are related to the interpretability. The first approach is most suitable for illustrating the completeness of a *maybe relation*, for example on the basis of graphics as seen in Figure 2 (property $\mathcal{A}$). In contrast, the two other approaches are more abstract. The benefit of the third approach is its simplicity (see complexity below), but from a probabilistic theory point of view the concept of the second one is still more apposite (property $\mathcal{B}$). However, in contrast to the other two approaches, the third one enables a decomposition of completeness into coverage and density (property $\mathcal{C}$), which in turn improves its interpretability. Additionally, both completeness metrics of approach 3, for the simple maybe indication as well as for an indication by individual tuple probabilities are comprehensible in an easy way (property $\mathcal{D}$). In the second approach the metric for a simple maybe indication can only be derived from those of the exact case by a substitution of the value 0.5 for every tuple probability. Hence approach 2 has a poor interpretability.

Another important factor is the complexity of the individual metrics. Given a relation $\mathcal{R}$ with n *definite* and m *maybe tuples*, w.r.t. the simple case, the complexity of all metrics is equal ($\mathcal{O}(n+m)$). In the exact case, at the worst in approach 1 each *maybe tuple* has another probability and the completeness of $m+1$ subrelations have to be calculated ($\mathcal{O}(max(m^2, nm))$). In approach 2 the completeness of 2^m possible instances is required if there are no tuple correlations ($\mathcal{O}(2^m(n+m))$). In approach 3 only the completeness of a single relation is needed ($\mathcal{O}(n+m)$). The complexities w.r.t. both cases (simple and exact) as well as the mentioned benefits and drawbacks of all approaches with respect to the interpretability are summarized in the following table:

	property $\mathcal{A}$	property $\mathcal{B}$	property $\mathcal{C}$	property $\mathcal{D}$	complexity: simple case	complexity: exact case
Approach 1:	+	∘	-	∘	$\mathcal{O}(n+m)$	$\mathcal{O}(max(m^2, nm))$
Approach 2:	∘	+	-	-	$\mathcal{O}(n+m)$	$\mathcal{O}(2^m(n+m))$
Approach 3:	∘	∘	+	+	$\mathcal{O}(n+m)$	$\mathcal{O}(n+m)$

Regarding its minor complexity, in databases with individual tuple probabilities, the metric of approach 3 is most suitable. At a first sight (without considerations on implementation- or application domain specific details), in databases with just a simple maybe indication all metrics can be assumed to be equivalently suitable.

5 Conclusion

Since current metrics of data completeness are not usable for estimating the completeness of *maybe relations*, we have used the metric defined by Naumann and

extended it for handling the vagueness resulting from the *maybe tuple* concept. Further, we have identified two cases. In the first case, *maybe tuples* are only indicated as 'maybe'. In the second, more exact case, every tuple is indicated by a probability of its own.

We have considered completeness from three different perspectives and have therefore introduced three corresponding approaches in order to measure this quality dimension. The resulting metrics supply the same results whether or not tuple correlations exist. In general, even though all resulting completeness values are an adequate representation of this quality dimension, each of the three approaches (and hence each of the corresponding metrics) has its benefits as well as its drawbacks. In contrast to the other two approaches, the approach based on tuple priorities enables a decomposition of completeness into coverage and density, which in turn increases the interpretability of the resulting values. Furthermore, its completeness metrics have by far the lowest complexity. Thus, we favor the usage of the metrics resulting from this approach.

So far, we have considered completeness only from a theoretical point of view. In reality such an exact calculation is often impossible because important information (e.g. $|\mathcal{E}|$) is missing. Thus, in future work these approaches have to be considered from a more practical (and hence vaguer) point of view, too.

Besides completeness, other quality dimensions are influenced by the possibility of *maybe tuples*. Especially quality dimensions for which the quality of a relation is derived from the qualities of its tuples (e.g. accuracy, currency) are affected. As for completeness, the *maybe tuples* have to be considered with a minor emphasis. The lower the probability of a tuple, the lower the influence of this tuple on the quality of the associated relation has to be.

References

1. Biskup, J.: Extending the Relational Algebra for Relations with Maybe Tuples and Existential and Universal Null Values. Fundam. Inform. 7(1), 129–150 (1984)
2. DeMichiel, L.G.: Resolving Database Incompatibility: An Approach to Performing Relational Operations over Mismatched Domains. IEEE Trans. Knowl. Data Eng. 1(4), 485–493 (1989)
3. Fuhr, N., et al.: A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems. ACM Trans. Inf. Syst. 15(1), 32–66 (1997)
4. Heinrich, B., et al.: Metrics for Measuring Data Quality - Foundations for an Economic Data Quality Management. In: ICSOFT (ISDM/EHST/DC), pp. 87–94 (2007)
5. Motro, A., Rakov, I.: Estimating the Quality of Databases. In: Andreasen, T., Christiansen, H., Larsen, H.L. (eds.) FQAS 1998. LNCS (LNAI), vol. 1495, pp. 298–307. Springer, Heidelberg (1998)
6. Naumann, F., et al.: Completeness of integrated information sources. Inf. Syst. 29(7), 583–615 (2004)
7. Scannapieco, M., Batini, C.: Completeness in the relational model: a comprehensive framework. In: IQ, pp. 333–345 (2004)
8. Sen, P., Deshpande, A.: Representing and Querying Correlated Tuples in Probabilistic Databases. In: ICDE, pp. 596–605 (2007)
9. Tseng, F.S.-C., et al.: Answering Heterogeneous Database Queries with Degrees of Uncertainty. Distributed and Parallel Databases 1(3), 281–302 (1993)