

---

# Relational Data Completeness in the Presence of Maybe-Tuples

**Fabian Panse — Norbert Ritter**

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

---

*RÉSUMÉ.*

*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 score this relation's quality in a meaningful way 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, confidence values can be specified. In this paper, the notion of data completeness is redefined w.r.t. both concepts. Thus, a more precise estimation of quality in databases with maybe tuples (e.g. probabilistic or fuzzy databases) is enabled.*

*MOTS-CLÉS :*

*KEYWORDS: Data completeness, quality composition, maybe-tupel, probabilistic database, fuzzy 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 clear whether or not a tuple must be considered to be part of a query result set. For indicating *possible* result tuples some data models (Biskup, 1984; DeMichiel, 1989) use the concept of *maybe tuples*. Furthermore, 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 confidence values is possible. Such confidence values can be interpreted as the probability by which the corresponding tuples belong to the considered relation (probabilistic databases (Barbará *et al.*, 1992; Tseng *et al.*, 1993)) or as their degree of membership (fuzzy databases (Galindo *et al.*, 2006)), respectively. On the whole, both types of models (with simple maybe indication as well as with exact confidence values) support the handling of tuples which may belong to a relation with less confidence.

For estimating a database's quality or to compare different databases containing information on the same issue in the last years various data quality dimensions have been defined. Since current metrics of these dimensions do not consider the uncertainty represented by *maybe tuples*, frequently these metrics are insufficient. Data completeness is one of the relevant quality dimensions. Therefore, in this paper new metrics for data completeness are defined. We present three different but each intuitive approaches and relate them to each other. We consider completeness from a theoretical point of view and define it as precise and exact as possible. Often some required information are not available and more approximate and hence more imprecise methods have to be used. But such a practical point of view is out of the scope of this paper and will be taken in future work.

The paper is structured as follows : In Section 2 related work is considered. Then, we introduce an initial set of completeness metrics defined for relations without *maybe tuples* in Section 3. After having presented relations with *maybe tuples* for more detail (Section 4), we introduce three approaches for extending our completeness metrics to the *maybe tuple*-concept in Section 5. In Section 6, we consider completeness composition w.r.t. relations with *maybe tuples*. Section 7 summarizes this paper and gives an outlook to future work.

## 2. Related Work

Metrics of data completeness are handled in different works (Motro *et al.*, 1997; Scannapieco *et al.*, 2004; Naumann *et al.*, 2004). None of these works, however, regards the uncertainty resulting from *maybe tuples*. In (Naumann *et al.*, 2004), completeness is considered on an extensional (data coverage) and an intensional (data density) level. Data coverage is the ratio of all stored to all existing entities of the modeled world. Data density represents the completeness of the stored entities and

is the proportion of non-null-values. Scannapieco et al. consider completeness w.r.t. several paradigms as the chosen world assumption (open world vs. closed world) and the absence or presence of null values. In contrast to Naumann’s approach, they consider completeness on multiple levels of granularity, e.g., relation, attribute and tuple. Furthermore, they distinguish between strong completeness (boolean : complete or incomplete) and weak completeness (percentage : different degrees of completeness). Both works consider completeness w.r.t. algebraic operations. Whereas Scannapieco stays within the relational algebra, Naumann introduces new operators for merging relational data and means for predicting the completeness of a merging result. Nevertheless, the metrics resulting from both approaches are similar to a large extent and only differ from the respective point of view.

Whereas the two before mentioned approaches assume sound relations, (Motro *et al.*, 1997) takes unsoundness into account. In general, unsoundness can result by mistake (e.g. an already fired employee has not been deleted from the database) or by the fact that the considered entity type does not completely match the actual scope of the considered relation (e.g. by integrating a relation *citizen* to get information on local *students*). Motro, however, mainly focuses on the proportion of the true extension that appears in the database extension (similar to the coverage). Intensional completeness can only be indirectly measured by so called *decomposed extensions*. Thus, for measuring the completeness of stored entities this approach is not convenient.

### 3. Completeness in Databases without *Maybe Tuples*

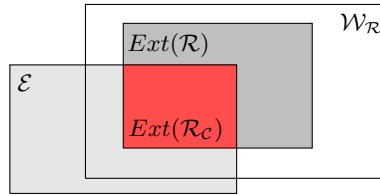
Our approaches for measuring completeness in databases with *maybe tuples* should base on completeness metrics for certain data. Unfortunately, as mentioned above, each set of metrics resulting from any previous work is not ideal. Thus, we initially define some metrics for measuring the completeness in databases without *maybe tuples*. In order to capture the best characteristics of all approaches presented in Section 2, we consider completeness as a mixture of the concepts defined in these works. To measure completeness as precise as possible, we use the open world assumption and take into account that attribute values can be incomplete. As (Scannapieco *et al.*, 2004), we consider completeness on different levels of granularity, where the completeness of one granularity based on the completeness of the granularity underlying beneath. For reasons of interpretability, as (Naumann *et al.*, 2004), we decompose completeness into coverage and density. As (Motro *et al.*, 1997), we take into account that relations can be unsound and hence can contain tuples which do not represent an entity of the considered world.

#### 3.1. Data Relevance

The completeness of data essentially depends on the intended use. A data set can be totally complete w.r.t. one context and totally incomplete w.r.t. another one. Thus, for measuring completeness only tuples and attributes which are relevant for the consid-

red entity type have to be taken into account. Since we consider databases with *maybe tuples*, in this paper we primarily focus on tuple relevance. The relevance of attributes is rudimentarily considered in the metric of tuple density (see Equation 3).

The extension of a relation  $\mathcal{R}$  (denoted as  $Ext(\mathcal{R})$ ) is defined as the set of real-world entities represented by this relation. The set of real-world entities which originally should be stored in  $\mathcal{R}$  (the original intended entity type) is denoted as reference extension  $\mathcal{W}_{\mathcal{R}}$ . The extension of an entity type  $\mathcal{E}$  is generally represented by  $\mathcal{E}$  itself.



**Figure 1.** Relevance of the extension of a relation  $\mathcal{R}$  for an entity type  $\mathcal{E}$

Given the regarded relation  $\mathcal{R}$ , the considered entity type  $\mathcal{E}$  and the mapping  $m : \mathcal{R} \rightarrow Ext(\mathcal{R})$  which maps tuples of  $\mathcal{R}$  on entities of  $Ext(\mathcal{R})$ , the subrelation of  $\mathcal{R}$  only containing all tuples which are relevant for the entity type  $\mathcal{E}$  is denoted as the ‘cleaned’ relation  $\mathcal{R}_C(\mathcal{E})$  (short  $\mathcal{R}_C$ ).

$$\mathcal{R}_C(\mathcal{E}) = \{t \in \mathcal{R} \mid m(t) \in \mathcal{E}\} \quad [1]$$

Thus, the extension of  $\mathcal{R}_C(\mathcal{E})$  is the intersection of  $Ext(\mathcal{R})$  and  $\mathcal{E}$  (see Figure 1) :

$$Ext(\mathcal{R}_C(\mathcal{E})) = Ext(\mathcal{R}) \cap \mathcal{E} \quad [2]$$

If  $\mathcal{R}$  is duplicate free, for every tuple of  $\mathcal{R}$  another real-world entity exists. Thus, the mapping  $m$  is injective and the size of  $\mathcal{R}$  is equal to the size of its extension ( $|\mathcal{R}| = |Ext(\mathcal{R})|$ ). In general, if no contrary information is available, a relation is usually assumed to be duplicate free. For simplification, in the following a relation is directly considered as its extension, if this fact is evident from the given context (e.g. in set-based graphics as Figure 16).

### 3.2. Initial Set of Completeness Metrics

The attribute value level is the lowest level of granularity. In existing approaches the density of an attribute value is either 1, if the value is specified, or 0 if the value is missing (null value). Since another notion of density is possible if partial information (e.g.  $age < 25$ ) is taken into account (Panse, 2009), we consider the density of a value  $v$  as  $d(v) \in [0, 1]$ .

The density of a tuple is defined as the average density of its attribute values relevant for the considered entity type. The density of a relation in turn is defined as the average density of all its relevant tuples. Given a relation  $\mathcal{R}$  having

the attributes  $\mathcal{A} = \{A_1, A_2, \dots, A_k\}$ , and an entity type  $\mathcal{E}$  having the attributes  $\mathcal{A}' = \{A_m, \dots, A_k, \dots, A_n\}$ , the density of each tuple  $t \in \mathcal{R}$  ( $d(t, \mathcal{E})$ ) and the density of  $\mathcal{R}$  itself ( $d(\mathcal{R}, \mathcal{E})$ ) w.r.t. the entity type  $\mathcal{E}$  can be defined as :

$$d(t, \mathcal{E}) = \frac{\sum_{i \in [m, k]} d(t.A_i)}{(n - m) + 1} \quad [3] \quad d(\mathcal{R}, \mathcal{E}) = \frac{\sum_{t \in \mathcal{R}_c} d(t, \mathcal{E})}{|\mathcal{R}_c|} \quad [4]$$

Furthermore, a relation's density can be defined as the average density of its attributes (Equation 6). The density of an attribute  $A$  in turn can be defined as the average density of its attribute values :

$$d(A, \mathcal{E}) = \frac{\sum_{t \in \mathcal{R}_c} d(t.A)}{|\mathcal{R}_c|} \quad [5] \quad d(\mathcal{R}, \mathcal{E}) = \frac{\sum_{i \in [m, k]} d(A_i, \mathcal{E})}{(n - m) + 1} \quad [6]$$

The coverage of a relation is the ratio of the number of its extension's relevant elements to the total number of actually entities of the considered entity type. If  $\mathcal{R}$  is considered to be duplicate free, the coverage of  $\mathcal{R}$  w.r.t.  $\mathcal{E}$  can be defined as :

$$c(\mathcal{R}, \mathcal{E}) = \frac{|Ext(\mathcal{R}_c)|}{|\mathcal{E}|} = \frac{|\mathcal{R}_c|}{|\mathcal{E}|} \quad [7]$$

As mentioned before, the completeness of a relation is composed of its extensional (coverage) and its intensional completeness (density) :

$$comp(\mathcal{R}, \mathcal{E}) = c(\mathcal{R}, \mathcal{E}) \cdot d(\mathcal{R}, \mathcal{E}) = \frac{\sum_{t \in \mathcal{R}_c} d(t, \mathcal{E})}{|\mathcal{E}|} \quad [8]$$

For simplification, in the following, we assume to consider the same entity type  $\mathcal{E}$  and hence use abbreviated notions of the defined completeness metrics (e.g.  $comp(\mathcal{R})$  instead of  $comp(\mathcal{R}, \mathcal{E})$ ).

### 3.3. Further Remarks

For the purpose of illustration, the coverage and density of a relation  $\mathcal{R}$  can be interpreted w.r.t. probability theory and w.r.t. set theory. In probability theory, coverage can be considered as the probability that an element of  $\mathcal{E}$  is represented by a tuple of  $\mathcal{R}$  (and hence belongs to the extension of  $\mathcal{R}$ ). Assuming a uniform distribution on the availability of all attribute values, the density of  $\mathcal{R}$  can be interpreted as the probability that an attribute value of  $\mathcal{R}$  is complete. In set theory, coverage can be considered as the relative size of the entity type covered by the extension of  $\mathcal{R}$ . Density can be considered as the average confidence of all elements in  $Ext(\mathcal{R})$ .

In this paper all metrics are generally defined from a theoretical point of view. Unfortunately, in many cases the size of  $\mathcal{E}$  is not known. Thus, this size often has

to be estimated or if available, a reference relation<sup>1</sup> has to be used (Scannapieco *et al.*, 2004). The same deficiency holds for information whether a tuple or an attribute is relevant for a specific entity type or not. As for many other expensive data analyzing operations, sampling techniques (Olken, 1993) have to be used. By using this approach, coverage can be estimated as :

$$c(\mathcal{R}, \mathcal{E}) = \frac{r_{\mathcal{R}}(\mathcal{E}) \cdot |Ext(\mathcal{R})|}{|\mathcal{E}|} = \frac{r_{\mathcal{R}}(\mathcal{E}) \cdot |\mathcal{R}|}{|\mathcal{E}|} \quad [9]$$

where  $r_{\mathcal{R}}(\mathcal{E})$  (relevance factor) represents the average relevance of entities of  $Ext(\mathcal{R})$  to the entity type  $\mathcal{E}$ . Using statistical techniques as sampling, the relevance factor is only measured for a small subrelation  $S \subset \mathcal{R}$  and results in :

$$r_{\mathcal{R}}(\mathcal{E}) = \frac{|Ext(S_C(\mathcal{E}))|}{|Ext(S)|} = \frac{|S_C(\mathcal{E})|}{|S|} \quad [10]$$

In the following, we generally consider completeness from the theoretical point of view. If some context information is missing (e.g.  $|\mathcal{E}|$ ), metrics for practical solutions are more approximate and hence more imprecise. Nevertheless, adaptations for such cases have to be based on the theoretical foundations defined in this work.

#### 4. Relations with Maybe-Tuples

In contrast to *definite tuples*, *maybe tuples* are tuples for which it is undefined whether they belong to the associated relation or not. A relation with *maybe tuples* is in the following denoted as *maybe relation*. As an example we consider the *maybe relation*  $m\_cAct$  representing male comedy movie actors (see Figure 2).  $m\_cAct$  results from selecting all tuples whose main genre is 'comedy' from a database relation  $m\_Act$  storing male actors (see Figure 3).

	ID	f_name	s_name	m_genre	M	$cf_{m\_cAct}$	
$t_1$	9	Ben	Stiller	comedy	NO	1.0	} $m\_cAct^D$
$t_2$	16	Jean	Reno	comedy	NO	1.0	
...	...	...	...	...	...	...	
$t_n$	22	Mel	Gibson	{comedy.action}	YES	0.5	} $m\_cAct^M$
$t_{n+1}$	126	Rene	Russo	comedy	YES	0.3	
...	...	...	...	...	...	...	

**Figure 2.** Sample maybe relation  $m\_cAct$  representing male comedy movie actors

*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

1. A relation which is known to be highly complete.

concluded, whether an entity is part of the extension of a specific entity type or not (e.g. during information elicitation it was ambiguous, whether Rene Russo is male or not). As a consequence, for representing this uncertainty, the corresponding tuple can neither be excluded from nor certainly included into the associated database relation. Thus, these tuples have to be indicated as 'maybe' (attribute 'M' for a simple maybe indication or attribute  $cf_{m\_Act}$ , if confidence values are available). Moreover, if a database contains null values or values representing partial information, during query evaluation for some tuples the query condition cannot be evaluated to TRUE or FALSE (e.g. the main genre of Mel Gibson is either 'comedy' or 'action'). Thus, these tuples are *possible* query results and have to be indicated as *maybe tuples*, too.

	ID	f_name	s_name	m_genre	M	$cf_{m\_Act}$
$t'_1$	1	Tom	Cruise	action	NO	1.0
$t'_2$	9	Ben	Stiller	comedy	NO	1.0
$t'_3$	16	Jean	Reno	comedy	NO	1.0
$t'_4$	22	Mel	Gibson	{comedy,action}	NO	1.0
...	...	...	...	...	...	...
$t'_k$	58	Jude	Law	drama	YES	0.8
$t'_{k+1}$	126	Rene	Russo	comedy	YES	0.3
...	...	...	...	...	...	...

**Figure 3.** *Maybe database relation  $m\_Act$  storing male movie actors*

A *maybe relation*  $\mathcal{R}$  can be divided into two subrelations : Relation  $\mathcal{R}^D$  contains all tuples which definitely belong to  $\mathcal{R}$  and relation  $\mathcal{R}^M$  contains all tuples which maybe belong to  $\mathcal{R}$ . Since this separation is lossless, always the equation  $\mathcal{R} = \mathcal{R}^D \cup \mathcal{R}^M$  holds. If  $\mathcal{R}$  does not contain duplicate entries (e.g. if  $\mathcal{R}$  is a database relation), the two subsets have to be disjoint ( $\mathcal{R}^D \cap \mathcal{R}^M = \emptyset$ ). Note, duplicate entries result from projections and hence a tuple can refer to multiple real-world entities. Nevertheless, the extensions of  $\mathcal{R}^D$  and  $\mathcal{R}^M$  are generally disjoint.

In databases with exact confidence specifications, for each tuple  $t \in \mathcal{R}$  an individual confidence value  $cf(t)_{\mathcal{R}}$  is defined, presenting the confidence that this tuple belongs to the associated relation. Since all tuples of the subrelation  $\mathcal{R}^D$  are definitely in  $\mathcal{R}$ , the individual confidence values of these tuples always have to be 1.0. In contrast, due to every *maybe tuple* only possibly belongs to the relation, its individual confidence value has to be lower than 1. However, because these tuples cannot certainly be excluded from this relation, their confidences have also to be greater than 0. Therefore, the confidence values of all *maybe tuples* are values within the range  $]0, 1[$ . The confidence of each tuple is related to the reference extension  $\mathcal{W}_{\mathcal{R}}$ . For example, given a relation  $\mathcal{R}$  for storing students, the confidence  $cf_{\mathcal{R}}(t) = 0.8$  means that the entity represented by tuple  $t$  is a student with a confidence of 0.8. If the completeness of  $\mathcal{R}$  is considered w.r.t. another entity type  $\mathcal{E}$ , e.g., citizens ( $\mathcal{E} \supset \mathcal{W}_{\mathcal{R}}$ ), the confidence values of its tuples and hence the resulting quality score become more unsound. One of the most intuitive interpretations of tuple membership in fuzzy databases is the certainty that the corresponding tuple belongs to the considered relation (Galindo *et*

*al.*, 2006). Since probability is a measure of certainty, we sometimes directly consider tuple confidence as tuple probability.

In the following,  $\mathcal{I}(\mathcal{R})$  represents the set of all possible instances and  $\mathcal{R}'$  represents the true instance of the relation  $\mathcal{R}$  under a closed world assumption<sup>2</sup>. Due to all tuples of  $\mathcal{R}^{\mathcal{D}}$  definitely belong to  $\mathcal{R}$ , each possible instance  $I \in \mathcal{I}(\mathcal{R})$  contains these tuples. In general, for every possible combination of *maybe tuples* (i.e. the power set ( $\mathcal{P}(\mathcal{R}^{\mathcal{M}})$ )) one possible instance of  $\mathcal{R}$  results. Thus, the set  $\mathcal{I}(\mathcal{R})$  is given as :

$$\mathcal{I}(\mathcal{R}) = \{\mathcal{R}^{\mathcal{D}} \cup M \mid M \in \mathcal{P}(\mathcal{R}^{\mathcal{M}})\} \quad [11]$$

If  $\mathcal{R}$  does not contain *maybe tuples*, all tuples of  $\mathcal{R}$  are known and the true set of tuples belonging to  $\mathcal{R}$  is completely described by  $\mathcal{R}$  itself. As a consequence,  $\mathcal{I}(\mathcal{R})$  contains just one element and the relations  $\mathcal{R}$  and  $\mathcal{R}'$  are equal. In contrast, if  $\mathcal{R}$  contains *maybe tuples*, the set of tuples which really belong to  $\mathcal{R}$  and hence the relation  $\mathcal{R}'$  are not completely known. If confidence is interpreted as probability, this uncertainty can be represented by a discrete probability distribution on  $\mathcal{R}'$  over the set  $\mathcal{I}(\mathcal{R})$ . For example, we assume a relation  $\mathcal{R}$  containing one definite tuple  $t_1$  and one *maybe tuple*  $t_2$  ( $cf(t_2)_{\mathcal{R}} = 0.6$ ). The set of all possible instances is  $\mathcal{I}(\mathcal{R}) = \{I_0 = \{t_1\}, I_1 = \{t_1, t_2\}\}$  and the true instance  $\mathcal{R}'$  is a random variable with the probability distribution  $P(\mathcal{R}' = I_0) = 0.4$  and  $P(\mathcal{R}' = I_1) = 0.6$ .

## 5. Problem Description

A *maybe tuple* (and hence the entity which is represented by this tuple) only possibly belongs to a relation (or entity type respectively). Thus, 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  $\mathcal{R}_3$  of Figure 4.  $\mathcal{R}_1$  and  $\mathcal{R}_3$  are relations without *maybe tuples*. Relation  $\mathcal{R}_2$  contains two definite (the same tuples as  $\mathcal{R}_1$ ) and one *maybe tuple*. We assume that all three tuples are relevant for the considered entity type. 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  $\mathcal{R}_2$ , however, 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 bounded by  $comp(\mathcal{R}_1) < comp(\mathcal{R}_2) < comp(\mathcal{R}_3)$ . Moreover, for adequate quality metrics a set of requirements exists (Heinrich *et al.*, 2007). For example, to satisfy the requirement of an interval scale, the completeness of  $\mathcal{R}_2$  has to increase linear from  $comp(\mathcal{R}_1)$  to  $comp(\mathcal{R}_3)$  with a growing confidence of  $t_3$ .

---

2. In this case, incompleteness resulting from missing values and totally missing tuples is ignored. The uncertain membership of *maybe tuples* is the only incomplete information and for all possible instances only the tuples of  $\mathcal{R}^{\mathcal{D}}$  and  $\mathcal{R}^{\mathcal{M}}$  are considered.



	<b>f_name</b>	<b>s_name</b>
$t_1$	Ben	Stiller
$t_2$	Jean	Reno

$\mathcal{R}_1$

	<b>f_name</b>	<b>s_name</b>	<b>M</b>	$cf_{\mathcal{R}_2}$
$t_1$	Ben	Stiller	NO	1.0
$t_2$	Jean	Reno	NO	1.0
$t_3$	Mel	Gibson	YES	$x$

$\mathcal{R}_2$

	<b>f_name</b>	<b>s_name</b>
$t_1$	Ben	Stiller
$t_2$	Jean	Reno
$t_3$	Mel	Gibson

$\mathcal{R}_3$

**Figure 4.** *Completeness classification of relations with maybe tuples*

## 6. Data Completeness Regarding Maybe-Tuples

In order to calculate exact completeness scores for *maybe relations*, we introduce three different but each intuitive approaches. The first approach is based on the possible world semantics and considers the completeness of a *maybe relation* as the expected completeness of its true instance. The second approach uses the average completeness of the relation's instances which can be the result of a so called  $\alpha$ -selection. The last approach is based on the fuzzy-set theory. Partially, we trace our new metrics to the initial ones, as defined in Section 3.2. For distinction, the newly defined metrics of completeness, coverage and density w.r.t. a relation  $\mathcal{R}$  and an approach  $A_i$  are denoted as  $comp'_{A_i}(\mathcal{R})$ ,  $c'_{A_i}(\mathcal{R})$  and  $d'_{A_i}(\mathcal{R})$ .

### 6.1. Approach 1 (Possible World Semantics)

As we think, the most intuitive way is to consider data completeness within the possible world semantics and to treat confidence as probability. In this case, the completeness of a *maybe relation*  $\mathcal{R}$  can be defined as the expected completeness over all possible instances of  $\mathcal{R}$  and hence as the expected completeness of  $\mathcal{R}'$ . Since the coverage and the density of each instance are not independent from each other, this approach unfortunately does not allow a decomposition into these two measures :

$$comp'_{A_1}(\mathcal{R}) = E(comp(\mathcal{R}')) = E(c(\mathcal{R}') \cdot d(\mathcal{R}')) \neq E(c(\mathcal{R}')) \cdot E(d(\mathcal{R}'))$$

#### 6.1.1. Individual Confidence Values :

In order to define the completeness of  $\mathcal{R}$  as the expectation value of  $comp(\mathcal{R}')$ , for each possible instance of  $\mathcal{R}$  the completeness<sup>3</sup> and the probability have to be known.

$$\begin{aligned} E(comp(\mathcal{R}')) &= \sum_{I_k \in \mathcal{I}(\mathcal{R}_c)} P(\mathcal{R}'_c = S_i) \cdot comp(I_k) & [12] \\ &= \frac{1}{|\mathcal{E}|} \sum_{I_k \in \mathcal{I}(\mathcal{R}_c)} P(\mathcal{R}'_c = I_k) \sum_{t \in I_k} d(t) \end{aligned}$$

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

In the exact case, the probability of a each possible instance  $I_k \in \mathcal{I}(\mathcal{R})$  results from the product of the probabilities (confidences) of all tuples in  $I_k$  and the inverse probabilities (confidences) of all tuples of  $\mathcal{R}_C$  which are not in  $I_k$  :

$$P(\mathcal{R}'_C = I_k) = \prod_{t \in I_k} cf(t)_{\mathcal{R}} \prod_{t \in \mathcal{R}_C \setminus I_k} (1 - cf(t)_{\mathcal{R}})$$

### 6.1.2. Simple Maybe Indication :

In the simple case, information on confidence values is not available. Thus, we assume that the possible instances are distributed uniformly. Without tuple dependencies there exist  $|\mathcal{I}(\mathcal{R}_C)| = |\mathcal{P}(\mathcal{R}^M)| = 2^{|\mathcal{R}_C^M|}$  possible instances. Therefore, the expectation value  $E(comp(\mathcal{R}'))$  and hence the completeness  $comp'_{A1}(\mathcal{R})$  defined in Equation 12 can be simplified to :

$$comp'_{A1}(\mathcal{R}) = E(comp(\mathcal{R}')) = \frac{1}{2^{|\mathcal{R}_C^M|}} \frac{1}{|\mathcal{E}|} \sum_{I_k \in \mathcal{I}(\mathcal{R}_C)} \sum_{t \in I_k} d(t) \quad [13]$$

If tuple dependencies exist, instead of  $2^{|\mathcal{R}_C^M|}$  the reduced number of possible instances  $|\mathcal{I}(\mathcal{R}_C)|$  has to be directly used.

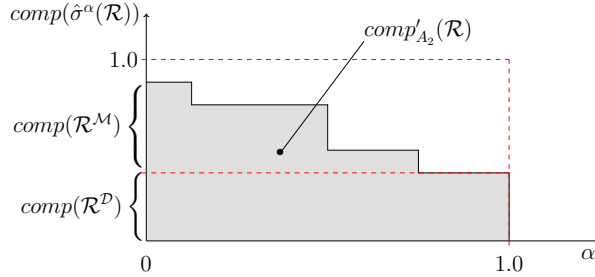
## 6.2. Approach 2 ( $\alpha$ -Selection)

The second approach is based on the  $\alpha$ -selection<sup>4</sup> introduced in (Tseng *et al.*, 1993). An  $\alpha$ -selection ( $\hat{\sigma}_P^\alpha(\mathcal{R})$ ) on a relation  $\mathcal{R}$  selects each tuple  $t \in \mathcal{R}$  which belongs to the result set  $\sigma_P(\mathcal{R})$  with a confidence greater or equal than  $\alpha$  ( $\alpha \in [0, 1]$ ) :

$$\hat{\sigma}_P^\alpha(\mathcal{R}) = \{t \mid t \in \sigma_P(\mathcal{R}) \wedge cf(t)_{\mathcal{R}} \geq \alpha\} \quad [14]$$

If no predicate  $P$  is defined, an  $\alpha$ -selection returns each tuple which belongs to  $\mathcal{R}$  with a confidence of  $\alpha$  or greater. In general,  $\alpha$ -selections can be used for a quality driven tuple-filtering. Thus, if for a simple relation output an  $\alpha$ -selection is used, the completeness of the resulting subrelation depends on the value of  $\alpha$ . The higher  $\alpha$ , the more tuple are filtered. Thus, the completeness  $comp(\hat{\sigma}^\alpha(\mathcal{R}))$  decreases monotonically (see Figure 5). Moreover, the completeness of a filtered relation  $\hat{\sigma}^\alpha(\mathcal{R})$  is always greater or equal than the completeness of the subrelation  $\mathcal{R}^D$  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 its subrelations resulting from all possible  $\alpha$ -selections applied on  $\mathcal{R}$ .

4. This is similar to the  $\alpha$ -cut known from fuzzy set theory.



**Figure 5.** Completeness of a maybe relation  $\mathcal{R}$  w.r.t. all possible  $\alpha$ -selections

### 6.2.1. Individual Confidence Values :

If individual confidence values are given, for each  $\alpha$  another subrelation can result from applying an  $\alpha$ -selection. Thus,  $\alpha$  has to be considered being within the continuous range  $[0, 1]$  and the completeness  $comp'_{A_2}(\mathcal{R})$  can be defined as the integral of  $comp(\hat{\sigma}^\alpha(\mathcal{R}))$  over  $\alpha$  (see the gray colored area in Figure 5) :

$$comp'_{A_2}(\mathcal{R}) = \int_{0.0}^{1.0} comp(\hat{\sigma}^\alpha(\mathcal{R}))d\alpha \quad [15]$$

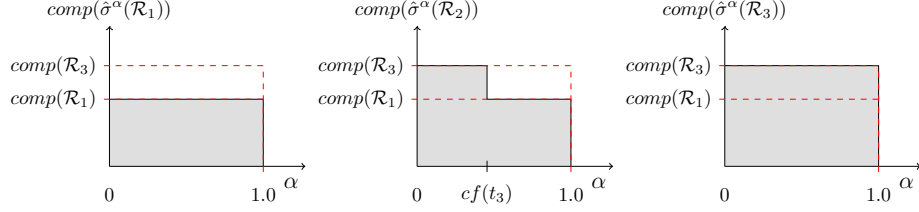
Since the tuples of  $\mathcal{R}^D$  belong to the result of each  $\alpha$ -selection on  $\mathcal{R}$ , the following equation is valid :

$$\begin{aligned} comp'_{A_2}(\mathcal{R}) &= \int_{0.0}^{1.0} comp(\hat{\sigma}^\alpha(\mathcal{R}^D))d\alpha + \int_{0.0}^{1.0} comp(\hat{\sigma}^\alpha(\mathcal{R}^M))d\alpha \\ &= comp(\mathcal{R}^D) + comp'_{A_2}(\mathcal{R}^M) \end{aligned}$$

Unfortunately, due to the inequality shown in Equation 16, a decomposition into coverage and density is not possible.

$$\int_{0.0}^{1.0} c(\hat{\sigma}^\alpha(\mathcal{R})) \cdot d(\hat{\sigma}^\alpha(\mathcal{R}))d\alpha \neq \int_{0.0}^{1.0} c(\hat{\sigma}^\alpha(\mathcal{R}))d\alpha \cdot \int_{0.0}^{1.0} d(\hat{\sigma}^\alpha(\mathcal{R}))d\alpha \quad [16]$$

For demonstrating the usability of this approach, we consider the example introduced in Section 5. The respective completeness of the three relations  $\mathcal{R}_1$ ,  $\mathcal{R}_2$  and  $\mathcal{R}_3$  calculated by the metric of Equation 15 is graphically represented in Figure 6. It illustrates that by using this approach the completeness of  $\mathcal{R}_2$  is always within the range  $[comp(\mathcal{R}_1), comp(\mathcal{R}_3)]$  and increases steadily from  $comp(\mathcal{R}_1)$  to  $comp(\mathcal{R}_3)$  with growing  $cf(t_3)_{\mathcal{R}_2}$ .



**Figure 6.** The completeness of  $\mathcal{R}_1$ ,  $\mathcal{R}_2$  and  $\mathcal{R}_3$  w.r.t. all possible  $\alpha$ -selections

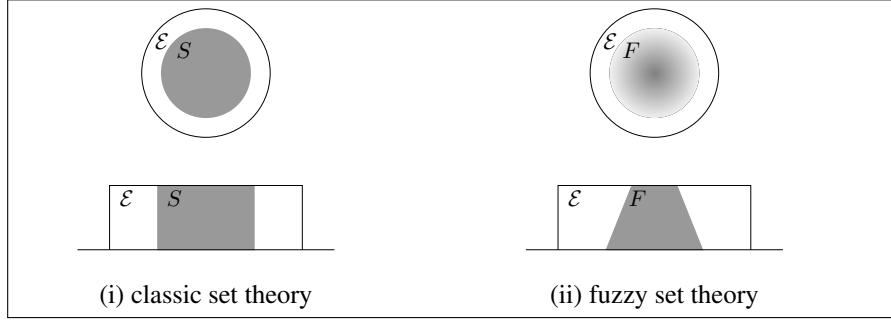
### 6.2.2. Simple Maybe Indication :

Intuitively, in the simple case, the confidence of each *maybe tuple* cannot be specified exactly and is assumed to be 0.5 (the possibility that such an element belongs to the set is equal to its contrary case). Therefore, from applying  $\alpha$ -selections only two subrelations can result : the subrelation  $\mathcal{R}^{\mathcal{D}}$ , if  $\alpha$  is within the range (0.5,1] and the whole relation  $\mathcal{R} = \mathcal{R}^{\mathcal{D}} \cup \mathcal{R}^{\mathcal{M}}$  otherwise. As a consequence, the completeness  $comp'_{A_2}(\mathcal{R})$  defined in Equation 15 can be simplified to :

$$\begin{aligned}
 comp'_{A_2}(\mathcal{R}) &= \int_0^{0.5} comp(\mathcal{R}^{\mathcal{D}} \cup \mathcal{R}^{\mathcal{M}})d\alpha + \int_{0.5}^{1.0} comp(\mathcal{R}^{\mathcal{D}})d\alpha \\
 &= comp(\mathcal{R}^{\mathcal{D}}) + \int_0^{0.5} comp(\mathcal{R}^{\mathcal{M}})d\alpha \\
 &= comp(\mathcal{R}^{\mathcal{D}}) + \frac{1}{2}comp(\mathcal{R}^{\mathcal{M}}) \tag{17}
 \end{aligned}$$

### 6.3. Approach 3 (Fuzzy Set Theory)

In databases without *maybe tuples*, completeness computation can be traced back to set theory. The relevant extension of a relation  $\mathcal{R}$  can be considered as a subset ( $S = Ext(\mathcal{R}_C) \subseteq \mathcal{E}$ ) of the whole entity type (or reference relation, respectively). In contrast, in databases with *maybe tuples*, the fuzzy set theory can be used. The relevant extension of a *maybe relation*  $\mathcal{R}$  can be considered as a fuzzy set, where all definite tuples represent crisp elements and all *maybe tuples* represent possible elements. Formally, the extension of a *maybe relation*  $\mathcal{R}_C$  can be interpreted as the fuzzy set  $F = (\mathcal{E}, \mu)$ , where  $Ext(\mathcal{R}_C^{\mathcal{D}})$  represents the kernel of  $F$  and  $cf_{\mathcal{R}}$  is the membership function  $\mu$ .



**Figure 7.** Relative size of a classic subset (left) and fuzzy subset (right)

### 6.3.1. Individual Confidence Values :

The cardinality (size) of a fuzzy set is the sum of all of its elements' memberships.

$$card(F) = \sum_{e \in F} \mu(e)$$

Thus, the coverage of a *maybe relation*  $\mathcal{R}$  can be still considered as the relevant amount of  $F = Ext(\mathcal{R}_C)$  to  $\mathcal{E}$  (see Figure 7) and can be defined as :

$$c'_{A3}(\mathcal{R}) = \frac{Card(Ext(\mathcal{R}_C))}{|\mathcal{E}|} = \frac{\sum_{t \in \mathcal{R}_C} cf(t)_{\mathcal{R}}}{|\mathcal{E}|} \quad [18]$$

The coverage can be also transformed into<sup>5</sup> :

$$\begin{aligned} c'_{A3}(\mathcal{R}) &= c(\mathcal{R}^D) + \frac{\sum_{t \in \mathcal{R}_C^M} cf(t)_{\mathcal{R}}}{|\mathcal{R}_C^M|} \cdot \frac{|\mathcal{R}_C^M|}{|\mathcal{E}|} \\ &= c(\mathcal{R}^D) + \underbrace{avg(cf(t)_{\mathcal{R}_C^M})}_{A_{\mathcal{R}}} \cdot c(\mathcal{R}^M) \end{aligned} \quad [19]$$

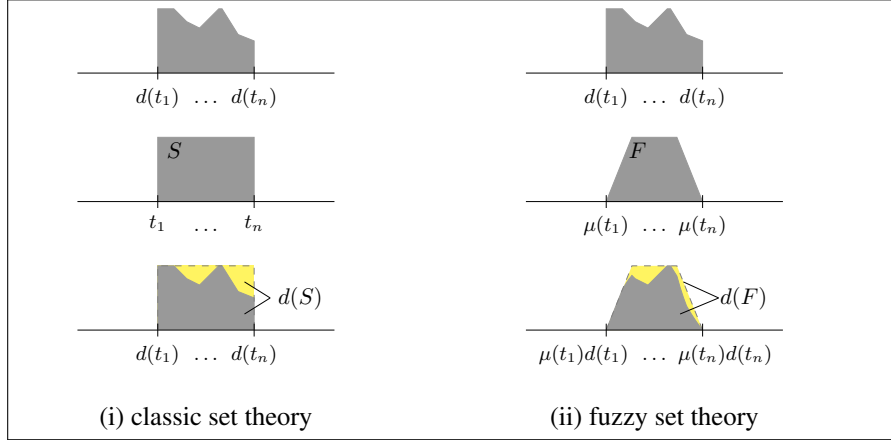
The density of a *maybe relation*  $\mathcal{R}$  in turn can be considered as the average density of the fuzzy set's elements (for illustration see Figure 8) :

$$d'_{A3}(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}_C} cf(t)_{\mathcal{R}} \cdot d(t)}{\sum_{t \in \mathcal{R}_C} cf(t)_{\mathcal{R}}} \quad [20]$$

Consequently, the completeness of  $\mathcal{R}$  results in :

$$comp'_{A3}(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}_C} cf(t)_{\mathcal{R}} \cdot d(t)}{|\mathcal{E}|} \quad [21]$$

5. Since in this notion the new coverage metric is reduced to the initial one, it is very useful for coverage estimation (see Section 7).



**Figure 8.** Average density for classic subset (left) and fuzzy subset (right)

### 6.3.2. Simple Maybe Indication :

In the simple maybe indication, once more the membership degree of possible fuzzy set's elements is considered to be 0.5. As a consequence, the coverage  $c'_{A3}(\mathcal{R})$  of  $\mathcal{R}$  can be directly calculated from the coverage scores of  $\mathcal{R}^{\mathcal{D}}$  and  $\mathcal{R}^{\mathcal{M}}$  :

$$c'_{A3}(\mathcal{R}) = c(\mathcal{R}^{\mathcal{D}}) + \frac{1}{2}c(\mathcal{R}^{\mathcal{M}}) = \frac{|\mathcal{R}_c^{\mathcal{D}}|}{|\mathcal{E}|} + \frac{1}{2} \frac{|\mathcal{R}_c^{\mathcal{M}}|}{|\mathcal{E}|} = \frac{|\mathcal{R}_c^{\mathcal{D}}| + \frac{1}{2}|\mathcal{R}_c^{\mathcal{M}}|}{|\mathcal{E}|} \quad [22]$$

Similarly to the coverage, the effect of the density  $d(\mathcal{R}^{\mathcal{M}})$  on  $d'_{A3}(\mathcal{R})$  is only half as high as the effect of the *definite tuples'* densities. Since single densities are only relative, the densities of both subrelations have to be correlated by taking into account their sizes :

$$\begin{aligned} d'_{A3}(\mathcal{R}) &= \frac{|\mathcal{R}_c^{\mathcal{D}}|d(\mathcal{R}^{\mathcal{D}}) + \frac{1}{2}|\mathcal{R}_c^{\mathcal{M}}|d(\mathcal{R}^{\mathcal{M}})}{|\mathcal{R}_c^{\mathcal{D}}| + \frac{1}{2}|\mathcal{R}_c^{\mathcal{M}}|} \\ &= \frac{\sum_{t \in \mathcal{R}_c^{\mathcal{D}}} d(t) + \frac{1}{2} \sum_{t \in \mathcal{R}_c^{\mathcal{M}}} d(t)}{|\mathcal{R}_c^{\mathcal{D}}| + \frac{1}{2}|\mathcal{R}_c^{\mathcal{M}}|} \end{aligned} \quad [23]$$

The completeness  $comp'_{A3}(\mathcal{R}) = c'_{A3}(\mathcal{R}) \cdot d'_{A3}(\mathcal{R})$  results in :

$$comp'_{A3}(\mathcal{R}) = comp(\mathcal{R}^{\mathcal{D}}) + \frac{1}{2}comp(\mathcal{R}^{\mathcal{M}}) \quad [24]$$

Thus, in the case of a simple maybe indication, the completeness metrics of Approach 1 and Approach 2 are identical (compare Equation 24 with Equation 17).

Note, the same metrics (for individual confidence values as well as for simple maybe indications) results from tracing completeness computation back to probability theory in the way already mentioned in Section 3.3.

#### 6.4. Incomplete Confidence Values

In some cases, for example resulting from a query on incomplete information, an individual confidence value is not completely known (e.g. the confidence that the tuple  $t$  belongs to the relation  $\mathcal{R}$  is lower than 0.5). To consider such cases, one possibility is to define the completeness as a partial value by using the minimal possible confidence  $cf(t)_{\mathcal{R}}^{min}$  of every tuple for calculating a lower and the maximal possible confidence  $cf(t)_{\mathcal{R}}^{max}$  of every tuple for calculating an upper bound. Another possibility is to use the expected confidence. By doing so, the metrics of the last approach have to be adapted to :

$$c'_{A3}(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}_c} E(cf(t)_{\mathcal{R}})}{|\mathcal{E}|} \quad [25] \quad d'_{A3}(\mathcal{R}) = \frac{\sum_{t \in \mathcal{R}_c} E(cf(t)_{\mathcal{R}}) \cdot d(t)}{\sum_{t \in \mathcal{R}_c} E(cf(t)_{\mathcal{R}})} \quad [26]$$

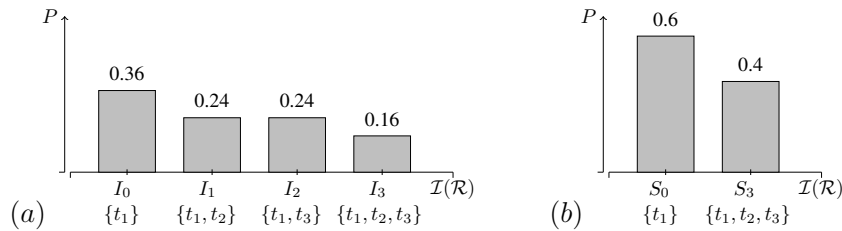
For calculating the expected confidence a distribution function is required ; if there is none, a distribution has to be assumed (e.g., a uniform distribution). In general, data quality has to appraise the data. Thus, we think completeness as a partial value is not feasible and using the expectation value is better suitable. In all three approaches, we derived the metric for the simple maybe indication from the general case by assuming a confidence of 0.5 for each *maybe tuple*. Since in this case, the confidence of the *maybe tuples* is completely unknown (within the range  $]0, 1[$ ), we already have used the expectation value intuitively.

#### 6.5. Tuple Dependencies

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 (Sen *et al.*, 2007) 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  $\mathcal{I}(\mathcal{R})$ .

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 ( $t_2 \in \mathcal{R} \Leftrightarrow t_3 \in \mathcal{R}$ ). As a consequence, instead of four possible instances  $\{\{t_1\}, \{t_1, t_2\}, \{t_1, t_3\}, \{t_1, t_2, t_3\}\}$  only two possible instances  $\mathcal{I}(\mathcal{R}) = \{\{t_1\}, \{t_1, t_2, t_3\}\}$  exist. Thus, it is obvious that our completeness metrics which are based on the possible world semantics (Approach 1) can be used in models with tuple dependencies without any adaptation.

In general, the total confidence of each tuple  $t$  is (independently of dependencies) always  $cf(t)_{\mathcal{R}}$ . If this confidence is interpreted as probability, it does not matter in which way this probability is distributed on the possible instances. In order to illustrate this fact, we consider the example mentioned above and assume that  $t_2$  as well as  $t_3$  has a probability of 0.4. The distribution of  $\mathcal{R}'$  with and without the tuple dependency is shown in Figure 9. The total probability of  $t_2$  is either the sum of  $P(I_1)$  and  $P(I_3)$



**Figure 9.** Distribution of  $\mathcal{R}'$  without (left) and with the tuple dependency (right)

(without dependency) or just  $P(I_3)$  (with dependency), but is always 0.4. For that reason, the metrics of the other two approaches are also independent to individual tuple dependencies and do not need to be adapted on such cases.

## 6.6. Comparison of Proposed Approaches

In the previously explained approaches, we defined metrics for measuring completeness of *maybe relations*. The next step is to compare these metrics to each other. As we proof (see Proofs 1-3), the metrics defined in all approaches supply the same completeness scores whether tuple dependencies are defined or not. We think, this fact indicates that the resulting scores are good representations of the actual relations' qualities. Regarding the requirements proposed in (Heinrich *et al.*, 2007) the metrics of all approaches satisfy the requirements of normalization, interval scale and adaptivity. Moreover, the interpretability of all three approaches is similar. Due to its possible decomposition of completeness into coverage and density, only the interpretability of the last approach gains a small edge of the interpretability of the other ones. Finally, the input parameters and hence the feasibilities of all approaches are equal.

Thus, the most severe difference of these approaches relates to the complexity of the individual metrics. In the following, we consider a relation  $\mathcal{R}$  with  $n$  *definite* and  $m$  *maybe tuples*. With respect to all metrics regarding individual confidence values (Equations 13, 15 and 21), a one-time calculation of all the tuples' densities is required ( $\mathcal{O}(n + m)$ ). In the metric of the first approach, the completeness of every possible instance of  $\mathcal{R}$  has to be calculated. If no tuple dependencies exist, the number of possible instances is  $2^m$ . Besides completeness<sup>6</sup>, for every possible instance its

6. The complexity of calculating a relation's completeness is always  $\mathcal{O}(n + m)$ .



probability has to be derived from the confidence values of  $(n+m)$  tuples. As a consequence, this metric has a complexity of  $2^m 2(n+m) = \mathcal{O}(2^m(n+m))$ . Using the metric of Approach 2 implies calculating the completeness of all possible subrelations which can result from any  $\alpha$ -selection. In the simplest case, all *maybe tuples* have the same confidence  $cf$  and only the completeness scores of two subrelations ( $\alpha = cf$  and  $\alpha = 1$ ) are required ( $2(n+m)$ ). In the most complex case, all *maybe tuples* have different confidences and the completeness scores of  $m+1$  subrelations have to be calculated ( $(m+1)(n+m)$ ). The integral of  $comp(\hat{\sigma}^\alpha(\mathcal{R}))$  over  $\alpha$  can be maximally split into  $m+1$  subintegrals of  $m+1$  different completeness scores. Thus, the minimal and maximal complexity of this metric is either  $2(n+m) + (m+1) = \mathcal{O}(n+m)$  or  $(m+1)(n+m) + (m+1) = \mathcal{O}(max(m^2, nm))$ . The metric of the last approach implies the calculation of only a single completeness score. Thus its complexity is just  $\mathcal{O}(n+m)$ . In conclusion, in databases with individual confidence values, the metric of Approach 3 has by far the lowest complexity (see Tableau 1). In the case of simple maybe indication, all three metrics have the same complexity ( $\mathcal{O}(n+m)$ ).

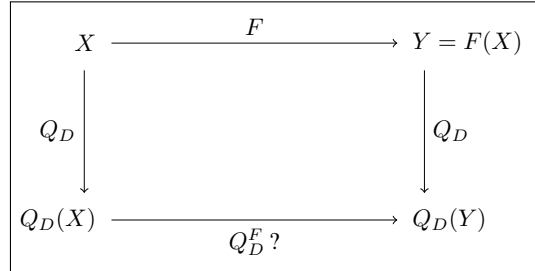
	complexity : simple case	complexity : exact case
Approach 1 :	$\mathcal{O}(n+m)$	$\mathcal{O}(2^m(n+m))$
Approach 2 :	$\mathcal{O}(n+m)$	$\mathcal{O}(max(m^2, nm))$
Approach 3 :	$\mathcal{O}(n+m)$	$\mathcal{O}(n+m)$

**Tableau 1.** Complexities of the three approaches

Regarding its minor complexity, in databases with individual confidence values, the metric of Approach 3 is most suitable. At a first sight (without considerations on implementation specific details as speed or storage requirements), in databases with just a simple maybe indication both metrics (Equations 17 and 13) can be assumed to be equivalently suitable.

## 7. Completeness Composition

Until now, we calculate completeness of a *maybe relation* starting from all its source data. In many applications, however, it is usual to combine data extracted from multiple sources (e.g. for quality improvement). Nevertheless, a recalculation of the resulting quality score is most often too inefficient or the required information for doing that is simply not available (e.g. for quality prediction in quality driven query answering). For that reason, it is important to be able to derive the quality of the resulting data from the quality of its source data. In (Batini *et al.*, 2006), this problem is stated as the quality composition problem which is illustrated in Figure 10. Given  $X$ , a set of source relations, which is processed by a generic composition function  $F$  (e.g. the join or the union), and a function  $Q_D$  calculates the score of the quality dimension  $D$ . The problem is to define a function  $Q_D^F(X)$  that calculates  $Q_D(Y)$  starting from  $Q_D(X)$ , instead of calculating such a score directly on  $Y$  by applying the function  $Q_D$ .



**Figure 10.** *The general problem of quality composition (Batini et al., 2006)*

Completeness composition w.r.t. algebraic operations on certain data is considered in several works (Naumann *et al.*, 2004; Scannapieco *et al.*, 2004). In general, multiple relations can be combined in different ways. As a representative, we consider the union merge operator<sup>7</sup> which is defined in (Naumann *et al.*, 2003). The relation  $\mathcal{T} = \mathcal{R} \sqcup \mathcal{S}$  resulting from the union merge of two relations  $\mathcal{R}$  and  $\mathcal{S}$  contains one tuple for each entity represented by a tuple in  $\mathcal{R}$  or  $\mathcal{S}$ . If an entity is represented by a tuple in both relations, the two corresponding tuples are merged to a single one. As a consequence, the union merge can be considered as a set union of the relations' extensions:  $Ext(\mathcal{R} \sqcup \mathcal{S}) = Ext(\mathcal{R}) \cup Ext(\mathcal{S})$  (see Figure 11). For an exact definition of the union merge, we refer the interested reader to (Naumann *et al.*, 2003).



**Figure 11.** *Union merge of the two relations  $\mathcal{R}$  and  $\mathcal{S}$*

### 7.1. Coverage Estimation for a Merge of Certain Data

Coverage is a measure for extensional completeness. If information is available, whether the extensions of the different sources overlap or not, the estimation's quality can be enhanced. The degree of extensional overlap may vary from no overlap to a complete overlap. In (Naumann *et al.*, 2004), four overlap situations of the extensions of two relations  $\mathcal{R}$  and  $\mathcal{S}$  are considered: disjointness, independence, quantified overlap and containment. If the extensions of both relations have no common entity, the relations are denoted to be disjoint ( $\mathcal{R} \cap^d \mathcal{S}$ ). Both relations are denoted to be independent ( $\mathcal{R} \cap^i \mathcal{S}$ ), if no information about any overlap is available, but the data of both relations is considered to be independent. If the exact degree of the extensions overlap is known ( $X$ ), a quantified overlap of both relations is given ( $\mathcal{R} \cap^X \mathcal{S}$ ). Finally, the

7. Also known as full outer join merge (Naumann *et al.*, 2004).

extension of one relation can be a subset of the extension of the other one ( $\mathcal{R} \cap^{\subseteq} \mathcal{S}$ ). These four overlap situations can be formalized as :

$$\text{(disjointness)} \quad \mathcal{R} \cap^d \mathcal{S} \equiv \text{Ext}(\mathcal{R}) \cap \text{Ext}(\mathcal{S}) = \emptyset \quad [27]$$

$$\text{(independence)} \quad \mathcal{R} \cap^i \mathcal{S} \equiv \text{Ext}(\mathcal{R}) \cap \text{Ext}(\mathcal{S}) = ? \quad [28]$$

$$\text{(quantified overlap)} \quad \mathcal{R} \cap^X \mathcal{S} \equiv \text{Ext}(\mathcal{R}) \cap \text{Ext}(\mathcal{S}) = X \quad [29]$$

$$\text{(containment)} \quad \mathcal{R} \cap^{\subseteq} \mathcal{S} \equiv \text{Ext}(\mathcal{R}) \subseteq \text{Ext}(\mathcal{S}) \quad [30]$$

Metrics for estimating the coverage of a merge of two certain relations (relations without *maybe tuples*) w.r.t. all four overlap situations are presented in Figure 12.

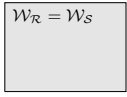

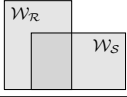
$(\mathcal{R} \cap^d \mathcal{S})$	$c(\mathcal{R} \sqcup \mathcal{S}) = c(\mathcal{R}) + c(\mathcal{S})$	(1)
$(\mathcal{R} \cap^i \mathcal{S})$	$c(\mathcal{R} \sqcup \mathcal{S}) = c(\mathcal{R}) + c(\mathcal{S}) - c(\mathcal{R}) \cdot c(\mathcal{S})$	(2)
$(\mathcal{R} \cap^X \mathcal{S})$	$c(\mathcal{R} \sqcup \mathcal{S}) = c(\mathcal{R}) + c(\mathcal{S}) -  X / \mathcal{E} $	(3)
$(\mathcal{R} \cap^{\subseteq} \mathcal{S})$	$c(\mathcal{R} \sqcup \mathcal{S}) = c(\mathcal{S})$	(4)

**Figure 12.** Coverage estimation w.r.t. the four extensional overlap situations of two certain relations  $\mathcal{R}$  and  $\mathcal{S}$

For instance, in case of independence, the resulting extension size is the addition of both relation sizes minus the estimated size of the relations' overlap. Since independence between the data of both relations is assumed, the estimated overlap is determined by the relative sizes of the relations' extensions. Using probability theory, the coverage of a merged relation can be considered as the probability that an entity of one of both reference extensions is represented by a tuple in the resulting relation. This is the case, if this entity is represented in at least one of the source relations. In case of independence, this probability is equal to the sum of probabilities that the entity is represented in one of both relations ( $c(\mathcal{R}) + c(\mathcal{S})$ ) minus the probability that this entity is represented in both relations ( $c(\mathcal{R}) \cdot c(\mathcal{S})$ ).

## 7.2. Coverage Estimation for Merged Maybe Relations

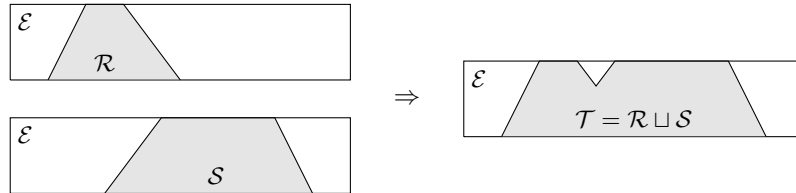
If for the concerned *maybe relations* only the scores of their whole coverage are known, the estimation of the resulting coverage score is equal to its estimation in certain data (only  $c$  is substitute by  $c'$ ). If in contrast for each *maybe relation*  $\mathcal{R}$  the coverage of its two subrelations  $\mathcal{R}^D$  and  $\mathcal{R}^M$  as well as the overlaps of the different subrelations are given, the estimation result can be determined more exactly.

Reference extensions	Confidence merging strategy
Equality <sup>8</sup>	 $cf_{\mathcal{R} \sqcup \mathcal{S}}(t) = \frac{q_{\mathcal{R}} \cdot cf_{\mathcal{R}}(t) + q_{\mathcal{S}} \cdot cf_{\mathcal{S}}(t)}{q_{\mathcal{R}} + q_{\mathcal{S}}}$
Disjointness	 $cf_{\mathcal{R} \sqcup \mathcal{S}}(t) = \min(1, cf_{\mathcal{R}}(t) + cf_{\mathcal{S}}(t))$
Overlap	 $cf_{\mathcal{R} \sqcup \mathcal{S}}(t) = \max(cf_{\mathcal{R}}(t), cf_{\mathcal{S}}(t))$

**Figure 13.** Three different overlap situations of reference extensions together with their corresponding confidence merging strategies

### 7.2.1. Confidence Merging

Coverage estimation essentially depends on the chosen strategy for confidence merging. Generally, the best strategy for merging confidence values can vary case-by-case. Three different strategies together with their corresponding cases are depicted in Figure 13. If the reference extensions of both sources are equal, for each entity the confidence values of all sources are either identical or the confidence value of at least one source has a minor quality. In this case, calculating a new confidence as the weighted average (weights  $q_{\mathcal{R}}$  and  $q_{\mathcal{S}}$ ) of the original confidence values seems most suitable. If the reference extensions of both sources are completely disjoint, the confidence values of both relations have to be added up (e.g. a person belonging to male actors as well as female actors with a certainty of 0.5 in each case is definitely an actor). Ultimately, if the original reference extensions of both sources partially overlap, an entity belongs to the resulting extension, if it belongs to one of the source extensions. Thus, confidence merging can be traced back to the union in fuzzy set theory by using the maximum operator (see Figure 14).

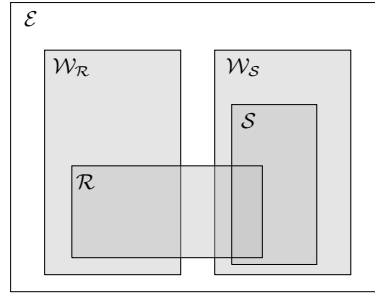


**Figure 14.** The relative size of a fuzzy set representing the union merge of  $\mathcal{R}$  and  $\mathcal{S}$

8. If this strategy is used for merging the confidences of all duplicates (not only *maybe tuples*), the extension of the definite result tuples is only  $Ext((\mathcal{R} \sqcup \mathcal{S})^D) = [Ext(\mathcal{R}^D) - Ext(\mathcal{S}^M)] \cup [Ext(\mathcal{S}^D) - Ext(\mathcal{R}^M)]$  instead of  $Ext((\mathcal{R} \sqcup \mathcal{S})^D) = Ext(\mathcal{R}^D) \cup Ext(\mathcal{S}^D)$ .

As an illustrating example, we consider a union merge of the sample *maybe relation*  $m\_cAct$  introduced in Section 4 and a second *maybe relation*  $euro\_Act$  storing european movie actors. The reference extensions of both relations partially overlaps (european male comedy movie actors). The goal of merging is to get as much as data on movie actors (every origin, gender or main genre) as possible. Thus, the considered entity type  $\mathcal{E}$  is a superset of both reference extensions ( $\mathcal{E} \supset \mathcal{W}_R \cup \mathcal{W}_S$ ). By assuming independence between the tuple memberships of the different source relations (e.g. an actor's origin does not influence his main genre or gender), a merge of these relations is applied by using the maximum operator. If a person is male and comedy movie actor with a confidence of  $cf_1$  and is an european movie actor with a confidence of  $cf_2$ , the confidence that this person is a movie actor is actually  $cf_3 \geq \max(cf_1, cf_2)$ . Nevertheless, since the exact confidence was not measurable during confidence merging, the minimal possible confidence ( $cf_3 = \max(cf_1, cf_2)$ ) has to be taken.

Although the reference extensions of two relations are disjoint, caused by incorrect entries, the extensions of both relations, however, can overlap (see Figure 15). A disjoint situation is given, if we merge the *maybe relation*  $m\_cAct$  with a second *maybe relation*  $f\_cAct$  representing female comedy movie actors. Since a person is either male or female, the confidence values of both relations have to be add up. For instance, if a person is a male comedy movie actor with a confidence of 0.5 and a female comedy movie actor with a confidence of 0.4, this person is an comedy movie actor ( $\mathcal{W}_R \cup \mathcal{W}_S$ ) with a confidence of 0.9.

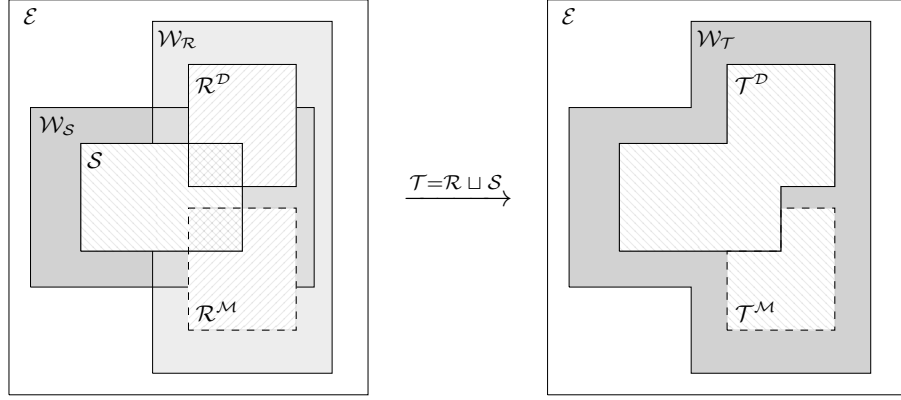


**Figure 15.** Situation of two overlapping relations with disjoint reference extensions

### 7.2.2. Union Merge of one Maybe Relation and one Certain Relation

In order to demonstrate coverage estimation, we firstly consider a union merge of a *maybe relation*  $\mathcal{R}$  and a certain relation  $\mathcal{S}$ . The reference extensions of both relations are assumed to be partially overlapping. This situation is graphically illustrated in Figure 16. Since  $Ext(\mathcal{R}^D)$  and  $Ext(\mathcal{R}^M)$  are per definition disjoint, the coverage of the union merge can be split into :

$$c'(\mathcal{R} \sqcup \mathcal{S}) = c'(\mathcal{R}^D \sqcup \mathcal{S}) + c'(\mathcal{R}^M \sqcup \mathcal{S}) - c(\mathcal{S}) \quad [31]$$



**Figure 16.** *Overlap situation of one maybe relation  $\mathcal{R}$ , one certain relation  $\mathcal{S}$  as well as their reference extensions  $\mathcal{W}_R$  and  $\mathcal{W}_S$  (left); the maybe relation  $\mathcal{T}$  resulting from  $\mathcal{R} \sqcup \mathcal{S}$  and its reference extension  $\mathcal{W}_T$  (right)*

The relation  $\mathcal{S}$  can be in five different overlap situations with  $\mathcal{R}^D$  as well as  $\mathcal{R}^M$  (see Figures 17 and 18). Altogether 18 different overlap situations exist (the combinations  $(D2, M5)$ ,  $(D3, M5)$ ,  $(D4, M5)$ ,  $(D5, M2)$ ,  $(D5, M3)$ ,  $(D5, M4)$ ,  $(D5, M5)$  are not possible).

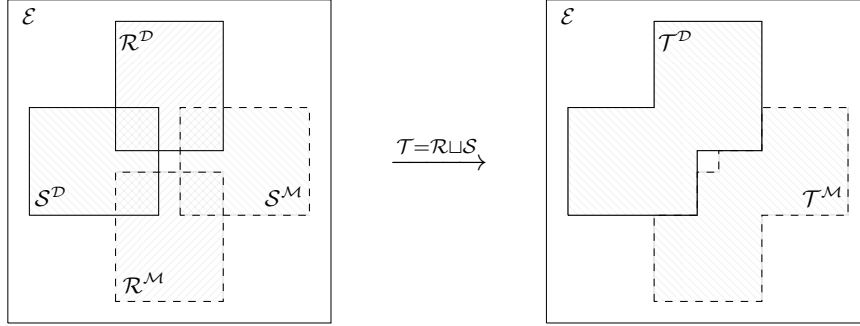
$\mathcal{R}^D \cap^d \mathcal{S}$	$c'(\mathcal{R}^D \sqcup \mathcal{S}) = c(\mathcal{R}^D) + c(\mathcal{S})$	(D1)
$\mathcal{R}^D \cap^i \mathcal{S}$	$c'(\mathcal{R}^D \sqcup \mathcal{S}) = c(\mathcal{R}^D) + c(\mathcal{S}) - c(\mathcal{R}^D) \cdot c(\mathcal{S})$	(D2)
$\mathcal{R}^D \cap^X \mathcal{S}$	$c'(\mathcal{R}^D \sqcup \mathcal{S}) = c(\mathcal{R}^D) + c(\mathcal{S}) -  X / \mathcal{E} $	(D3)
$\mathcal{R}^D \cap^{\supseteq} \mathcal{S}$	$c'(\mathcal{R}^D \sqcup \mathcal{S}) = c(\mathcal{R}^D)$	(D4)
$\mathcal{R}^D \cap^{\subseteq} \mathcal{S}$	$c'(\mathcal{R}^D \sqcup \mathcal{S}) = c(\mathcal{S})$	(D5)

**Figure 17.** *Coverage estimation w.r.t. the five overlap situations of  $\mathcal{R}^D$  and  $\mathcal{S}$*

$\mathcal{R}^M \cap^d \mathcal{S}$	$c'(\mathcal{R}^M \sqcup \mathcal{S}) = c(\mathcal{S}) + A_{\mathcal{R}} c(\mathcal{R}^M)$	(M1)
$\mathcal{R}^M \cap^i \mathcal{S}$	$c'(\mathcal{R}^M \sqcup \mathcal{S}) = c(\mathcal{S}) + A_{\mathcal{R}} [c(\mathcal{R}^M) - c(\mathcal{R}^M) \cdot c(\mathcal{S})]$	(M2)
$\mathcal{R}^M \cap^X \mathcal{S}$	$c'(\mathcal{R}^M \sqcup \mathcal{S}) = c(\mathcal{S}) + A_{\mathcal{R}} [c(\mathcal{R}^M) -  X / \mathcal{E} ]$	(M3)
$\mathcal{R}^M \cap^{\supseteq} \mathcal{S}$	$c'(\mathcal{R}^M \sqcup \mathcal{S}) = c(\mathcal{S}) + A_{\mathcal{R}} [c(\mathcal{R}^M) - c(\mathcal{S})]$	(M4)
$\mathcal{R}^M \cap^{\subseteq} \mathcal{S}$	$c'(\mathcal{R}^M \sqcup \mathcal{S}) = c(\mathcal{S})$	(M5)

**Figure 18.** *Coverage estimation w.r.t. the five overlap situations of  $\mathcal{R}^M$  and  $\mathcal{S}$*

As a representative, we consider the most complex situation, in which  $\mathcal{R}^D$  and  $\mathcal{S}$  as well as  $\mathcal{R}^M$  and  $\mathcal{S}$  are assumed to be independent. Since both reference extensions partially overlap, the extension of the resulting relation definitely contains the extension of  $\mathcal{R}^D$  and  $\mathcal{S}$  and possibly contains this part of the extension of  $\mathcal{R}^M$  which does not belong to the extension of  $\mathcal{S}$  ( $Ext(\mathcal{R}^M) - Ext(\mathcal{S})$ ). The coverage  $c'(\mathcal{R} \sqcup \mathcal{S})^D$



**Figure 19.** *Overlap situation of two maybe relations  $\mathcal{R}$  and  $\mathcal{S}$  (left); the maybe relation  $\mathcal{T}$  resulting from  $\mathcal{S} \sqcup \mathcal{R}$  (right)*

is equal to the sum of the coverage of  $\mathcal{R}^{\mathcal{D}}$  and  $\mathcal{S}$  minus the coverage of the estimated overlap ( $c(\mathcal{R}^{\mathcal{D}}) \cdot c(\mathcal{S})$ ). The coverage  $c'(\mathcal{R} \sqcup \mathcal{S})^{\mathcal{M}}$  is the relative size of the extension only covered by  $\mathcal{R}^{\mathcal{M}}$  weighted with its average tuple membership ( $A_{\mathcal{R}} \cdot [c(\mathcal{R}^{\mathcal{M}}) - c(\mathcal{S}) \cdot c(\mathcal{R}^{\mathcal{M}})]$ ). As a consequence, the coverage of the union merge  $\mathcal{R} \sqcup \mathcal{S}$  results in :

$$c'(\mathcal{R} \sqcup \mathcal{S}) = c(\mathcal{R}^{\mathcal{D}}) + c(\mathcal{S}) - c(\mathcal{R}^{\mathcal{D}}) \cdot c(\mathcal{S}) + A_{\mathcal{R}} \cdot [c(\mathcal{R}^{\mathcal{M}}) - c(\mathcal{S}) \cdot c(\mathcal{R}^{\mathcal{M}})]$$

### 7.2.3. Union Merge of two Maybe Relations

Instead of 18, far more different overlap situations exist, if two *maybe relations* are merged. As an example, we consider the situation that the subrelations of both *maybe relations* are pairwise independent from each other and assume that their reference extensions partially overlap (see Figure 19).

An entity of  $\mathcal{E}$  is represented by a tuple in  $\mathcal{T}^{\mathcal{D}}$ , if there exists an associated tuple belonging to  $\mathcal{R}^{\mathcal{D}}$  or  $\mathcal{S}^{\mathcal{D}}$ . Thus, the coverage of  $\mathcal{T}^{\mathcal{D}}$  results in :

$$c(\mathcal{T}^{\mathcal{D}}) = c(\mathcal{R}^{\mathcal{D}}) + c(\mathcal{S}^{\mathcal{D}}) - c(\mathcal{R}^{\mathcal{D}})c(\mathcal{S}^{\mathcal{D}}) \quad [32]$$

An entity is represented by a tuple in  $\mathcal{T}^{\mathcal{M}}$ , if there exists an associated tuple belonging to  $\mathcal{R}^{\mathcal{M}}$  or  $\mathcal{S}^{\mathcal{M}}$  which does not belong to  $\mathcal{T}^{\mathcal{D}}$ . Thus, the coverage of  $\mathcal{T}^{\mathcal{M}}$  results in :

$$c(\mathcal{T}^{\mathcal{M}}) = (c(\mathcal{R}^{\mathcal{M}}) + c(\mathcal{S}^{\mathcal{M}}) - c(\mathcal{R}^{\mathcal{M}})c(\mathcal{S}^{\mathcal{M}})) \cdot (1 - c(\mathcal{T}^{\mathcal{D}})) \quad [33]$$

In databases with only a simple maybe indication, the coverage of  $\mathcal{T}$  is :

$$c'(\mathcal{T}) = c(\mathcal{T}^{\mathcal{D}}) + \frac{1}{2}c(\mathcal{T}^{\mathcal{M}}) \quad [34]$$

In databases with exact confidence values, the average confidence of the resulting *maybe tuples* need to be estimated. The average confidence of all tuples only belonging to  $\mathcal{R}^{\mathcal{M}}$  is  $A_{\mathcal{R}}$  and the average confidence of all tuples only belonging to  $\mathcal{S}^{\mathcal{M}}$  is  $A_{\mathcal{S}}$ . The confidence  $cf_{\mathcal{T}}$  of a tuple belonging to both source relations depends on the

chosen confidence merging strategy. In databases with exact confidence values, the average confidence of the tuples in  $\mathcal{T}^{\mathcal{M}}$  need to be estimated. The average confidence of all tuples only belonging to  $\mathcal{R}^{\mathcal{M}}$  is  $A_{\mathcal{R}}$  and the average confidence of all tuples only belonging to  $\mathcal{S}^{\mathcal{M}}$  is  $A_{\mathcal{S}}$ . The confidence  $cf_{\mathcal{T}}$  of a tuple belonging to both source relations depends on the chosen confidence merging strategy. The average confidence of these tuples cannot always be exactly derived from the average confidences  $A_{\mathcal{R}}$  and  $A_{\mathcal{S}}$ , but sometimes only can be restricted by an lower and an upper bound (see Figure 20). In our example both reference extensions partially overlap. In this case the confidence of all common tuples is at least greater than their confidence in  $\mathcal{R}$  (or  $\mathcal{S}$  respectively). Consequently, for the average confidence  $A_{\mathcal{R} \cap \mathcal{S}}$  applies :

$$A_{\mathcal{R} \cap \mathcal{S}} \geq A_{\mathcal{R}}, A_{\mathcal{R} \cap \mathcal{S}} \geq A_{\mathcal{S}} \Rightarrow A_{\mathcal{R} \cap \mathcal{S}} \geq \max(A_{\mathcal{R}}, A_{\mathcal{S}}) \quad [35]$$

On the other hand, since each of these confidence values is within the range  $]0, 1[$  (the considered tuples belong to  $\mathcal{T}^{\mathcal{M}}$ ), the maximum confidence of each tuple is definitely lower than the sum of both confidences :

$$A_{\mathcal{R} \cap \mathcal{S}} < \min(1, A_{\mathcal{R}} + A_{\mathcal{S}}) \quad [36]$$

Using the estimated confidence  $A_{\mathcal{R} \cap \mathcal{S}}$ , the coverage  $c'(\mathcal{T})$  can be predicted as :

$$c'(\mathcal{T}) = c(\mathcal{T}^{\mathcal{D}}) + A_{\mathcal{T}} \cdot [c(\mathcal{T}^{\mathcal{M}})] \quad [37]$$

where  $A_{\mathcal{T}}$  is defined as :

$$\begin{aligned} A_{\mathcal{T}} = & [c(\mathcal{R}^{\mathcal{M}}) - c(\mathcal{R}^{\mathcal{M}})c(\mathcal{S})] \cdot A_{\mathcal{R}} + [c(\mathcal{S}^{\mathcal{M}}) - c(\mathcal{R})c(\mathcal{S}^{\mathcal{M}})] \cdot A_{\mathcal{S}} \\ & + [c(\mathcal{R}^{\mathcal{M}})c(\mathcal{S}^{\mathcal{M}})] \cdot A_{\mathcal{R} \cap \mathcal{S}} \end{aligned} \quad [38]$$

### 7.3. Density Estimation

Density estimation for the union merge of certain relations is defined in (Naumann *et al.*, 2004). This estimation can be adapted to *maybe relations* in a similar fashion as we did it for coverage in this paper.

## 8. Conclusion

Current metrics of data completeness are not useable for estimating the completeness of relations with *maybe tuples*. For that reason, we extended these metrics for

---

9. If this strategy is used for merging the confidences of all duplicates (not only *maybe tuples*), the extension of the definite result tuples is reduced from  $Ext((\mathcal{R} \sqcup \mathcal{S})^{\mathcal{D}}) = Ext(\mathcal{R}^{\mathcal{D}}) \cup Ext(\mathcal{S}^{\mathcal{D}})$  to  $Ext((\mathcal{R} \sqcup \mathcal{S})^{\mathcal{D}}) = [Ext(\mathcal{R}^{\mathcal{D}}) - Ext(\mathcal{S}^{\mathcal{M}})] \cup [Ext(\mathcal{S}^{\mathcal{D}}) - Ext(\mathcal{R}^{\mathcal{M}})]$ .



Average confidence estimation		
$A_T = [c(\mathcal{R}^M) - c(\mathcal{R}^M)c(\mathcal{S})] \cdot A_{\mathcal{R}} + [c(\mathcal{S}^M) - c(\mathcal{R})c(\mathcal{S}^M)] \cdot A_{\mathcal{S}}$		
+	Equality	$[c(\mathcal{R}^M)c(\mathcal{S}^D)] \cdot \frac{q_{\mathcal{R}} \cdot A_{\mathcal{R}} + q_{\mathcal{S}}}{q_{\mathcal{R}} + q_{\mathcal{S}}} + [c(\mathcal{R}^D)c(\mathcal{S}^M)] \cdot \frac{q_{\mathcal{R}} + q_{\mathcal{S}} \cdot A_{\mathcal{S}}}{q_{\mathcal{R}} + q_{\mathcal{S}}}$ $+ [c(\mathcal{R}^M)c(\mathcal{S}^M)] \cdot \frac{q_{\mathcal{R}} \cdot A_{\mathcal{R}} + q_{\mathcal{S}} \cdot A_{\mathcal{S}}}{q_{\mathcal{R}} + q_{\mathcal{S}}}$
	Disjointness	$[c(\mathcal{R}^M)c(\mathcal{S}^M)] \cdot A_{\mathcal{R} \cap \mathcal{S}}$ <hr style="border-top: 1px dashed black;"/> $A_{\mathcal{R} \cap \mathcal{S}} > \max(A_{\mathcal{R}}, A_{\mathcal{S}}); A_{\mathcal{R} \cap \mathcal{S}} \leq \min(1, A_{\mathcal{R}} + A_{\mathcal{S}})$
	Overlap	$[c(\mathcal{R}^M)c(\mathcal{S}^M)] \cdot A_{\mathcal{R} \cap \mathcal{S}}$ <hr style="border-top: 1px dashed black;"/> $A_{\mathcal{R} \cap \mathcal{S}} \geq \max(A_{\mathcal{R}}, A_{\mathcal{S}}); A_{\mathcal{R} \cap \mathcal{S}} < \min(1, A_{\mathcal{R}} + A_{\mathcal{S}})$

**Figure 20.** Estimation of average confidence value for merged *maybe tuples* w.r.t. the three different overlap situations of reference extensions

handling the vagueness resulting from the *maybe tuple*-concept. Moreover, we have distinguished two cases of handling *maybe tuples*. In the first case, *maybe tuples* are only indicated as 'maybe'. In the second more exact case every tuple is indicated by an own confidence value. We considered completeness from three different perspectives and hence introduced three corresponding approaches to measure this quality dimension. We compared these metrics w.r.t. different requirements and properties as interpretability or complexity and showed that the metrics of all three approaches supply the same completeness scores whether or not tuple dependencies exist. Finally, we considered completeness composition and proposed strategies for predicting the completeness score of a merged *maybe relation* based on the completeness scores of its sources. Altogether, these results enable an adequate scoring of completeness in databases with *maybe tuples*. Moreover, the presented methods for completeness composition enable an enhanced quality prediction of relations with *maybe tuples* in quality driven query answering (e.g. for discovering the best sources in data integration).

So far, by regarding *maybe tuples*, we only considered uncertainty on tuple level. Nevertheless, in order to enable an adequate completeness scoring of probabilistic- or fuzzy data, uncertainty on attribute value level (e.g. probabilistic- or possibilistic distributions on single values) also has to be taken into account. Consequently, defi-

ning completeness for uncertain attribute values, among others, is an important issue of future work.

Moreover, in this work, we 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 reflections these approaches have to be considered from a more practical (but also more vague) 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 confidence value of a tuple the lower the influence of this tuple on the quality of the associated relation has to be.

Ultimately, for data models using the concept of *maybe tuples*, we think a new quality dimension is required. Different distributions of confidence values and hence different distributions on a relation's true instance can lead to the same completeness score. Thus, the uncertainty resulting from relations with a high number of *maybe tuples* is neither considered by data completeness nor by any other existing quality dimension. Therefore, for estimating the quality of such relations we need a quality dimension which represents the vagueness resulting from a high number of *maybe tuples*.

## 9. Bibliographie

- Barbará D., Garcia-Molina H., Porter D., « The Management of Probabilistic Data », *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 4, n° 5, p. 487-502, 1992.
- Batini C., Scannapieco M., *Data Quality : Concepts, Methodologies and Techniques*, Data-Centric Systems and Applications, Springer, 2006.
- Biskup J., « Extending the Relational Algebra for Relations with Maybe Tuples and Existential and Universal Null Values », *Fundamenta Informaticae*, vol. 7, n° 1, p. 129-150, 1984.
- DeMichiel L. G., « Resolving Database Incompatibility : An Approach to Performing Relational Operations over Mismatched Domains », *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 1, n° 4, p. 485-493, 1989.
- Galindo J., Urrutia A., Piattini M., *Fuzzy Databases - Modeling, Design and Implementation*, Idea Group Publishing, 2006.
- Heinrich B., Kaiser M., Klier M., « Metrics for Measuring Data Quality - Foundations for an Economic Data Quality Management », *International Conference on Software and Data Technologies (ICSOFIT)*, vol. ISDM/EHST/DC, p. 87-94, 2007.
- Motro A., Rakov I., *Not all answers are equally good : Estimating the Quality of Database Answers*, Flexible Query-Answering Systems, Kluwer Academic Publishers, chapter 1, p. 1-21, 1997.
- Naumann F., Freytag J. C., Leser U., « Completeness of Information Sources », *International Workshop on Data Quality in Cooperative Information Systems (DQCIS)*, 2003.

- Naumann F., Freytag J. C., Leser U., « Completeness of Integrated Information Sources », *Information Systems*, vol. 29, n° 7, p. 583-615, 2004.
- Olken F., Random Sampling from Databases, PhD thesis, University of California, 1993.
- Panse F., « Completeness of Attribute Values Representing Partial Information », , <http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:gbv:18-228-7-1478>, 2009.
- Scannapieco M., Batini C., « Completeness in the Relational Model : a Comprehensive Framework », *MIT Conference on Information Quality (IQ)*, p. 333-345, 2004.
- Sen P., Deshpande A., « Representing and Querying Correlated Tuples in Probabilistic Databases », *International Conference on Data Engineering (ICDE)*, p. 596-605, 2007.
- Tseng F. S.-C., Chen A. L. P., Yang W.-P., « Answering Heterogeneous Database Queries with Degrees of Uncertainty », *Distributed and Parallel Databases (DPD)*, vol. 1, n° 3, p. 281-302, 1993.

**A. Proofs :****Preliminaries :**

Given two relations  $\mathcal{R}_1$  and  $\mathcal{R}_2$  with disjoint extensions ( $Ext(\mathcal{R}_1) \cap Ext(\mathcal{R}_2) = \emptyset$ ).

**Theorem 1** *The completeness of their union is equal to the sum of their individual completeness scores.*

**Proof**

$$\begin{aligned}
 comp(\mathcal{R}_1 \cup \mathcal{R}_2) &= \frac{\sum_{t \in \mathcal{R}_1 \cup \mathcal{R}_2} d(t)}{|\mathcal{E}|} \\
 &= \frac{\sum_{t \in \mathcal{R}_1} d(t) + \sum_{t \in \mathcal{R}_2} d(t)}{|\mathcal{E}|} \\
 &= \frac{\sum_{t \in \mathcal{R}_1} d(t)}{|\mathcal{E}|} + \frac{\sum_{t \in \mathcal{R}_2} d(t)}{|\mathcal{E}|} \\
 &= comp(\mathcal{R}_1) + comp(\mathcal{R}_2)
 \end{aligned}$$

**Proof 1 : Equality of Approach 1 and Approach 3 (individual confidence values)**

Given a *maybe relation*  $\mathcal{R}$  with  $\mathcal{I}(\mathcal{R}_C) = \{I_1, I_2, \dots, I_n\}$  possible instances.

**Theorem 2** *The completeness scores of  $\mathcal{R}$  resulting from the metrics of Approach 1 and Approach 3 are equal.*

**Proof**

$$\begin{aligned}
comp'_{A1}(\mathcal{R}) &= \sum_{I_k \in \mathcal{I}(\mathcal{R}_C)} P(\mathcal{R}'_C = I_k) \cdot comp(I_k) \\
&= \frac{1}{|\mathcal{E}|} \sum_{I_k \in \mathcal{I}(\mathcal{R}_C)} P(\mathcal{R}'_C = I_k) \cdot \sum_{t \in I_k} d(t) \\
&=^{10} \frac{1}{|\mathcal{E}|} \sum_{t \in \mathcal{R}_C} cf(t)_{\mathcal{R}} d(t) \\
&= comp'_{A3}(\mathcal{R})
\end{aligned}$$

---

10. In probabilistic databases, per definition, the sum of probabilities of all possible instances a tuple  $t$  belongs to is equal to the probability of  $t$  itself :

$$(\forall t \in \mathcal{R}_C) : cf(t)_{\mathcal{R}} = \sum_{I_k \in \mathcal{I}(\mathcal{R}_C)} P(\mathcal{R}'_C = I_k) \cdot mem(t, I_k)$$

, where  $mem(t, I_k)$  is a boolean function resulting to 1 if  $t$  belongs to  $I_k$  and 0 otherwise.

**Proof 2 : Equality of Approach 2 and Approach 3 (individual confidence values)**

Given a *maybe relation*  $\mathcal{R}$  with  $\mathcal{R}^{\mathcal{M}} = \{t_1, t_2, t_3, \dots, t_n\}$ . For simplification, we assume that each *maybe tuple* has another confidence and that the tuple indexes are ordered by the tuples' confidences ( $(\forall i, j \in \{1, 2, \dots, n\}) : i < j \Rightarrow cf(t_i) < cf(t_j)$ ). For the sake of convenience the confidence of tuple  $t_i$  is abbreviated as  $cf_i$ . The relation  $\mathcal{R}_i$  is defined to be a one-tuple relation only containing the tuple  $t_i$ .

**Theorem 3** *The completeness scores of  $\mathcal{R}$  resulting from the metrics of Approach 2 and Approach 3 are equal.*

**Proof**

$$\begin{aligned}
comp'_{A3}(\mathcal{R}) &= \int_{0.0}^{1.0} comp(\hat{\sigma}^\alpha(\mathcal{R}))d\alpha \\
&= \int_{0.0}^{cf_1} comp(\hat{\sigma}^\alpha(\mathcal{R}))d\alpha + \int_{cf_1}^{cf_2} comp(\hat{\sigma}^\alpha(\mathcal{R}))d\alpha \\
&+ \int_{cf_2}^{cf_3} comp(\hat{\sigma}^\alpha(\mathcal{R}))d\alpha + \dots + \int_{cf_n}^{1.0} comp(\hat{\sigma}^\alpha(\mathcal{R}))d\alpha \\
&= \int_{0.0}^{cf_1} comp(\mathcal{R})d\alpha + \int_{cf_1}^{cf_2} comp(\mathcal{R} \setminus \{t_1\})d\alpha \\
&+ \int_{cf_2}^{cf_3} comp(\mathcal{R} \setminus \{t_1, t_2\})d\alpha + \dots + \int_{cf_n}^{1.0} comp(\mathcal{R}^{\mathcal{D}})d\alpha \\
&= cf_1 \cdot comp(\mathcal{R}) + (cf_2 - cf_1) \cdot comp(\mathcal{R} \setminus \{t_1\}) \\
&+ (cf_3 - cf_2) \cdot comp(\mathcal{R} \setminus \{t_1, t_2\}) + \dots + (1 - cf_n) \cdot comp(\mathcal{R}^{\mathcal{D}}) \\
&= cf_1 \cdot comp(\mathcal{R}_1) + cf_2 \cdot comp(\mathcal{R} \setminus \{t_1\}) \\
&+ (cf_3 - cf_2) \cdot comp(\mathcal{R} \setminus \{t_1, t_2\}) + \dots + (1 - cf_n) \cdot comp(\mathcal{R}^{\mathcal{D}}) \\
&= cf_1 \cdot comp(\mathcal{R}_1) + cf_2 \cdot comp(\mathcal{R}_2) + cf_3 \cdot comp(\mathcal{R}_3) \\
&+ \dots + cf_n \cdot comp(\mathcal{R}_n) + comp(\mathcal{R}^{\mathcal{D}}) \\
&= cf_1 \frac{d(t_1)}{|\mathcal{E}|} + cf_2 \frac{d(t_2)}{|\mathcal{E}|} + cf_3 \frac{d(t_3)}{|\mathcal{E}|} + \dots + cf_n \frac{d(t_n)}{|\mathcal{E}|} + \frac{\sum_{t \in \mathcal{R}^{\mathcal{D}}} d(t)}{|\mathcal{E}|} \\
&= \frac{cf_1 \cdot d(t_1) + cf_2 \cdot d(t_2) + cf_3 \cdot d(t_3) + \dots + cf_n \cdot d(t_n) + \sum_{t \in \mathcal{R}^{\mathcal{D}}} d(t)}{|\mathcal{E}|} \\
&= \frac{\sum_{t \in \mathcal{R}_c} cf(t)d(t)}{|\mathcal{E}|} = comp'_{A1}(\mathcal{R})
\end{aligned}$$

**Proof 3 : Equality of Approach 1 and Approach 3 (simple maybe indication)**

Given a *maybe relation*  $\mathcal{R}$  with  $\mathcal{I}(\mathcal{R}_c) = \{I_1, I_2, \dots, I_n\}$  possible instances.

**Theorem 4** *The completeness scores of  $\mathcal{R}$  resulting from the metrics of Approach 1 and Approach 3 are equal.*

**Proof**

$$\begin{aligned}
 comp'_{A1}(\mathcal{R}) &= \frac{1}{2^{|\mathcal{R}_c^{\mathcal{M}}|}} \frac{1}{|\mathcal{E}|} \sum_{S_i \in \mathcal{I}(\mathcal{R}_c)} \sum_{t \in S_i} d(t) \\
 &=^{11} \frac{1}{2^{|\mathcal{R}_c^{\mathcal{M}}|}} \frac{1}{|\mathcal{E}|} (|\mathcal{I}(\mathcal{R}_c)| \sum_{t \in \mathcal{R}_c^{\mathcal{D}}} d(t) + \frac{1}{2} |\mathcal{I}(\mathcal{R}_c)| \sum_{t \in \mathcal{R}_c^{\mathcal{M}}} d(t)) \\
 &= \frac{|\mathcal{I}(\mathcal{R}_c)|}{2^{|\mathcal{R}_c^{\mathcal{M}}|}} \frac{1}{|\mathcal{E}|} \left( \sum_{t \in \mathcal{R}_c^{\mathcal{D}}} d(t) + \frac{1}{2} \sum_{t \in \mathcal{R}_c^{\mathcal{M}}} d(t) \right) \\
 &=^{12} \frac{\sum_{t \in \mathcal{R}_c^{\mathcal{D}}} d(t)}{|\mathcal{E}|} + \frac{\frac{1}{2} \sum_{t \in \mathcal{R}_c^{\mathcal{M}}} d(t)}{|\mathcal{E}|} \\
 &= comp(\mathcal{R}^{\mathcal{D}}) + \frac{1}{2} comp(\mathcal{R}^{\mathcal{M}}) \\
 &= comp'_{A3}(\mathcal{R})
 \end{aligned}$$

---

11. Each possible instance contains all tuples belonging to  $\mathcal{R}^{\mathcal{D}}$ . In contrast, without tuple dependencies each *maybe tuple* belongs to exact the half of all possible instances. Thus, the density of every definite tuple is added up for  $|\mathcal{I}(\mathcal{R}_c)|$  times and the density of every definite tuple is added up for  $1/2 |\mathcal{I}(\mathcal{R}_c)|$  times.

12. Without tuple dependencies, the number of possible instances is  $2^{|\mathcal{R}_c^{\mathcal{M}}|}$ .