# Parsimonious Representation of Knowledge Uncertainty using Metadata about Validity and Completeness

Célia da Costa Pereira, Didier Dubois, Henri Prade, Andrea G. B. Tettamanzi

# Parsimonious Representation of Knowledge Uncertainty Using Metadata About Validity and Completeness

Célia da Costa Pereira[1] [a], Didier Dubois[2] [b], Henri Prade[2] [c] and Andrea G. B. Tettamanzi[3] [d]

[1] *Université Côte d'Azur, CNRS, I3S, Sophia Antipolis, France*
[2] *IRIT – CNRS, 118, route de Narbonne, Toulouse, France*
[3] *Université Côte d'Azur, Inria, CNRS, I3S, Sophia Antipolis, France*
*celia.pereira@unice.fr, dubois@irit.fr, prade@irit.fr, andrea.tettamanzi@univ-cotedazur.fr*

Abstract: We investigate how metadata about the uncertainty of knowledge contained in a knowledge base can be expressed parsimoniously and used for reasoning. We propose an approach based on possibility theory, whereby a classical knowledge base plus metadata about the degree of validity and completeness of some of its portions are used to represent a possibilistic belief base. We show how reasoning on such belief base can be done using a classical reasoner.

## 1 INTRODUCTION AND RELATED WORK

In general, the process of getting a piece of information from a Knowledge Base (KB) is driven by practical purposes—such a piece of information will be used to justify certain decisions, for example. Its quality has therefore an important role to play in the success of the decisions made. The quality of a piece of information can be measured by considering different dimensions. Most contributions in the literature concentrate exclusively on the amount of true (known) facts in a KB for assessing its quality. (Wick et al., 2013), for example, propose several algorithms for estimating a value of confidence based on the probability of a fact in a KB being true. In (Dong et al., 2014), the authors studied the applicability of data fusion techniques to solve the problem of knowledge base feeding. The criterion they used to construct quality knowledge bases was to identify the true values of data items among multiple observed values provided from different (and maybe unknown) sources with different reliabilities. However, as it has been pointed out for example by (Razniewski et al., 2016),

> While quite some facts are known about the

world, little is known about how much is unknown.

In other words, a knowledge base is in general incomplete. This obviously has an impact on the overall quality of a KB—the more it is incomplete the lesser is its quality and the more the pieces of information extracted from it have to be considered (used) with caution.

The problem of representing *both* validity and completeness has begun to be dealt with for information stored in databases many years ago before being addressed for knowledge bases (KBs). For example, we can consider the model of database integrity proposed by (Motro, 1989) and the work by (Demolombe, 1996), who used modal logic for reasoning about validity and completeness of information stored in relational databases as precursors of some ideas that have later been adopted for KBs. However, the representation of the incompleteness in information stored in the databases has been inspired by earlier works on the representation of incompleteness in knowledge bases as the ones proposed by (Levesque, 1980; Levesque, 1982) and the one proposed by (Collins et al., 1975) for reasoning with this kind of knowledge bases.

Recent work on annotating KBs with metadata about their completeness has been done, in the context of the semantic Web, by (Darari et al., 2013; Razniewski et al., 2016), who studied the way in which statements about completeness can be used

---

[a] [ID] https://orcid.org/0000-0001-6278-7740
[b] [ID] https://orcid.org/0000-0002-6505-2536
[c] [ID] https://orcid.org/0000-0003-4586-8527
[d] [ID] https://orcid.org/0000-0002-8877-4654

when answering queries. According to their approach, it is then possible, given a statement about a topic, to specify if information about it in the base is complete or not. However, the gradual view of completeness in data sources has not been considered.

Solutions to construct a possibilistic belief base from a crisp KB using topical validity and completeness metadata, like (da Costa Pereira et al., 2017), suffer from some limitations, mainly due to the fact that, in order to guarantee consistency, they have to sacrifice much of the expressive power of the knowledge representation language. In particular, the "facts" that can be recorded in the knowledge base are restricted only to ground formulas without negation and disjunction. Indeed, *negative information* (i.e., facts that do not hold) is critical for the correctness of queries involving negation (Razniewski et al., 2016). Completeness and negative information are closely related: if we know that a portion of a KB is complete, it is as if we knew an infinity of negated facts (all those relevant to that portion that are not in the KB).

Motivated by the above considerations, we want to answer the following research question: *is it possible that a classical knowledge base plus metadata information on the (gradual) validity and completeness with respect to a few configurations (groups, portions, subjects, topics of statements it contains), enables one to represent a possibilistic belief base and perform possibilistic inferences by using a classical reasoner?* This research question leads us to the following sub-question: *which should be a suitable definition for such a configuration which in turn will allow us to define appropriate* validity *and* completeness *functions?*

We propose a framework based on possibility theory to represent and reason about gradual notions of validity and completeness in KBs. Since it would be impractical to associate values of possibility and necessity to each single assertion in a knowledge base (KB), we show that, thanks to validity and completeness metadata, it is possible to express degrees of possibility/necessity for formulas entailed by the KB in a parsimonious way (i.e., without having to associate a weight to each single formula) and to perform possibilistic inferences on top of a classical KB, considering, in addition to possibilistic uncertainty, also negative information.

In particular, we present a way to represent validity and completeness information (with respect to particular *slices* of information) in a knowledge base in a way that permits said validity and completeness information to simulate the knowledge base as a possibilistic knowledge base, where the possibilistic information is derived exclusively from the validity and completeness. The advantage of this approach is that the validity and completeness is only assessed at the slice level, where in a possibilistic knowledge base, the possibility distribution needs to be defined for all facts. Therefore, when the number of slices is much smaller than the number of facts, the proposed representation is much more parsimonious.

The paper is organized as follows: Section 2 gives then some background about the formal tools we use. We present our proposal in Section 3, which explains how (gradual) validity and completeness are related to the beliefs of an agent. Finally, Section 4 discusses some possible applications of our proposal.

## 2 BACKGROUND

We first provide a brief refresher on possibility theory, before recalling the basics of possibilistic logic, a logic where classical formulas are weighted in terms of certainty.

### 2.1 Possibility Theory

Fuzzy sets (Zadeh, 1965) are sets whose elements have degrees of membership in $[0,1]$. Possibility theory (Dubois and Prade, 1988) is a mathematical theory of uncertainty that relies upon fuzzy set theory, in that the (fuzzy) set of possible values for a variable of interest is used to describe the uncertainty as to its precise value. At the semantic level, the membership function of such set, $\pi$, is called a *possibility distribution* and its range is $[0,1]$. A possibility distribution can represent the available knowledge of an agent. $\pi(I)$ represents the degree of compatibility of the interpretation $I$ with the available knowledge about the real world if we are representing uncertain pieces of knowledge. By convention, $\pi(I) = 1$ means that it is totally possible for $I$ to be the real world, $1 > \pi(I) > 0$ means that $I$ is only somehow possible, while $\pi(I) = 0$ means that $I$ is certainly not the real world.

A possibility distribution $\pi$ is said to be normalized if there exists at least one interpretation $I_0$ s.t. $\pi(I_0) = 1$, i.e., there exists at least one possible situation which is consistent with the available knowledge.

**Definition 1.** *(Possibility and Necessity Measures) A possibility distribution $\pi$ induces a* possibility measure *and its dual* necessity measure*, denoted by $\Pi$ and $N$ respectively. Both measures apply to a classical set $S \subseteq \Omega$ and are defined as follows:*

$$\Pi(S) = \max_{I \in S} \pi(I); \tag{1}$$

$$N(S) = 1 - \Pi(\bar{S}) = \min_{I \in \bar{S}}\{1 - \pi(I)\}. \tag{2}$$

In words, $\Pi(S)$ expresses to what extent $S$ is consistent with the available knowledge. Conversely, $N(S)$ expresses to what extent $S$ is entailed by the available knowledge. It is equivalent to the impossibility of its complement $\bar{S}$—the more $\bar{S}$ is impossible, the more $S$ is certain. A few properties of $\Pi$ and $N$ induced by a normalized possibility distribution on a finite universe of discourse $\Omega$ are the following. For all subsets $A, B \subseteq \Omega$:

1. $\Pi(A \cup B) = \max\{\Pi(A), \Pi(B)\}$;
2. $\Pi(A \cap B) \leq \min\{\Pi(A), \Pi(B)\}$;
3. $\Pi(\emptyset) = N(\emptyset) = 0; \quad \Pi(\Omega) = N(\Omega) = 1$;
4. $N(A \cap B) = \min\{N(A), N(B)\}$;
5. $N(A \cup B) \geq \max\{N(A), N(B)\}$;
6. $\Pi(A) = 1 - N(\bar{A})$ (duality);
7. $N(A) > 0 \Rightarrow \Pi(A) = 1; \quad \Pi(A) < 1 \Rightarrow N(A) = 0$;

A consequence of these properties is that $\max\{\Pi(A), \Pi(\bar{A})\} = 1$. In case of complete ignorance on $A$, $\Pi(A) = \Pi(\bar{A}) = 1$.

## 2.2 Possibilistic Logic

Before going into details about possibilistic logic, we would like to put forward our motivation for such a logic for handling uncertainty in this work. Information is often pervaded with uncertainty, and it may be convenient to associate pieces of information with certainty levels. These certainty levels can often be qualitatively assessed only using a finite completely ordered scale ranging from "fully certain" to "not certain at all", with intermediary levels such as "almost certain", or "somewhat certain". Possibility theory offers such a qualitative setting, when a finite subset of $[0,1]$ including 0 and 1 is used and then only the ordering of the degrees in $[0,1]$ is meaningful, in agreement with the use of max and min operators. Moreover, the inverse mapping $1 - (\cdot)$ exchanges the necessity scale with a possibility scale, such as "fully possible", "quite possible", "somehow possible", "not possible at all (= impossible)". In the following, the pieces of information are associated with certainty levels which are viewed as lower bounds of necessity measures. Then, the min-decomposability of necessity measures with respect to conjunction acknowledges the fact that to be certain at least at some level $\alpha$ that a conjunction of facts is true, we should be certain at least at level $\alpha$ that the truth of each fact is certain at least at level $\alpha$.

Possibilistic logic (Dubois et al., 1994) has been originally motivated by the need to manipulate syntactic expressions of the form $(\phi, \alpha)$ where $\phi$ is a classical logic formula, and $\alpha$ is a certainty level, with the intended semantics that $N(\phi) \geq \alpha$, where $N$ is a necessity measure. It is then possible to consider that all the propositions of the considered language can be totally ordered on a given scale. In our case, propositions are formulas. Besides, in possibilistic logic, a level of inconsistency can be associated with a knowledge base as recalled now.

A possibilistic knowledge base $B$ is a set of possibilistic logic formulas $\{(\phi_i, \alpha_i) \mid i = 1, \ldots, m\}$. Clearly, $B$ can be layered into a set of nested classical bases $B_\alpha = \{\phi_i \mid (\phi_i, \alpha_i) \in B \text{ and } \alpha_i \geq \alpha\}$ such that $B_\alpha \subseteq B_\beta$ if $\alpha \geq \beta$. Proving syntactically $B \vdash (\phi, \alpha)$ amounts to proceeding by refutation and proving $B \cup \{(\neg\phi, 1)\} \vdash (\bot, \alpha)$ by repeated application of the resolution rule $(\neg\phi \vee \psi, \alpha), (\phi \vee \nu, \beta) \vdash (\psi \vee \nu, \min(\alpha, \beta))$. Moreover, $B \vdash (\phi, \alpha)$ if and only if $B_\alpha \vdash \phi$ and $\alpha > inc(B)$, where $inc(B)$ is inconsistency level of $B$ defined as $inc(B) = \max\{\alpha \mid B \vdash (\bot, \alpha)\}$. It can be shown that $inc(B) = 0$ iff $B^*$ is consistent, with $B^* = \{\phi_i \mid (\phi_i, \alpha_i) \in B\}$. Thus reasoning from a possibilistic base just amounts to reasoning classically with subparts of the base whose formulas are strictly above the certainty level.

A possibilistic knowledge base $B = \{(\phi_i, \alpha_i) \mid i = 1, \ldots, m\}$ encodes the constraints $N(\phi_i) \geq \alpha_i$. $B$ is thus semantically associated with a possibility distribution (Dubois et al., 1994)

$$\pi_B(I) = \min_{i=1,\ldots,m} \max(\phi_i^I, 1 - \alpha_i),$$

where $\phi_i^I = 1$ if $I$ is a model of $\phi_i$, and $\phi_i^I = 0$ otherwise. As it can be seen, $\pi_B(I)$ is all the larger as the interpretation $I$ makes false only formulas with low certainty levels. $\pi_B$ is the largest possibility distribution, i.e., the least committed distribution assigning the largest possibility levels in agreement with the constraints $N(\phi_i) \geq \alpha_i$ for $i = 1, \ldots, m$. The distribution $\pi_B$ rank-orders the interpretations $I$ of the language induced by the $\phi_i$'s according to their plausibility on the basis of the strength of the pieces of information in $B$. If the set of formulas $B^*$ is consistent then the distribution $\pi_B$ is normalized (i.e., $\exists I, \pi_B(I) = 1$). The semantic entailment is defined by $B \models (\phi, \alpha)$ iff $\forall I, \pi_B(I) \leq \pi_{\{(\phi, \alpha)\}}(I)$. Reasoning by refutation in propositional possibilistic logic is sound and complete, applying the syntactic resolution rule. Namely, it can be shown that $B \models (\phi, \alpha)$ iff $B \vdash (\phi, \alpha)$ and $inc(B) = 1 - \max_I \pi_B(I)$.

Algorithms for reasoning in possibilistic logic and an analysis of their complexity, which is similar to the one of classical logic, multiplied by the logarithm of the number of levels used in the necessity scale, can be found in (Lang, 2001).

# 3 REPRESENTING AND REASONING WITH VALIDITY AND COMPLETENESS

As it was hypothesized by (Motro, 1989) for the case of relational databases, here, to formalize the concepts of validity and completeness in the the case of knowledge bases, we shall assume the existence of a *hypothetical knowledge base* that captures a designated environment of the real world perfectly. The knowledge base $K$ mentioned in the paper is then an approximation of such hypothetical knowledge base.

When dealing with relational databases, only the statements explicitly present in the database are considered as true (valid). The others are considered as false (closed world assumption). When dealing with sets of formulas, the true statements are those explicitly represented in the dataset, plus those which can be inferred thanks to a reasoner. However, due to the open world assumption, we cannot suppose that the other statements are false—the truth status of some statements may be unknown in case of incomplete knowledge.

In this section, we recall the notions of validity and completeness, first introduced in (Demolombe, 1996), and made gradual in the setting of possibilistic logic (Dubois and Prade, 1997) for dealing with relational databases and adapt them to the more general setting of knowledge bases, where (i) the open world assumption holds, (ii) implicit knowledge can be inferred by logical deduction, and (iii) negative information is also taken into account unlike what was proposed in (da Costa Pereira et al., 2017).

It is often the case that the knowledge contained in a knowledge base is not all certain to the same degree. There will be statements whose truth is absolutely certain. This might be the case of ontological axioms or integrity constraints. Other groups of statements, obtained for example from the same source or covering the same subject, might have the same degree of certainty, but statements from different portions of the knowledge base might be believed with greater or lower certainty.

Our working hypothesis is that, as suggested by (da Costa Pereira et al., 2017), the degree of certainty of every piece of information depends on the degree to which the knowledge base is valid and complete with respect to all groups, portions, subjects, topics (or whatever else we wish to call them) of statements it contains. We think that an intuitive name for this notion of a semantically determined homogeneous portion of a knowledge base may be a *slice* and we will stick to this term from now on.

While it is true that the term *slice* might lead to confusion with the same term as used in the hypercube data model (Gray et al., 1997), as a matter of fact, the suggestion that a *slice* may be a subset defined by fixing one or more dimensions constitutes a good and useful intuition. Indeed, if we interpret slices as knowledge "topics" or "domains", then this is exactly what slices are, with the specificity that here every "dimension" can be viewed as a binary truth assignment to a formula, e.g., in a hypotetical knowledge base about travel, "$x$ is a flight and $x$ departs from London", thus giving the slice of the knowledge base that provides information about flights departing from London.

## 3.1 Postulates

To be able to talk about the validity and completeness of information stored in a knowledge base with respect to a particular *slice*, we need a formal way of defining the latter. We begin with the most general and neutral definition, whereby a slice $T$ is just a set of formulas. A more precise definition is deferred to when we will have discussed the properties that a slice must satisfy.

A few basic postulates for slices, based on common sense arguments, are the following.

**P1** Slices are non-empty. We assume slices are defined by the designer or a user of a knowledge base in order to state metadata about the validity and completeness of portions of knowledge in the base; defining an empty slice would defeat its purpose.

**P2** Slices are all distinct. Defining two equivalent slices would be redundant and of no practical use; therefore, we can safely bar this possibility.

**P3** For every slice not contained in another slice (we may call it a "top-level" slice), there exists a formula entailed by the knowledge base that only belongs in that slice and in no other slice.

What justifies stating Postulate P3 is that for every portion of a knowledge base a knowledge engineer might want to define in order to state metadata on it, one would expect that either that portion is a proper subset of another portion (i.e., a sort of sub-topic or sub-domain), or, if it is not, then its very definition is motivated by the existence of some facts that are not covered by other slices. For instance, in a knowledge base about travel, I might want to define a slice about "airports" because there are assertions involving airports, like "London Heathrow (LHR) has four operational terminals", that do not deal with any other possible slices, like "flights", "airlines", "aircraft", and so

on. Or I might choose to define "aviation", which includes all of them, including the assertion about LHR, which is not covered by any other existing slice.

Let $K$ be a set of formulas in a decidable logical language $\mathcal{L}$, for which there exists a reasoner capable of performing inferences and deduce other formulas which are not explicitly contained in $K$.

Under the closed-world hypothesis typical of databases, which is the setting in which (Dubois and Prade, 1997) was stated, it would be reasonable to admit that what cannot be deduced from an agent's knowledge base corresponds to what the agent believes to be false.

However, in the case of a knowledge base, the open-world assumption holds and the agent is capable of performing logical inferences (e.g., thanks to a reasoner). Therefore, we must think in terms of logical entailment of formulas.

Without loss of generality, we will assume a Herbrand semantics for $\mathcal{L}$.

**Definition 2.** *The Herbrand base of $\mathcal{L}$ is the set $H_{\mathcal{L}}$ of all ground atoms in $\mathcal{L}$. An interpretation (or model) is a function $I : H_{\mathcal{L}} \to \{0, 1\}$, which can also be viewed as a subset of the Herbrand base, $I \subseteq H_{\mathcal{L}}$ (the set of all atoms $\phi$ such that $\phi^I = 1$). We denote $\Omega = 2^{H_{\mathcal{L}}}$ the set of all interpretations.*

We write $K \models \phi$ to denote the fact that formula $\phi$ is a logical consequence of all the formulas in $K$. Assuming the usual definition of satisfaction (given an interpretation $I \in \Omega$ and a formula $\phi \in \mathcal{L}$, $I \models \phi$ if and only if $\phi$ evaluates to *true* in $I$), we define the notion of entailment as follows: $K \models \phi$ if and only if, for every interpretation $I \in \Omega$, $I \models K$ implies $I \models \phi$.

Using a sound and complete reasoner, if $K \models \phi$, $\phi$ can also be deduced from $K$ by the agent (which we write $K \vdash \phi$), whereas if $K \nvdash \phi$ ($\phi$ cannot be deduced from $K$), this means that $K \nvDash \phi$ ($\phi$ is not a logical consequence of $K$). Finally, given a set $S$ of formulas, $K \models S$ if and only if $\forall \phi \in S, K \models \phi$.

## 3.2 Graded Validity and Completeness

The purpose of a knowledge base is to store axioms and assertions that summarize an agent's knowledge about the world or, at least, a limited portion of the world, which is relevant to the problem the agent is intended to deal with. We will take an objectivist stance by assuming that there exists, among all the possible interpretations on language $\mathcal{L}$, one that reflects the actual state of affairs. Let us denote such interpretation by $I^*$. Then we may say that the objective truth of any formula $\phi \in \mathcal{L}$ is given by $\phi^{I^*}$. To be absolutely clear about that, we are assuming that $I^*$ is real and

an objective truth exists for every formula, independently of a knowledge base $K$ and of the agent using it. As a matter of fact, the knowledge represented in (a slice of) $K$ might (and will, in general) not reflect reality perfectly or accurately.

Given this premise, the notions of *validity* and *completeness* of a knowledge base $K$ with respect to a slice may be defined as follows:

- $K$ is *valid* with respect to a slice iff, for every formula $\phi$ in that slice, $K \models \phi$ implies that $I^* \models \phi$, i.e., $\phi$ is objectively true;

- $K$ is *complete* with respect to a slice iff, for every formula $\phi$ in that slice, $K \nvDash \phi$ implies that $I^* \nvDash \phi$, i.e., $\phi$ is objectively false.

As pointed out in Section 1, we will use the term *beliefs* to refer to (possibly partial, incomplete, or invalid) information held by an agent. An agent may then believe something to different degrees. We suppose that these degrees depend on both the degree of completeness of the sets of statements and on the reliability or trustworthiness of the information source. For example (da Costa Pereira et al., 2017), information related to an Air France flight should be complete if the source is the Air France carrier itself. However, the completeness could be lower if the source is a private travel agency with a partial coverage about the current flights from the different companies including those of Air France. Similarly, the degree of trust to be associated with information fed by a clerk should be lower than the one to be associated with information fed by a supervisor. Still, we would like to stress that the way in which such degrees are obtained is out of the scope of this paper.

As pointed out in Section 2, possibility theory is well suited to model degrees of certainty or, dually, degrees of possibility. Besides, possibility theory, unlike other theories of uncertainty like probability theory, is well -suited to model total ignorance which is necessary to represent situations in which we have, for example, both $K \nvDash \phi$ and $K \nvDash \neg\phi$. This is the reason why, here, we adopt this theory to represent the gradual property of both the reliability of an information source as well as the completeness of information regarding a particular slice.

We assume that $K$ is a consistent, classical (as opposed to possibilistic) knowledge base, i.e., $K$ contains statements (such as axioms and assertions) expressed in one of the decidable logical languages usually employed to represent knowledge in practical applications (examples might be Datalog, description logics, RDF + RDFS, or one of the profiles of OWL).

We assume that, in addition to $K$, metadata about validity and completeness of information stored in $K$ are given in the form of two functions, **Val** and **Comp**,

which associate a degree of validity and completeness, respectively, to a number of slices defined on $K$. Let $\mathcal{S}_K \subset 2^L$ the set of such slices. In practice, these two functions might be implemented by a lookup table, listing their values for each defined slice.

**Definition 3.** *Let* $\mathbf{Val} : \mathcal{S}_K \to [0,1]$ *be such that, for each slice* $T \in \mathcal{S}_K$, $\mathbf{Val}(T)$ *is the degree to which $K$ contains valid information about slice $T$, which means, for all formulas $\phi$ such that $K \models \phi$ and $\phi \in T$,*

$$N(\phi) \geq \mathbf{Val}(T).$$

Intuitively, if we can deduce $\phi$ from the knowledge base $K$, and $\phi$ belongs in slice $T$, and we know that the source who fed $K$ is (somehow) reliable for domain of $T$ then the agent should believe $\phi$ at least as much as the degre to which the source is reliable. If the source is fully reliable then $\phi$ will be certain for the agent.

**Definition 4.** *Let* $\mathbf{Comp} : \mathcal{S}_K \to [0,1]$ *be such that, for each slice* $T \in \mathcal{S}_K$, $\mathbf{Comp}(T)$ *is the degree to which $K$ contains complete information about slice $T$, which means, for all formulas $\phi$ such that $K \not\models \phi$ and $\phi \in T$,*

$$\Pi(\phi) \leq 1 - \mathbf{Comp}(T).$$

Intuitively, if we cannot deduce $\phi$ from the knowledge base $K$, and $\phi$ belongs in slice $T$, and we know that information in $K$ about slice $T$ is (somehow) complete, then $\phi$ should be certainly false.

It is reasonable to assume that, given two slices $T$ and $T'$,

$$T \subseteq T' \quad \Rightarrow \quad \mathbf{Val}(T) \geq \mathbf{Val}(T'), \qquad (3)$$
$$T \subseteq T' \quad \Rightarrow \quad \mathbf{Comp}(T) \geq \mathbf{Comp}(T'). \qquad (4)$$

Indeed, if we are told that $K$ contains reliable (resp. complete) information about a broader domain $T'$ to a given degree $\alpha$, then $K$ cannot be less reliable (complete) about a narrower (i.e., more specific) domain $T$; if anything, it might be *more* reliable (complete) about $T$ if a more reliable (complete) source is available just for $T$.

The extent to which the agent believes $\phi$ depends on (i) what is supposed to be known about $\phi$ — can we deduce $\phi$ from $K$? —, and (ii) on the validity and completeness of $K$ with respect to the slices that contain $\phi$. That being the case, $K$, together with the metadata provided by $\mathbf{Val}$ and $\mathbf{Comp}$, should allow us to compute the degree of possibility and necessity for any arbitrary formula $\phi$, as follows:

$$\Pi(\phi) = \begin{cases} 1, & \text{if } K \models \phi, \\ \min_{T:\phi \in T}\{1 - \mathbf{Comp}(T)\}, & \text{otherwise}; \end{cases} \quad (5)$$

$$N(\phi) = \begin{cases} \max_{T:\phi \in T} \mathbf{Val}(T), & \text{if } K \models \phi, \\ 0, & \text{otherwise}. \end{cases} \quad (6)$$

Let us call $\pi$ the hypothetical possibility distribution that induces the possibility and necessity measures

of Equations 5 and 6 and let $B$ a hypothetical possibilistic belief base corresponding to it. Furthermore, among all possibility distributions compatible with $\Pi$ and $N$, we will select the one that makes the least commitment, i.e., the maximal (most general) one.

## 3.3 Existence Conditions

We now derive a necessary condition for the existence of such a possibility distribution $\pi$.

Let $\phi$ be such that $K \models \phi$; if $K$ is consistent, $K \not\models \neg\phi$. By the duality property of the possibility and necessity measures, it must be $\Pi(\phi) = 1 - N(\neg\phi)$; this is satisfied by Equations 5 and 6, since $\Pi(\phi) = 1$ and $N(\neg\phi) = 0$.

**Proposition 1.** *Let us assume that $K$ is consistent and that there exists a possibility distribution $\pi$ that induces the possibility and necessity measures of Equations 5 and 6. Then, for all $\phi$ such that $K \models \phi$,*

$$\max_{T:\phi \in T} \mathbf{Val}(T) = \max_{T':\neg\phi \in T'} \mathbf{Comp}(T'). \qquad (7)$$

*Proof.* If measures $N$ and $\Pi$ are induced by the same possibility distribution $\pi$, it must be $N(\phi) = 1 - \Pi(\neg\phi)$; therefore, by Equations 5 and 6, we can write

$$\begin{aligned} \max_{T:\phi \in T} \mathbf{Val}(T) &= N(\phi) = 1 - \Pi(\neg\phi) \\ &= 1 - \min_{T':\neg\phi \in T'}\{1 - \mathbf{Comp}(T')\} \\ &= \max_{T':\neg\phi \in T'} \mathbf{Comp}(T'), \end{aligned}$$

which proves the thesis. $\square$

A formula $\phi$ such that $K \not\models \phi$ and $K \not\models \neg\phi$ poses no problem, because $\mathbf{Comp}$ and $\mathbf{Val}$ do not interact:

$$\Pi(\phi) = \min_{T:\phi \in T}\{1 - \mathbf{Comp}(T)\},$$
$$\Pi(\neg\phi) = \min_{T':\neg\phi \in T'}\{1 - \mathbf{Comp}(T')\},$$
$$N(\phi) = N(\neg\phi) = 0.$$

We now prove that a formula and its negation cannot belong in the same slice, unless the two functions $\mathbf{Val}$ and $\mathbf{Comp}$ are identical.

**Proposition 2.** *Either $\mathbf{Val}(T) = \mathbf{Comp}(T)$ for all slice $T$, or, for all formula $\phi$ and for all slice $T$, $\phi \in T \Rightarrow \neg\phi \notin T$.*

*Proof.* By contradiction: we show that if $\phi \in T$ and $\neg\phi \in T$, a contradiction can be derived. Let $\phi$ be a formula belonging in just one slice $T$. If the slices for which $\mathbf{Val}$ or $\mathbf{Comp}$ are defined as all distinct, there will always be at least one such formula. If not, the equivalent slices can be merged together so that all slices are distinct. By the assumption, $\neg\phi \in T$

too. Now, we apply Equations 5 and 6 to compute the possibility and necessity of both $\phi$ and $\neg\phi$: without loss of generality, let us assume $K \models \phi$, and therefore, $K \not\models \neg\phi$; then

$$\begin{array}{llll} \Pi(\phi) & = & 1, & \Pi(\neg\phi) & = & 1 - \mathbf{Comp}(T), \\ N(\phi) & = & \mathbf{Val}(T), & N(\neg\phi) & = & 0. \end{array}$$

By the duality property,

$$\mathbf{Val}(T) = N(\phi) = 1 - \Pi(\neg\phi) = \mathbf{Comp}(T).$$

$\square$

We can observe that the case in which $\mathbf{Val} = \mathbf{Comp}$ defeats the very purpose of having the two complementary notions of validity and completeness; if that were the case, it would suffice to call the function into which those two notions would confound themselves "trust", because it would reflect a general notion of reliability of information about a given slice. Therefore, since we are interested in investigating the use of validity and completeness as two distinct notions, in what follows, we will make the assumption that, in general, $\mathbf{Val} \neq \mathbf{Comp}$. As a consequence, we now know that any acceptable definition of what a slice is will have to satisfy the postulate that a formula and its negation cannot belong in the same slice: formally, for every slice $T$ and formula $\phi$, $\phi \in T \Rightarrow \neg\phi \notin T$. For instance, a slice might be defined as the set of (ground) formulas that are satisfied by a formula with free variables (i.e., a query).

With this notion of slice (i.e., such that if a formula belongs to a slice, then the negation of that formula does not belong in the slice), we are able to prove possibilistic generalizations of results obtained by (Demolombe, 1999) in a KD doxastic logic.

First of all, the fact that a formula and its negation cannot belong to the same slice motivates the definition of the dual of a slice.

**Definition 5.** *Let $T$ be a slice. The dual of $T$, denoted $\neg T$, is the slice such that $\phi \in T$ iff $\neg\phi \in \neg T$.*

The dual of a slice is thus a sort of complement, but not in the set-theoretic sense, because there may exist a formula $\psi$ such that $\psi \notin T$ and $\psi \notin \neg T$; therefore, in general, $\neg T \neq \overline{T}$. A straightforward consequence of Definition 5 is that $\neg(\neg T) = T$.

The intuition behind the next proposition is that if the knowledge base is consistent and if information is complete concerning both $T$ and $\neg T$, then, for each formula $\phi$ in $T$ or in $\neg T$, either the agent believes $\phi$ or it believes $\neg\phi$.

**Proposition 3.** *If $K$ is consistent, for all slice $T$,*

$$\min\{\mathbf{Comp}(T), \mathbf{Comp}(\neg T)\} \leq \min_{\phi \in T} \max\{N(\phi), N(\neg\phi)\}.$$

*Proof.* We prove this proposition by showing that, for all formula $\phi \in T$,

$$\min\{\mathbf{Comp}(T), \mathbf{Comp}(\neg T)\} \leq \max\{N(\phi), N(\neg\phi)\},$$

from which the thesis follows. Given a formula $\phi \in T$, there are just three mutually exclusive cases:

**Case I.** $K \models \phi$ and, therefore, by the consistency of $K$, $K \not\models \neg\phi$; we thus have, by Def. 4,

$$N(\phi) = 1 - \Pi(\neg\phi) \geq \mathbf{Comp}(\neg T);$$

hence,

$$\begin{array}{rll} \max\{N(\phi), N(\neg\phi)\} & \geq \\ \geq & N(\phi) & \geq & \mathbf{Comp}(\neg T) \\ \geq & \min\{\mathbf{Comp}(T), \mathbf{Comp}(\neg T)\}. \end{array}$$

**Case II.** $K \not\models \phi$ and $K \not\models \neg\phi$; we thus have, by Def. 4,

$$\begin{array}{lll} N(\phi) = 1 - \Pi(\neg\phi) & \geq & \mathbf{Comp}(\neg T); \\ N(\neg\phi) = 1 - \Pi(\phi) & \geq & \mathbf{Comp}(T); \end{array}$$

hence,

$$\begin{array}{rll} \max\{N(\phi), N(\neg\phi)\} & \geq \\ \geq & \max\{\mathbf{Comp}(\neg T), \mathbf{Comp}(T)\} \\ \geq & \min\{\mathbf{Comp}(T), \mathbf{Comp}(\neg T)\}. \end{array}$$

**Case III.** $K \models \neg\phi$ and, therefore, by the consistency of $K$, $K \not\models \phi$; we thus have, by Def. 4,

$$N(\neg\phi) = 1 - \Pi(\phi) \geq \mathbf{Comp}(T);$$

hence,

$$\begin{array}{rll} \max\{N(\phi), N(\neg\phi)\} & \geq \\ \geq & N(\neg\phi) & \geq & \mathbf{Comp}(T) \\ \geq & \min\{\mathbf{Comp}(T), \mathbf{Comp}(\neg T)\}. \end{array}$$

Since the three above cases are exhaustive for every formula $\phi \in T$, this concludes the proof. $\square$

**Proposition 4.** *If $\pi$ is the least commitment possibility distribution such that Equations 5 and 6 hold, for all slice $T$,*

$$\mathbf{Val}(T) = \min_{\phi \in T : K \models \phi} N(\phi). \qquad (8)$$

*Proof.* By Def. 3, $\mathbf{Val}(T) \leq N(\phi)$ for all $\phi \in T$ such that $K \models \phi$; therefore, we can write

$$\mathbf{Val}(T) \leq \min_{\phi \in T : K \models \phi} N(\phi). \qquad (9)$$

On the other hand, by Equation 6, we can write

$$\min_{\phi \in T : K \models \phi} N(\phi) = \min_{\phi \in T : K \models \phi} \max_{T' : \phi \in T'} \mathbf{Val}(T').$$

Now, we observe that the set of slices $\{T' : \phi \in T'\}$ always includes $T$ as one of its elements, because $\phi \in T$; therefore,

$$\max_{T' : \phi \in T'} \mathbf{Val}(T') \geq \mathbf{Val}(T);$$

furthermore,

- either $T$ is a top-level slice and, by Postulate **P3**, there exists a formula $\phi^* \in T$ that does not belong to any other slice, in which case $\max_{T':\phi^* \in T'} \textbf{Val}(T') = \max\{\textbf{Val}(T)\} = \textbf{Val}(T)$, whence we can conclude that

$$\min_{\phi \in T:K \models \phi} \max_{T':\phi \in T'} \textbf{Val}(T') = \textbf{Val}(T),$$

  which is the thesis;

- or there exists another slice $\hat{T}$ such that $T \subset \hat{T}$, for which, by Equation 3, $\textbf{Val}(\hat{T}) \leq \textbf{Val}(T)$, which leads us to conclude that

$$\min_{\phi \in T:K \models \phi} \max_{T':\phi \in T'} \textbf{Val}(T') \leq \textbf{Val}(T),$$

  which, together with Equation 9, yields the thesis.

No other case being possible, this concludes the proof. $\square$

It has been proven (Demolombe, 1999) that, for a consistent base $K$, if $K$ is complete about $\neg T$ (the complement of slice $T$), then $K$ is valid about slice $T$; the following an extension of that result to the gradual case.

**Proposition 5.** *If $K$ is consistent, for all slice $T$,* $\textbf{Comp}(\neg T) \leq \textbf{Val}(T)$.

*Proof.* By Def. 4, $\textbf{Comp}(\neg T) \leq 1 - \Pi(\neg \psi) = N(\psi)$, for all $\neg \psi \in \neg T$ (and, therefore, $\psi \in T$) such that $K \not\models \neg \psi$; let

$$\beta = \min_{\psi \in T:K \not\models \neg \psi} N(\psi);$$

then we can write $\textbf{Comp}(\neg T) \leq \beta$. Clearly, $\beta \leq \min_{\phi \in T:K \models \phi} N(\phi)$, because $\{\phi \in T : K \models \phi\} \subseteq \{\psi \in T : K \not\models \neg \psi\}$, since $K \models \phi \Rightarrow K \not\models \neg \phi$. Therefore, by Proposition 4, we have

$$\textbf{Comp}(\neg T) \leq \beta \leq \min_{\phi \in T:K \models \phi} N(\phi) = \textbf{Val}(T),$$

which proves the thesis. $\square$

## 3.4 Complexity of Reasoning

The results that have been derived above show that one can "simulate", as it were, a possibilistic belief base by means of a crisp base $K$ together with metadata about the validity and completeness of $K$ with respect to a number of "slices" (i.e., sets of formulas).

It is not important to know a least-commitment possibility distribution that induces the possibility and necessity measures of Equations 5 and 6 or to represent one of its corresponding possibilistic bases $B$ explicitly, since $K$, together with its associated metadata **Val** and **Comp**, is sufficient to compute any possibilistic inference using any available classical reasoner, as demonstrated by the algorithm shown in Figure 1, adapted from (da Costa Pereira et al., 2017).

**Require:** $K \subset \mathcal{L}$: a consistent KB; $\phi \in \mathcal{L}$: a formula;
**Ensure:** $N(\phi)$.
1: $\alpha \leftarrow 0$
2: **if** $K \models \phi$ **then**
3:     **for all** slice $T \in \mathcal{S}_K$ **do**
4:         **if** $\phi \in T$ **and** $\alpha < \textbf{Val}(T)$ **then**
5:             $\alpha \leftarrow \textbf{Val}(T)$
6:         **end if**
7:     **end for**
8: **else if** $K \not\models \neg \phi$ **then**
9:     **for all** slice $T$ **do**
10:        **if** $\neg \phi \in T$ **and** $\alpha < \textbf{Comp}(T)$ **then**
11:           $\alpha \leftarrow \textbf{Comp}(T)$
12:        **end if**
13:     **end for**
14: **end if**
15: **return** $\alpha$.

Figure 1: An algorithm that "simulates" a possibilistic inference from $B$ using $K$, **Val**, and **Comp**.

**Property 1.** *Algorithm 1 is correct (i.e., it computes $N(\phi)$).*

*Proof.* If $K \models \phi$, Equation 6 is applied; otherwise, Equation 5 together with duality: $N(\phi) = 1 - \Pi(\neg \phi)$. $\square$

**Property 2.** *The cost of Algorithm 1, is at most $\|\mathcal{S}_K\| + 2$ classical inferences, where $\|\mathcal{S}_K\|$ is the number of slices defined for $K$.*

*Proof.* Algorithm 1 needs first of all to execute at most two classical inferences: the one in Line 2 and, in case $K \not\models \phi$, the one in Line 8. Then, checking whether a formula belongs in a topic costs at most one classical inference and it has to be done for all the slices defined for $K$. $\square$

Notice that, according to this result, while the cost of a possibilistic inference is higher then the cost of a classical inference, it is so only by a factor which depends on the number of slices defined on the KB. It is to be expected that this number will, in general, be much smaller than the number of facts contained in the KB. In other words, the overall complexity of possibilistic inference will be in the same class as classical inference.

## 4 DISCUSSION AND CONCLUSION

We have shown that a classical knowledge base plus metadata information on the (gradual) validity and completeness of its "slices" enables one to represent

a possibilistic belief base and perform possibilistic inferences by using a classical reasoner at a cost which, albeit larger than the classical counterpart by a multiplicative factor proportional to the number of slices, lies in the same complexity class.

All of our results are valid for the general case of a decidable fragment of first-order logic and thus they can be readily transferred to state-of-the-art and popular knowledge representation languages, like Datalog and RDF + OWL and their reasoners. This also means that our suggestion to use gradual metadata about validity and completeness may be applied to representing and reasoning with possibilistic uncertainty on top of the standard infrastructure of the semantic Web, without requiring any *ad hoc* extension and at a reasonable cost. In that setting, one way of implementing the notion of a slice might be through RDF named graphs.

Future work includes demonstrating how our proposal can be deployed on the semantic Web infrastructure to represent and reason about uncertain knowledge with a proof-of-concept implementation.

# ACKNOWLEDGEMENTS

# REFERENCES

Collins, A., Warnock, E. H., Aiello, N., and Miller, M. L. (1975). Reasoning from incomplete knowledge. In BOBROW, D. G. and COLLINS, A., editors, *Representation and Understanding*, pages 383 – 415. Morgan Kaufmann, San Diego.

da Costa Pereira, C., Dubois, D., Prade, H., and Tettamanzi, A. G. B. (2017). Handling topical metadata regarding the validity and completeness of multiple-source information: A possibilistic approach. In *SUM*, volume 10564 of *Lecture Notes in Computer Science*, pages 363–376, Berlin. Springer.

Darari, F., Nutt, W., Pirrò, G., and Razniewski, S. (2013). Completeness statements about RDF data sources and their use for query answering. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, pages 66–83, Berlin. Springer.

Demolombe, R. (1996). Answering queries about validity and completeness of data: From modal logic to relational algebra. In *FQAS*, volume 62 of *Datalogiske Skrifter (Writings on Computer Science)*, pages 265–276, Roskilde. Roskilde University.

Demolombe, R. (1999). Database validity and completeness: Another approach and its formalisation in modal logic. In *KRDB*, volume 21 of *CEUR Workshop Proceedings*, pages 11–13, Aachen. CEUR-WS.org.

Dong, X. L., Gabrilovich, E., Heitz, G., Horn, W., Murphy, K., Sun, S., and Zhang, W. (2014). From data fusion to knowledge fusion. *Proc. VLDB Endow.*, 7(10):881–892.

Dubois, D., Lang, J., and Prade, H. (1994). Possibilistic logic. In *Handbook of logic in artificial intelligence and logic programming (vol. 3): nonmonotonic reasoning and uncertain reasoning*, pages 439–513. Oxford University Press, New York, NY, USA.

Dubois, D. and Prade, H. (1988). *Possibility Theory—An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York.

Dubois, D. and Prade, H. (1997). Valid or complete information in databases - A possibility theory-based analysis. In *DEXA*, volume 1308 of *Lecture Notes in Computer Science*, pages 603–612, Berlin. Springer.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. (1997). Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub totals. *Data Min. Knowl. Discov.*, 1(1):29–53.

Lang, J. (2001). Possibilistic logic: complexity and algorithms. In Kohlas, J. and Moral, S., editors, *Algorithms for Uncertainty and Defeasible Reasoning*, Vol. 5 of Handbook of Defeasible Reasoning and Uncertainty Management Systems (Gabbay, D. M. and Smets, Ph., eds.), pages 179–220. Kluwer Acad. Publ., Dordrecht.

Levesque, H. J. (1980). Incompleteness in knowledge bases. *SIGART Bull.*, 74:150–152.

Levesque, H. J. (1982). The logic of incomplete knowledge bases. In *On Conceptual Modelling (Intervale)*, pages 165–189, New York. Springer.

Motro, A. (1989). Integrity = validity + completeness. *ACM Trans. Database Syst.*, 14(4):480–502.

Razniewski, S., Suchanek, F. M., and Nutt, W. (2016). But what do we actually know? In *AKBC@NAACL-HLT*, pages 40–44, Stroudsburg, PA. The Association for Computer Linguistics.

Wick, M. L., Singh, S., Kobren, A., and McCallum, A. (2013). Assessing confidence of knowledge base content with an experimental study in entity resolution. In *Proceedings of the 2013 workshop on Automated knowledge base construction, AKBC@CIKM 13, San Francisco, California, USA, October 27-28, 2013*, pages 13–18, New York. ACM.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.