Unifying Knowledge Base Completion with PU Learning to Mitigate the Observation Bias

Jonas Schouterden, Jessa Bekker, Jesse Davis, Hendrik Blockeel

KU Leuven, Department. of Computer Science, B-3000 Leuven, Belgium
Leuven.AI - KU Leuven Institute for AI, B-3000 Leuven, Belgium
{jonas.schouterden, jessa.bekker, jesse.davis, hendrik.blockeel}@kuleuven.be

Abstract

Methods for Knowledge Base Completion (KBC) reason about a knowledge base (KB) in order to derive new facts that should be included in the KB. This is challenging for two reasons. First, KBs only contain positive examples. This complicates model evaluation which needs both positive and negative examples. Second, those facts that were selected to be included in the knowledge base, are most likely not an i.i.d. sample of the true facts, due to the way knowledge bases are constructed. In this paper, we focus on rule-based approaches, which traditionally address the first challenge by making assumptions that enable identifying negative examples, which in turn makes it possible to compute a rule's confidence or precision. However, they largely ignore the second challenge, which means that their estimates of a rule's confidence can be biased. This paper approaches rule-based KBC through the lens of PU learning, which can cope with both challenges. We make three contributions. (1) We provide a unifying view that formalizes the relationship between multiple existing confidences measures based on (i) what assumption they make about and (ii) how their accuracy depends on the selection mechanism. (2) We introduce two new confidence measures that can mitigate known biases by using propensity scores that quantify how likely a fact is to be included the KB. (3) We show through theoretical and empirical analysis that taking the bias into account improves the confidence estimates, even when the propensity scores are not known exactly.

1 Introduction

Knowledge Bases (KBs) such as Wikidata (Vrandecic and Krötzsch 2014), YAGO (Rebele et al. 2016) and DBpedia (Auer et al. 2007) are large collections of knowledge about the world. They contain entities, such as *Audrey*, *Belgium* and *BestActress1960*, and facts about those entities such as $\langle$*Audrey, wasBornIn, Belgium*$\rangle$ and $\langle$*Audrey, wonOscar, BestActress1960*$\rangle$. KBs are typically constructed through crowdsourcing or automatically extracting information from the web (Rebele et al. 2016). Consequently, these KBs are incomplete as they do not contain all facts.

Knowledge Base Completion (KBC) (Nickel et al. 2015) aims to address the issue of incompleteness by reasoning over the knowledge base in order to derive new facts that should be included in the KB. This is typically achieved by learning a model from the initial incomplete KB. One common way to do this is to take a rule-based approach (Galárraga et al. 2013; Pellissier Tanon et al. 2017; Zupanc and Davis 2018). This would result in learning rules like $\langle X, wonOscar, Y\rangle \wedge \langle X, isMarriedTo, Z\rangle \Rightarrow \langle Z, livesIn, USA\rangle$, meaning that partners of Oscar winners usually live in the USA. A common measure for the quality of (intermediate) models is *confidence* (precision), which is the fraction of correctly predicted facts.

However, learning from the incomplete KB is challenging for two important reasons. First, KBs operate under *Open World* semantics which means that the truth value of any triple not in the KB is unknown. These triples could be true (i.e., they belong in the KB) or false (i.e., they should be excluded from the KB). Practically, this means the data only contains positive examples, whereas most learners require both positive and negative examples. It also implies that a rule's confidence cannot be computed without making additional assumptions. Second, the way knowledge bases are constructed makes it highly unlikely that the facts included in the observed KB are an i.i.d sample of the facts in the *ideal knowledge base*, aka the ground truth. Indeed, studies have shown that knowledge bases suffer from *observation biases*: They contain cultural biases, contain more facts about famous people and represents men and women differently (Callahan and Herring 2011; Wagner et al. 2015; Soulet et al. 2018). If knowledge base completion is applied while ignoring the observation bias, then the newly inferred facts are likely to strengthen the bias. Yet, to the best of our knowledge, this is what all current KBC methods do.

PU learning (learning from positive and unlabeled examples) (Bekker and Davis 2020), which is concerned with learning a binary classifier while only having access to positive and unlabeled examples, is well-equipped for addressing both these challenges. First, it perfectly matches the type of data available for KBC: the positive examples are the facts in the KB whereas the unlabeled data is any potential fact that is not included in the KB. Second, recent work in PU learning (Kato, Teshima, and Honda 2018; Bekker, Robberechts, and Davis 2019; Gong et al. 2021) has explicitly modeled the selection mechanism that determines the probability of observing a positive example's label, i.e., the observation bias.

Motivated by this, we view the KBC task as a PU Learning

problem, which enables us to explicitly consider the selection mechanism. We consider rule-based approaches to KBC and make three contributions. (1) We provide a unifying view that formalizes the relationship between multiple existing confidences measures based on (i) what assumption they make about and (ii) how their accuracy depends on the selection mechanism. (2) We introduce two new confidence measures that can mitigate known biases by using propensity scores that quantify how likely a fact is to be included the KB. (3) We show through theoretical and empirical analysis that taking the bias into account improves the confidence estimates, even when the propensity scores are not known exactly.

2 Preliminaries

Knowledge Bases (KBs) store interlinked information about entities in the form of relations between the entities, often as RDF triple stores (WWW Consortium 2004). Using this format, the KB is a triple $(\mathcal{E}, \mathcal{P}, F)$, with $\mathcal{E}$ the set of entities, $\mathcal{P}$ the set of predicates and the F set of facts, denoted by $\langle s, p, o\rangle$ triples with subject $s \in \mathcal{E}$ and object $o \in \mathcal{E}$ and predicate $p \in \mathcal{P}$, meaning that a relation of type p holds between entities s and o. The triples in a KB are a subset of the Cartesian product $\mathcal{E} \times \mathcal{P} \times \mathcal{E}$ and each predicate and entity in $\mathcal{E}$ and $\mathcal{P}$ occurs at least once in a triple in the KB.

A knowledge base models a certain part of the world. We call a knowledge base *ideal* if it has a triple for each relevant fact, and *incomplete* if it contains only a subset of those triples. We will consistently use I to refer to the ideal knowledge in some context, and K to refer to a given "known" (incomplete) knowledge base. The task of *Knowledge Base Completion* (KBC) is then: given a knowledge base K, reconstruct the ideal knowledge base I.

The KBC task is often approached as follows: given an incomplete knowledge base, rules are derived of the form $Body(s, o) \Rightarrow \langle s, p, o\rangle$, with the semantics that if $Body(s, o)$ (some condition on s and o) is fulfilled in K, then $\langle s, p, o\rangle$ is in I. These rules can then be used to derive new facts (facts that are not in K, but are in I). We follow these semantics: $Body(s, o)$ is always applied to K, predicting $\langle s, p, o\rangle$ to be in I. $Body(s, o)$ is typically a conjunctive condition (Galárraga et al. 2013; Fürnkranz, Gamberger, and Lavrač 2014; Pellissier Tanon et al. 2017; Zupanc and Davis 2018; Lajus, Galárraga, and Suchanek 2020), though this is not essential for this paper.

In the remainder of this paper, we will use the following notation. In the context of a specific rule, R refers to the rule itself, and p refers to the (fixed) predicate of the rule's prediction. We use the following indicator functions (which return 1 if the associated condition is true and 0 otherwise):

- $R(s, o)$: $\langle s, o\rangle$ fulfills the rule's conditions $Body(s, o)$
- $y(s, o) = y(\langle s, p, o\rangle)$: $\langle s, p, o\rangle$ is in I ("is a fact")
- $l(s, o) = l(\langle s, p, o\rangle)$: $\langle s, p, o\rangle$ is in K ("is observed")
- $y(s) = y(\langle s, p\rangle) = \max_o y(s, o)$
- $l(s) = l(\langle s, p\rangle) = \max_o l(s, o)$

For readability, we use the short versions $y(s, o)$, $l(s, o)$, $y(s)$, $l(s)$ when p is implied (e.g., in the context of a single rule). Note that $y(s)$ and $l(s)$ indicate that, for this specific s, at least one triple of the form $\langle s, p, \cdot\rangle$ is respectively in I / in K.

Based on the above functions, we define the following sets:

- $\mathbf{R} = \{\langle s, o\rangle \mid R(s, o) = 1\}$
- $\mathbf{R}^+ = \{\langle s, o\rangle \in \mathbf{R} \mid y(s, o) = 1\}$
- $\mathbf{R^l} = \{\langle s, o\rangle \in \mathbf{R} \mid l(s, o) = 1\}$
- $\mathbf{R_s^+} = \{\langle s, o\rangle \in \mathbf{R} \mid y(s) = 1\}$
- $\mathbf{R_s^l} = \{\langle s, o\rangle \in \mathbf{R} \mid l(s) = 1\}$

That is, $\mathbf{R}$ is the rule's *coverage* containing all $\langle s, o\rangle$ triples for which the rule fires (i.e. the rule's predictions); among those, $\mathbf{R}^+$ and $\mathbf{R^l}$ contain respectively the true and observed ones. $\mathbf{R_s^+}$ and $\mathbf{R_s^l}$ respectively restricts $\mathbf{R}$ to triples with an s for which at least one $\langle s, p, o\rangle$ is true or observed.

This paper focuses on *evaluating rules* of the format just described. *Confidence measures* are typically used to evaluate the quality of rules, during rule induction and model evaluation. The *confidence* of a rule is

$$\mathrm{conf}(R) = \frac{\sum_{\langle s,o\rangle \in \mathbf{R}} y(s, o)}{|\mathbf{R}|} = \frac{|\mathbf{R}^+|}{|\mathbf{R}|}.$$

This definition reflects the fact that rules are executed on K but their predictions are considered correct if the predicted triple is in I.

When constructing a rule set from a knowledge base, a learner typically repeatedly tries to pick the highest-confidence rule from a number of options. As the learner has access to K but not I, it cannot compute $\mathrm{conf}(R)$, so it must estimate it. Before discussing existing and novel estimators in section 4, we first discuss how current KBC approaches deal with this.

3 Assumptions on the Selection Mechanism

If K were an i.i.d. sample from the set of all triples, labeled positive or negative according to whether they are in I, and each triple had the same probability of being included in K, then simply counting how many of the predicted triples are labeled positive or negative would yield an unbiased estimator of $\mathrm{conf}(R)$ (just like test set accuracy is an unbiased estimator for population accuracy). But K contains no negative examples at all. This poses the following challenge:

> How can one estimate the confidence of a rule without access to negative examples?

3.1 Typical assumptions in KBC

In general, evaluating models without access to negative examples remains an open problem (Speranskaya, Schmitt, and Roth 2020; Pezeshkpour, Tian, and Singh 2020). A common approach in KBC is to make assumptions that allow deriving negative examples. Two prominent assumptions are:

The closed-world assumption (CWA) (naively) assumes that all facts not included in K are false. Hence, if a rule derives a fact not in K, that corresponds to a false positive.

The partial-completeness assumption (PCA) (a.k.a. *local closed-world assumption*) assumes that if a $\langle s, p, o\rangle$ triple

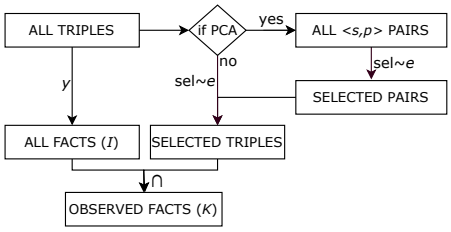


Figure 1: Which facts of I are observed in K depends the selection mechanism.

is observed, then all triples $\langle s, p, \cdot \rangle$ in I with the same subject and predicate are observed (Galárraga et al. 2013; Dong et al. 2014). It follows from this that if $\langle s, p, o \rangle$ is predicted and K contains $\langle s, p, o' \rangle$ for some $o' \neq o$, but not $\langle s, p, o \rangle$, then this must be a false positive.

The key insight in our paper is that these assumptions fail to account for the underlying mechanism that determines how the KB is populated, which results in biased estimates of a rule's confidence. In the following, we explicitly look at possible selection mechanisms and how the CWA and PCA assumptions connect with them.

3.2 Selection Mechanisms

Which facts are observed in K depends on how the KB was populated. Conceptually, this can be modeled by assuming that whether a fact is included or not in the KB depends on a *selection mechanism* $\text{sel}(\langle s, p, o \rangle)$ which selects triples $\langle s, p, o \rangle$ from $\mathcal{E} \times \mathcal{P} \times \mathcal{E}$ (see Figure 1). If a selected triple is a fact in I then it becomes part of K: $l(\langle s, p, o \rangle) = \text{sel}(\langle s, p, o \rangle) y(\langle s, p, o \rangle)$.

The selection mechanism can operate in different ways. From a probabilistic point of view, the simplest version is that K is an independent and identically distributed (i.i.d.) sample from I. CWA is then the special case where each triple has probability 1 of being included in K. More realistically, groups of triples might be selected together (not independent) and some triples might be more likely to be selected than others (not identically distributed). PCA implies one particular type of dependence: It assumes a hierarchical selection mechanism that first selects pairs $\langle s, p \rangle$, then selects *all* triples $\langle s, p, \cdot \rangle$ of the selected pairs, as depicted in figure 1.

We next focus on the actual selection probabilities. While existing rule-based KBC work neither explicitly states nor considers the selection mechanism, the field of PU Learning makes such assumptions very explicit. Therefore, we follow their terminology. The probability with which a positive example is selected for inclusion is called its *propensity score* e (Bekker, Robberechts, and Davis 2019):

$$e(\langle s, p, o \rangle) = \Pr(\text{sel}(\langle s, p, o \rangle) = 1) \quad \textit{\# no PCA}$$
$$e(\langle s, p \rangle) = e(\langle s, p, o \rangle) = \Pr(\text{sel}(\langle s, p \rangle) = 1) \quad \textit{\# PCA}$$

From this, the probability that a triple appears in K follows:

$$\Pr(l(\langle s, p, o \rangle) = 1) = e(\langle s, p, o \rangle) y(\langle s, p, o \rangle)$$

We use shorthands $e(s, o) = e(\langle s, p, o \rangle)$ and $e(s) = e(\langle s, p \rangle)$.

Assumptions in PU Learning about the selection probabilities range from *Selected Completely At Random (SCAR)*, where each positive example has the constant probability $e(\cdot) = c$ to be selected, to *Selected At Random (SAR)*, where the propensity score is a function of the example's features (Bekker, Robberechts, and Davis 2019).

Based on this, we propose the following taxonomy for assumptions about KBC selection probabilities:

- *Closed World Assumption (CWA):* All facts are observed: $e(\langle s, p, o \rangle) = 1$.
- *SCAR* All facts have the same probability to be selected: $e(\langle s, p, o \rangle) = c$
- *SCAR-per-predicate (SCAR$_p$):* All facts about the same predicate p have the same probability to be selected: $e(\langle s, p, o \rangle) = c_p$
- *SCAR-per-rule (SCAR$_R$):* All facts predicted by a rule R have the same probability to be selected: $R(s, o) = 1 \Rightarrow e(\langle s, p, o \rangle) = c_R$
- *SAR:* The probability that a fact gets selected, depends on its characteristics in the incomplete KB K.[1]

Note that this categorization is largely orthogonal to any dependence structures in the selection mechanism. In particular, all five levels are compatible with PCA, though additional restrictions may apply (e.g., SCAR-per-rule under PCA implies that rules with different c_R cannot cover the same subject s).

Only SAR, the least strict assumption, can in general represent common observation biases such as higher propensity scores for famous entities, and certain predicates having different propensity scores for women and men (Callahan and Herring 2011; Wagner et al. 2015). SCAR-per-rule can include such biases, but only if each rule covers one group exclusively (famous or plebeian, man or woman).

Note that all except the SAR assumption consider the observed facts covered by a certain rule to be unbiased. The next section will show that this assumption is made implicitly by all existing confidence measures.

4 Confidence Estimators

We now return to the problem of estimating the confidence of a rule R, $\text{conf}(R)$. First, We discuss and analyze estimators that have been used in the KBC field. Second, we will introduce new estimators that account for possible observation bias.

4.1 Existing confidence estimators

CWA-based estimator Under the closed-world assumption, a prediction is considered correct if it is known to be correct (i.e., the predicted triple is observed in K), and incorrect otherwise. The confidence calculated as such is usually referred to as the *standard confidence* (Galárraga et al. 2013), but for clarity, we call it the CWA-based estimator.

$$\textit{CWA}(R) = \frac{\sum_{\langle s,o \rangle \in \mathbf{R}} l(s, o)}{|\mathbf{R}|} = \frac{|\mathbf{R^1}|}{|\mathbf{R}|} = \text{conf}(R) \frac{|\mathbf{R^1}|}{|\mathbf{R^+}|}$$

[1] More realistically, the probability depends on its true (possibly unobserved) characteristics. However, in the KBC task, only the characteristic that are observable are relevant.

CWA(R) generally *underestimates* $\text{conf}(R)$ because $K \subset I$ and therefore $|\mathbf{R^1}| \leq |\mathbf{R}^+|$, yielding *CWA*$(R) \leq \text{conf}(R)$.

Now consider different possible realizations of K given some I. Because, the probability for a fact to be included in K is $Pr(l(s,o)=1 | y(s,o)=1) = e(s,o)$, the expected value for *CWA*(R) over all K is

$$\mathop{\mathbb{E}}_{\text{sel} \sim e} [\textit{CWA}(R)] = \text{conf}(R) \frac{1}{|\mathbf{R}^+|} \sum_{\langle s,o \rangle \in \mathbf{R}^+} e(s,o).$$

Under SCAR-per-predicate, with a constant $e(s,o) = c_p$ per predicate p, it has the same constant multiplicative bias c_p for all rules for a predicate p, meaning that the ranking of rules is still expected to be correct. To better analyse the problem for the setting where only the relative ranking of rules matter, we introduce an *inverse* c_p-*weighted* version of the CWA-based estimator $ICW(R) = \frac{1}{c_p} CWA(R)$, which is indeed unbiased under SCAR-per-predicate: $\mathbb{E}_{\text{sel} \sim c_p} [ICW(R)] = \text{conf}(R)$.

PCA-based estimator To solve the above-mentioned underestimation problem, the PCA-based estimator only considers the subset of predictions $\mathbf{R^1_s}$ assumed to have a known truth value under the PCA assumption. That is, the PCA-based estimator only considers predictions $\mathbf{R^1_s}$ for which the subject appears in an observed fact in K ($l(s) = 1$). For all triples in $\mathbf{R^1_s}$, if the specific triple is observed ($l(s,o) = 1$) then the prediction is considered correct, if it is not observed ($l(s,o) = 0$) then the prediction is considered incorrect (Galárraga et al. 2013):

$$\textit{PCA}(R) = \frac{\sum_{\langle s,o \rangle \in \mathbf{R}} l(s,o)}{\sum_{\langle s,o \rangle \in \mathbf{R}} l(s)} = \frac{|\mathbf{R^1}|}{|\mathbf{R^1_s}|}$$

While the PCA-based estimator is a commonly-used confidence estimator, we are, to the best of our knowledge, the first to study under which conditions it is expected to perform well. The PCA-based estimator can suffer from biases induced by three factors, which are mathematically derived and interpreted in appendix A[2]:

$$\mathop{\mathbb{E}}_{\text{sel} \sim e} [\textit{PCA}(R)] \approx \frac{\mathbb{E}_{\text{sel} \sim e} \left[|\mathbf{R^1}| \right]}{\mathbb{E}_{\text{sel} \sim e} \left[|\mathbf{R^1_s}| \right]} \quad \begin{matrix} \textit{first-order} \\ \textit{Taylor approximation} \end{matrix}$$

$$= \text{conf}(R) \cdot \textit{bias}_{\not{PCA}}(R) \cdot \textit{bias}_{y(s)=0}(R) \cdot \textit{bias}_{e(s)}(R)$$

$$= \text{conf}(R) \frac{\sum_{\langle s,o \rangle \in \mathbf{R}^+} e(s,o)}{\sum_{\langle s,o \rangle \in \mathbf{R}^+} e(s)} \frac{|\mathbf{R}|}{|\mathbf{R^+_s}|} \frac{\frac{1}{|\mathbf{R}^+|} \sum_{\langle s,o \rangle \in \mathbf{R}^+} e(s)}{\frac{1}{|\mathbf{R^+_s}|} \sum_{\langle s,o \rangle \in \mathbf{R^+_s}} e(s)},$$

The three biases can cancel each other out, making it hard to predict how *PCA*(R) will perform, when the problem is not well understood. We explain the three biases using an example rule that predicts which people won an Oscar, e.g., $\langle$*Audrey*, *wonOscar*, *BestActress1954*$\rangle$. Under PCA, K is assumed to contain either all or none of the Oscars that each person has won.

[2]The appendices can be found at: https://github.com/ML-KULeuven/KBC-as-PU-Learning

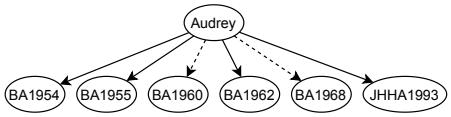


Figure 2: Correctly predicted facts $\mathbf{R}^+$ for subject *Audrey* by the example rule. Dotted arrows indicate predicted facts not in K, which the PCA considers to be false positives.

To illustrate when $\textit{bias}_{\not{PCA}}$ arises, consider a KB that only contains facts denoting four of Audrey Hepburn's six Oscars meaning the PCA assumption is violated for the $\langle$*Audrey*, *wonOscar*$\rangle$ pair. Suppose a rule is learned that correctly derives all six of Audrey Hepburn's Oscars (Figure 2). Making the PCA assumption results in the two Oscar wins not in the KB being incorrectly denoted as false positives, yielding an *underestimate* of the rule's confidence just like in the CWA case. More formally, this bias arises whenever a rule fires for a $\langle s,p \rangle$ pair for which the PCA assumption is violated and the rule derives a $\langle s,p,o \rangle$ s.t. $l(\langle s,p,o \rangle) = 0$ and $y(\langle s,p,o \rangle) = 1$ (an unobserved fact). This bias never arises for functional predicates (where each subject appears in at most 1 fact) because the PCA trivially holds for such predicates.

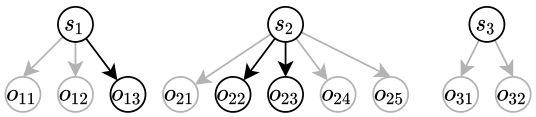
To illustrate when $\textit{bias}_{y(s)=0}$ arises, consider a rule that predicts that Alan Rickman won an Oscar, which is a false positive since he has never won an Oscar. However, because there is no fact $\langle$*AlanRickman*, *wonOscar*, $\cdot\rangle$ in K, since no such facts exist, the PCA estimator disregards the prediction in its confidence. In this case the PCA estimator *overestimates* the rule's confidence. More generally, such an overestimate occurs whenever a rule predicts a triple $\langle s,p,o' \rangle$ where $\forall o : y(\langle s,p,o \rangle) = 0$ (s occurs in no facts). In these cases, these false positive predictions $\mathbf{R} \setminus \mathbf{R^+_s}$ are ignored by the PCA.

The third bias factor $\textit{bias}_{e(s)}$ is the mean $e(s)$ over all correct predictions $\mathbf{R}^+$ made by the rule divided by the mean $e(s)$ over all its predictions $\mathbf{R^+_s}$ (restricting s to those s where $y(s) = 1$) (Figure 3). This bias is > 1 when correct predictions tend to have higher $e(s)$, or, put differently, when there are more correct predictions for high-propensity subjects. Vice versa, this bias is < 1 when there are fewer correct predictions for high-propensity subjects. In our Oscars example, when a rule happens to give more accurate predictions for high-propensity Oscar winners than for low-propensity ones, the confidence of this rule is overestimated.

In our experiments in Section 5, **Q3** illustrates how *PCA*(R) can vary due to this bias when $e(s)$ varies for different subjects.

4.2 Observation Bias Aware Confidence Estimators

In this section we propose two novel confidence estimators that can counteract observation biases by explicitly taking the selection mechanism into account. The difference between the estimators is whether or not they make the PCA assumption. The proposed estimators need propensity scores as input, therefore we additionally analyze their bias when using imperfect propensity scores and show that rough estimates are



$$bias_{e(s)} = \frac{\frac{1}{3}[1{\cdot}e(s_1)+2{\cdot}e(s_2)]}{\frac{1}{10}[3{\cdot}e(s_1)+5{\cdot}e(s_2)+2{\cdot}e(s_3)]}$$

Figure 3: Example $bias_{e(s)}$ calculation. The arrows indicate predictions $\mathbf{R_s^+}$ for which the subject occurs in at least one fact. Black arrows are correct predictions $\mathbf{R}^+$, grey arrows are incorrect predictions $\mathbf{R_s^+} \setminus \mathbf{R}^+$.

better than assuming that there is no observation bias.

The **Inverse Propensity Weighted estimator (IPW)** aims to debias $CWA(R)$ by weighting the observed triples with inverse propensity score estimates $\hat{e}(s,o)$:

$$IPW(R) = \frac{1}{|\mathbf{R}|} \sum_{\langle s,o\rangle \in \mathbf{R}} \frac{l(s,o)}{\hat{e}(s,o)}$$

From its expected value over all possible K, it is clear that the estimator is *unbiased* when $\hat{e}(s,o) = e(s,o)$:

$$\mathop{\mathbb{E}}_{\text{sel}\sim e}[IPW(R)] = \text{conf}(R) \frac{1}{|\mathbf{R}^+|} \sum_{\langle s,o\rangle \in \mathbf{R}^+} \frac{e(s,o)}{\hat{e}(s,o)}.$$

Similarly, the **Inverse Propensity Weighted PCA-based estimator (IPW-PCA)** aims to debias $PCA(R)$:

$$IPW\text{-}PCA(R) = \frac{\sum_{\langle s,o\rangle \in \mathbf{R}} \frac{l(s,o)}{\hat{e}(s)}}{\sum_{\langle s,o\rangle \in \mathbf{R}} \frac{l(s)}{\hat{e}(s)}}$$

The first-order Taylor approximation of its expected value is:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\text{sel}\sim e}[IPW\text{-}PCA(R)] \\
&\approx \text{conf}(R) \cdot bias_{\cancel{PCA}}^{IPW\text{-}PCA}(R) \cdot bias_{y(s)=0}(R) \cdot bias_{e(s)}^{IPW\text{-}PCA}(R) \\
&\approx \text{conf}(R) \frac{\sum_{\langle s,o\rangle \in \mathbf{R}^+} \frac{e(s,o)}{\hat{e}(s)}}{\sum_{\langle s,o\rangle \in \mathbf{R}^+} \frac{e(s)}{\hat{e}(s)}} \frac{|\mathbf{R}|}{|\mathbf{R_s^+}|} \frac{\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle \in \mathbf{R}^+} \frac{e(s)}{\hat{e}(s)}}{\frac{1}{|\mathbf{R_s^+}|}\sum_{\langle s,o\rangle \in \mathbf{R_s^+}} \frac{e(s)}{\hat{e}(s)}}.
\end{aligned}$$

Here, 2 of the 3 bias factors are inverse propensity weighted versions of the corresponding $PCA(R)$ biases. Note that when $\hat{e}(s){=}e(s)$, the $bias_{e(s)}^{IPW\text{-}PCA}(R)$ related to the selection mechanism completely disappears.

Most often, the exact propensity scores $e(s,o)$ cannot be used for $\hat{e}(s,o)$, as they are unknown. When the propensity scores are not known exactly, using reasonable estimates for $\hat{e}(s,o)$ can still result in a better confidence estimator than not using any $\hat{e}(s,o)$ and making a SCAR assumption.

To investigate how accurate the propensity score estimates $\hat{e}(\cdot)$ should be, we compare confidence estimators when the PCA does or does not hold, respectively: $IPW\text{-}PCA(R)$ vs $PCA(R)$ and $IPW(R)$ vs $CWA(R)$[3].

[3]To allow for rule comparison, we considered the calibrated version $ICW(R) = \frac{1}{c_p} CWA(R)$.

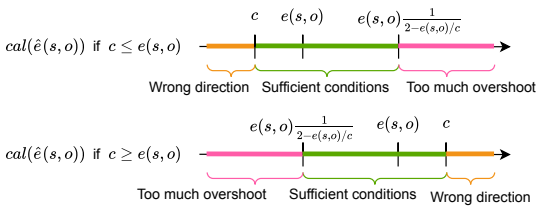


Figure 4: If $cal\,(\hat{e}(s,o))$ is between c and $e(s,o) \cdot \frac{1}{2-e(s,o)/c}$, then $\langle s,o\rangle$ has a lower error contribution in the IPW(-PCA) estimator than in the CWA/PCA estimator.

In appendix B, we show that an individual triple $\langle s,o\rangle$ has a smaller error contribution in the IPW(-PCA) estimator than in the CWA/PCA estimators when its propensity score estimate $\hat{e}(s,o)$ satisfies:

$$\begin{cases} c \le cal\,(\hat{e}(s,o)) \le e(s,o)\frac{1}{2-e(s,o)/c} & \text{, when } c \le e(s,o), \\ c \ge cal\,(\hat{e}(s,o)) \ge e(s,o)\frac{1}{2-e(s,o)/c} & \text{, when } c \ge e(s,o), \end{cases}$$

with $c{=}c_R$ under PCA and $c{=}c_p$ otherwise. Here, $cal(\hat{e}(s,o))$ multiplicatively calibrates $\hat{e}(s,o)$, so that $\mathbb{E}\,[e(s,o)/\,cal(\hat{e}(s,o))] = \mathbb{E}\,[e(s,o)/c] = 1$ (note that $\hat{e}(s,o)$ itself need not be calibrated). In other words, the IPW-(PCA) confidence estimator is preferable over the CWA/PCA confidence estimator as soon as $cal\,(\hat{e}(s,o))$ deviates from c "in the right direction", that is, towards $e(s,o)$, and this up to the point where it overshoots by a certain factor (see Figure 4). The allowed overshoot increases with $e(s,o)/c$.

As shown, reasonable estimates $\hat{e}$ can be used in the IPW(-PCA) estimators that do not need be calibrated; only their relative values matter. In practice, these relative $\hat{e}$ could be derived from domain knowledge, e.g., from research on KB bias (Callahan and Herring 2011; Wagner et al. 2015), or estimated through incompleteness estimation (Razniewski, Suchanek, and Nutt 2016; Galárraga et al. 2017). For example, in a movie-recommendation setting, Saito et al. (2020) use a movie's popularity as its propensity score.

5 Experiments

We aim to empirically answer the following research questions: can we effectively account for observation biases (i.e., obtain more accurate confidence estimates) using the newly proposed propensity-based estimators, **(Q1)** when the propensities are known, **(Q2)** when propensities are guessed ("noisy" propensities), **(Q3)** even when the PCA assumption holds?

5.1 Experimental Setup

Evaluating a confidence estimator requires knowledge of I, which is generally unavailable for real-world KBs. We therefore equate I to a real-world KB from which K's are generated by applying different selection mechanisms. Our I is the popular KBC benchmark dataset Yago3-10 (Mahdisoltani, Biega, and Suchanek 2015). Rules predicting any $p \in \mathcal{P}$ are mined from I with AMIE (Galárraga et al. 2015) with its default settings and a minimum $CWA(R) \ge 0.1$. This set

of rules serves as the testbed for our confidence estimators (thus, the same rules are used over all K and estimators). See Appendix E for the rule list.

The **applied selection mechanisms** differ in two ways. First, they either explicitly uphold the PCA (by selecting subjects s in **Q3**) or not (by selecting triples in **Q1, Q2**). For functional p, the PCA always holds by definition. Second, the mechanisms differ in which assumptions hold for the propensity scores: CWA, SCAR_p or SAR. Under SCAR_p, c_p is varied. Two SAR mechanisms are considered: 1) $\text{SAR}_{\text{group}}$ where the subjects of the triples are divided into two groups $S_q, S_{\neg q}$ (e.g., actors and non-actors), each with a constant propensity score c_q, $c_{\neg q}$, and 2) SAR_{pop} where a triple's propensity score is a logistic function of the number of facts in which the subject occurs, thus reflecting the subject's *popularity*:

$$\#(s,p) = |\{\langle s,q,\cdot\rangle \in I\} \cup \{\langle \cdot,q,s\rangle \in I\}|\,,\ q \neq p$$

$$e(\langle s,p,o\rangle) = \max\left[\frac{2}{1+e^{-k\cdot\#(s,p)}} - 1,\ e_{\min}\right]$$

More popular s have a higher e. The scaling factor k determines how often s must occur for a given $e(\langle s,p,o\rangle)$. Choosing $e_{\min} > 0$ allows unpopular s to be selected.

When a rule's coverage $\mathbf{R}$ changes by applying the rule to different K, not only the estimators but also the rule's actual confidence conf can change. In order to keep conf constant, the chosen selection mechanisms should not affect $\mathbf{R}$. Therefore, 1) only non-recursive rules are considered, and 2) each selection mechanism is applied to a single $p \in \mathcal{P}$ at a time; the facts of $\mathcal{P} \setminus \{p\}$ are completely included in K (cfr. CWA). This way, we can vary $e(\cdot)$ for p while keeping the confidence $\text{conf}(R)$ constant.

As **evaluation metric**, the Brier score $\mathbb{E}_R[\widehat{\text{conf}}(R) - \text{conf}(R)]^2$ is chosen (with $\widehat{\text{conf}}$ any estimator); this is a standard way of evaluating probability estimates.

For **Q1** and **Q2**, we compare $\mathit{ICW}(R)$[4] and $\mathit{IPW}(R)$ to $\mathit{CWA}(R)$ and $\mathit{PCA}(R)$. For **Q3**, we compare $\mathit{IPW\text{-}PCA}(R)$ to $\mathit{PCA}(R)$, as the former modifies the latter to consider propensity scores.

Propensity scores are required to calculate $\mathit{IPW(\text{-}PCA)}(R)$. We use correct propensity scores $e(\cdot)$ for the idealized scenarios in **Q1** and **Q3** and noisy versions $\hat{e}$ for **Q2**.

More details about the exact setup can be found in appendix C. Our source code is publicly available.[5]

5.2 Results

(Q1) Does using the ground truth propensity scores lead to a better confidence estimate? Table 1 shows the Brier scores for the estimators under SCAR_p, $\text{SAR}_{\text{group}}$ and SAR_{pop}. Only the leftmost $\mathit{IPW}(R)$ column is relevant for **Q1**, i.e., the column with superscript **Q1**. The table shows that using correct propensity scores under SCAR_p and SAR results in a much better conf estimate: the Brier score for $\mathit{IPW}(R)$ is often orders of magnitude lower than for the other estimators. (See also Tables 3, 4 and 5 in appendix D.)

[4] For c_p the average $e(\cdot)$ over all p-triples in K is used. Note that under SCAR_p, $\mathit{ICW}(R) = \mathit{IPW}(R)$.

[5] https://github.com/ML-KULeuven/KBC-as-PU-Learning

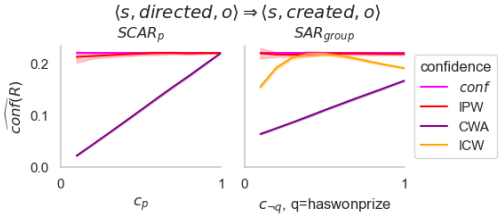


Figure 5: For a single rule under SCAR_p (left), dividing $\mathit{CWA}(R)$ by c_p is (trivially) as good as $\mathit{ICW}(R) = \mathit{IPW}(R) \approx \text{conf}(R)$ for all c_p. Under $\mathit{SAR}_{\mathit{group}}$ (right), dividing $\mathit{CWA}(R)$ by c_p fails to account for the bias for most $c_{\neg q}$ [4]. Correct propensity scores are used, and $c_q = 0.5$ for $\text{SAR}_{\text{group}}$. $\mathit{conf}(R)$ is hidden by $\mathit{IPW}(R)$.

Figure 5 zooms in on a single rule. It shows how, under SCAR_p, both $\mathit{IPW}(R)$ and $\mathit{ICW}(R)$ almost perfectly compensate for $\mathit{CWA}(R)$'s underestimation. However, under $\text{SAR}_{\text{group}}$, $\mathit{ICW}(R)$ does not recover $\text{conf}(R)$ for most $c_{\neg q}$, while $\mathit{IPW}(R)$ still does.

This illustrates how bad $\mathit{CWA}(R)$ can be by ignoring the observation bias. If a learner merely *ranks* rules predicting the same predicate p under the simple SCAR_p, all $\mathit{CWA}(R)$ are biased with the same constant c_p. Under this simple setting, $\mathit{CWA}(R)$ works as well as $\mathit{ICW}(R)$ (and hence $\mathit{IPW}(R)$). However, $\mathit{CWA}(R)$ fails under more complex settings (e.g., $\text{SAR}_{\text{group}}$), where using propensity scores allows $\mathit{IPW}(R)$ to be clearly superior to $\mathit{CWA}(R)$.

Figure 6 (left) shows $\mathit{PCA}(R)$ and $\mathit{IPW\text{-}PCA}(R)$ for a single rule under SCAR_p, for both p (*person*, *diedin*, *place*) and its inverse p^{-1} (*place*, *wheredied*, *person*). Here, the PCA holds for p (a person dies in at most 1 place), but not for p^{-1} (only a fraction c_p of all the people who died somewhere are in K). Therefore, $\mathit{PCA}_p(R)$ remains constant for most c_p, differing from $\text{conf}(R)$ with a constant factor $\mathit{bias}_{y(s)=0}(R) = |\mathbf{R}|/|\mathbf{R}^+_{\mathbf{s}}|$. In contrast, $\mathit{bias}_{\not PCA}$ causes $\mathit{PCA}_{p^{-1}}(R)$ to vary with c_p.

The 3 factors that can cause $\mathit{PCA}(R)$ to be biased interact; if they are unknown in advance, it is difficult to say how well $\mathit{PCA}(R)$ will perform. For example in Figure 6 (left), $\mathit{PCA}_{p^{-1}}(R)$ is equal to $\text{conf}(R)$ at approximately $c = 0.7$ due to an 'accidental' combination of these dimensions. However, if the PCA holds and $|\mathbf{R}|/|\mathbf{R}^+_{\mathbf{s}}|$ is close to 1, $\mathit{PCA}(R)$ will be close to conf for all c_p, as shown with $\mathit{PCA}_p(R)$. Figure 6 (right) illustrates $\mathit{PCA}(R)$ under SAR. The difference between $\mathit{PCA}_p(R)$ and $\mathit{conf}(R)$ is equal to $|\mathbf{R}|/|\mathbf{R}^+_{\mathbf{s}}|$ for the SCAR point ($c_q = c_{\neg q}$), but changes when varying the relative number of triples in S_q and $S_{\neg q}$ in K (see also the results for **Q3**).

This illustrates how complex the behavior of the $\mathit{PCA}(R)$ is. Its disadvantage is its dependence on different interacting biases. In contrast, using propensity scores allows $\mathit{IPW}(R)$ to be close to $\text{conf}(R)$ in all settings.

(Q2) Can noisy propensity scores be used to improve confidence estimates? The rightmost IPW columns in Table 1 (with superscript **Q2**) show Brier scores for $\mathit{IPW}(R)$ with respectively 0.1 and -0.1 added as noise to the correct propen-

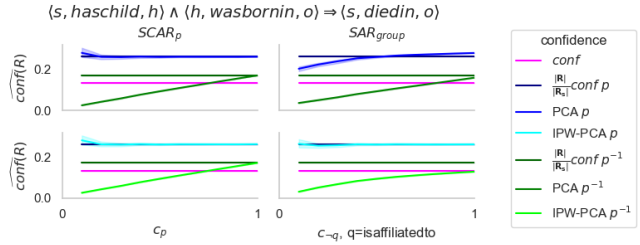


Figure 6: PCA holds for $p = \textit{diedin}$ but not for $p^{-1} = \textit{wheredied}$. Under SCAR_p (left), $\textit{PCA}_p(R)$ for the example R differs from conf with $|\mathbf{R}|/|\mathbf{R}_\mathbf{s}^+|$ for most c_p, while $\textit{PCA}_{p^{-1}}(R)$ also varies with c_p. Under $\text{SAR}_{\text{group}}$ with $c_q = 0.5$ (right), $\textit{PCA}_p(R)$ also varies with $c_{\neg q}$, which $\textit{IPW-PCA}_p(R)$ corrects (see also figure 7).

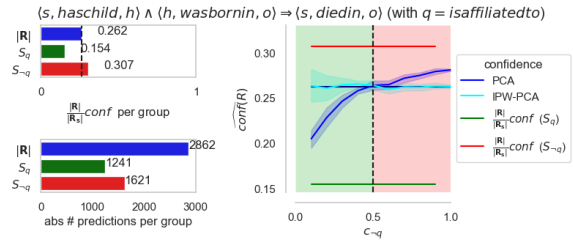


Figure 7: If the PCA holds under $\text{SAR}_{\text{group}}$, $\textit{PCA}(R)$ for a rule with different group-local confidences (upper left) but a similar number of predictions per group (lower left) changes with $c_{\neg q}$ towards $|\mathbf{R}|/|\mathbf{R}_\mathbf{s}^+| \cdot \text{conf}(R)$ of the overrepresented group, while $\textit{IPW-PCA}(R)$ remains more constant (right). $c_q = 0.5$.

sity scores for SCAR_p and $\text{SAR}_{\text{group}}$. For SAR_{pop}, the noisy propensity scores are obtained by increasing/decreasing k by 10%. The results show that $\textit{IPW}(R)$ is generally the best estimator for $\text{conf}(R)$ if the noise is not too large. The exact differences between $\textit{IPW}(R)$, $\textit{CWA}(R)$ and $\textit{PCA}(R)$ depend on 1) the specific selection mechanism affecting $\textit{CWA}(R)$ and $\textit{PCA}(R)$ as seen in **Q1**, and 2) the noisy propensity scores affecting $\textit{IPW}(R)$.

(Q3) When the PCA holds, $\textit{PCA}(R)$ becomes a better estimate as $\textit{bias}_{\textit{PCA}}(R) = 1$. Are there then still situations in which $\textit{PCA}(R)$ can be improved by using propensity scores? Here, we compare $\textit{IPW-PCA}(R)$ and $\textit{PCA}(R)$ under $\text{SAR}_{\text{group}}$ and explicitly uphold the PCA for non-functional p by selecting subjects. e.g., if a person is a *citizenof* multiple countries, then either all or none of its triples are selected. We consider rules with predictions in both subject groups S_q and $S_{\neg q}$, with neither group dominating in size: $0.3 \leq (|\mathbf{R} \cap S_q|)/|\mathbf{R}| \leq 0.7$. We only include rules for which the group-local confidence (the confidence considering only the predictions in a group) differs by at least 0.1. The total confidence is the weighted mean of the group-local confidences where the weights are the fraction of predictions per group. By varying $c_{\neg q}$ for a fixed c_q, $\textit{bias}_{e(s)}(R)$ is varied (while $\textit{bias}_{y(s)=0}(R) = |\mathbf{R}|/|\mathbf{R}_\mathbf{s}^+|$ remains constant): the relative number of subjects (and thus triples) in K belonging to each group varies, and is different from I for $c_q \neq c_{\neg q}$. Consequently, the total $\textit{PCA}(R)$ moves towards the group-local $\textit{PCA}(R)$ of the overrepresented group (Figure 7). However, $\textit{IPW-PCA}(R)$ remains relatively constant. Table 2 shows the Brier scores for this specific scenario. The results highlight how well $\textit{PCA}(R)$ works under PCA without needing propensity scores: although $\textit{IPW-PCA}(R)$ is mostly better, its improvement is rather small.

In conclusion, $\textit{CWA}(R)$ and $\textit{PCA}(R)$ fail to account for general observation biases in contrast to $\textit{IPW}(R)$ and $\textit{IPW-PCA}(R)$, which make them explicit through propensity scores.

6 Related Work

Several **confidence measures for KBC** have been introduced to address the problem of dealing with the lack of negative examples (Galárraga et al. 2013; Pellissier Tanon et al. 2017; Zupanc and Davis 2018), but none of them handle general observation biases. The confidence measures from Zupanc and Davis and Pellissier Tanon et al. were omitted from the discussion, because, the former consists of an ad-hoc pipeline, and the latter consistently underestimates the confidence (see Appendix F). Other confidences were introduced with different goals: xconf (Zhou, Sadeghian, and Wang 2019) to limit computation effort and smooth confidence to cope with rules with low support (Meilicke et al. 2019).

Estimating propensity scores is closely related to estimating where the KB is more and less complete. The limited work on this topic combines several simple **completeness oracles** such as popularity, cardinality and change over time (Razniewski, Suchanek, and Nutt 2016; Galárraga et al. 2017). Soulet et al. (2018) estimate **representativeness** as deviation from an i.i.d. sample, but do not estimate which bias is causing the deviation nor propose a method for mitigating the bias at learning time.

Most of the work in **PU Learning** has been conducted under the SCAR assumption, where the labeled examples are an i.i.d. sample from the true positive examples (Elkan and Noto 2008). This is clearly violated by KBs. The more general SAR assumption allows for non-i.i.d. selection mechanisms, but needs additional assumptions to enable learning. Only a handful of such assumptions have been proposed (Kato, Teshima, and Honda 2018; Bekker, Robberechts, and Davis 2019; Gong et al. 2021). While none of these assumptions are sufficient for the KBC setting, the notion of explicitly modeling the selection mechanism inspired this paper.

Recommender systems solve a problem similar to KBC, but with only 1 predicate type ($|\mathcal{P}| = 1$). Similar to our approach, Saito et al. (2020) and Gupta et al. (2021) adapt the PU learning loss function from Bekker, Robberechts, and Davis (2019). Amongst others, our paper differs from these works in 1) our unifying view on how KBs are constructed (including the taxonomy of assumptions) that allows analyzing the conditions under which evaluation metrics can be evaluated, 2) our analysis of the commonly used CWA and PCA estimators and 3) our newly proposed IPW(-PCA)

		# rules	CWA^{Q1,2}	PCA^{Q1,2}	ICW^{Q1,2}	IPW^{Q1,2}	IPW $-\Delta^{Q2}$	IPW $+\Delta^{Q2}$
SCAR$_p$	$c_p = 0.3$	47	292.5	192.7	**4.6**	**4.6**	154.2	40.9
	$c_p = 0.7$	47	53.8	173.3	**1.1**	**1.1**	18.1	10.0
SAR$_{\text{group}}$	$c_q = 0.5, c_{\neg q} = 0.3$	33	189.5	155.6	8.1	**3.4**	50.3	14.0
	$c_q = 0.5, c_{\neg q} = 0.7$	33	83.8	155.1	4.0	**1.8**	6.9	4.6
SAR$_{\text{pop}}$	$k = 0.01$	47	458.6	264.2	168.5	62.8	81.8	**56.6**
	$k = 0.1$	47	172.3	182.7	51.0	**3.5**	7.4	6.2

Table 1: Results for **Q1** and **Q2** (see superscript). $\left[\widehat{\text{conf}} - \text{conf}\right]^2 \cdot 10^4$ under SCAR$_p$, SAR$_{\text{group}}$ and SAR$_{\text{pop}}$. Results are averaged over p, the rules and (for SAR$_{\text{group}}$) q. The 3 IPW confidence columns: 1 with correct $\hat{e} = e$ (left) and 2 with noisy $\hat{e} \neq e$ (middle and right). For SCAR$_p$, noisy $\hat{c}_p = c_p \pm 0.1$. For SAR$_{\text{group}}$, noisy $\hat{c}_{\neg q} = c_{\neg q} \pm 0.1$. For SAR$_{\text{pop}}$, the noisy $\hat{e}$ are obtained by using $\hat{k} = k \pm 0.1k$.

p	# R	$c_{\neg q} = 0.3$				$c_{\neg q} = 0.7$			
	PCA	IPW-PCA	IPW-PCA($-\Delta$)	IPW-PCA($+\Delta$)	PCA	IPW-PCA	IPW-PCA($-\Delta$)	IPW-PCA($+\Delta$)	
dealswith	1	22.4	**16.9**	13.5	19.9	**9.1**	12.4	10.7	14.0
diedin	1	3.9	**1.6**	3.7	2.0	1.4	**0.5**	0.7	0.7
happenedin	1	6.6	**1.7**	4.0	3.3	2.3	**0.7**	1.0	1.1
iscitizenof	2	**13.1**	14.5	20.0	13.2	11.0	**9.9**	10.2	10.0
isleaderof	1	58.1	**55.5**	58.0	56.6	74.1	**72.0**	72.7	71.6
ispoliticianof	3	**8.0**	9.3	16.9	7.8	9.2	**8.3**	8.4	8.5
livesin	1	7.0	**6.8**	8.5	6.7	4.7	**4.0**	4.1	4.1
participatedin	1	16.0	**11.6**	8.1	14.1	**10.9**	13.2	12.1	14.2

Table 2: Results for **Q3**. $\left[\widehat{\text{conf}} - |\mathbf{R}|/|\mathbf{R}^+_{\mathbf{s}}| \cdot \text{conf}\right]^2 \cdot 10^4$ for SAR$_{\text{group}}$ with PCA upheld (avg. over q and rules predicting p). $c_{\neg q} \in \{0.3, 0.7\}$ with $c_q = 0.5$ Bold is best per $c_{\neg q}$ and p. Three *IPW-PCA*(R) columns for $\hat{c}_{\neg q} = c_{\neg q} + \Delta$, $\Delta \in \{0, \pm 0.1\}$. Rules are included if $0.3 \leq (|\mathbf{R} \cap S_q|)/|\mathbf{R}| \leq 0.7$ and the difference in group-local $|\mathbf{R}|/|\mathbf{R}^+_{\mathbf{s}}| \cdot$ *conf*(R) is at least 0.1.

confidence measures.

Biases are mostly studied in the **fairness** literature (Mehrabi et al. 2019; Barocas, Hardt, and Narayanan 2019), with the aim to learn bias-free models. This paper, in contrast, does not enforce certain ideals, but rather aims to increase model quality by being conscious of observation biases. This is a recent perspective in fairness literature (Blum and Stangl 2020).

7 Conclusion

We investigated rule evaluation, specifically confidence estimation, for knowledge base completion in the face of observation biases. Our theoretical and empirical analysis has shown that ignoring the observation bias results in biased confidence estimates. Yet, this is exactly what existing methods do. We have proposed two new confidence estimators that can mitigate known biases by using propensity scores that quantify how likely a fact is to be included the KB. We have shown that these estimators are unbiased with respect to the observation bias. Our experiments showed that the Brier score of our *IPW*(R) measure is often orders of magnitude lower than those of the other estimators when observation biases are present. Our metric even outperforms the others when it has inexact values for the propensity scores.

Acknowledgements

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. Jonas Schouterden is supported by the KU Leuven Research Fund (C14/17/070). Jessa Bekker is also supported by the Research Foundation - Flanders under the Data- driven logistics project (FWO-S007318N), and Jesse Davis is also supported by the Research Foundation - Flanders (G0D8819N).

References

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, 722–735. Springer.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Bekker, J.; and Davis, J. 2020. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4): 719–760.

Bekker, J.; Robberechts, P.; and Davis, J. 2019. Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data. In *ECML PKDD: Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer.

Blum, A.; and Stangl, K. 2020. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? In Roth, A., ed., *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, volume 156 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 3:1–3:20. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. ISBN 978-3-95977-142-9.

Callahan, E.; and Herring, S. 2011. Cultural bias in Wikipedia content on famous persons. *J. Assoc. Inf. Sci. Technol.*, 62: 1899–1915.

Dong, X. L.; Gabrilovich, E.; Heitz, G.; Horn, W.; Lao, N.; Murphy, K.; Strohmann, T.; Sun, S.; and Zhang, W. 2014. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, 601–610. Association for Computing Machinery. ISBN 9781450329569.

Elkan, C.; and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 213–220. ACM.

Fürnkranz, J.; Gamberger, D.; and Lavrač, N. 2014. *Foundations of Rule Learning*. Springer Publishing Company, Incorporated. ISBN 3642430465, 9783642430466.

Galárraga, L.; Razniewski, S.; Amarilli, A.; and Suchanek, F. M. 2017. Predicting Completeness in Knowledge Bases. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*.

Galárraga, L.; Teflioudi, C.; Hose, K.; and Suchanek, F. M. 2015. Fast rule mining in ontological knowledge bases with AMIE+. *The VLDB Journal*, 24: 707–730.

Galárraga, L. A.; Teflioudi, C.; Hose, K.; and Suchanek, F. 2013. AMIE: Association Rule Mining Under Incomplete Evidence in Ontological Knowledge Bases. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, 413–422. New York, NY, USA: ACM. ISBN 978-1-4503-2035-1.

Gong, C.; Wang, Q.; Liu, T.; Han, B.; You, J. J.; Yang, J.; and Tao, D. 2021. Instance-Dependent Positive and Unlabeled Learning with Labeling Bias Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Gupta, S.; Wang, H.; Lipton, Z. C.; and Wang, Y. 2021. Correcting Exposure Bias for Link Recommendation. In *ICML*.

Kato, M.; Teshima, T.; and Honda, J. 2018. Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*.

Lajus, J.; Galárraga, L.; and Suchanek, F. M. 2020. Fast and Exact Rule Mining with AMIE 3. *The Semantic Web*, 12123: 36 – 52.

Mahdisoltani, F.; Biega, J. A.; and Suchanek, F. M. 2015. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Meilicke, C.; Chekol, M. W.; Ruffinelli, D.; and Stuckenschmidt, H. 2019. Anytime Bottom-Up Rule Learning for Knowledge Graph Completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 3137–3143. International Joint Conferences on Artificial Intelligence Organization.

Nickel, M.; Murphy, K.; Tresp, V.; and Gabrilovich, E. 2015. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1): 11–33.

Pellissier Tanon, T.; Stepanova, D.; Razniewski, S.; Mirza, P.; and Weikum, G. 2017. Completeness-Aware Rule Learning from Knowledge Graphs. In *The Semantic Web – ISWC 2017*, volume 1, 507–525. ISBN 978-3-319-68288-4.

Pezeshkpour, P.; Tian, Y.; and Singh, S. 2020. Revisiting Evaluation of Knowledge Base Completion Models. In *Automated Knowledge Base Construction*.

Razniewski, S.; Suchanek, F.; and Nutt, W. 2016. But what do we actually know? In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, 40–44.

Rebele, T.; Suchanek, F.; Hoffart, J.; Biega, J.; Kuzey, E.; and Weikum, G. 2016. YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *International semantic web conference*, 177–185. Springer.

Saito, Y.; Yaginuma, S.; Nishino, Y.; Sakata, H.; and Nakata, K. 2020. Unbiased Recommender Learning from Missing-Not-At-Random Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 501–509.

Soulet, A.; Giacometti, A.; Bouchou-Markhoff, B.; and Suchanek, F. M. 2018. Representativeness of Knowledge Bases with the Generalized Benford's Law. In *International Semantic Web Conference*.

Speranskaya, M.; Schmitt, M.; and Roth, B. 2020. Ranking vs. Classifying: Measuring Knowledge Base Completion Quality. In *Automated Knowledge Base Construction*.

Vrandecic, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57: 78–85.

Wagner, C.; Garcia, D.; Jadidi, M.; and Strohmaier, M. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media*.

WWW Consortium. 2004. RDF Primer (W3C Recommendation 2004-02-10).

Zhou, X.; Sadeghian, A.; and Wang, D. 2019. Mining Rules Incrementally over Large Knowledge Bases. *ArXiv*, abs/1904.09399.

Zupanc, K.; and Davis, J. 2018. Estimating rule quality for knowledge base completion with the relationship between coverage assumption. In *Proceedings of the Web Conference 2018*, 1–9.

A PCA Confidence and its Expected Value

A.1 Rewriting the expected value of *PCA*(R) in function of its biases and the true confidence

We approximate the expected value of the PCA-based confidence estimator using the fact that the first-order Taylor approximation of $\mathbb{E}[X/Y]$ is $\mathbb{E}(X)/\mathbb{E}(Y)$:

$$\begin{aligned}\mathop{\mathbb{E}}_{\text{sel}\sim e}[\textit{PCA}(R)] &= \mathop{\mathbb{E}}_{\text{sel}\sim e}\left[\frac{\sum_{\langle s,o\rangle\in\mathbf{R}} l(s,o)}{\sum_{\langle s,o\rangle\in\mathbf{R}} l(s)}\right] \\ &\approx \frac{\mathbb{E}_{\text{sel}\sim e}\left[\sum_{\langle s,o\rangle\in\mathbf{R}} l(s,o)\right]}{\mathbb{E}_{\text{sel}\sim e}\left[\sum_{\langle s,o\rangle\in\mathbf{R}} l(s)\right]} \\ &= \frac{\sum_{\langle s,o\rangle\in\mathbf{R}} y(s,o)e(s,o)}{\sum_{\langle s,o\rangle\in\mathbf{R}} y(s)e(s)}\end{aligned}$$

This approximation can be rewritten as follows:

$$\begin{aligned}\mathop{\mathbb{E}}_{\text{sel}\sim e}[\textit{PCA}(R)] &\approx \text{conf}(R) && \textit{bias}_{\not{\textit{PCA}}}(R) && \textit{bias}_{y(s)=0}(R) && \textit{bias}_{e(s)}(R) \\ &= \text{conf}(R) && \frac{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s,o)}{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)} && \frac{|\mathbf{R}|}{|\mathbf{R}_\mathbf{s}^+|} && \frac{\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)}{\frac{1}{|\mathbf{R}_\mathbf{s}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}_\mathbf{s}^+} e(s)}.\end{aligned}$$

Proof.

$$\begin{aligned}&\text{conf}(R)\frac{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s,o)}{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)}\frac{|\mathbf{R}|}{|\mathbf{R}_\mathbf{s}^+|}\frac{\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)}{\frac{1}{|\mathbf{R}_\mathbf{s}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}_\mathbf{s}^+} e(s)} \\ =\ &\frac{|\mathbf{R}^+|}{|\mathbf{R}|}\frac{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s,o)}{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)}\frac{|\mathbf{R}|}{|\mathbf{R}_\mathbf{s}^+|}\frac{\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)}{\frac{1}{|\mathbf{R}_\mathbf{s}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}_\mathbf{s}^+} e(s)} \\ &\textit{Remove terms that cancel each other out} \\ =\ &\frac{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s,o)}{\sum_{\langle s,o\rangle\in\mathbf{R}_\mathbf{s}^+} e(s)} \\ &\textit{Sum over } \mathbf{R} \textit{ and select triples from } \mathbf{R}^+ \textit{ and } \mathbf{R}_\mathbf{s}^+ \textit{ using } y(s,o) \textit{ and } y(s) \\ =\ &\frac{\sum_{\langle s,o\rangle\in\mathbf{R}} y(s,o)e(s,o)}{\sum_{\langle s,o\rangle\in\mathbf{R}} y(\langle s,p\rangle)e(s)}\end{aligned}$$

□

bias$_{\not{\textit{PCA}}}(R)$**: Bias due to PCA being violated**

$$\textit{bias}_{\not{\textit{PCA}}}(R) = \frac{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s,o)}{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)}$$

When the PCA assumption holds, $e(s) = e(s,o)$, therefore the *bias*$_{\not{\textit{PCA}}}$ disappears as it reduces to 1.

When the PCA assumption is violated, then $e(s) > e(s,o)$[6] and consequently *bias*$_{\not{\textit{PCA}}}(R) < 1$. In this case, *bias*$_{\not{\textit{PCA}}}(R)$ leads to an underestimation, as it represents how much is deviated from PCA. When the triples are selected independently, the probability for selecting a subject is: $e(s) = 1 - \prod_{o:\langle s,o\rangle\in\mathbf{R}^+}(1 - e(s,o))$.

bias$_{y(s)=0}$**: Bias due to subjects that appear in no facts**

$$\textit{bias}_{y(s)=0} = \frac{|\mathbf{R}|}{|\mathbf{R}_\mathbf{s}^+|}$$

This bias does not depend on the selection mechanism. It will always play a role when $\mathbf{R}_\mathbf{s}^+ \subset \mathbf{R}$.

[6] $e(s)$ is the probability of *at least one positive triple* being selected, thus larger than the probability of a specific triple being selected.

bias$_{e(s)}(R)$**: Bias due to the pair-selection mechanism** $e(s)$

$$\textit{bias}_{e(s)}(R) = \frac{\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)}{\frac{1}{|\mathbf{R}_\mathbf{s}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}_\mathbf{s}^+} e(s)}.$$

When $e(s)$ is a **constant** for a rule (i.e. under CWA, SCAR-per-predicate, and SCAR-per-rule), then $\textit{bias}_{e(s)(R)} = 1$, i.e. this bias does not play a role. If under these selection mechanisms the PCA assumption holds (i.e. $\textit{bias}_{\cancel{PCA}}(R) = 1$), then the PCA-based confidence is only biased by $\textit{bias}_{y(s)=0}$:

$$\begin{aligned}\mathop{\mathbb{E}}_{\text{sel}\sim e}[\textit{PCA}(R)] &\approx \text{conf}(R)\cdot \textit{bias}_{y(s)=0}(R)\\ &= \text{conf}(R)\cdot\frac{|\mathbf{R}|}{|\mathbf{R}_\mathbf{s}^+|}\end{aligned}$$

When $e(s)$ is **not a constant** (general SAR case), then $\textit{PCA}(R)$ can suffer from this observation bias. To see when this bias occurs, note how the numerator and denominator of $\textit{bias}_{e(s)}$ sum over respectively $\mathbf{R}^+$ and $\mathbf{R}_\mathbf{s}^+$, with $\mathbf{R}^+ \subseteq \mathbf{R}_\mathbf{s}^+$. The numerator of $\textit{bias}_{e(s)}$ represents the fraction of predicted facts that are *observed*, while the denominator represents the fraction of triples that have an *observed subject*. Their summations can be rewritten as summations over the subjects of the triples in $\mathbf{R}^+$ and $\mathbf{R}_\mathbf{s}^+$:

$$\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s) = \frac{\sum_{s:\exists\langle s,o\rangle\in\mathbf{R}^+} w^+(s)\cdot e(s)}{\sum_{s:\exists\langle s,o\rangle\in\mathbf{R}^+} w^+(s)}$$

$$\frac{1}{|\mathbf{R}_\mathbf{s}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}_\mathbf{s}^+} e(s) = \frac{\sum_{s:\exists\langle s,o\rangle\in\mathbf{R}_\mathbf{s}^+} w_s^+(s)\cdot e(s)}{\sum_{s:\exists\langle s,o\rangle\in\mathbf{R}_\mathbf{s}^+} w_s^+(s)}$$

Thus, the numerator and denominator are the weighted means of $e(s)$ over the subjects s occurring in respectively $\mathbf{R}^+$ and $\mathbf{R}_\mathbf{s}^+$. Here, $w^+(s)$ is the number of true facts of s in $\mathbf{R}^+$, while $w_s^+(s)$ is the number of predicted triples with s in $\mathbf{R}_\mathbf{s}^+$. The fraction

$$\frac{w^+(s)}{w_s^+(s)} \in [0,\ 1]$$

is the *subject-local* confidence of rule R for s.

For no bias to occur, i.e. for $\textit{bias}_{e(s)} = 1$, the numerator and denominator should be equal. Note that for a given s, should the number of true facts among s's triples in $\mathbf{R}_\mathbf{s}^+$ increase, i.e. should the subject-specific confidence $w^+(s)/w_s^+(s)$ increase, then the weight for s in the numerator increases relative to the denominator. Consequently, should there be a correlation between the propensity scores $e(s)$ and the subject-specific confidences $w^+(s)/w_s^+(s)$, the numerator and denominator of $\textit{bias}_{e(s)}$ are not equal, and this bias occurs.

B How good do the propensity scores estimates need to be to result in better confidence estimates than methods assuming SCAR?

The IPW(-PCA) confidence requires propensity score estimates $\hat{e}(s,o)$ as input. In this section, we investigate how much well the confidence $\text{conf}(R)$ can be estimated if we have an *imperfect* description of the observation bias, i.e. $\hat{e}(s,o) \neq e(s,o)$ for some $\langle s,o\rangle$. In other words, how accurate do the propensity score estimates $\hat{e}(s,o)$ need to be for inverse propensity weighted confidence estimators to be better than confidence estimators assuming SCAR? Section B.1 investigates the setting where the PCA assumption does not hold by comparing $\textit{IPW}(R)$ with $\textit{CWA}(R)$ and $\textit{ICW}(R)$. Section B.2 investigates the setting where the PCA assumption does hold by comparing $\textit{IPW-PCA}(R)$ with *PCA*.

B.1 SAR without PCA

First, we compare the IPW confidence estimator with the CWA-based estimator. Then, we compare the IPW estimator with the ICW estimator, and look at the effect of calibrating propensity score estimates.

SAR without PCA: $\textit{IPW}(R)$ **vs** $\textit{CWA}(R)$ Our goal is to investigate for which estimated propensity scores $\hat{e}(s,o)$ the IPW estimator is a better estimator for $\text{conf}(R)$ than the CWA-based estimator:

$$\left|\mathop{\mathbb{E}}_{\text{sel}\sim e}[\textit{IPW}(R)] - \text{conf}(R)\right| \leq \left|\mathop{\mathbb{E}}_{\text{sel}\sim e}[\textit{CWA}(R)] - \text{conf}(R)\right|$$

This can be rewritten as:

$$\left|\text{conf}(R)\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+}\frac{e(s,o)}{\hat{e}(s,o)}-\text{conf}(R)\right| \leq \left|\text{conf}(R)\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+}e(s,o)-\text{conf}(R)\right|$$

$$\left|\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+}\frac{e(s,o)}{\hat{e}(s,o)}-1\right| \leq \left|\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+}e(s,o)-1\right|$$

$$\left|\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+}\left[\frac{e(s,o)}{\hat{e}(s,o)}-1\right]\right| \leq \left|\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+}\left[e(s,o)-1\right]\right|$$

These two errors are the sums of the error terms of the individual triples $\langle s,o\rangle \in \mathbf{R}^+$:

$$\left[\frac{e(s,o)}{\hat{e}(s,o)}-1\right] \text{ for the IPW estimator}$$

$$\left[e(s,o)-1\right] \text{ for the CWA-based estimator}$$

The error terms of the individual triples can be positive or negative, so they can cancel each other out. Still, it is interesting to see that a *sufficient (but not necessary)* condition for the IPW error to be smaller than the CWA error is when for each triple, its contribution to the IPW error is smaller than its contribution to the CWA error:

$$\forall\langle s,o\rangle\in\mathbf{R}^+ : \left|\frac{e(s,o)}{\hat{e}(s,o)}-1\right| \leq \left|e(s,o)-1\right| \tag{1}$$

When rewriting this, this corresponds to the following condition:

$$\frac{e(s,o)}{2-e(s,o)} \leq \hat{e}(s,o) \leq 1$$

Note that as $\frac{e(s,o)}{2-e(s,o)} \leq e(s,o)$, this condition also holds if $e(s,o) \leq \hat{e}(s,o) \leq 1$, i.e. if the propensity score estimate $\hat{e}(s,o)$ is an *overestimate* of the true value $e(s,o)$. If $\hat{e}(s,o)$ is an underestimate, than the error contribution of this triple is still smaller for the IPW estimator than the CWA-based estimator as long as $\frac{e(s,o)}{2-e(s,o)} \leq \hat{e}(s,o)$

Thus, for the IPW estimator to do better than the CWA confidence, a sufficient condition is that the propensity score estimates are either overestimates of their true values or ***reasonable*** *underestimates*. Here, $\frac{e(s,o)}{2-e(s,o)}$ defines what is "reasonable".

SAR without PCA: *IPW*(R) vs *ICW*(R) Under SCAR-per-predicate (SCAR$_p$), the CWA-based estimator has a constant multiplicative bias c_p. Thus, under SCAR$_p$, the quality of the CWA-based estimator is not affected by this bias c_p when used to compare rules using ranking, as the CWA confidence for all rules[7] have this constant multiplicative bias. Consequently, to investigate how good the propensity scores have to be for the IPW confidence to be better than estimators assuming SCAR, it is more fair to compare the IPW confidence to the ICW confidence (than the CWA confidence), as the ICW confidence corrects for this multiplicative bias; $\mathit{ICW}(R) = \frac{1}{c_p}\mathit{CWA}(R)$. Under the more general SAR, c_p is defined as $c_p = \mathbb{E}_{\langle s,o\rangle\in\mathcal{E}\times p\times\mathcal{E}}[e(s,o)]$.

Similarly to the ICW estimator being a calibrated version of the CWA-based estimator, we can *calibrate* the estimated propensity scores $\hat{e}(s,o)$ that we use in the IPW estimator. To calibrate the estimated propensity scores, they are multiplied with a positive constant $\alpha > 0$:

$$cal(\hat{e}(s,o)) = \alpha \cdot \hat{e}(s,o)$$

The calibration constant α is chosen such that:

$$\mathop{\mathbb{E}}_{\langle s,p,o\rangle\in F|p=p}\left[\frac{e(s,o)}{cal(\hat{e}(s,o))}\right] = \mathop{\mathbb{E}}_{\langle s,p,o\rangle\in F|p=p}\left[\frac{e(s,o)}{c_p}\right] = 1.$$

with F being the set of facts in I. That is, on average, the calibrated propensity scores

Recall, the goal is to investigate conditions under which the (calibrated) estimated propensity scores $cal(\hat{e}(s,o))$ result in the IPW estimator being a better estimator for $\text{conf}(R)$ than the ICW estimator, i.e. when the expected IPW estimator is closer to the true confidence than the ICW estimator:

$$\left|\mathop{\mathbb{E}}_{\text{sel}\sim e}[\mathit{IPW}(R)] - \text{conf}(R)\right| \leq \left|\mathop{\mathbb{E}}_{\text{sel}\sim e}[\mathit{ICW}(R)] - \text{conf}(R)\right|$$

[7] Assuming the rules all predict the same predicate p.

This can be rewritten as:

$$\left|\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+}\left[\frac{e(s,o)}{cal(\hat{e}(s,o))}-1\right]\right| \leq \left|\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+}\left[\frac{e(s,o)}{c_p}-1\right]\right|$$

These two errors are the sums of the error terms of the individual triples $\langle s,o\rangle \in \mathbf{R}^+$:

$$\left[\frac{e(s,o)}{\hat{e}(s,o)}-1\right] \text{ for the IPW estimator}$$
$$\left[\frac{e(s,o)}{c_p}-1\right] \text{ for the ICW estimator}$$

The error terms of the individual triples can be positive or negative, so they can cancel each other out. Still, it is interesting to see that a *sufficient (but not necessary)* condition for the IPW error to be smaller than the CWA (ICW) error is when for each triple, its contribution to the IPW error is smaller than its contribution to the CWA (ICW) error:

$$\forall\langle s,o\rangle \in \mathbf{R}^+ : \left|\frac{e(s,o)}{\hat{e}(s,o)}-1\right| \leq \left|\frac{e(s,o)}{c_p}-1\right| \tag{2}$$

Or when rewritten:

$$\begin{aligned}
\left|\frac{e(s,o)}{cal(\hat{e}(s,o))}-1\right| &\leq \left|\frac{e(s,o)}{c_p}-1\right| \\
\left|\frac{e(s,o)-cal(\hat{e}(s,o))}{cal(\hat{e}(s,o))}\right| &\leq \left|\frac{e(s,o)-c_p}{c_p}\right| \\
\frac{|e(s,o)-cal(\hat{e}(s,o))|}{cal(\hat{e}(s,o))} &\leq \frac{|e(s,o)-c_p|}{c_p} \qquad c_p>0,\ cal(\hat{e}(s,o))>0
\end{aligned}$$

Rewriting the numerators $|e(s,o)-c_p|$ and $|e(s,o)-cal(\hat{e}(s,o))|$ results in 4 different cases:

$$\begin{cases}
\begin{aligned} c_p\cdot[e(s,o)-cal(\hat{e}(s,o))] &\leq cal(\hat{e}(s,o))\cdot[e(s,o)-c_p] \\ c_p &\leq cal(\hat{e}(s,o)) \end{aligned} & \text{, when } e(s,o)\geq c_p \text{ and } e(s,o)\geq cal(\hat{e}(s,o)) \\
\begin{aligned} c_p\cdot[e(s,o)-cal(\hat{e}(s,o))] &\geq -cal(\hat{e}(s,o))\cdot[e(s,o)-c_p] \\ c_p e(s,o) &\geq -cal(\hat{e}(s,o))e(s,o)+2c_p\,cal(\hat{e}(s,o)) \\ cal(\hat{e}(s,o)) &\leq e(s,o)\frac{1}{2-e(s,o)/c_p} \end{aligned} & \text{, when } e(s,o)\geq c_p \text{ and } e(s,o)\leq cal(\hat{e}(s,o)) \\
\begin{aligned} c_p\cdot[e(s,o)-cal(\hat{e}(s,o))] &\geq cal(\hat{e}(s,o))\cdot[e(s,o)-c_p] \\ c_p &\geq cal(\hat{e}(s,o)) \end{aligned} & \text{, when } e(s,o)\leq c_p \text{ and } e(s,o)\leq cal(\hat{e}(s,o)) \\
\begin{aligned} c_p\cdot[e(s,o)-cal(\hat{e}(s,o))] &\leq -cal(\hat{e}(s,o))\cdot[e(s,o)-c_p] \\ cal(\hat{e}(s,o)) &\geq e(s,o)\frac{1}{2-e(s,o)/c_p} \end{aligned} & \text{, when } e(s,o)\leq c_p \text{ and } e(s,o)\geq cal(\hat{e}(s,o))
\end{cases}$$

These cases are summarized as the following intervals for the (calibrated) estimated propensity scores $cal(\hat{e}(s,o))$:

$$\begin{cases}
e(s,o)\frac{1}{2-e(s,o)/c_p} \geq cal\left(\hat{e}(s,o)\right) \geq c_p & \text{, when } e(s,o)\geq c_p, \\
e(s,o)\frac{1}{2-e(s,o)/c_p} \leq cal\left(\hat{e}(s,o)\right) \leq c_p & \text{, when } e(s,o)\leq c_p,
\end{cases}$$

These intervals are also depicted in Figure 8. If the (calibrated) propensity scores $cal(\hat{e}(s,o))$ fall within these intervals, this is sufficient for the IPW estimator to be better than the ICW estimator.

These intervals show that using propensity score estimate $\hat{e}(s,o)$ corresponding to a *weaker* version of the observation bias or a *"reasonable" exaggeration* of the observation bias is preferable over not taking the bias into account at all. A **weaker** version of the observation bias means that $cal\left(\hat{e}(s,o)\right)$ is in between c and $e(s,o)$, i.e.,

$$\begin{cases}
c \leq cal\left((s,o)\right) \leq e(s,o) & \text{if } c\leq e(s,o) \\
c \geq cal\left((s,o)\right) \geq e(s,o) & \text{if } c\geq e(s,o)
\end{cases}$$

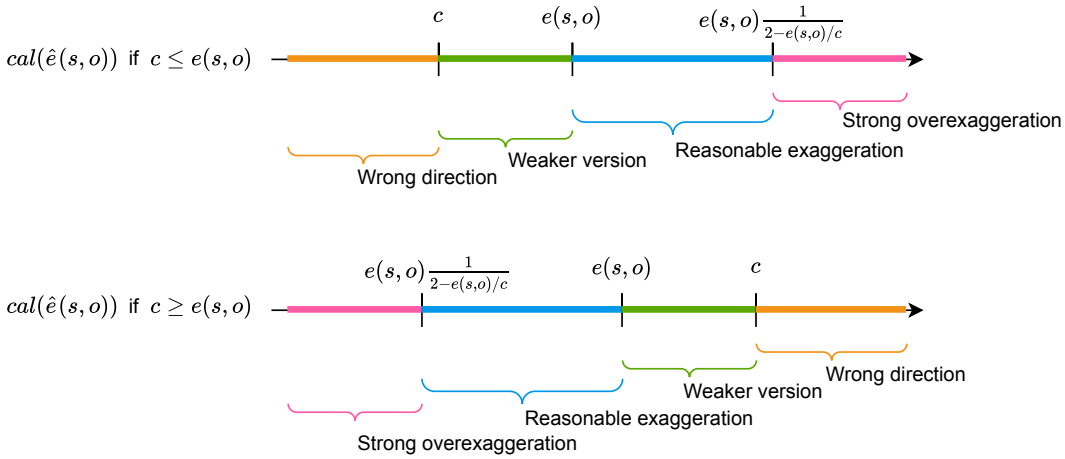


Figure 8: Possible values of the propensity score estimate $cal\,(\hat{e}(s,o))$ for a triple $\langle s,o\rangle$ in relation to the true value $e(s,o)$ and c. If $cal\,(\hat{e}(s,o))$ is a weaker version of the observation mechanism or a reasonable exaggeration, then the error contribution of the triple $\langle s,o\rangle$ is smaller for the IPW estimator than for the CWA (ICW) estimator.

An **exaggeration** (**stronger** version of the observation bias) means that $cal\,((s,o))$ is in the same direction as $e(s,o)$, but further away from c, i.e.,

$$\begin{cases} c \leq e(s,o) \leq cal\,((s,o)) & \text{if } c \leq e(s,o) \\ c \geq e(s,o) \geq cal\,((s,o)) & \text{if } c \geq e(s,o) \end{cases}$$

How much exaggeration is **"reasonable"** is defined by $\frac{1}{2-e(s,o)/c}$: the stronger the actual observation bias is (i.e., the more $e(s,o)/c$ deviates from 1), the stronger of an exaggeration is acceptable.

In conclusion, these sufficient conditions show that, even when the exact propensity scores $e(s,o)$ are unknown, a reasonable estimate $cal(\hat{e}(s,o))$ can still result in the IPW estimator being a better estimator than confidence estimators making the SCAR assumption.

B.2 SAR with PCA: *IPW-PCA*(R) vs *PCA*(R)

Our goal is to investigate for which estimated propensity scores $\hat{e}(s,o)$ the IPW-PCA estimator is a better estimator for $\mathrm{conf}(R)$ than the PCA estimator when the PCA is upheld bu the selection mechanism:

$$\left| \mathop{\mathbb{E}}_{\mathrm{sel}\sim e}[\textit{IPW-PCA}(R)] - \mathrm{conf}(R) \right| \leq \left| \mathop{\mathbb{E}}_{\mathrm{sel}\sim e}[\textit{PCA}(R)] - \mathrm{conf}(R) \right|$$

Recall the first-order Taylor approximations:

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathrm{sel}\sim e}[\textit{PCA}(R)] &\approx \mathrm{conf}(R) \cdot \textit{bias}_{\cancel{PCA}}(R) \cdot \textit{bias}_{y(s)=0}(R) \cdot \textit{bias}_{e(s)}(R) \\ &\approx \mathrm{conf}(R) \cdot 1 \cdot \frac{|\mathbf{R}|}{|\mathbf{R_s^+}|} \cdot \frac{\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)}{\frac{1}{|\mathbf{R_s^+}|}\sum_{\langle s,o\rangle\in\mathbf{R_s^+}} e(s)} \\ \mathop{\mathbb{E}}_{\mathrm{sel}\sim e}[\textit{IPW-PCA}(R)] &\approx \mathrm{conf}(R) \cdot \textit{bias}_{\cancel{PCA}}^{\textit{IPW-PCA}}(R) \cdot \textit{bias}_{y(s)=0}(R) \cdot \textit{bias}_{e(s)}^{\textit{IPW-PCA}}(R) \\ &\approx \mathrm{conf}(R) \cdot 1 \cdot \frac{|\mathbf{R}|}{|\mathbf{R_s^+}|} \cdot \frac{\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+} \frac{e(s)}{\hat{e}(s)}}{\frac{1}{|\mathbf{R_s^+}|}\sum_{\langle s,o\rangle\in\mathbf{R_s^+}} \frac{e(s)}{\hat{e}(s)}} \end{aligned}$$

In this comparison, we assume that PCA holds, and thus $\textit{bias}_{\cancel{PCA}}(R) = \textit{bias}_{\cancel{PCA}}^{\textit{IPW-PCA}}(R) = 1$. Note how under PCA, the *IPW-PCA*(R) and *PCA*(R) only differ in their $\textit{bias}_{e(s)}$, as they have the same $\cdot\textit{bias}_{y(s)=0}(R)$. We will disregard the effect of the $\textit{bias}_{y(s)=0}$, to concentrate fully on the biases introduced by the selection mechanism: $\textit{bias}_{e(s)}$ and $\textit{bias}_{e(s)}^{\textit{IPW-PCA}}$.

To simplify the mathematics, we will write $\textit{bias}_{e(s)}$ for both estimators as if they are *IPW-PCA*(R), but we will restrict the *PCA*(R) to a constant value for $\hat{e}(s)$. This is equivalent to *PCA*(R), because constant multiplicative biases in the propensity

scores are canceled out in any case. Furthermore, we calibrate $\hat{e}(s)$ to remove any multiplicative bias in the propensity scores. To calibrate the estimated propensity scores, they are multiplied with a constant $cal\left(\hat{e}(s)\right) = \alpha \cdot \hat{e}(s)$, with α chosen such that

$$\begin{aligned} &\mathop{\mathbb{E}}_{\langle s,o\rangle \in \mathbf{R_s^+}} \left[\frac{e(s)}{cal(\hat{e}(s))}\right] = \mathop{\mathbb{E}}_{\langle s,o\rangle \in \mathbf{R_s^+}} \left[\frac{e(s)}{c_R}\right] = 1 \\ \Leftrightarrow\ & c_R = \frac{1}{|\mathbf{R_s^+}|} \sum_{\langle s,o\rangle \in \mathbf{R_s^+}} e(s). \end{aligned}$$

The constant $\hat{e}(s)$ used for *PCA*(R) is c_R to also have it calibrated. The effect of the calibrations is that the denominator of $bias_{e(s)}$ of for the PCA-based estimators both become 1:

$$\begin{aligned} bias_{e(s)}^{\textit{IPW-PCA}}(R) &= \frac{\frac{1}{|\mathbf{R^+}|}\sum_{\langle s,o\rangle \in \mathbf{R^+}} \frac{e(s)}{cal(\hat{e}(s))}}{\frac{1}{|\mathbf{R_s^+}|}\sum_{\langle s,o\rangle \in \mathbf{R^+}} \frac{e(s)}{cal(\hat{e}(s))}} &&= \frac{1}{|\mathbf{R^+}|} \sum_{\langle s,o\rangle \in \mathbf{R^+}} \frac{e(s)}{cal(\hat{e}(s))} \\ bias_{e(s)}^{\textit{PCA}}(R) &= \frac{\frac{1}{|\mathbf{R^+}|}\sum_{\langle s,o\rangle \in \mathbf{R^+}} \frac{e(s)}{c_R}}{\frac{1}{|\mathbf{R_s^+}|}\sum_{\langle s,o\rangle \in \mathbf{R^+}} \frac{e(s)}{c_R}} &&= \frac{1}{|\mathbf{R^+}|} \sum_{\langle s,o\rangle \in \mathbf{R^+}} \frac{e(s)}{c_R} \end{aligned}$$

The expected estimate of the calibrated Inverse Propensity Weighted PCA-based estimator is better than the one from the calibrated PCA-based estimator when it their $bias_{e(s)}(R)$ is closer to 1. The error is the sum of the errors of individual triples. The errors can positive or negative, so they can cancel each other out. Still, it is interesting to see that for an individual positive triple, its contributed error to *IPW-PCA*(R) is smaller than to *PCA*(R) when:

$$\left|\frac{e(s)}{cal(\hat{e}(s))} - 1\right| \leq \left|\frac{e(s)}{c_R} - 1\right|$$

A derivation analogous to the one for *IPW*(R) vs *CWA*(R), shows that the inequality holds when:

$$\begin{cases} e(s)\frac{1}{2-e(s)/c_R} \geq cal\left(\hat{e}(s)\right) \geq c_R & \text{, when } e(s) \geq c_R, \\ e(s)\frac{1}{2-e(s)/c_R} \leq cal\left(\hat{e}(s)\right) \leq c_R & \text{, when } e(s) \leq c_R, \end{cases}$$

C Extended experimental setup

In this section, we describe the experimental setup as used in section 5 in more detail.

C.1 Computing infrastructure: software and hardware

Software All software was completely written in Python3.8, using Numpy, Pandas and Pylo as main libraries. All source code will be made publicly available upon acceptance.

Hardware All experiments were run in parallel using the Dask Python library over a cluster of machines each having an *Intel Core i7-2600 (3.40GHz)* CPU with 16 GB 4x4GB DDR3 and *Ubuntu 20.04.3 LTS (Focal Fossa)* as operating system.

C.2 Accounting for randomness

To account for randomness, at the beginning of each experiment, a *Random* Python object was initialized with seed 3. This *Random* object was used for all random choices, e.g., those made by the selection mechanism. Every experimental setting was run 10 times in a row, initializing the *Random* object only once and reusing it over multiple iterations without reinitialization. By setting a fixed seed, the random choices by the selection mechanism become in a sense deterministic, as the same results are obtained every time. Thus note that every result for a specific setting of the parameters for a selection mechanism to generate a K from I is actually the mean over 10 random trials.

C.3 The selection mechanisms and their chosen parameters

Each of our selection mechanisms selecting a K from I is specified using a set of parameters.

The first parameter is the **choice of the predicate** p to which the selection mechanism is applied. That is, the facts with predicate p in K are obtained by applying a selection mechanism to the facts with predicate p in I. All other facts with predicate $q \neq p$ are completely included in K. We also call p the *predicted* predicate, as we only apply rules predicting p to K. For the motivation behind applying the selection mechanisms to one p at a time, see section 5.

The second (set of) parameter(s) is a **specification of the correct propensity scores** e that are used to generate K from I. This specification is different for SCAR_p, $\text{SCAR}_{\text{group}}$ and SCAR_{pop}.

The third parameter is **whether or not the PCA is explicitly enforced** for non-functional p, i.e. whether the selection mechanism selects pairs $\langle s, p\rangle$ or triples $\langle s, p, o\rangle$. When the PCA is not explicitly enforced, the selection mechanism decides for each p-triple in K whether it should be in I. When the PCA is explicitly enforced, the selection mechanism does not look at the p-triples individually, but at the *subjects* s of those triples. That is, the selection mechanism decides for each subject s whether it is selected to be in K. If s is selected, all p-triples in I are included in K.

We now go over the types of selection mechanism used in this paper, together with the parameters that specify them and the parameter choices used.

SCAR$_p$ The SCAR$_p$ selection mechanism is defined by three parameters: 1) the choice of predicted predicate p, 2) the constant label frequency (i.e. selection probability) $e(\cdot) = c_p$, and 3) whether or not the PCA is explicitly enforced.

For the illustrative examples in figures 5 (left) and 6 (left), c_p is varied between 0.1 and 1 in steps of 0.1 and triples are selected.

In table 1 and table 3, we include results for all p for which rules predicting p were mined with AMIE. We include two choices of c_p ($c_p = 0.3$ and $c_p = 0.7$) and select triples. We also include results for the IPW confidence using noisy propensity scores $\hat{e} = \hat{c}_p$: for each choice of c_p, the noisy propensity scores used in the IPW confidence are $\hat{c}_p = c \pm 0.1$.

SCAR$_{\text{group}}$ Under SAR$_{\text{group}}$, triples with predicate p are partitioned into two sets $S_q, S_{\neg q}$ based on their subject s and a second predicate $q \neq p$: $\langle s, p, o\rangle \in S_q$ if $y(\langle s, q\rangle) = 1$, else $\langle s, p, o\rangle \in S_{\neg q}$. E.g., triples $\langle \textit{Parent}, \textit{hasChild}, \textit{Child}\rangle$ can be partitioned using $q = \textit{isPoliticianOf}$ to reflect whether or not the parent of a child is a politician. Each group as ha constant propensity score: $e(\langle s, p, o\rangle) = c_q$ if $\langle s, p, o\rangle \in S_q$, else $e(\langle s, p, o\rangle) = c_{\neg q}$. Each combination $(p, q, c_q, c_{\neg q})$, together with the choice of selecting triples or subjects of p, corresponds to a different selection mechanism. Note that when $c_q = c_{\neg q}$, this actually becomes a SCAR setting.

To make sure p and q share enough s-entities without completely overlapping, our results only include p, q-combinations that share at least 10 s-entities, and the fraction of shared s-entities for p is between 0.1 and 0.9.

To generate the examples in figure 5 (right), figure 6 (right) and figure 7 (right), c_q is fixed at $c_q = 0.5$, while $c_{\neg q}$ is varied between 0.1 and 1 in steps of 0.1. For figure 5 (right) and figure 6 (right), triples are selected, while for figure 7, the PCA is explicitly upheld by selecting subjects.

In table 1, table 2 and table 4, we fix c_q at $c_q = 0.5$ and include two choice of $c_{\neg q}$ ($c_{\neg q} = 0.3$ and $c_{\neg q} = 0.7$). We also include results for the IPW confidence using noisy propensity scores by adding noise to $c_{\neg q}$: for each choice of c_p, the noisy propensity scores used in the IPW confidence are $\hat{c}_{\neg q} = c_{\neg q} \pm 0.1$, with $\hat{c}_q = c_q = 0.5$.

For table 1 and table 4, the selection mechanisms select triples. For table 2, the selection mechanism explicitly upheld the PCA by selecting subjects.

SCAR$_{\text{pop}}$ Under the popularity-based SCAR$_{\text{pop}}$, the propensity score $e(\langle s, p, o\rangle)$ is a reflection how popular the subject s is in I. As popularity metric, we count in how many times entity s occurs as as either subject or object in facts with predicate $p \neq q$ in I. This count is indicated as $\#(s, p)$:

$$\#(s, p) = \left|\{\langle s, q, \cdot\rangle \in I\} \cup \{\langle \cdot, q, s\rangle \in I\}\right|,\ q \neq p$$

A propensity score in the interval $[0, 1]$ is obtained by feeding the count as input into a scaled logistic function:

$$e(\langle s, p, o\rangle) = \max\left[\frac{2}{1 + e^{-k \cdot \#(s,p)}} - 1,\ e_{\text{min}}\right]$$

Here, setting $e_{\text{min}} > 0$ allows for unpopular entities s (that occur in very few facts) to have a minimum probability to be selected. In our experiments, we set $e_{\text{min}} = 0.01$. Using this propensity score definition, more popular s have a higher e. The scaling factor k determines how often s must occur in facts to have a specific selection probability $e(\langle s, p, o\rangle)$; for the same probability e, a lower k requires s to occur in more facts. e.g., when setting $k \approx 2.94$, subjects s occurring in at least one triple ($\#(s, p) = 1$) already have a selection probability $e(s, p, o) \geq 0.9$, while for the lower $k \approx 0.5$, subjects s should occur in five triples ($\#(s, p) = 5$) for the same selection probability.

For our experimental results in table 1 and table 5, we include two choices of k: $k = 0.01$ and $k = 0.1$ and set $e_{\text{min}} = 0.01$. The noisy propensity scores $\hat{e}$ used with the IPW confidence are obtained by increasing/decreasing each k with 10%.

D Extended experimental results

Here, we provide more detail about the results discussed in section 5. For the conclusions drawn from these results, we refer to section 5.

Obtaining our results In general, we obtain our results in our figures and tables as follows. First, a selection mechanism is specified using a specific set of parameters, as described in section C.3. e.g., for SCAR$_p$, we choose: 1) a predicate p to which the selection mechanism is applied, 2) the constant propensity score c_p and 3) whether or not the PCA is explicitly upheld. Secondly, the selection mechanism is applied to I to select an incomplete KB K. We do this 10 times, resulting in 10 different K. The

random choices of the selection mechanism are obtained using a single *Random* Python object. Our results are always averaged over these 10 random trials of the selection mechanism.

We apply to each K the rules predicting p (the same p as chosen for K), generating predictions per rule. Given I and the predictions of a rule R, the confidence measures for that rule can be calculated. The true confidence $\mathrm{conf}(R)$ for a rule is calculated by evaluating the rule body on K and then checking which of its predictions are in I. Each experimental setup (generating K and applying rules) was given 12 hours to complete; only those rules that finished generating predictions for all 10 random trials are considered in our results.

For each rule R and related K, the squared difference between the confidence estimators $\widehat{\mathrm{conf}}(R)$ (i.e. *CWA*(R), *PCA*(R), *ICW*(R) and *IPW*(R)) and the true confidence $\mathrm{conf}(R)$ these measures approximate is calculated; this is averaged over the 10 randomly selected K. To answer **Q3**, we also calculate the squared difference between *(IPW-)PCA*(R) and the scaled true confidence $|\mathbf{R}|/|\mathbf{R}_{\mathbf{s}}^{+}| \cdot \mathrm{conf}(R)$.

Description of the result tables Our results are listed in table 1 and table 2 in the body of the paper. Table 1 is a summary of the results shown in table 3 for SCAR_p, table 4 for $\mathrm{SCAR}_{\mathrm{group}}$ and table 5 for $\mathrm{SCAR}_{\mathrm{pop}}$. The parameter setup for these tables is described in section C. Each of these tables shows the results for two specific parameter settings for a specific selection mechanism.

Table 3 shows the Brier score results for the $\mathbf{SCAR}_p$ selection mechanism selecting triples; i.e. the PCA is not explicitly upheld. The table has two large vertical parts corresponding to two choices for c_p: $c_p = 0.3$ for the left part and $c_p = 0.7$ for the right part. Every row corresponds to one choice of p for the SCAR_p selection mechanism; on each row, the Brier scores listed for each confidence estimator are the means over the rules predicting that p. The number of rules used per p is listed as well.

Similarly, table 4 shows the Brier scores for the $\mathbf{SCAR}_{\mathbf{group}}$ selection mechanism selecting triples. It has columns for $c_{\neg q} \in \{0.3, 0.7\}$ with $c_q = 0.5$ fixed. Each row is averaged both over the rules predicting the specific p and the q's defining the groups. Not every combination of predicates is used for p and q; see section C for a description of which p,q-combinations are included.

Table 5 shows the Brier scores for the $\mathbf{SAR}_{pop}$ selection mechanism selecting triples. It has columns columns for $k = 0.01$ and $k = 0.1$. Each row shows the mean over the rules predicting p.

The results in table 1 are obtained by taking per confidence estimator the mean over the averaged squared error per p, as shown in table 3 for SCAR_p, table 4 for $\mathrm{SCAR}_{\mathrm{group}}$ and table 5 for $\mathrm{SCAR}_{\mathrm{pop}}$.

Table 4 shows the results for the experimental setup for **Q3**. It also is a $\mathbf{SCAR}_{\mathbf{group}}$ selection mechanism, but here, *the PCA is explicitly upheld for non-functional* p by selecting triples. For this experimental setup, we only considered a specific scenario. That is, we only included rules for which the group-local scaled confidences $|\mathbf{R}|/|\mathbf{R}_{\mathbf{s}}^{+}| \cdot \mathit{conf}(R)$ differ with at least 0.1. Here, 'group-local' indicates that the only the predictions belonging to group S_q or $S_{\neg q}$ are considered. Also, we only include rules such that neither group S_q or $S_{\neg q}$ dominates in the predictions, by requiring

$$0.3 \leq \frac{|\mathbf{R} \cap S_q|}{|\mathbf{R}|} \leq 0.7$$

The two directions of the PCA-based confidence estimator The intuitive idea behind the PCA-based confidence is that "the more the PCA holds" for a predicted relation p of a given rule R, the better the PCA-based confidence works as an approximation for $\mathrm{conf}(R)$. This corresponds to $\mathit{bias}_{\not{PCA}}$ being closer to one. One specific type of relation are functional relations. A relation p is functional if it is a function in the mathematical sense: mapping every subject s in its domain onto at most one object o. An example is the *diedin* relation: every person dies in at most one place. As the PCA holds by definition for functional relations, the PCA-based confidence is often a good confidence estimator in practice for functional relations. (Galárraga et al. 2013) define a 'functionality' measure which measures how much a predicate p acts as a function in K. Measuring the functionality of p is based on the idea that the PCA is more likely to hold for predicates that act more like functions. However, this requires choosing what the domain and range of the function are; i.e. if the subjects represent the function domain and the objects the range, or vice versa. Consider for example the *diedin* relation (*person*, *diedin*, *place*). While every person (domain) dies in at most one place (range), a place (domain) can have multiple people (range) who died there. Thus, a relation p can be a function in one direction but not in the other.

In this paper, the PCA is defined based on the **subjects** s being the domain of the relation p predicted by the considered rule. However, this choice is in a sense arbitrary. The PCA could also be assumed to be based on the **objects** of p. As mentioned by (Galárraga et al. 2013), assuming the PCA for the objects of p has the same effect as adding its inverse relation p^{-1} to the knowledge base and assuming the PCA for the subjects of p^{-1} (e.g., if $p =$ *haschild* is part of K, $p^{-1} =$ *hasparent* could be added).

As the PCA-based confidence differs in performance when predicting p or p^{-1} for a given p, (Galárraga et al. 2013) propose to always calculate the PCA-based confidence estimator as follows: find out in which direction p is the most functional, and use that direction to calculate the PCA-based confidence estimator. However, we choose to include the PCA-based confidences for both p and p^{-1} for a given p in our experimental results as listed in table 3 for SCAR_p, table 4 for $\mathrm{SCAR}_{\mathrm{group}}$ and table 5 for $\mathrm{SCAR}_{\mathrm{pop}}$. For table 1, which summarizes the previous tables, the listed PCA-based confidence is the mean of the two PCA-based confidences (p and p^{-1}).

Table 3: $\left[\widehat{conf} - \text{conf}\right]^2 \cdot 10^4$ for SCAR_p (no PCA upheld), avg. over the rules predicting p. Entries for $c_p \in \{0.3, 0.7\}$. Bold is best per c_p and p. Three IPW confidence columns per c_p: $\hat{c_p} = c_p$ and $\hat{c_p} = c \pm 0.1$.

p	# R	$c_p = 0.3$						$c_p = 0.7$					
		CWA	PCA_p	$\text{PCA}_{p^{-1}}$	IPW	IPW $(-\Delta)$	IPW $(+\Delta)$	CWA	PCA_p	$\text{PCA}_{p^{-1}}$	IPW	IPW $(-\Delta)$	IPW $(+\Delta)$
actedin	1	50.4	43.3	36.3	**0.2**	27.7	6.3	8.7	275.9	6.5	**0.1**	3.6	1.3
created	2	475.6	43.3	41.9	**4.1**	236.5	65.8	85.1	103.3	118.4	**1.6**	30.9	15.1
dealswith	7	178.7	129.5	61.5	**2.7**	109.2	22.1	34.0	13.5	0.8	**0.5**	10.5	6.4
diedin	1	87.8	164.5	54.2	**0.6**	38.7	12.9	15.7	167.6	0.6	**0.1**	5.0	2.7
directed	2	561.3	85.9	36.3	**6.7**	347.9	66.6	106.3	519.0	44.9	**0.6**	32.3	19.2
exports	1	115.3	38.2	97.7	**6.1**	70.1	18.5	22.4	4.1	15.7	**1.6**	8.0	5.2
graduatedfrom	2	255.4	30.5	227.3	**3.0**	147.6	32.3	46.1	2.8	32.1	**0.6**	16.8	7.8
happenedin	1	406.9	276.9	257.7	**0.7**	204.8	53.2	76.0	39.5	5.0	**0.1**	21.8	13.7
hascapital	5	147.5	791.8	544.0	**8.2**	64.3	28.7	28.3	769.5	824.2	**1.2**	7.7	6.5
hasneighbor	3	232.4	82.4	87.5	**11.2**	170.9	30.6	39.0	**3.6**	4.4	4.5	26.0	7.4
imports	2	94.9	38.2	59.7	**2.1**	36.5	16.4	18.5	5.6	3.5	**0.4**	4.2	4.1
iscitizenof	3	242.7	150.1	162.2	**8.9**	133.7	37.7	42.8	212.5	99.8	**1.4**	17.1	7.7
isconnectedto	1	65.0	5.1	6.4	**1.4**	39.9	8.4	13.3	76.9	58.8	**0.2**	3.0	2.8
isleaderof	1	61.6	383.1	110.2	**13.7**	27.0	21.9	8.9	987.6	352.9	**2.3**	7.8	2.5
islocatedin	2	1517.3	748.7	565.4	**3.7**	711.3	208.9	281.1	173.4	4.0	**1.0**	81.6	51.2
ismarriedto	1	56.9	108.3	106.9	**0.2**	26.4	8.0	10.7	141.8	146.9	**0.1**	2.9	2.0
ispoliticianof	5	678.5	431.8	632.3	**16.1**	336.6	103.8	128.3	504.6	628.6	**4.6**	40.7	26.7
livesin	3	78.0	483.0	70.1	**3.0**	53.5	10.3	14.0	719.0	10.3	**0.3**	5.7	2.4
participatedin	1	418.3	172.7	293.7	**0.8**	229.1	51.2	77.2	21.4	49.1	**0.1**	24.3	13.3
wasbornin	2	130.8	56.6	116.0	**0.9**	67.8	17.3	24.8	51.9	17.2	**0.1**	6.8	4.6
worksat	1	287.8	14.9	248.7	**2.9**	158.2	37.4	48.2	29.3	32.0	**1.7**	23.9	7.5

Table 4: $\left[\widehat{conf} - conf\right]^2 \cdot 10^4$ for $\text{SAR}_{\text{group}}$ (no PCA upheld), avg. over the rules predicting p and q. Entries for $c_{\neg q} \in \{0.3, 0.7\}$ with $c_q = 0.5$ Bold is best per $c_{\neg q}$ and p. Three IPW confidence columns per $c_{\neg q}$: $\hat{c}_{neg} = c_{neg}$ and $\hat{c}_{neg} = c_{neg} \pm 0.1$.

p	# rules	$c_{\neg q} = 0.3$							$c_{\neg q} = 0.7$						
		CWA	ICW	PCA_p	$\text{PCA}_{p^{-1}}$	IPW	IPW $(-\Delta)$	IPW $(+\Delta)$	CWA	ICW	PCA_p	$\text{PCA}_{p^{-1}}$	IPW	IPW $(-\Delta)$	IPW $(+\Delta)$
created	2	383.8	4.9	27.3	59.0	**4.4**	107.9	31.1	132.3	3.9	63.2	88.9	**2.3**	15.7	8.0
dealswith	7	100.4	**1.0**	57.9	18.5	1.1	4.1	1.1	83.1	1.7	49.6	13.4	**1.0**	1.3	1.1
diedin	1	78.9	3.1	117.0	46.4	**0.6**	27.3	8.8	19.6	0.1	193.5	1.6	**0.1**	3.5	1.9
directed	2	301.9	**2.4**	165.5	36.6	3.1	113.7	23.4	92.8	1.2	516.7	31.4	**0.5**	11.3	6.5
exports	1	76.4	6.9	19.7	62.3	**6.1**	21.2	6.9	44.7	4.0	14.4	35.7	**3.6**	5.0	4.0
graduatedfrom	2	255.1	25.9	30.2	230.6	**2.9**	145.5	32.1	46.1	4.3	2.8	31.6	**0.6**	16.9	7.7
happenedin	1	287.4	1.3	171.5	143.7	**0.5**	41.6	10.6	137.9	1.4	87.4	34.5	**0.2**	5.4	2.4
imports	2	69.8	**2.3**	23.7	36.1	2.6	12.3	6.9	33.2	2.2	13.7	12.4	**1.1**	2.2	2.2
iscitizenof	3	153.8	13.7	195.8	84.2	**7.3**	22.9	10.7	91.4	9.2	180.4	93.3	**3.6**	5.2	4.6
isleaderof	1	29.0	4.6	640.0	183.9	**4.5**	6.0	4.0	25.5	4.5	723.5	190.2	**4.4**	4.6	4.2
ispoliticianof	5	440.4	14.0	441.0	518.8	**11.2**	65.2	26.1	259.8	14.4	471.0	602.4	**6.4**	10.1	11.5
livesin	3	58.6	6.4	550.0	51.0	**2.0**	22.9	5.2	24.4	1.7	688.3	19.7	**0.8**	3.0	1.7
participatedin	1	294.3	4.8	67.7	202.5	**0.6**	60.2	14.9	153.1	1.5	9.2	106.8	**0.3**	6.9	4.0
wasbornin	2	122.9	22.5	64.8	110.1	**0.7**	53.2	14.4	29.0	5.4	48.2	20.3	**0.1**	5.3	3.9

For answering **Q3**, the PCA was explicitly upheld for non-functional relations by selecting the subjects of a given p instead of triples. As such, table 2 shows the PCA-based confidence for p (and not p^{-1}).

E Mining non-recursive rules mined from Yago3-10 using AMIE

Yago3-10 (Mahdisoltani, Biega, and Suchanek 2015) is a benchmark KBC dataset we used as I from which different K are selected using various selection mechanisms. Yago3-10 is a subset of Yago3.0.2 obtained by only including the facts from Yago3.0.2 for which the entities from Yago3.0.2 occur in at least 10 facts. Yago3-10 is freely available online through multiple channels, e.g., through the AmpliGraph API [8].

AMIE3 (Lajus, Galárraga, and Suchanek 2020) was used to mine rules. Amie is freely available online [9]. We used its default settings, with a minimum CWA-based confidence (a.k.a. *standard confidence*) $\mathit{CWA}(R) \geq 0.1$. The non-recursive rules obtained from this mining process are listed in table 6. This table is generated based on AMIE's output. Its columns (from the left to the right) are the rule R, the rule's 'head coverage' (= the rule's support divided by the number of facts for relation p in K), the CWA-based confidence for R, the PCA-based confidence in the most functional direction of p in K, the number of labeled predictions of R in K, the total number of predictions, the number of predictions for which a subject is labeled in K (or object for p^{-1}), and whether the subject or object is considered to be the domain of the most functional direction of p.

F Completeness-aware confidence estimator (CARL) Consistently Underestimates the Confidence

The CARL estimator was motivated by the observation that cardinality information, i.e. the number of objects for a $\langle s, p \rangle$ pair, might be available (Pellissier Tanon et al. 2017). From these cardinalities one could estimate the number of

[8] https://github.com/Accenture/AmpliGraph/

[9] https://github.com/lajus/amie/

Table 5: $\left[\widehat{conf} - conf\right]^2 \cdot 10^4$ for SAR_{pop} (no PCA upheld), avg. over the rules predicting p. Entries for $k = \in \{0.01, 0.1\}$. Bold is best per k and p. Three IPW confidence columns per k: $\hat{k} = k$, $\hat{k} = k \cdot 0.9$ and $\hat{k} = k \cdot 1.1$.

p	# R	$k = 0.01$							$k = 0.1$						
		CWA	ICW	PCA_p	PCA_{p-p}	IPW	IPW (0.9)	IPW (1.1)	CWA	ICW	PCA_p	PCA_{p-p}	IPW	IPW (0.9)	IPW (1.1)
actedin	1	90.0	55.1	**1.4**	15.8	1.5	4.2	1.4	21.7	44.2	**183.7**	2.6	0.2	1.3	0.4
created	2	831.1	187.9	231.6	144.6	**61.0**	63.7	73.8	197.9	8.7	32.6	372.7	**3.1**	10.2	6.5
dealswith	7	123.0	43.2	75.0	17.9	**9.3**	19.1	6.5	0.8	0.4	13.3	33.6	**0.2**	0.3	0.2
diedin	1	160.5	32.9	155.2	141.6	**12.6**	14.2	14.1	59.9	5.5	139.3	27.5	**0.3**	2.1	1.2
directed	2	1000.1	**121.7**	289.1	266.4	192.7	284.5	146.9	247.2	**43.1**	206.2	232.3	3.5	15.2	5.9
exports	1	31.3	**1.6**	11.6	20.8	13.1	17.4	11.2	0.1	**0.1**	6.3	0.1	0.1	0.2	0.1
graduatedfrom	2	476.8	37.8	70.6	449.4	**18.4**	23.9	23.1	162.2	1.9	17.5	136.0	**1.7**	8.4	3.5
happenedin	1	761.5	78.1	560.4	644.7	**14.2**	28.7	16.4	346.3	27.7	224.5	196.1	**1.1**	8.5	6.7
hascapital	5	72.2	49.1	297.3	512.2	**6.7**	9.4	6.6	19.7	51.8	500.3	808.1	**0.5**	1.2	0.6
hasneighbor	3	110.1	26.0	9.5	16.1	**9.4**	15.9	8.7	0.4	0.2	51.3	51.2	**0.2**	0.2	0.2
imports	2	35.7	6.0	**2.6**	12.7	3.1	3.7	3.8	0.2	0.1	**1.7**	7.6	0.1	0.1	0.1
iscitizenof	3	434.4	**69.6**	178.2	403.1	77.4	102.0	67.7	129.0	**7.9**	187.0	63.5	7.4	12.6	9.2
isconnectedto	1	129.2	22.9	92.7	88.0	**18.4**	22.3	14.1	93.6	4.3	24.4	29.8	**3.5**	6.4	3.6
isleaderof	1	**106.0**	106.0	106.0	106.0	106.0	106.0	106.0	**52.9**	23.1	478.0	209.6	12.6	15.4	12.0
islocatedin	2	2649.1	1903.7	1504.1	1326.0	**226.0**	235.6	243.4	1500.1	797.9	781.9	561.1	**14.3**	2.8	40.6
ismarriedto	1	105.2	21.0	57.6	130.8	**3.1**	5.7	3.2	45.3	7.1	102.8	132.2	**0.2**	1.3	0.8
ispoliticianof	5	1200.9	**349.1**	410.4	1080.5	425.9	591.1	330.7	417.8	**18.8**	412.0	453.7	16.9	22.9	28.3
livesin	3	145.9	**19.7**	289.1	139.2	22.2	28.4	20.4	42.8	**4.3**	593.8	35.8	2.0	5.3	1.6
participatedin	1	399.2	275.5	96.4	239.1	**5.6**	17.1	7.3	36.0	7.3	107.4	13.1	**0.1**	1.3	1.0
wasbornin	1	135.3	23.8	39.0	129.8	**1.6**	2.6	3.2	48.2	12.8	32.2	40.7	**0.1**	1.2	1.0
worksat	1	526.6	**80.4**	125.2	479.7	82.8	112.2	71.5	148.6	**16.2**	2.1	111.8	4.2	12.8	4.7

R	Head Coverage	$CWA(R)$	$PCA(R)$	$\lvert\mathbf{R}^l\rvert$	$\lvert\mathbf{R}\rvert$	$\lvert\mathbf{R}^{\mathbf{l}}_{\mathbf{s}}\rvert$	PCA domain
$\langle s, directed, o\rangle \Rightarrow \langle s, actedin, o\rangle$	0.017	0.102	0.104	558	5481	5365	o
$\langle s, actedin, o\rangle \wedge \langle s, directed, o\rangle \Rightarrow \langle s, created, o\rangle$	0.031	0.380	0.567	212	558	374	o
$\langle s, directed, o\rangle \Rightarrow \langle s, created, o\rangle$	0.173	0.219	0.326	1202	5481	3683	o
$\langle s, hasneighbor, o\rangle \Rightarrow \langle s, dealswith, o\rangle$	0.126	0.295	0.337	164	555	486	s
$\langle s, hasneighbor, h\rangle \wedge \langle o, hasneighbor, h\rangle \Rightarrow \langle s, dealswith, o\rangle$	0.177	0.129	0.148	231	1792	1562	s
$\langle o, hasneighbor, s\rangle \Rightarrow \langle s, dealswith, o\rangle$	0.127	0.297	0.340	165	555	485	s
$\langle g, hasneighbor, s\rangle \wedge \langle g, hasneighbor, o\rangle \Rightarrow \langle s, dealswith, o\rangle$	0.177	0.128	0.146	230	1795	1573	s
$\langle h, hasneighbor, s\rangle \wedge \langle o, hasneighbor, h\rangle \Rightarrow \langle s, dealswith, o\rangle$	0.177	0.128	0.147	230	1792	1566	s
$\langle h, hasneighbor, o\rangle \wedge \langle s, hasneighbor, h\rangle \Rightarrow \langle s, dealswith, o\rangle$	0.177	0.129	0.147	231	1792	1567	s
$\langle g, iscitizenof, s\rangle \wedge \langle g, iscitizenof, o\rangle \Rightarrow \langle s, dealswith, o\rangle$	0.068	0.132	0.218	89	676	409	s
$\langle s, haschild, h\rangle \wedge \langle h, wasbornin, o\rangle \Rightarrow \langle s, diedin, o\rangle$	0.041	0.132	0.262	379	2862	1446	s
$\langle s, actedin, o\rangle \wedge \langle s, created, o\rangle \Rightarrow \langle s, directed, o\rangle$	0.039	0.451	0.542	212	470	391	o
$\langle s, created, o\rangle \Rightarrow \langle s, directed, o\rangle$	0.219	0.173	0.237	1202	6933	5082	o
$\langle s, imports, o\rangle \Rightarrow \langle s, exports, o\rangle$	0.159	0.153	0.177	60	393	339	s
$\langle s, hasacademicadvisor, h\rangle \wedge \langle h, worksat, o\rangle \Rightarrow \langle s, graduatedfrom, o\rangle$	0.051	0.272	0.301	374	1376	1243	s
$\langle s, worksat, o\rangle \Rightarrow \langle s, graduatedfrom, o\rangle$	0.082	0.177	0.193	599	3384	3109	s
$\langle o, participatedin, s\rangle \Rightarrow \langle s, happenedin, o\rangle$	0.293	0.288	0.367	1482	5150	4042	o
$\langle g, diedin, o\rangle \wedge \langle g, iscitizenof, s\rangle \Rightarrow \langle s, hascapital, o\rangle$	0.035	0.116	0.120	89	766	741	s
$\langle g, diedin, o\rangle \wedge \langle g, ispoliticianof, s\rangle \Rightarrow \langle s, hascapital, o\rangle$	0.051	0.242	0.282	130	537	461	s
$\langle o, islocatedin, s\rangle \wedge \langle s, islocatedin, o\rangle \Rightarrow \langle s, hascapital, o\rangle$	0.351	0.157	0.788	900	5740	1142	s
$\langle g, ispoliticianof, s\rangle \wedge \langle g, livesin, o\rangle \Rightarrow \langle s, hascapital, o\rangle$	0.026	0.171	0.189	66	386	349	s
$\langle g, ispoliticianof, s\rangle \wedge \langle g, wasbornin, o\rangle \Rightarrow \langle s, hascapital, o\rangle$	0.051	0.129	0.142	131	1017	924	s
$\langle s, dealswith, o\rangle \Rightarrow \langle s, hasneighbor, o\rangle$	0.295	0.126	0.164	164	1302	997	o
$\langle o, dealswith, s\rangle \Rightarrow \langle s, hasneighbor, o\rangle$	0.297	0.127	0.175	165	1302	944	o
$\langle o, dealswith, s\rangle \wedge \langle s, dealswith, o\rangle \Rightarrow \langle s, hasneighbor, o\rangle$	0.097	0.338	0.458	54	160	118	o
$\langle s, dealswith, h\rangle \wedge \langle h, exports, o\rangle \Rightarrow \langle s, imports, o\rangle$	0.501	0.109	0.127	197	1814	1546	s
$\langle s, exports, o\rangle \Rightarrow \langle s, imports, o\rangle$	0.153	0.159	0.171	60	378	351	s
$\langle s, playsfor, o\rangle \Rightarrow \langle s, isaffiliatedto, o\rangle$	0.746	0.869	0.946	278848	321024	294723	s
$\langle s, hasacademicadvisor, h\rangle \wedge \langle h, livesin, o\rangle \Rightarrow \langle s, iscitizenof, o\rangle$	0.029	0.254	0.390	101	398	259	s
$\langle g, hasacademicadvisor, s\rangle \wedge \langle g, livesin, o\rangle \Rightarrow \langle s, iscitizenof, o\rangle$	0.025	0.251	0.368	86	343	234	s
$\langle s, livesin, o\rangle \Rightarrow \langle s, iscitizenof, o\rangle$	0.120	0.139	0.358	415	2980	1159	s
$\langle g, owns, s\rangle \wedge \langle g, owns, o\rangle \Rightarrow \langle s, isconnectedto, o\rangle$	0.010	0.116	0.276	325	2806	1177	o
$\langle s, livesin, o\rangle \wedge \langle s, wasbornin, o\rangle \Rightarrow \langle s, isleaderof, o\rangle$	0.015	0.103	0.311	14	136	45	o
$\langle s, hascapital, o\rangle \Rightarrow \langle s, islocatedin, o\rangle$	0.011	0.387	0.418	993	2563	2378	s
$\langle o, hascapital, s\rangle \Rightarrow \langle s, islocatedin, o\rangle$	0.020	0.680	0.682	1742	2563	2554	s
$\langle s, haschild, h\rangle \wedge \langle o, haschild, h\rangle \Rightarrow \langle s, ismarriedto, o\rangle$	0.236	0.107	0.239	886	8284	3705	s
$\langle s, haschild, h\rangle \wedge \langle h, iscitizenof, o\rangle \Rightarrow \langle s, ispoliticianof, o\rangle$	0.057	0.532	0.831	123	231	148	s
$\langle s, haschild, h\rangle \wedge \langle h, isleaderof, o\rangle \Rightarrow \langle s, ispoliticianof, o\rangle$	0.049	0.211	0.397	106	502	267	s
$\langle g, haschild, s\rangle \wedge \langle g, iscitizenof, o\rangle \Rightarrow \langle s, ispoliticianof, o\rangle$	0.065	0.560	0.778	140	250	180	s
$\langle g, haschild, s\rangle \wedge \langle g, isleaderof, o\rangle \Rightarrow \langle s, ispoliticianof, o\rangle$	0.021	0.128	0.260	46	358	177	s
$\langle s, isleaderof, o\rangle \Rightarrow \langle s, ispoliticianof, o\rangle$	0.065	0.146	0.458	140	957	306	s
$\langle s, hasacademicadvisor, h\rangle \wedge \langle h, iscitizenof, o\rangle \Rightarrow \langle s, livesin, o\rangle$	0.034	0.137	0.437	101	735	231	s
$\langle g, hasacademicadvisor, s\rangle \wedge \langle g, iscitizenof, o\rangle \Rightarrow \langle s, livesin, o\rangle$	0.025	0.124	0.381	74	598	194	s
$\langle s, iscitizenof, o\rangle \Rightarrow \langle s, livesin, o\rangle$	0.139	0.120	0.471	415	3453	881	s
$\langle o, happenedin, s\rangle \Rightarrow \langle s, participatedin, o\rangle$	0.288	0.293	0.310	1482	5052	4779	o
$\langle s, isaffiliatedto, o\rangle \Rightarrow \langle s, playsfor, o\rangle$	0.869	0.746	0.825	278848	373721	337858	s
$\langle s, diedin, o\rangle \Rightarrow \langle s, wasbornin, o\rangle$	0.025	0.122	0.174	1132	9244	6499	s
$\langle g, diedin, o\rangle \wedge \langle g, haschild, s\rangle \Rightarrow \langle s, wasbornin, o\rangle$	0.010	0.196	0.284	454	2321	1597	s
$\langle g, graduatedfrom, o\rangle \wedge \langle g, hasacademicadvisor, s\rangle \Rightarrow \langle s, worksat, o\rangle$	0.079	0.243	0.365	268	1104	735	s

Table 6: Non-recursive rules mined from Yago3-10 using AMIE.

non-observed facts predicted by a rule $\sum_{\langle s,o\rangle\in\mathbf{R}^+}(1-l(s,o))$. This estimate is represented by $npi(R)$. In the ideal case $npi(R)=\sum_{\langle s,o\rangle\in\mathbf{R}^+}(1-l(s,o))$.

The proposed CARL confidence estimator is:

$$CARL(R)=\frac{\sum_{\langle s,o\rangle\in\mathbf{R}}l(s,o)}{|\mathbf{R}|-npi(R)}.$$

However, even with a perfect $npi(R)$, this estimator is biased and will consistently underestimate the confidence. The more non-observed facts the rule predicts, the worse the bias will be.

Assuming a perfect estimate for the number of missing facts $npi(R)=\sum_{\langle s,o\rangle\in\mathbf{R}^+}(1-l(s,o))$, then the CARL confidence estimator is always smaller than the true confidence $CARL(R)\leq\mathrm{conf}(R)$. The two are equal if and only if CWA holds, or when all predicted triples are facts in I: $CARL(R)=\mathrm{conf}(R)$ iff $npi(R)=0$ or $|\mathbf{R}^+|=|\mathbf{R}|$.

Proof. $|\mathbf{R}^+|>0$ and $|\mathbf{R}|>0$

1. Inequality, given $|\mathbf{R}^+|<|\mathbf{R}|$ and $npi(R)>0$:

$$\begin{aligned}
\left|\mathbf{R}^+\right| &< |\mathbf{R}| \\
\left|\mathbf{R}^+\right| npi(R) &< |\mathbf{R}|\, npi(R) \\
|\mathbf{R}|\left|\mathbf{R}^+\right| - \left|\mathbf{R}^+\right| npi(R) &> |\mathbf{R}|\left|\mathbf{R}^+\right| - |\mathbf{R}|\, npi(R) \\
\left|\mathbf{R}^+\right|(|\mathbf{R}| - npi(R)) &> |\mathbf{R}|\left(\left|\mathbf{R}^+\right| - npi(R)\right) \\
\frac{|\mathbf{R}^+|}{|\mathbf{R}|} &> \frac{|\mathbf{R}^+| - npi(R)}{|\mathbf{R}| - npi(R)}
\end{aligned}$$

2. Equality, given $|\mathbf{R}^+|=|\mathbf{R}|$

$$\begin{aligned}
\left|\mathbf{R}^+\right| &= |\mathbf{R}| \\
\left|\mathbf{R}^+\right| - npi(R) &= |\mathbf{R}| - npi(R) \\
\frac{|\mathbf{R}^+| - npi(R)}{|\mathbf{R}| - npi(R)} &= 1 = \frac{|\mathbf{R}^+|}{|\mathbf{R}|}
\end{aligned}$$

3. Equality, given $npi(R)=0$

$$\frac{|\mathbf{R}^+| - npi(R)}{|\mathbf{R}| - npi(R)} = \frac{|\mathbf{R}^+|}{|\mathbf{R}|}$$

□

Unbiased estimator using *npi*(R) If $npi(R)$ would be known, an unbiased estimator for the confidence would be

$$\frac{npi(R)+\sum_{\langle s,o\rangle\in\mathbf{R}}l(s,o)}{|\mathbf{R}|}=\frac{|\mathbf{R}^+|}{|\mathbf{R}|}.$$