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 $\mathbb{E}(X)/\mathbb{E}(Y)$:

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

This approximation can be rewritten as follows:

$$\begin{aligned}
\underset{\text{sel}\sim e}{\mathbb{E}}[\mathit{PCA}(R)] \approx \text{conf}(R) \qquad & \mathit{bias}_{\cancel{PCA}}(R) \qquad && \mathit{bias}_{y(s)=0}(R) \qquad && \mathit{bias}_{e(s)}(R)\\
= \text{conf}(R) \qquad & \frac{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s,o)}{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)} \qquad && \frac{|\mathbf{R}|}{|\mathbf{R}_\mathbf{s}^+|} \qquad && \frac{\frac{1}{|\mathbf{R}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)}{\frac{1}{|\mathbf{R}_\mathbf{s}^+|}\sum_{\langle s,o\rangle\in\mathbf{R}_\mathbf{s}^+} e(s)}.
\end{aligned}$$

Proof.

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

□

bias$_{\cancel{PCA}}(R)$**: Bias due to PCA being violated**

$$\mathit{bias}_{\cancel{PCA}}(R) = \frac{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s,o)}{\sum_{\langle s,o\rangle\in\mathbf{R}^+} e(s)}$$

When the PCA assumption holds, $e(s) = e(s,o)$, therefore the $\mathit{bias}_{\cancel{PCA}}$ disappears as it reduces to 1.

When the PCA assumption is violated, then $e(s) > e(s,o)$[6] and consequently $\mathit{bias}_{\cancel{PCA}}(R) < 1$. In this case, $\mathit{bias}_{\cancel{PCA}}(R)$ 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**

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

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 $\mathbf{R}^+_\mathbf{s}$:

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

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

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: ***IPW***(R) **vs** ***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:

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

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

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

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

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

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

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:

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

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; $\textit{ICW}(R)=\frac{1}{c_p}\textit{CWA}(R)$. 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 $\alpha>0$:

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

The calibration constant α is chosen such that:

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

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 $\textit{cal}(\hat{e}(s,o))$ result in the IPW estimator being a better estimator for $\operatorname{conf}(R)$ than the ICW estimator, i.e. when the expected IPW estimator is closer to the true confidence than the ICW estimator:

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

[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.,

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

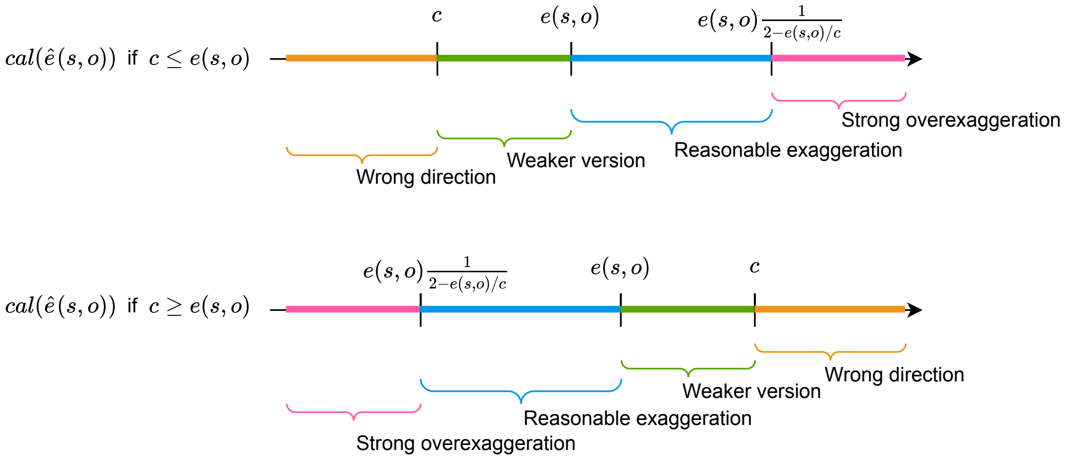

Figure 8: Possible values of the propensity score estimate $cal\,(\hat{e}(s,o))$ for a triple $\langle s,o\rangle$ in relation to the true value $e(s,o)$ and c. If $cal\,(\hat{e}(s,o))$ is a weaker version of the observation mechanism or a reasonable exaggeration, then the error contribution of the triple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

$$\left| \underset{\mathrm{sel}\sim e}{\mathbb{E}}[\mathit{IPW\text{-}PCA}(R)] - \mathrm{conf}(R) \right| \leq \left| \underset{\mathrm{sel}\sim e}{\mathbb{E}}[\mathit{PCA}(R)] - \mathrm{conf}(R) \right|$$

Recall the first-order Taylor approximations:

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

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

To simplify the mathematics, we will write $\mathit{bias}_{e(s)}$ 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

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

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:

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

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$_p$, SCAR$_\text{group}$ and SCAR$_\text{pop}$.

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 $\langle$*Parent*, *hasChild*, *Child*$\rangle$ can be partitioned using $q =$ *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 $\#(s, p)$:

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

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 SCAR$_{\text{group}}$ and table 5 for SCAR$_{\text{pop}}$. 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 **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 **SCAR**$_{\textbf{group}}$ 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 **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 SCAR$_{\text{group}}$ and table 5 for SCAR$_{\text{pop}}$.

Table 4 shows the results for the experimental setup for **Q3**. It also is a **SCAR**$_{\textbf{group}}$ 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 $|\mathbf{R}|/|\mathbf{R}_{\mathbf{s}}^{+}| \cdot$ *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 $\textit{bias}_{\cancel{PCA}}$ 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 SCAR$_{\text{group}}$ and table 5 for SCAR$_{\text{pop}}$. 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$_p$	PCA$_{p^{-1}}$	IPW	IPW $(-\Delta)$	IPW $(+\Delta)$	CWA	PCA$_p$	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 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$_p$	PCA$_{p^{-1}}$	IPW	IPW $(-\Delta)$	IPW $(+\Delta)$	CWA	ICW	PCA$_p$	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*) *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)	$\lvert \mathbf{R}^l \rvert$	$\lvert \mathbf{R} \rvert$	$\lvert \mathbf{R}_\mathbf{s}^\mathbf{l} \rvert$	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 $\mathit{npi}(R)$. In the ideal case $\mathit{npi}(R) = \sum_{\langle s,o\rangle \in \mathbf{R}^+} (1 - l(s,o))$.

The proposed CARL confidence estimator is:

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

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

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

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

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

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

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

3. Equality, given $\mathit{npi}(R) = 0$

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

□

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

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