Supplementary Material for Bayesian Networks for Data Integration in the Absence of Foreign Keys

Bohan Zhang, Scott Sanner, Mohamed Reda Bouadjenek, and Shagun Gupta

APPENDIX

A. ADDITIONAL EXPERIMENTS

Experiment: Shared Variables (Synthetic)

This experiment is designed to investigate the effect of the number of shared variables. Since real-world datasets with a large number of shared variables are largely unavailable, $1,000,000$ rows of synthetic data are generated using the Bayesian network shown in Fig. 1, with 6 random variables, and the following projection onto two local relations: (1) `A`, `B`, `C`, `D`, `E` and (2) `B`, `C`, `D`, `E`, `F`. To reduce the number of shared variables in experiments, we simply remove `B`, `C`, `D`, `E` to achieve the desired amount of sharing.

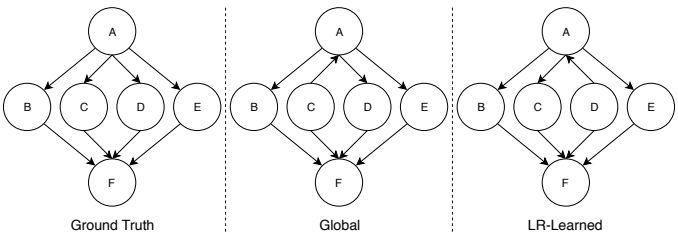

Fig. 1. Bayesian Networks, Experiment 2 (Shared Variables)

We evaluate the effect of removing shared variables in Fig. 1 by measuring the mean absolute deviation and KL divergence of two cross-table queries, $P(F|A = 0)$ and $P(F|A = 1)$. In brief, the performance worsens in both cases as we reduce the number of shared variables from 4 to 1 indicating that *more shared variables promote increased accuracy in LR-BN inference* since there are more paths (i.e., effectively more bandwidth) for information flow.

• *B. Zhang, S. Sanner, and M. R. Bouadjenek are with the Department of Mechanical and Industrial Engineering, 27 King's College Cir, Toronto, ON M5S 3H7, Canada.*
E-mail: see http://d3m.mie.utoronto.ca/

Manuscript received XXX YY, ZZZZ; revised XXX YY, ZZZZ.

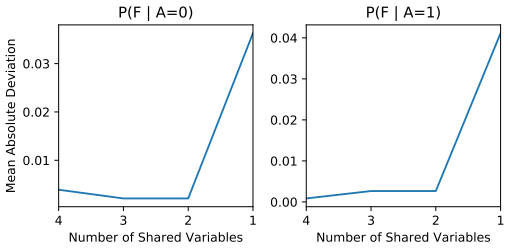

Fig. 2. Mean Absolute Deviation vs. Number of Shared Variables

Experiment: Shared Variable Cardinality (Synthetic)

In this experiment, we use a simple model, $A \rightarrow B \rightarrow C$, and $1,000,000$ rows of synthetic data, to explore the impact of shared variables' cardinality. The model is projected onto two local relations: (1) `A`, `B` and (2) `B`, `C`. Here, `B` is the shared variable with a cardinality of 5. During our experiment, we reduce `B`'s cardinality by one at a time and observe the change in mean absolute deviation to determine the significance of the shared variable's cardinality.

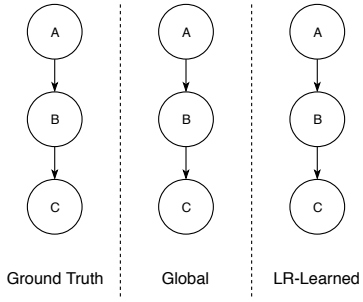

Fig. 3. Bayesian Networks, Experiment 3 (Cardinality)

The model, $A \rightarrow B \rightarrow C$, as shown in Fig. 3., is used to validate the hypothesis that accuracy will increase with shared variables' cardinality. In this model, `B` has a an original cardinality of 5, and we collapse its cardinality by one at a time, until it has a cardinality of 1. The mean absolute deviation of two cross-table queries, $P(C|A = 0)$ and $P(C|A = 1)$, is plotted with respect to different cardinalities in Fig. 4. Unsurprisingly, the mean absolute

deviation increases with decreasing shared-variable cardinality as suggested by the $> 12.5\%$ mean absolute deviation shown in $P(C|A = 1)$. In short, fewer values in shared variables limits the information that can be transmitted through shared variables.

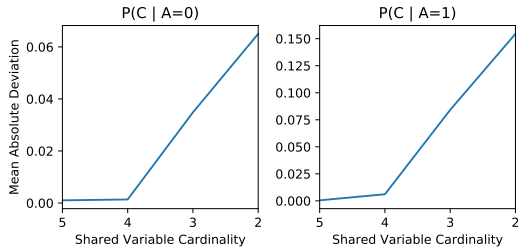

Fig. 4. Mean Absolute Deviation vs. Shared Variable's Cardinality

B. Sample Bias in Local Relations

We address here the specific case of data integration when local relations come from different distributions.

Consider a simplified version of the targeted marketing example discussed in the main article, where a company would like to target advertising based on consumer education level. Let us suppose that there are two datasets to integrate: a company dataset that relates age range $x \in X$ with consumer purchase behavior $y \in Y$ in Relation 1 (R_{L_1}) and a census dataset that relates age range $x \in X$ with education level $z \in Z$ in Relation 2 (R_{L_2}). In this case, the two datasets are drawn from different distributions and can be seen as projections of a global relation followed by a randomized subsampling procedure according to the distribution of each respective dataset.

In this Appendix, we demonstrate that when two local binary relations arise from different distributions and contain a single shared parent attribute in a Bayesian network structure, (a) one only needs to know which of the distributions represents the true (intended) prior of the shared parent attribute while learning the Bayesian network parameters, and very conveniently, (b) the distributions conditioned on this shared parent attribute remain unaffected. While this is just one case of many possible scenarios, it does suggest that there are straightforward ways to resolve issues of sample bias and mismatch in local relations within the LR-BNLEARN framework proposed in this article.

Problem Setup: Given two local relations, $R_{L_1}(X, Y)$ and $R_{L_2}(X, Z)$, our goal is to query Y given Z as evidence. To this end, let us assume we wish to learn the parameters for the LR-Learnable Bayesian network with $(X \rightarrow Y)$ and $(X \rightarrow Z)$ as shown for BN_1 in Fig. 5. However, in this particular case R_{L_1} and R_{L_2} are drawn from different distributions and thus have different marginals over X:

- Table 1 (R_{L_1}) : $q(x, y) \rightarrow q(x)$
- Table 2 (R_{L_2}) : $p(x, z) \rightarrow p(x)$

How can we then learn the parameters for the CPDs of the Bayesian Network in Fig. 5?

Solution: Suppose that we know the true marginal distribution over ages corresponding to the sample population we care about is $p(x)$ of Table 2. That is, the sample space of individuals that we are concerned with is the sample space of individuals from the census — we might consider Table 1 from a company to not be as representative of the general population distribution as Table 2 should be according to census policy. As a result, considering $q(x, y)$ to be Table 1's distribution, we want a modified version of Table 1's distribution $\tilde{q}(x, y)$ that uses the correct prior $p(x)$ of the shared variable X from Table 2:

TABLE 1
Local Relation 1

X	Y
1	0
1	0
0	1
0	0
0	1

TABLE 2
Local Relation 2

X	Z
0	1
1	0
1	0
1	0
0	1

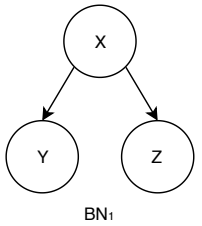

Fig. 5. An illustration showing two local relations in Tables 1 and 2 arising from different distributions. We want to learn the DAG for BN_1.

$$q(x, y) = q(y|x)q(x) \quad \text{under Table 1's distribution}$$
$$\tilde{q}(x, y) = q(y|x)\underbrace{\tilde{q}(x)}_{p(x)} \quad \text{introduce modified } q \text{ to match Table 1's } p(x)$$

Clearly, now $\tilde{q}(x, y)$ has a marginal distribution over X of $\tilde{q}(x) = p(x)$ as intended while retaining the conditional $q(y|x)$ from $q(x, y)$:

$$\tilde{q}(x) = \sum_y \tilde{q}(x, y) = \sum_y q(y|x)p(x) = p(x)\sum_y q(y|x) \overset{1}{=} p(x)\,.$$

This scheme suggests the overall graphical model can be learned using empirical distributions for the following:

- $X : p(x)$
- $X \rightarrow Y : q(y|x)$
- $X \rightarrow Z : p(z|x)$

We make two key observations about this above solution:

Claim 1: Given the choice of two marginals $p(x)$ and $q(x)$, we choose the marginal $p(x)$ corresponding to the table that provides our target sampling distribution (Table 2 in this case).

Claim 2: We use the empirical $q(y|x)$ and $p(z|x)$ to estimate their respective CPDs, which is what we would have done in the original methodology if both tables had the same sampling distribution.

Another more mathematical treatment to justify the above claims would be to estimate $q(x)$ and $q(y|x)$ under an *importance sampling* correction that corrects the biased distribution $q(x)$ to $p(x)$.

A justification of Claims 1 and 2 is given below based on this importance sampling approach, but first we briefly review the importance sampling estimator for completeness. In general, given a function $f(x)$ where x has distribution p, importance sampling allows us to estimate the expectation of $f(x)$ by sampling $\sim$ from an alternate distribution q:

$$\begin{aligned}\mathbb{E}_p\big[f(x)\big] &= \mathbb{E}_q\left[\frac{f(x)\cdot p(x)}{q(x)}\right] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)\cdot p(x_i)}{q(x_i)}, \quad x_i \sim q(x) \\ &= \frac{1}{n}\sum_{i=1}^{n} f(x_i)\cdot w_i\,,\end{aligned}$$

where $w_i = p(x_i)/q(x_i)$ is the importance sampling weight.

Importance Sampling Justification of Claims 1 and 2: We wish to learn the maximum likelihood parameters $\boldsymbol{\theta}$ for the Bayesian network edge $X \to Y$ given n data samples $\langle x_i, y_i\rangle \sim q(x,y)$, corrected via importance sampling to have marginal $p(x)$. Because marginal constraint $p(x)$ states nothing about $p(y|x)$, we will assume $p(y|x) = q(y|x)$:

$$\begin{aligned}\arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}:D) &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(x_i,y_i:\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \log\left(\prod_{i=1}^{n} p(x_i,y_i:\boldsymbol{\theta})\right) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\Big(p(x_i,y_i:\boldsymbol{\theta})\Big) \qquad (1)\end{aligned}$$

We can view the main expression of (1) as a Monte Carlo estimate of the following expectation:

$$\sum_{i=1}^{n}\log p(x_i,y_i:\boldsymbol{\theta}) = \mathbb{E}_{p(x,y)}\Big[\log p(x,y:\boldsymbol{\theta})\Big]$$

Next, we can apply the importance sampling correction to reweight samples $\langle x_i, y_i\rangle \sim q(x,y)$:

$$\begin{aligned}&\mathbb{E}_{q(y,x)}\left[\frac{p(y,x)}{q(y,x)}\log p(x,y:\boldsymbol{\theta})\right] \\ &= \sum_{i=1}^{n}\frac{p(y_i,x_i)}{q(y_i,x_i)}\ \log\Big(p(x_i,y_i:\boldsymbol{\theta})\Big) \\ &= \sum_{i=1}^{n}\frac{p(y_i|x_i)p(x_i)}{q(y_i|x_i)q(x_i)}\ \log\Big(p(x_i,y_i:\boldsymbol{\theta})\Big) \\ &= \sum_{i=1}^{n}\frac{p(x_i)}{q(x_i)}\ \log\Big(p(x_i,y_i:\boldsymbol{\theta})\Big) \\ &= \sum_{i=1}^{n} w_i\ \log\Big(p(x_i,y_i:\boldsymbol{\theta})\Big) \\ &= \sum_{i=1}^{n} w_i\ \log\Big(p(x_i:\boldsymbol{\theta})\cdot p(y_i\mid x_i:\boldsymbol{\theta})\Big) \\ &= \sum_{i=1}^{n}\left[w_i\ \log\Big(p(x_i:\boldsymbol{\theta})\Big) + w_i\ \log\ \Big(p(y_i\mid x_i:\boldsymbol{\theta})\Big)\right] \\ &= \underbrace{\sum_{i=1}^{n}\left[w_i\ \log\Big(p(x_i:\boldsymbol{\theta})\Big)\right]}_{\text{A}} + \underbrace{\sum_{i=1}^{n}\left[w_i\ \log\Big(p(y_i\mid x_i:\boldsymbol{\theta})\Big)\right]}_{\text{B}} \qquad (2)\end{aligned}$$

We see that the likelihood decomposes into two separate terms, one for estimation of the parameters of the prior distribution $p(x)$ (A) and the other for the estimation of parameters of the conditional distribution $p(y|x)$ (B). We note that because terms A and B involve disjoint parameter sets, they can be maximized separately.

For Claim 1, we consider maximization of term A from equation (2) for the prior parameters:

$$\sum_{i=1}^{n}\left[w_i\ \log\Big(p(x_i:\boldsymbol{\theta})\Big)\right]$$

Without loss of generality, we assume that X is a Bernoulli random variable taking on two values: $x_i = 1$ with probability θ_X and $x_i = 0$ with probability $1-\theta_X$. Below, we let $\{\cdot\}$ denote the 0-1 indicator function that takes value 1 when its argument $\cdot$ is true. We also let $w^{x=j} = p(x=j)/q(x=j)$.

$$\begin{aligned}&= \sum_{i=1}^{n}\left[w_i\ \log\Big(\theta_X^{\{x_i=1\}}\cdot(1-\theta_X)^{\{x_i=0\}}\Big)\right] \\ &= \sum_{i=1}^{n}\Big[w_i\{x_i=1\}\log\theta_X + w_i\{x_i=0\}\log(1-\theta_X)\Big] \\ &= \log\theta_X \sum_{\{i|x_i=1\}} w^{x=1} + \log(1-\theta_X)\sum_{\{i|x_i=0\}} w^{x=0}\end{aligned}$$

Since the log likelihood for the exponential family is concave, we can solve for the maximizing θ_X by differentiating w.r.t. θ_X and setting it equal to 0:

$$\begin{aligned}&\Rightarrow \frac{\sum_{\{i|x_i=1\}} w^{x=1}}{\theta_X} - \frac{\sum_{\{i|x_i=0\}} w^{x=0}}{(1-\theta_X)} = 0 \\ &\Rightarrow \sum_{\{i|x_i=1\}} w^{x=1}(1-\theta_X) = \sum_{\{i|x_i=0\}} w^{x=0}\theta_X \\ &\theta_X = \frac{\sum_{\{i|x_i=1\}} w^{x=1}}{\sum_{\{i|x_i=1\}} w^{x=1} + \sum_{\{i|x_i=0\}} w^{x=0}}\end{aligned}$$

Letting $\#[\cdot]$ denote the count of data i meeting the specified criteria of its argument $\cdot$, recalling the definition of importance weight $w^{x=j}$, and recalling that the total number of samples is n, we complete the derivation:

$$\begin{aligned}\theta_X &= \frac{\frac{p(x=1)}{q(x=1)}\#[x_i=1]}{\frac{p(x=1)}{q(x=1)}\#[x_i=1] + \frac{p(x=0)}{q(x=0)})\#[x_i=0]} \\ &= \frac{\frac{p(x=1)}{q(x=1)}\frac{\#[x_i=1]}{n}}{\frac{p(x=1)}{q(x=1)}\frac{\#[x_i=1]}{n} + \frac{p(x=0)}{q(x=0)}\frac{\#[x_i=0]}{n}} \\ &= \frac{\frac{p(x=1)}{q(x=1)}q(x=1)}{\frac{p(x=1)}{q(x=1)}q(x=1) + \frac{p(x=0)}{q(x=0)}q(x=0)} \\ &= \underbrace{\frac{p(x=1)}{p(x=1)+p(x=0)}}_{1} \quad = \quad \boxed{p(x=1)}\end{aligned}$$

Here we arrive at the intuitive result of **Claim 1** that the estimate of the maximum likelihood parameters for the prior

over X using data samples $\langle x_i, y_i \rangle \sim q(x, y)$ matches $p(x)$ from Table 2 when using importance sampling to correct the marginal sample bias of Table 1 to match Table 2 on their shared variable X.

We now proceed to address Claim 2. Without loss of generality, we assume X and Y are Bernoulli random variables. Let $\theta_{Y|x^0}$ ($\theta_{Y|x^1}$) be the probability of $y_i = 1$ if conditioned on $x_i = 0$ ($x_i = 1$). Decomposing term B from equation (2) in a similar manner as for Claim 1, we arrive at a different result for the maximum likelihood parameters of the importance sampling corrected conditional:

$$\begin{aligned}
&\sum_{i=1}^{n} \left[w_i \ \log \Big(p(y_i \mid x_i : \boldsymbol{\theta}) \Big) \right] \\
&= \underbrace{\sum_{\{i|x_i=0\}} \left[w_i \ \log \Big(p(y_i \mid x_i : \theta_{Y|x^0}) \Big) \right]}_{\text{C}} \\
&\quad + \underbrace{\sum_{\{i|x_i=1\}} \left[w_i \ \log \Big(p(y_i \mid x_i : \theta_{Y|x^1}) \Big) \right]}_{\text{D}} \qquad (3)
\end{aligned}$$

Now we consider only term C from (3) since the derivation for term D is identical and independently maximized:

$$\begin{aligned}
&= \sum_{\{i|x_i=0\}} \left[w_i \ \log \Big(\theta_{Y|x^0}^{\{x_i=0,y_i=1\}} (1 - \theta_{Y|x^0})^{\{x_i=0,y_i=0\}} \Big) \right] \\
&= \sum_{\{i|x_i=0\}} \Big[w_i \{x_i = 0, y_i = 1\} \ \log \theta_{Y|x^0} + \\
&\qquad\qquad w_i \{x_i = 0, y_i = 0\} \ \log(1 - \theta_{Y|x^0}) \Big] \\
&= \log \theta_{Y|x^0} \sum_{\{i|y_i=1,x_i=0\}} \Big[w_i \{x_i = 0, y_i = 1\} \Big] + \\
&\quad \log(1 - \theta_{Y|x^0}) \sum_{\{i|y_i=0,x_i=0\}} \Big[w_i \{x_i = 0, y_i = 0\} \Big] \\
&= \log \theta_{Y|x^0} \sum_{\{i|y_i=1,x_i=0\}} \big[w_i \big] + \log(1 - \theta_{Y|x^0}) \sum_{\{i|y_i=0,x_i=0\}} \big[w_i \big]
\end{aligned}$$

Differentiating w.r.t. $\theta_{Y|x^0}$ and setting it equal to 0:

$$\begin{aligned}
&\Rightarrow \frac{\sum_{\{i|y_i=1,x_i=0\}} \big[w_i \big]}{\theta_{Y|x^0}} - \frac{\sum_{\{i|y_i=0,x_i=0\}} \big[w_i \big]}{(1 - \theta_{Y|x^0})} = 0 \\
&\Rightarrow \sum_{\{i|y_i=1,x_i=0\}} \big[w_i \big] (1 - \theta_{Y|x^0}) = \sum_{\{i|y_i=0,x_i=0\}} \big[w_i \big] \theta_{Y|x^0} \\
&\theta_{Y|x^0} = \frac{\sum_{\{i|y_i=1,x_i=0\}} \big[w_i \big]}{\sum_{\{i|y_i=1,x_i=0\}} \big[w_i \big] + \sum_{\{i|y_i=0,x_i=0\}} \big[w_i \big]}
\end{aligned}$$

Finally, since w_i only depends on x_i, and $x_i = 0$ throughout term C, we can factor it out of the numerator and denominator and thus arrive at a fortuitous cancellation:

$$\begin{aligned}
\theta_{Y|x^0} &= \frac{\cancel{w_i}^{1} \sum_{\{i|y_i=1,x_i=0\}} 1}{\cancel{w_i}^{1} \Big(\sum_{\{i|y_i=1,x_i=0\}} 1 + \sum_{\{i|y_i=0,x_i=0\}} 1 \Big)} \\
&= \frac{\#[x_i = 0, y_i = 1]}{\#[x_i = 0, y_i = 1] + \#[x_i = 0, y_i = 0]}
\end{aligned}$$

Noting that the bottom term is just the total count $\#[x_i = 0]$ since Y is a binary variable and both $y = 1$ and $y = 0$ are considered, we can easily identify this as an empirical estimate of $q(y = 1|x = 0)$:

$$\theta_{Y|x^0} = \frac{\#[x_i = 0, y_i = 1]}{\#[x_i = 0]} = \boxed{q(y = 1|x = 0)}$$

Here we arrive at the final result of **Claim 2** that the estimate of the maximum likelihood parameters for the edge $X \rightarrow Y$ are simply the **unweighted** empirical conditional probabilities $q(y|x)$ from Table 1 data samples $\langle x_i, y_i \rangle \sim q(x, y)$ since the importance weights cancel when conditioning on X. As for $p(z|x)$ being the correct empirical conditional distribution for $X \rightarrow Z$ — this follows trivially from the fact that Table 2 is already the target sampling distribution and requires no importance sampling correction.