MANAGEMENT SCIENCE
DOI 10.1287/mnsc.1070.0745ec
pp. ec1–ec9

e-companion
ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"A Framework for Reconciling Attribute Values from Multiple Data Sources" by Zhengrui Jiang, Sumit Sarkar, Prabuddha De, and Debabrata Dey, *Management Science*, DOI 10.1287/mnsc.1070.0745.

Proofs of Propositions, Derivations, and Additional Analyses

EC.1. Proofs of Propositions

PROPOSITION 1. *For a given entity instance, the likelihood ratio for any two values of an attribute, which are not recorded in any of the data sources, is the same as the ratio of their prior probabilities.*

PROOF OF PROPOSITION 1. Suppose an attribute value a_i is not recorded in any of the data sources S_1 through S_n for an entity instance. Then, from Assumptions 1 and 2 in the paper, we have

$$\begin{aligned}
&\mathrm{P}(A=a_i \mid A_{S_1}=a_k, A_{S_2}=a_l, \ldots, A_{S_n}=a_t W) \quad (i\neq k,\ i\neq l, \ldots, i\neq t)\\
&\quad= \frac{\mathrm{P}(A_{S_1}=a_k \mid A=a_i)\mathrm{P}(A_{S_2}=a_l \mid A=a_i)\times\cdots\times\mathrm{P}(A_{S_n}=a_t \mid A=a_i)}{\mathrm{P}(A_{S_1}=a_k, A_{S_2}=a_l, \ldots, A_{S_n}=a_t)}\mathrm{P}(A=a_i)\\
&\quad= \frac{[(1-R^A_{S_1})/(m-1)][(1-R^A_{S_2})/(m-1)]\times\cdots\times[(1-R^A_{S_n})/(m-1)]}{\mathrm{P}(A_{S_1}=a_k, A_{S_2}=a_l, \ldots, A_{S_n}=a_t)}\mathrm{P}(A=a_i).
\end{aligned}$$

Clearly, the above probability expression is proportional to the prior probability $\mathrm{P}(A=a_i)$, $\forall\, i$. Therefore, we must have

$$\frac{\mathrm{P}(A=a_i \mid A_{S_1}\neq a_i, A_{S_2}\neq a_i, \ldots, A_{S_n}\neq a_i)}{\mathrm{P}(A=a_j \mid A_{S_1}\neq a_j, A_{S_2}\neq a_j, \ldots, A_{S_n}\neq a_j)} = \frac{\mathrm{P}(A=a_i)}{\mathrm{P}(A=a_j)}, \quad \forall\, i, j.$$

PROPOSITION 2. *For an attribute with an infinitely large domain and a finite ratio of prior probabilities for any two of its values, the value that appears in most* (*and at least two*) *data sources for an entity instance is true with probability one, and all other values have a probability of zero. In case there are more than one such most frequent attribute values, the probabilities for these values can be calculated by ignoring all the less frequent ones and their associated data sources.*

PROOF OF PROPOSITION 2. We consider the two cases separately:

Case 1. There is only one most frequent value. Suppose h unique values of an attribute A are recorded in n data sources (S_1 through S_n) for an entity instance. For expositional convenience, we denote the set of recorded values of A by $\mathbf{V}$. Among them, the most frequent value V_1 appears p ($p\geq 2$) times in S_1 through S_p, the second most frequent value V_2 is recorded q ($q<p$) times in S_{p+1} through $S_{p+q}, \ldots$, and the least frequent value V_h is recorded r ($r<p$) times in S_{n-r+1} through S_n. Let m be the size of the domain A. From (1) and (2) in the paper, we have

$$\mathrm{P}(A=V_1 \mid A_{S_1}, \ldots, A_{S_n}) = \frac{N}{N+D},$$

where

$$N=\mathrm{P}(A=V_1)\prod_{k=1}^{p}R^A_{S_k}\prod_{k=p+1}^{n}\frac{1-R^A_{S_k}}{m-1} = \mathrm{P}(A=V_1)(m-1)^{-(n-p)}\prod_{k=1}^{p}R^A_{S_k}\prod_{k=p+1}^{n}(1-R^A_{S_k}), \qquad \text{and}$$

$$\begin{aligned}D &= \mathrm{P}(A=V_2)\prod_{k=p+1}^{p+q} R_{S_k}^A \prod_{k\notin[p+1,\,p+q]} \frac{1-R_{S_k}^A}{m-1} + \cdots + \mathrm{P}(A=V_h)\prod_{k=n-r+1}^{n} R_{S_k}^A \prod_{k=1}^{n-r} \frac{1-R_{S_k}^A}{m-1} \\ &\quad + \left[1-\sum_{i=1}^{h}\mathrm{P}(A=V_i)\right]\prod_{k=1}^{n}\frac{1-R_{S_k}^A}{m-1} \\ &= \mathrm{P}(A=V_2)(m-1)^{-(n-q)}\prod_{k=p+1}^{p+q} R_{S_k}^A \prod_{k\notin[p+1,\,p+q]}(1-R_{S_k}^A) + \cdots + \mathrm{P}(A=V_h)(m-1)^{-(n-r)} \\ &\quad \cdot \prod_{k=n-r+1}^{n} R_{S_k}^A \prod_{k=1}^{n-r}(1-R_{S_k}^A) + \left[1-\sum_{i=1}^{h}\mathrm{P}(A=V_i)\right](m-1)^{-n}\prod_{k=1}^{n}(1-R_{S_k}^A).\end{aligned}$$

Therefore

$$\begin{aligned}&\mathrm{P}(A=V_1 \mid A_{S_1},\ldots,A_{S_n}) \\ &= \frac{\prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{n}(1-R_{S_k}^A)}{\prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{n}(1-R_{S_k}^A) + \mu_2\cdot\prod_{k=p+1}^{p+q} R_{S_k}^A \prod_{k\notin[p+1,\,p+q]}(1-R_{S_k}^A) + \cdots + \mu_h\cdot\prod_{k=n-r+1}^{n} R_{S_k}^A \prod_{k=1}^{n-r}(1-R_{S_k}^A) + \nu\cdot\prod_{k=1}^{n}(1-R_{S_k}^A)},\end{aligned} \tag{EC1}$$

where

$$\mu_2 = \frac{\mathrm{P}(A=V_2)(m-1)^{-(n-q)}}{\mathrm{P}(A=V_1)(m-1)^{-(n-p)}} = \frac{\mathrm{P}(A=V_2)}{\mathrm{P}(A=V_1)}(m-1)^{-(p-q)},$$

$$\mu_h = \frac{\mathrm{P}(A=V_h)(m-1)^{-(n-r)}}{\mathrm{P}(A=V_1)(m-1)^{-(n-p)}} = \frac{\mathrm{P}(A=V_h)}{\mathrm{P}(A=V_1)}(m-1)^{-(p-r)}, \qquad \text{and}$$

$$\nu = \frac{\left[1-\sum_{i=1}^{h}\mathrm{P}(A=V_i)\right](m-1)^{-n}}{\mathrm{P}(A=V_1)(m-1)^{-(n-p)}} = \frac{\left[1-\sum_{i=1}^{h}\mathrm{P}(A=V_i)\right]}{\mathrm{P}(A=V_1)}(m-1)^{-p}.$$

Since $\mathrm{P}(A=V_j)/\mathrm{P}(A=V_1)$, $j=2,\ldots,h$, is finite and $p>q\geq\cdots\geq r$, we must have

$$\lim_{m\to\infty}\mu_j = 0, \quad j=2,\ldots,h. \tag{EC2}$$

We next examine ν. Letting

$$L_t = \frac{\mathrm{P}(A=a_t)}{\mathrm{P}(A=V_1)}, \quad \forall\, t, \qquad \text{and} \qquad \theta = \max_t L_t,$$

we have

$$\nu < (m-h)\theta(m-1)^{-p}. \tag{EC3}$$

Since $p\geq 2$, the RHS of (EC3) goes to zero as m approaches infinity.[1] Therefore

$$\lim_{m\to\infty}\nu = 0. \tag{EC4}$$

From (EC2) and (EC4), we conclude

$$\lim_{m\to\infty}\mathrm{P}(A=V_1\mid A_{S_1},\ldots,A_{S_n}) = \frac{\prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{n}(1-R_{S_k}^A)}{\prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{n}(1-R_{S_k}^A) + 0 + \cdots + 0 + 0} = 1.$$

Case 2. There are two or more most frequent values. We show below the case with two most frequent values. The case where there are more than two most frequent values can be proved in an analogous manner.

[1] It can be easily verified that with $p=1$, the limit of ν is greater than zero. Therefore, Proposition 2 requires that the most frequent value appear in at least two data sources.

We still use $\mathbf{V}$ to denote the set of recorded values of A for an entity instance. Among them, the two most frequent values V_1 and V_2 appear p ($p \geq 2$) times in S_1 through S_p and S_{p+1} through S_{2p}, respectively; the second most frequent value V_3 appears q ($q < p$) times in S_{2p+1} through $S_{2p+q}, \ldots,$ and the least frequent value V_h is recorded r ($r < p$) in times in S_{n-r+1} through S_n. Similar to (EC1), we have

$$\mathrm{P}(A = V_1 \mid A_{S_1}, \ldots, A_{S_n}) = \frac{\prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{n} (1 - R_{S_k}^A)}{\prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{n} (1 - R_{S_k}^A) + (\mathrm{P}(A = V_2)/\mathrm{P}(A = V_1)) \prod_{k=p+1}^{2p} R_{S_k}^A \prod_{k \notin [p+1,\,2p]} (1 - R_{S_k}^A) + \omega},$$

where

$$\omega = \mu_3 \cdot \prod_{k=2p+1}^{2p+q} R_{S_k}^A \prod_{k \notin [2p+1,\,2p+q]} (1 - R_{S_k}^A) + \cdots + \mu_h \cdot \prod_{k=n-r+1}^{n} R_{S_k}^A \prod_{k=1}^{n-r} (1 - R_{S_k}^A) + \nu \cdot \prod_{k=1}^{n} (1 - R_{S_k}^A),$$

$$\mu_3 = \frac{\mathrm{P}(A = V_3)(m-1)^{-(n-q)}}{\mathrm{P}(A = V_1)(m-1)^{-(n-p)}} = \frac{\mathrm{P}(A = V_3)}{\mathrm{P}(A = V_1)} (m-1)^{-(p-q)},$$

$$\mu_h = \frac{\mathrm{P}(A = V_h)(m-1)^{-(n-r)}}{\mathrm{P}(A = V_1)(m-1)^{-(n-p)}} = \frac{\mathrm{P}(A = V_h)}{\mathrm{P}(A = V_1)} (m-1)^{-(p-r)}, \qquad \text{and}$$

$$\nu = \frac{\left[1 - \sum_{i=1}^{h} \mathrm{P}(A = V_i)\right](m-1)^{-n}}{\mathrm{P}(A = V_1)(m-1)^{-(n-p)}} = \frac{\left[1 - \sum_{i=1}^{h} \mathrm{P}(A = V_i)\right]}{\mathrm{P}(A = V_1)} (m-1)^{-p}.$$

Once again, it can be shown that: $\lim_{m \to \infty} \mu_j = 0$, $j = 3, \ldots, h$, and $\lim_{m \to \infty} \nu = 0$, which lead to

$$\lim_{m \to \infty} \omega = 0.$$

Therefore

$$\begin{aligned}
&\lim_{m \to \infty} \mathrm{P}(A = V_1 \mid A_{S_1}, \ldots, A_{S_n}) \\
&\quad = \frac{\prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{n} (1 - R_{S_k}^A)}{\prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{n} (1 - R_{S_k}^A) + (\mathrm{P}(A = V_2)/\mathrm{P}(A = V_1)) \prod_{k=p+1}^{2p} R_{S_k}^A \prod_{k \notin [p+1,\,2p]}^{n} (1 - R_{S_k}^A) + 0} \\
&\quad = \frac{\mathrm{P}(A = V_1) \prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{2p} (1 - R_{S_k}^A)}{\mathrm{P}(A = V_1) \prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{2p} (1 - R_{S_k}^A) + \mathrm{P}(A = V_2) \prod_{k=p+1}^{2p} R_{S_k}^A \prod_{k=1}^{p} (1 - R_{S_k}^A)}.
\end{aligned}$$

Similarly, we can obtain the probability that V_2 is the true attribute value:

$$\lim_{m \to \infty} \mathrm{P}(A = V_2 \mid A_{S_1}, \ldots, A_{S_n}) = \frac{\mathrm{P}(A = V_2) \prod_{k=p+1}^{2p} R_{S_k}^A \prod_{k=1}^{p} (1 - R_{S_k}^A)}{\mathrm{P}(A = V_1) \prod_{k=1}^{p} R_{S_k}^A \prod_{k=p+1}^{2p} (1 - R_{S_k}^A) + \mathrm{P}(A = V_2) \prod_{k=p+1}^{2p} R_{S_k}^A \prod_{k=1}^{p} (1 - R_{S_k}^A)}.$$

Clearly, $\mathrm{P}(\mathrm{AOV}) = 0$.

Proposition 3. *In a group of attribute values satisfying either of the following conditions*—(i) *none of the values in the group appears in any query, and* (ii) *if one value appears in a query, then all other values in the group also appear in the query in exactly the same manner—the total expected error cost associated with the most likely value in the group is always the lowest.*

Proof of Proposition 3. A group of attribute values satisfying the condition specified in Proposition 3 has the following property: if one value results in an entity instance being selected by a query, so does every other value in the group; if one value does not lead to the selection of the entity instance by a query, neither do other values in the group. Therefore, a query that includes one of the corresponding attribute values always falls into one of the following two cases:

Case 1. The entity instance is not selected by the query, no matter which value is picked from the group. Under this scenario, we only need to consider the expected cost of type I error, which, as we show in §4, depends on the probability of those values that lead to the entity instance being selected, not the probability of the value picked. Therefore, the expected cost of type I error is the same regardless of which value from the group is chosen.

Case 2. The entity instance is selected by the query, no matter which value is picked from the group. Under this scenario, we need to consider the costs of type II error and misrepresentations error(s). As we have shown in §4, the expected cost of type II error depends on the probability that the selection condition is violated, regardless of the value picked. Therefore, the expected cost of type II is also the same for all values in the group.

From the above analyses, we can see that the total expected error cost depends only on the total expected cost of misrepresentation errors. As we have shown in §4, the expected cost of misrepresentation error for an attribute is proportional to one minus the probability of the picked value. Therefore, selecting the most likely value for every attribute always results in the lowest expected misrepresentation cost, which leads to the lowest total expected error cost since the expected costs of type I and type II errors are the same for all value combinations.

EC.2. Probability Derivation for Two Data Sources

In this part, we show how the desired probability estimates are obtained for one attribute in two data sources. Equation (3) mentioned below refers to the same equation in the paper.

Case 1a. When $A_{S_1} = A_{S_2} = a_i$, (3) can be rewritten as

$$\mathrm{P}(A = a_i \mid A_{S_1} = a_i, A_{S_2} = a_i) = \frac{\mathrm{P}(A_{S_1} = a_i \mid A = a_i)\mathrm{P}(A_{S_2} = a_i \mid A = a_i)\mathrm{P}(A = a_i)}{\sum_{v=1}^{m} \mathrm{P}(A_{S_1} = a_i \mid A = a_v)\mathrm{P}(A_{S_2} = a_i \mid A = a_v)\mathrm{P}(A = a_v)}. \tag{EC5}$$

The numerator of the RHS of (EC5) equals $\mathrm{P}(A = a_i)R_{S_1}^A R_{S_2}^A$.

The denominator of the RHS of (EC5) equals

$$\begin{aligned}
&\mathrm{P}(A_{S_1} = a_i \mid A = a_i)\mathrm{P}(A_{S_2} = a_i \mid A = a_i)\mathrm{P}(A = a_i) + \sum_{v \neq i} \mathrm{P}(A_{S_1} = a_i \mid A = a_v)\mathrm{P}(A_{S_2} = a_i \mid A = a_v)\mathrm{P}(A = a_v)\\
&\quad = \mathrm{P}(A = a_i)R_{S_1}^A R_{S_2}^A + \sum_{v \neq i} [(1 - R_{S_1}^A)/(m-1)][(1 - R_{S_2}^A)/(m-1)]\mathrm{P}(A = a_v)\\
&\quad = \mathrm{P}(A = a_i)R_{S_1}^A R_{S_2}^A + [1 - \mathrm{P}(A = a_i)](1 - R_{S_1}^A)(1 - R_{S_2}^A)/(m-1)^2.
\end{aligned}$$

Therefore

$$\mathrm{P}(A = a_i \mid A_{S_1} = a_i, A_{S_2} = a_i) = \frac{\mathrm{P}(A = a_i)R_{S_1}^A R_{S_2}^A}{\mathrm{P}(A = a_i)R_{S_1}^A R_{S_2}^A + [1 - \mathrm{P}(A = a_i)](1 - R_{S_1}^A)(1 - R_{S_2}^A)/(m-1)^2}.$$

Case 1b. When $A_{S_1} = A_{S_2} = a_j \neq a_i$, (3) can be rewritten as

$$\mathrm{P}(A = a_i \mid A_{S_1} = a_j, A_{S_2} = a_j) = \frac{\mathrm{P}(A_{S_1} = a_j \mid A = a_i)\mathrm{P}(A_{S_2} = a_j \mid A = a_i)\mathrm{P}(A = a_i)}{\sum_{v=1}^{m} \mathrm{P}(A_{S_1} = a_j \mid A = a_v)\mathrm{P}(A_{S_2} = a_j \mid A = a_v)\mathrm{P}(A = a_v)}. \tag{EC6}$$

The numerator of the RHS of (EC6) equals $\mathrm{P}(A = a_i)(1 - R_{S_1}^A)(1 - R_{S_2}^A)/(m-1)^2$.

The denominator of the RHS of (EC6) equals

$$\begin{aligned}
&\mathrm{P}(A_{S_1} = a_j \mid A = a_j)\mathrm{P}(A_{S_2} = a_j \mid A = a_j)\mathrm{P}(A = a_j) + \sum_{v \neq j} \mathrm{P}(A_{S_1} = a_j \mid A = a_v)\mathrm{P}(A_{S_2} = a_j \mid A = a_v)\mathrm{P}(A = a_v)\\
&\quad = \mathrm{P}(A = a_j)R_{S_1}^A R_{S_2}^A + \sum_{v \neq j} [(1 - R_{S_1}^A)/(m-1)][(1 - R_{S_2}^A)/(m-1)]\mathrm{P}(A = a_v)\\
&\quad = \mathrm{P}(A = a_j)R_{S_1}^A R_{S_2}^A + [1 - \mathrm{P}(A = a_j)](1 - R_{S_1}^A)(1 - R_{S_2}^A)/(m-1)^2.
\end{aligned}$$

Therefore

$$\mathrm{P}(A = a_i \mid A_{S_1} = a_j, A_{S_2} = a_j) = \frac{\mathrm{P}(A = a_i)(1 - R_{S_1}^A)(1 - R_{S_2}^A)/(m-1)^2}{\mathrm{P}(A = a_j)R_{S_1}^A R_{S_2}^A + [1 - \mathrm{P}(A = a_j)](1 - R_{S_1}^A)(1 - R_{S_2}^A)/(m-1)^2}.$$

Case 2a. When $A_{S_1}=a_i$, $A_{S_2}=a_j\neq a_i$, (3) can be rewritten as

$$\mathrm{P}(A=a_i\mid A_{S_1}=a_i, A_{S_2}=a_j)=\frac{\mathrm{P}(A_{S_1}=a_i\mid A=a_i)\mathrm{P}(A_{S_2}=a_j\mid A=a_i)\mathrm{P}(A=a_i)}{\sum_{v=1}^{m}\mathrm{P}(A_{S_1}=a_i\mid A=a_v)\mathrm{P}(A_{S_2}=a_j\mid A=a_v)\mathrm{P}(A=a_v)}. \tag{EC7}$$

The numerator of the RHS of (EC7) equals $\mathrm{P}(A=a_i)R_{S_1}^A(1-R_{S_2}^A)/(m-1)$.

The denominator of the RHS of (EC7) equals

$$\begin{aligned}&\mathrm{P}(A_{S_1}=a_i\mid A=a_i)\mathrm{P}(A_{S_2}=a_j\mid A=a_i)\mathrm{P}(A=a_i)+\mathrm{P}(A_{S_1}=a_i\mid A=a_j)\mathrm{P}(A_{S_2}=a_j\mid A=a_j)\mathrm{P}(A=a_j)\\&\quad+\sum_{v\neq i,j}\mathrm{P}(A_{S_1}=a_j\mid A=a_v)\mathrm{P}(A_{S_2}=a_j\mid A=a_v)\mathrm{P}(A=a_v)\\&\quad=\mathrm{P}(A=a_i)R_{S_1}^A(1-R_{S_2}^A)/(m-1)+\mathrm{P}(A=a_j)(1-R_{S_1}^A)/(m-1)R_{S_2}^A\\&\qquad+\sum_{v\neq i,j}[(1-R_{S_1}^A)/(m-1)][(1-R_{S_2}^A)/(m-1)]\mathrm{P}(A=a_v)\\&\quad=\mathrm{P}(A=a_i)R_{S_1}^A(1-R_{S_2}^A)/(m-1)+\mathrm{P}(A=a_j)(1-R_{S_1}^A)/(m-1)R_{S_2}^A\\&\qquad+[1-\mathrm{P}(A=a_i)-\mathrm{P}(A=a_j)](1-R_{S_1}^A)(1-R_{S_2}^A)/(m-1)^2.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathrm{P}(A=a_i\mid A_{S_1}=a_i, A_{S_2}=a_j)\\&\quad=\frac{\mathrm{P}(A=a_i)R_{S_1}^A(1-R_{S_2}^A)}{\mathrm{P}(A=a_i)R_{S_1}^A(1-R_{S_2}^A)+\mathrm{P}(A=a_j)(1-R_{S_1}^A)R_{S_2}^A+[1-\mathrm{P}(A=a_i)-\mathrm{P}(A=a_j)](1-R_{S_1}^A)(1-R_{S_2}^A)/(m-1)}.\end{aligned}$$

Case 2b. When $A_{S_1}=a_j\neq a_i$, $A_{S_2}=a_k\neq a_i$, $a_j\neq a_k$, (3) can be rewritten as

$$\mathrm{P}(A=a_i\mid A_{S_1}=a_j, A_{S_2}=a_k)=\frac{\mathrm{P}(A_{S_1}=a_j\mid A=a_i)\mathrm{P}(A_{S_2}=a_k\mid A=a_i)\mathrm{P}(A=a_i)}{\sum_{v=1}^{m}\mathrm{P}(A_{S_1}=a_j\mid A=a_v)\mathrm{P}(A_{S_2}=a_k\mid A=a_v)\mathrm{P}(A=a_v)}. \tag{EC8}$$

The numerator of the RHS of (EC8) equals $\mathrm{P}(A=a_i)(1-R_{S_1}^A)(1-R_{S_2}^A)/(m-1)^2$.

The denominator of the RHS of (EC8) equals

$$\begin{aligned}&\mathrm{P}(A_{S_1}=a_j\mid A=a_j)\mathrm{P}(A_{S_2}=a_k\mid A=a_j)\mathrm{P}(A=a_j)+\mathrm{P}(A_{S_1}=a_j\mid A=a_k)\mathrm{P}(A_{S_2}=a_k\mid A=a_k)\mathrm{P}(A=a_k)\\&\quad+\sum_{v\neq j,k}\mathrm{P}(A_{S_1}=a_j\mid A=a_v)\mathrm{P}(A_{S_2}=a_j\mid A=a_v)\mathrm{P}(A=a_v)\\&\quad=\mathrm{P}(A=a_j)R_{S_1}^A(1-R_{S_2}^A)/(m-1)+\mathrm{P}(A=a_k)(1-R_{S_1}^A)/(m-1)R_{S_2}^A\\&\qquad+[1-\mathrm{P}(A=a_j)-\mathrm{P}(A=a_k)](1-R_{S_1}^A)(1-R_{S_2}^A)/(m-1)^2.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathrm{P}(A=a_i\mid A_{S_1}=a_j, A_{S_2}=a_k)\\&\quad=\frac{\mathrm{P}(A=a_i)(1-R_{S_1}^A)(1-R_{S_2}^A)/(m-1)}{\mathrm{P}(A=a_j)R_{S_1}^A(1-R_{S_2}^A)+\mathrm{P}(A=a_k)(1-R_{S_1}^A)R_{S_2}^A+[1-\mathrm{P}(A=a_j)-\mathrm{P}(A=a_k)](1-R_{S_1}^A)(1-R_{S_2}^A)/(m-1)}.\end{aligned}$$

EC.3. Implications of the Threshold Rule

Recall that the threshold rule to select Robert Black as a target customer is given by

$$\mathrm{P}(\text{married})>\frac{\gamma_{\mathrm{II}}}{\gamma_{\mathrm{I}}+\gamma_{\mathrm{II}}-\gamma_{\mathrm{m}}(AD)[1-\mathrm{P}(ad_2)]}.$$

Assuming, as before, $\gamma_{\mathrm{I}}=2$, $\gamma_{\mathrm{II}}=1$, and $\gamma_{\mathrm{m}}(AD)=0.6$, the RHS of the above threshold rule equals 0.364. Since P(married) = 0.308 for Robert Black, as before, he should not be selected for direct

Table EC.1 Probability Distribution of Marital Status

Marital status in S_1	Marital status in S_2	P(married)
M	M	0.973
N	M	0.692
M	N	0.308
N	N	0.027

marketing. From Table EC.1, we further conclude that if the observed *Marital_Status* for a customer is "M" in S_2 (regardless of the value in S_1), the customer should be selected for direct marketing.

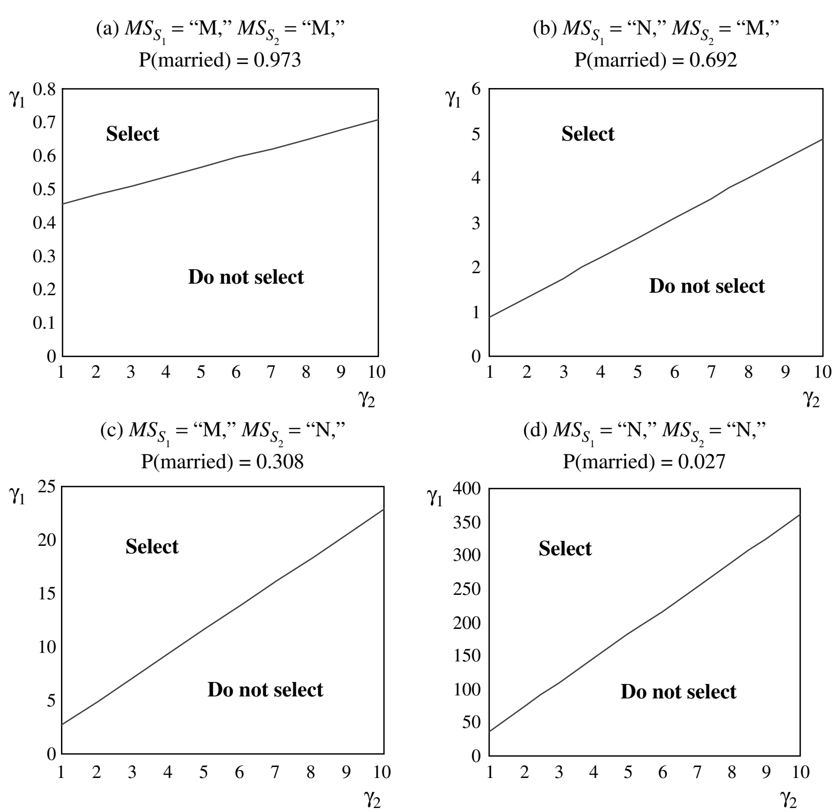
Similarly, it is easy to see that, if γ_{II} equals 0.7, then the RHS reduces to 0.286. In this case, as long as the observed *Marital_Status* in one of the data sources is "M," the customer should be selected. On the other hand, if the net benefit of direct-marketing γ_{I} is 0.5 instead of 2, and $\gamma_{\mathrm{m}}(AD)$ is 0.3 instead of 0.6, the RHS becomes 0.729. In this case, a customer should not be selected unless both data sources show a *Marital_Status* of "M" for the customer.

To further examine the impact of the costs of type I and type II errors on decision-making, we rewrite the threshold rule as

$$\gamma_{\mathrm{I}} > \gamma_{\mathrm{m}}(AD)[1-\mathrm{P}(ad_2)] + \frac{1-\mathrm{P(married)}}{\mathrm{P(married)}}\gamma_{\mathrm{II}}. \tag{EC9}$$

Figure EC.1 is created based on (EC9). In each of the four sub-figures, a straight line divides the plane into two regions. The decision should be to select the entity instance if the point corresponding to a given set of γ_{I} and γ_{II} values falls into the region above the straight line. Note that the y-axes of the four sub-figures are based on different scales. We can see from Figure EC.1 that, when the probability of being married is higher, a lower γ_{I} value would be sufficient for selecting the entity instance.

Figure EC.1 Graphical Representation of Threshold Rule for Selection of an Entity Instance, $\gamma_{\mathrm{m}}(AD) = 1.0$



EC.4. Costs of Errors Associated with Queries with Multiple Attributes in Selection Condition

The analyses presented in §§4.3 and 4.4 involve queries with a single attribute in the selection condition. Here we examine queries with multiple attributes in their selection condition. For example:

Q7: *Display the IDs and Names of those customers whose Employer is "GTE" OR whose Profession is "programmer."*

Q8: *Display the Names, Addresses, and Marital_Statuses of those customers whose Employer is "Walmart" AND whose Profession is "electrician" OR "accountant."*

Because all the attributes in the selection condition of a query jointly determine whether or not a particular entity instance is selected by that query, all attributes in the selection condition are placed into a single interrelated group for simultaneous resolution of their values. For this reason, in the graph-based procedure, we simply connect each attribute in a selection condition of a query to every other attribute in that selection condition. We first examine the costs of type I and type II errors associated with Q7 and Q8; the results are summarized in Table EC.2. With respect to Q7, as long as the stored *Employer* value is "GTE," or the stored *Profession* value is "programmer," the entity instance will be selected by Q7. Given that the entity instance is selected, if the true Employer value is not "GTE," and the true *Profession* value is not "programmer," then a type II error occurs. The frequency of its occurrence equals the product of f(Q7), and the probability that neither "GTE" nor "programmer" represents the true value of the respective attribute, given by (1 − P(GTE) − P(programmer) + P(GTE, programmer)). On the other hand, if neither "GTE" nor "programmer" is stored, the entity instance will not be selected by Q7. In this case, if "GTE" is the correct employer, or "programmer" is the true profession, then a type I error occurs. The frequency of this occurrence equals f(Q7)(P(GTE) + P(programmer) − P(GTE, programmer)). The cost derived with respect to Q8 is similar to that of Q7. The only difference is that the probability of "GTE" OR "programmer" for Q7 is replaced by the probability of "Walmart" AND ("electrician" OR "accountant") for Q8.

By comparing the results in Table EC.2 with those in Tables 5 and 6, we find that if the selection condition is considered as a whole, irrespective of the level of complexity of the selection condition, the same rules always apply: If the entity instance is selected based on the stored attribute values, the cost of type II error equals the product of the unit type II error cost, the frequency of the query, and the probability that the selection condition is violated. If the entity instance is not selected, then the cost of type I error equals the product of the unit type I error cost, the frequency of the query, and the probability that the selection condition is satisfied. Although the derivation of the probability that a selection condition is satisfied is somewhat more involved when multiple attributes are considered, as shown in Table EC.2, its estimation is conceptually straightforward.

Table EC.2 Costs of Type I and Type II Errors with Queries Having Multiple Attributes in Their Selection Condition

Query	Values to chose	Query result	If true values are	Type I error cost	Type II error cost
Q7 (OR)	(GTE, −) or (−, programmer)	Select	(GTE, −) or (−, programmer)	N/A	0
			Others	N/A	γ_{II}(Q7)f(Q7)(1 − P(GTE) − P(programmer) + P(GTE, programmer))
	Others	Do not select	(GTE, −) or (−, programmer)	γ_{I}(Q7)f(Q7)(P(GTE) − P(GTE, programmer)) + P(programmer)	N/A
			Others	0	N/A
Q8 (AND)	(Walmart, accountant) or (Walmart, electrician)	Select	(Walmart, accountant) or (Walmart, electrician)	N/A	0
			Others	N/A	γ_{II}(Q8)f(Q8)(1 − (Walmart, accountant) − P(Walmart, electrician))
	Others	Do not select	(Walmart, accountant) or (Walmart, electrician)	γ_{I}(Q8)f(Q8)(P(Walmart, accountant) + P(Walmart, electrician))	N/A
			Others	0	N/A

With respect to the misrepresentation cost involving multiple attributes, the rules are similar to the ones for a single attribute: (i) misrepresentation costs are counted only if the particular entity instance is selected by query q and a type II error does not occur for q; (ii) the cost of misrepresenting a particular attribute A with respect to query q equals the product of $\gamma_{\mathrm{m}}(q, A)$, $f(q)$, the probability that the selection condition of q is truly satisfied, and the probability that the stored value for A is incorrect; and (iii) if multiple attributes from the projection list of the same query are misrepresented, their overall misrepresentation costs must be considered.

EC.5. Theoretical Comparison of the Proposed Framework with the Other Four Approaches

Our proposed framework first estimates the probabilities of the possible attribute values, then calculates the expected error costs associated with these values, and finally selects the value with the minimum error cost. Therefore, by construction, our framework always leads to the minimum expected error cost. For this reason, unless one of the other four approaches *always* produces the same result as the one given by our framework (in which case that approach is as good as ours), our framework will outperform the other four approaches in minimizing the total expected error cost. We next show that none of the other four approaches guarantees the same decision as the one provided by our framework for all possible observations.

We illustrate this with the direct marketing example discussed in §3, considering one decision problem and two data sources. As shown in §2, given the data source reliabilities and the set of observed values in the two sources, we can compute the probability that a customer is truly married. From the threshold condition (10) in the paper, we conclude that our framework would select a particular customer (or, in other words, select the value "married" to store in the master record), if the probability that the customer is married is greater than the *threshold value* $\gamma_{\mathrm{II}}/(\gamma_{\mathrm{I}}+\gamma_{\mathrm{II}}-\gamma_{\mathrm{m}}(AD)[1-\mathrm{P}(ad_2)])$. The other four approaches would make their decision without considering the error costs or the probability associated with the attribute *Address* (AD). Without loss of generality, we assume that the reliability of the attribute *Marital_Status* (MS) in data source S_1 is lower than that in data source S_2. The selections by the four approaches for different realizations of the stored values are summarized in Table EC.3. As before, MS_{S_1} and MS_{S_2} in this table represent the observed *Marital_Status* (MS) values in data sources S_1 and S_2, respectively.

Let us first compare our framework with MLV. Assume that the stored values for a given customer in S_1 and S_2 are "N" and "M," respectively. With these two observations, $\mathrm{P}(MS=\text{“M”} \mid MS_{S_1}=\text{“N”}, MS_{S_2}=\text{“M”})$ is greater than 0.5, and consequently MLV selects the value "M." The decision made by our proposed framework, on the other hand, depends on both the probability that the customer is married and the threshold value $\gamma_{\mathrm{II}}/(\gamma_{\mathrm{I}}+\gamma_{\mathrm{II}}-\gamma_{\mathrm{m}}(AD)[1-\mathrm{P}(ad_2)])$. For instance, if this threshold value is greater than $P(MS=\text{“M”} \mid MS_{S_1}=\text{“N,”}\ MS_{S_2}=\text{“M”})$, then our framework would select the value "N." In this case, the decision made by our framework and MLV would be different. With a closer look at the threshold value condition, we conclude that unless the condition $\mathrm{P}(MS=\text{“M”} \mid MS_{S_1}=\text{“M,”} MS_{S_2}=\text{“N”}) \le \gamma_{\mathrm{II}}/(\gamma_{\mathrm{I}}+\gamma_{\mathrm{II}}-\gamma_{\mathrm{m}}(AD)[1-\mathrm{P}(ad_2)]) \le \mathrm{P}(MS=\text{“M”} \mid MS_{S_1}=\text{“N,”} MS_{S_2}=\text{“M”})$ is satisfied, the values chosen by our proposed framework and MLV will be different for one or more of the four possible realizations.

To illustrate, assume, as in §2, that MS is accurate 80% of the time in S_1 and 90% in S_2. From §2, we have

$$\mathrm{P}(MS=\text{“M”} \mid MS_{S_1}=\text{“M,”} MS_{S_2}=\text{“M”})=0.973,$$
$$\mathrm{P}(MS=\text{“M”} \mid MS_{S_1}=\text{“N,”} MS_{S_2}=\text{“M”})=0.692,$$
$$\mathrm{P}(MS=\text{“M”} \mid MS_{S_1}=\text{“M,”} MS_{S_2}=\text{“N”})=0.308,$$
$$\mathrm{P}(MS=\text{“M”} \mid MS_{S_1}=\text{“N,”} MS_{S_2}=\text{“N”})=0.027.$$

From condition (10), we can see that when the threshold value is in the range (0.308, 0.692), our proposed framework will select the values "M," "M," "N," and "N" for the four realizations shown in Table EC.3, respectively. Therefore, the selections made by our framework and MLV are always the same for this range of the threshold value. On the other hand, when the threshold value lies in the range (0.692, 0.973), our framework will select "M," "N," "N," and "N," respectively, for the four

Table EC.3 Values Chosen by Four Ad Hoc Approaches

MS_{S_1}	MS_{S_2}	$P(MS = \text{"M"} \mid MS_{S_1}, MS_{S_2})$	MLV	MRS	SMV	WMV
M	M	$P(MS = \text{"M"} \mid MS_{S_1} = \text{"M,"} MS_{S_2} = \text{"M"}) > 0.5$	M	M	M	M
N	M	$P(MS = \text{"M"} \mid MS_{S_1} = \text{"N,"} MS_{S_2} = \text{"M"}) > 0.5$	M	M	M/N	M
M	N	$P(MS = \text{"M"} \mid MS_{S_1} = \text{"M,"} MS_{S_2} = \text{"N"}) < 0.5$	N	N	M/N	N
N	N	$P(MS = \text{"M"} \mid MS_{S_1} = \text{"N,"} MS_{S_2} = \text{"N"}) < 0.5$	N	N	N	N

realizations. The decision made using MLV will be different if we observe $MS_{S_1} =$ "N" and $MS_{S_2} =$ "M." Similarly, with the threshold value in the range (0.027, 0.308), the selections made by our framework will be "M," "M," "M," and "N," respectively. Here, the decisions made by our framework and MLV will differ when $MS_{S_1} =$ "M" and $MS_{S_2} =$ "N" are observed. Similarly, we find that our framework always selects "N" when the threshold value is greater than 0.973, and "M" when the threshold value is less than 0.027. In these two cases, the selections made by the two approaches are different for two of the four feasible observations.

For this example, comparing our approach to MRS and WMV leads to the same conclusions as with MLV. When the stored values are different in the two data sources, SMV randomly selects one of the two possible values. As a result, 50% of the time the result selected by SMV for such observations will be different from the one selected by our approach.

From the above analyses, we see that, in general, none of the other approaches provides the same results as the ones produced by our framework for all possible realizations. Therefore, for the reason stated before, our framework will strictly outperform the other four approaches in minimizing the total expected error cost.