APPENDIX

A. Addressing multi-value cases

In this subsection, we will show how to extend the methodology of CBP in binary case into multi-value cases. At first, we replace the RR perturbation algorithm by k-RR [25], which is an extended version of RR for multi-value perturbation. For the case where each attribute has m possible values, the corresponding k-RR perturbation algorithm is given as follows:

$$P(b'|b) = \begin{cases} Q_1, & if\ b' = b \\ Q_2, & if\ b' \neq b \end{cases}$$

where $Q_1 + (m-1)Q_2 = 1$.

In the previous discussion, the two possible results are denoted by s and $\overline{s}$ in the context of a binary case. Now, for general multi-value cases, we redefine the η and θ as the value of two different outputs, which shall replace the corresponding positions of s and $\overline{s}$ in Theorem 4.1. Hence, Theorem 4.1 is still valid, but the deduction of term $P(E_{b_i=s}^{b'_j=s})$ needs to be revised:

$$\begin{aligned} &P(E_{b_i=\theta}^{b'_j=\eta}) = P(b'_j = \eta | b_i = \theta) \\ =&P(b_j = \eta, b'_j = b_j | b_i = \theta) + \sum_{\sigma \neq \eta} P(b_j = \sigma, b'_j \neq b_j | b_i = \theta) \\ =&\frac{P(b'_j = b_j)P(b_j = \eta, b_i = \theta)}{P(b_i = \theta)} \\ &+ \sum_{\sigma \neq \eta} \frac{P(b'_j \neq b_j)P(b_j = \sigma, b_i = \theta)}{P(b_i = \theta)} \end{aligned} \tag{8}$$

When there are only two attributes a and b, their correlation is already shown in Table II. Therefore, Equation 8 can be represented as:

$$\begin{aligned} &P(E_{a=\theta}^{b'=\eta}) \\ =&\frac{P(b' = b)P(b = \eta, a = \theta)}{P(a = \theta)} + \sum_{\sigma \neq \eta} \frac{P(b' \neq b)P(b = \sigma, a = \theta)}{P(a = \theta)} \\ =&Q_1 \frac{Pa_\theta b_\eta}{\sum_\sigma Pa_\theta b_\sigma} + Q_2 \frac{\sum_{\sigma \neq \eta} Pa_\theta b_\sigma}{\sum_\sigma Pa_\theta b_\sigma} \end{aligned}$$

B. Proof of Lemma 5.1

Proof A.1: (Lemma 5.1) Let $\pi_{b_{ki}}$ denote the true frequency of attribute b_{ki} in group G_i, $Var[\widetilde{\pi} b_{ki}(j)]$ denote the estimated variance of the j-th user, where $b_{ki}(j)$ yields to the Bernoulli distribution. The variance of estimated frequency of b_{ki} turns out to be:

$$\begin{aligned} Var[\widetilde{\pi}(b_{ki})] =& \frac{N_i Var[b_{ki}(j)]}{(\mathbb{Q}_i - (1 - \mathbb{Q}_i))^2 N_i^2} \\ =& \frac{\mathbb{Q}_i(1 - \mathbb{Q}_i) + \pi_{b_{ki}}(1 - \mathbb{Q}_i - (1 - \mathbb{Q}_i))}{N_i(\mathbb{Q}_i - (1 - \mathbb{Q}_i))^2} \\ =& \frac{\mathbb{Q}_i(1 - \mathbb{Q}_i)}{N_i(2\mathbb{Q}_i - 1)^2} \end{aligned}$$

The variance of estimated frequency of G_i turns out to be:

$$Var[\widetilde{\pi}(G_i)] = \sum_{k=1}^{s} \frac{\mathbb{Q}_i(1 - \mathbb{Q}_i)}{N_i(2\mathbb{Q}_i - 1)^2} = \frac{s\mathbb{Q}_i(1 - \mathbb{Q}_i)}{N_i(2\mathbb{Q}_i - 1)^2}$$

Let

$$f(\mathbb{Q}_i) = s\frac{\mathbb{Q}_i(1 - \mathbb{Q}_i)}{(2\mathbb{Q}_i - 1)^2} \tag{9}$$

The variance of estimated frequency of all groups can be written as:

$$Var[\widetilde{\pi}(B)] = \sum_{i=1}^{g} \frac{f(\mathbb{Q}_i)}{N_i} \tag{10}$$

Now the mission is to find the minimum value of the above function, which can be solved using Cauchy inequality. Note that $N = \sum_{i=1}^{g} N_i$, multiply equation 10 by N, we have:

$$\begin{aligned} N \cdot Var[\widetilde{\pi}(B)] = N \cdot \sum_{i=1}^{g} \frac{f(\mathbb{Q}_i)}{N_i} &= \sum_{i=1}^{g} \frac{f(\mathbb{Q}_i)}{N_i} \sum_{i=1}^{g} N_i \\ &\geq (\sum_{i=1}^{g} \sqrt{\frac{f(\mathbb{Q}_i)}{N_i}} \sqrt{N_i})^2 \\ &= (\sum_{i=1}^{g} \sqrt{f(\mathbb{Q}_i)})^2 \end{aligned} \tag{11}$$

Plug equation 9 into equation 11 and we have:

$$Min(Var[\widetilde{\pi}(B)]) = (\sum_{i=1}^{g} \sqrt{s\frac{\mathbb{Q}_i(1 - \mathbb{Q}_i)}{(2\mathbb{Q}_i - 1)^2}})^2 / N$$

If and only if vector $\sqrt{\frac{s\mathbb{Q}_i(1-\mathbb{Q}_i)}{N_i(2\mathbb{Q}_i-1)^2}}$ and $\sqrt{N_i}$ are linearly dependent, the inequality holds as an equality.

C. Proof of Theorem 5.1

Proof A.2: (Theorem 5.1) As stated in the proof of Lemma 5.1, if and only if vector $\sqrt{\frac{s\mathbb{Q}_i(1-\mathbb{Q}_i)}{(2\mathbb{Q}_i-1)^2}}$ and $\sqrt{N_i}$ are linearly dependent, the minimum variance of estimation in CBPS is reached. Hence, there exist a λ, such that:

$$\lambda\sqrt{N_i} = \sqrt{\frac{s\mathbb{Q}_i(1 - \mathbb{Q}_i)}{(2\mathbb{Q}_i - 1)^2}}$$

Hence,

$$\lambda N_i = \sqrt{\frac{s\mathbb{Q}_i(1 - \mathbb{Q}_i)}{N_i(2\mathbb{Q}_i - 1)^2}}\ i \in \{1, 2, ..., g\}$$

Notice that $\Sigma N_i = N$, with $g + 1$ independent equations, the $g + 1$ parameters (N_i, λ) can be solved.