D. Appendix: Formulas used in El Emam et al., "A Globally Optimal k-Anonymity Method for the De-identification of Health Data"

Derviation 1: Calculation of DM metric

$DM = \sum_{f_i \geq k} (f_i)^2 + \sum_{f_i < k} (n \times f_i)$ where f_i is the size of the equivalence class i, i = 1 … Z, where Z is the total number of equivalence classes in the dataset, and n is the total number of records in the dataset.

Derivation 2: Calculation of DM* metric

$DM^* = \sum_{i=1}^{Z} f_i^2$. The DM^* value for Figure 3, table (a) is 16 and for table (b), it is 28.

Derivation 3: Calculation of non-uniform entropy metric

Let V_j be a quasi-identifier, with $1 \leq j \leq J$ and J is the total number of quasi-identifiers, and $V_j = \{a_1, \ldots, a_m\}$ where m is the total number of possible values that V_j can take. For example, in table (a) in Figure 3, if V_j is the gender quasi-identifier, then $m = 2$, and $a_1 = "Male"$ and $a_2 = "Female"$. When the dataset is generalized the quasi-identifiers are denoted by V'_j and $V'_j = \{b_1, \ldots, b_{m'}\}$ where $m' \leq m$. For example, in the gender case $m' = 1$, and $a_1 \in b_1$ and $a_2 \in b_1$.

Let each cell in the original dataset be denoted by R_{ij} where $1 \leq i \leq n$ and the cells in the generalized dataset denoted by R'_{ij}. The conditional probability that the value on a randomly selected record in the original dataset is a_r given that the new generalized value is b_r (where $a_r \in b_r$) is given by:

$$\Pr\left(a_r \middle| b_{r'}\right) = \frac{\sum_{i=1}^{n} I\left(R_{ij} = a_r\right)}{\sum_{i=1}^{n} I\left(R'_{ij} = b_r\right)} \quad \text{.................................... (1)}$$

Where $I(\cdot)$ is the indicator function. The non-uniform entropy information loss is then given by:

$$-\sum_{i=1}^{n} \sum_{j=1}^{J} \log_2 \left(\Pr\left(R_{ij} \middle| R'_{ij}\right)\right) \quad \text{.................................... (2)}$$

Derivation 4: Calculation of size of equivalence classes for generalized frequency set

The complexity of an efficient merge sort is given by $N'_{L,h,p} \times \log\left(N'_{L,h,p}\right)$ [81] and the additional pass through the generalized frequency set has computations proportional to $N'_{L,h,p}$.

Therefore, the total computation to determine whether a particular node is k-anonymous is given by:

$$\left(J'_{L,h,p} \times N_{L,h,p}\right) + \left(N'_{L,h,p} \times \log\left(N'_{L,h,p}\right)\right) + N'_{L,h,p} \quad \text{..... (3)}$$

Derivation 5: Calculation of weighted non-uniform entropy

Assuming entropy is used as the information loss metric, weighting of quasi-identifiers can be achieved by using a weighted non-uniform entropy as follows:

$$-\sum_{i=1}^{n}\sum_{j=1}^{J}\left(w_j \times \log_2\left(\Pr\left(R_{ij}\middle|R'_{ij}\right)\right)\right) \quad \text{..... (4)}$$

where w_j is a weight between zero and one assigned by the data recipient to reflect variable importance.