Online Appendix to: Anonymization-Based Attacks in Privacy-Preserving Data Publishing

RAYMOND CHI-WING WONG
The Hong Kong University of Science and Technology
ADA WAI-CHEE FU
The Chinese University of Hong Kong
and
KE WANG and JIAN PEI
Simon Fraser University

A. DYNAMIC PROGRAMMING FOR EFFICIENT CREDIBILITY COMPUTATION

In this section, we will describe how we compute the credibility efficiently by dynamic programming. Specifically, Prob($|C_i(s)| = j \mid K_{ad}^{min}$) can be calculated by a dynamic programming approach. Before describing how to make use of a dynamic programming approach, we define the following events. Let F_i be the event that $0 \leq |C_i(s)| \leq \lfloor \frac{n_i}{l} \rfloor$. Let G_i be the event that $\lfloor \frac{n_i}{l} \rfloor + 1 \leq |C_i(s)| \leq n_i$. Let H_i be the event that $0 \leq |C_i(s)| \leq n_i$.

We illustrate the events in Figure 12. We can see that $F_i \cup G_i = H_i$.

The aim is to evaluate Prob($|C_i(s)| = j \mid K_{ad}^{min}$).

$$\begin{aligned} & Prob(|C_i(s)| = j \mid K_{ad}^{min}) \\ = \; & Prob(|C_i(s)| = j \mid \text{at least one } C_k \text{ among } C_1, C_2, \ldots, C_p \text{ violates } l\text{-diversity}) \end{aligned}$$

Since the event that at least one C_k among $C_1, C_2, \ldots, C_p$ violates l-diversity is equal to the event that at least one G_k occurs among $G_1, G_2, \ldots, G_p$, we have

$$\begin{aligned} & Prob(|C_i(s)| = j \mid K_{ad}^{min}) \\ = \; & Prob(|C_i(s)| = j \mid \text{at least one } G_k \text{ occurs among } G_1, G_2, \ldots, G_p) \\ = \; & \frac{\text{total no. of cases that } |C_i(s)| = j \text{ and at least one } G_k \text{ occurs among } G_1, \ldots, G_p}{\text{total no. of cases that at least one } G_k \text{ occurs among } G_1, G_2, \ldots, G_p}. \end{aligned}$$

Let A be the numerator (i.e., total number of cases that $|C_i(s)| = j$ and at least one G_k occurs among $G_1, \ldots, G_p$). Let B be the denominator (i.e., total number of cases that at least one G_k occurs among $G_1, G_2, \ldots, G_p$).

In the following, we consider the number of cases in the records in $C_1, C_2, \ldots, C_p$ only. Let there be x sensitive values in $C_1, C_2, \ldots, C_p$. Suppose that from dynamic programming, the total number of cases in the records in $C_1, C_2, \ldots, C_p$ is equal to Q. We can easily obtain the total number of cases in the records in all classes (i.e., $C_1, C_2, \ldots, C_p, C_{p+1}, \ldots, C_u$) by multiplying Q by $C_{n_s - x}^{N}$, where N is the total number of records in $C_{p+1}, C_{p+2}, \ldots, C_u$ and n_s is the total number of sensitive values in the total dataset.

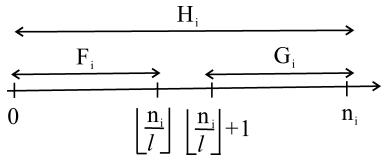

Fig. 12. Illustration of some events.

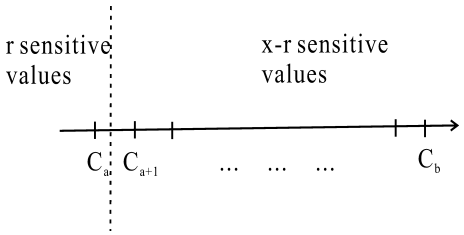

Fig. 13. Illustration of $n([a, b], x)$, $m([a, b], x)$, and $u([a, b], x)$.

For dynamic programming, we make use of three variables for the computation of A and B.

(1) $n([a, b], x)$ is the number of cases where at least one G_k occurs among G_a, G_{a+1}, ..., G_b when there are x sensitive values in C_a, C_{a+1}, ..., C_b, for $a, b = 1, 2, \ldots, p$ and $x = 1, 2, \ldots, n_s$.
(2) $m([a, b], x)$ is the number of cases where H_a, H_{a+1}, ... and H_b occur when there are x sensitive values in C_a, C_{a+1}, ..., C_b, for $a, b = 1, 2, \ldots, p$ and $x = 1, 2, \ldots, n_s$.
(3) $u([a, b], x)$ is the number of cases where F_a, F_{a+1}, ... and F_b occur when there are x sensitive values in C_a, C_{a+1}, ..., C_b, for $a, b = 1, 2, \ldots, p$ and $x = 1, 2, \ldots, n_s$.

Consider $m([a, b], x)$. Among $C_a, C_{a+1}, \ldots, C_b$, we divide the classes into two parts, $\{C_a\}$ and $\{C_{a+1}, \ldots, C_b\}$. See Figure 13. Suppose we allocate r sensitive values to C_a and $x - r$ sensitive values to $C_{a+1}, \ldots, C_b$. The number of cases where there are r sensitive values in class C_a of size n_a is equal to $C_r^{n_a}$. The number of cases where $G_{a+1}, \ldots, G_b$ occur when the number of sensitive values allocated to them is equal to $x - r$ is equal to $m([a + 1, b], x - r)$. Thus, for a given r, the total number of cases is equal to $C_r^{n_a} \times m([a + 1, b], x - r)$.

$$m([a, b], x) = \sum_{r=0}^{n_a} C_r^{n_a} \times m([a + 1, b], x - r)$$

We define the base cases of $m([a, b], x)$ as follows. The base case happens when $a = b$. It is impossible that the number of sensitive values allocated to C_a is greater than the class size of C_a or smaller than 0. Thus the term should be set to 0 in both cases. If the number of sensitive values allocated to C_a ranges from 0 to n_a, the term is the number of possible combinations where there are

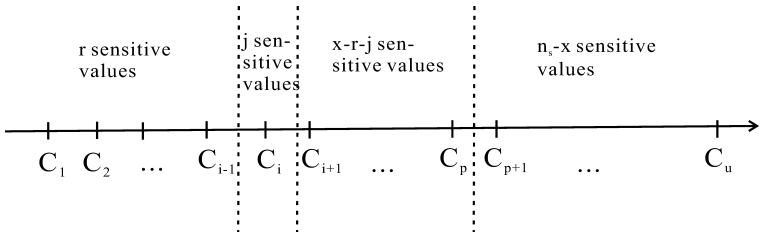

Fig. 14. Illustration of $\mathcal{A}(x)$.

x sensitive values in class C_a of size n_a (i.e., $C_x^{n_a}$).

$$m([a,a],x) = \begin{cases} 0 & \text{if} \quad x > n_a \\ 0 & \text{if} \quad x < 0 \\ C_x^{n_a} & \text{if } 0 \le x \le n_a \end{cases}$$

The term $u([a,b],x)$ is the same as $m([a,b],x)$ except that the upper boundary of term F_i is equal to $\lfloor \frac{n_a}{l} \rfloor$, instead of n_a. Similarly, we have the following formula.

$$u([a,b],x) = \sum_{r=0}^{\lfloor \frac{n_a}{l} \rfloor} C_r^{n_a} \times u([a+1,b],x-r)$$

$$u([a,a],x) = \begin{cases} 0 & \text{if } x \ge \lfloor \frac{n_a}{l} \rfloor + 1 \\ 0 & \text{if} \quad x < 0 \\ C_x^{n_a} & \text{if } 0 \le x \le \lfloor \frac{n_a}{l} \rfloor \end{cases}$$

Next consider $n([a,b],x)$. Let r be the number of tuples with s in C_a. We can also derive $n([a,b],x)$ as follows similarly by considering two cases: (1) $\lfloor \frac{n_a}{l} \rfloor + 1 \le r \le n_a$ and (2) $0 \le r \le \lfloor \frac{n_a}{l} \rfloor$.

$$n([a,b],x) = \sum_{r=\lfloor \frac{n_a}{l} \rfloor + 1}^{n_a} C_r^{n_a} \times m([a+1,b],x-r) + \sum_{r=0}^{\lfloor \frac{n_a}{l} \rfloor} C_r^{n_a} \times n([a+1,b],x-r)$$

The base cases of $n([a,b],x)$ can also be easily derived as follows.

$$n([a,a],x) = \begin{cases} 0 & \text{if} \quad x > n_a \\ 0 & \text{if} \quad 0 \le x \le \lfloor \frac{n_a}{l} \rfloor \\ C_x^{n_a} & \text{if } \lfloor \frac{n_a}{l} \rfloor + 1 \le x \le n_a \end{cases}$$

Now, consider A. Recall that A is the total number of cases that; (1) at least one G_k occurs among $G_1, G_2, \ldots, G_p$ and (2) there are j sensitive values in C_i.

Let $\mathcal{A}(x)$ be the total number of aforesaid cases provided that there are x sensitive values in $C_1, C_2, \ldots, C_p$.

We consider the number of cases involving all classes (i.e., $C_1, \ldots, C_p$, $C_{p+1}, \ldots, C_u$). Suppose we allocate x sensitive values in $C_1, C_2, \ldots, C_p$ and $n_s - x$ sensitive values in $C_{p+1}, C_{p+2}, \ldots, C_u$. Recall that C_i contains j sensitive values. Within $C_1, C_2, \ldots, C_p$, we further allocate: (1) r sensitive values to $C_1, C_2, \ldots, C_{i-1}$, (2) j sensitive values to C_i, and (3) $x - r - j$ sensitive values to $C_{i+1}, C_{i+2}, \ldots, C_p$. See Figure 14.

There are two cases.

Case 1. $\lfloor \frac{n_i}{l} \rfloor + 1 \leq j \leq n_i$, that is, G_i occurs. This means that the number of sensitive values in a class C_k ($|C_k(s)|$) of $C_1, C_2, \ldots, C_{i-1}$ or $C_{i+1}, C_{i+2}, \ldots, C_p$ ranges from 0 to n_k.

The number of cases that $|C_k(s)|$ for a class C_k of $C_1, C_2, \ldots, C_{i-1}$ ranges from 0 to n_k is equal to $m([1, i-1], r)$. Similarly, the number of cases that $|C_k(s)|$ for a class C_k of $C_{i+1}, C_{i+2}, \ldots, C_p$ ranges from 0 to n_k is equal to $m([i+1, p], x-r-j)$. Thus, if we consider all possible values of r from 0 to $x-j$, the total number of these cases is equal to $\sum_{r=0}^{x-j} m([1, i-1], r) \times m([i+1, p], x-r-j)$.

Note that the number of cases that there are j sensitive values in C_i of size n_i is equal to $C_j^{n_i}$. Also, the number of cases that there are $n_s - x$ sensitive values in N tuples in classes $C_{p+1}, C_{p+2}, \ldots, C_u$ is equal to $C_{n_s-x}^N$. Thus, $\mathcal{A}(x) = C_{n_s-x}^N \times C_j^{n_i} \times \sum_{r=0}^{x-j} m([1, i-1], r) \times m([i+1, p], x-r-j)$ in this case.

Case 2. $0 \leq j \leq \lfloor \frac{n_i}{l} \rfloor$, that is, G_i does not occur. There are the following subcases.

Case 2(a). At least one G_k occurs among $G_1, G_2, \ldots, G_{i-1}$, In this case, the number of sensitive values in a class C_k of $C_{i+1}, C_{i+2}, \ldots, C_p$ ranges from 0 to n_k.

The number of cases that at least one G_k occurs among $G_1, G_2, \ldots, G_{i-1}$ is equal to $n([1, i-1], r)$. The number of cases that the number of sensitive values in a class C_k of $C_{i+1}, C_{i+2}, \ldots, C_p$ ranges from 0 to n_k is equal to $m([i+1, p], x-r-j)$. Thus, the total number of these cases is equal to $n([1, i-1], r) \times m([i+1, p], x-r-j)$.

Case 2(b). All G_k among $G_1, G_2, \ldots, G_{i-1}$ does not occur. In other words, all $F_1, F_2, \ldots, F_{i-1}$ occur. Besides, we should also know that there is at least one G_k occuring among $G_{i+1}, G_{i+2}, \ldots, G_p$.

The number of cases that all $F_1, F_2, \ldots, F_{i-1}$ occur is equal to $u([1, i-1], r)$. The number of cases that at least one G_k occurs among $G_{i+1}, G_{i+2}, \ldots, G_p$ is equal to $n([i+1, p], x-r-j)]$. Thus, the total number of these cases is equal to $u([1, i-1], r) \times n([i+1, p], x-r-j)$.

By combining Case 2(a) and Case 2(b) and considering all possible values r from 0 to $x-j$, we obtain the total number of cases equal to $\sum_{r=0}^{x-j} [n([1, i-1], r) \times m([i+1, p], x-r-j) + u([1, i-1], r) \times n([i+1, p], x-r-j)$.

Similarly, the number of cases that there are j sensitive values in C_i of size n_i is equal to $C_j^{n_i}$. Also, the number of cases that there are $n_s - x$ sensitive values in N tuples in classes $C_{p+1}, C_{p+2}, \ldots, C_u$ is equal to $C_{n_s-x}^N$. Thus, the total number of cases in Case (2) is equal to $C_{n_x-x}^N \times C_j^{n_i} \times \sum_{r=0}^{x-j} [n([1, i-1], r) \times m([i+1, p], x-r-j) + u([1, i-1], r) \times n([i+1, p], x-r-j)]$.

We obtain $\mathcal{A}(x)$ as follows.

$$\mathcal{A}(x) = \begin{cases} C_{n_s-x}^N \times C_j^{n_i} \times \sum_{r=0}^{x-j} m([1, i-1], r) \\ \quad \times m([i+1, p], x-r-j) & \text{if } \lfloor \frac{n_i}{l} \rfloor + 1 \leq j \leq n_i \\ C_{n_s-x}^N \times C_j^{n_i} \times \sum_{r=0}^{x-j} [n([1, i-1], r) \\ \quad \times m([i+1, p], x-r-j) & \text{if } 0 \leq j \leq \lfloor \frac{n_i}{l} \rfloor \\ \quad + u([1, i-1], r) \times n([i+1, p], x-r-j)] \end{cases}$$

By considering all possible values of x from $\lfloor \frac{n_1}{l} \rfloor + 1$ to n_s, A is equal to the following. (Note that it is impossible that $x < \lfloor \frac{n_1}{l} \rfloor + 1$ because it means that there is no need for generalization.)

$$A = \sum_{x=\lfloor \frac{n_1}{l} \rfloor + 1}^{n_s} \mathcal{A}(x)$$

Consider B where B is the total number of cases where at least one G_k occurs among $G_1, G_2, \ldots, G_p$. By considering all possible values x from $\lfloor \frac{n_1}{l} \rfloor + 1$ to n_s, we obtain the following formula.

$$B = \sum_{x=\lfloor \frac{n_1}{l} \rfloor}^{n_s} n([1, p], x) \times C_{n_s - x}^{N}$$

Algorithm. Algorithm 2 shows the computation of the credibility by dynamic programming. It involves two phases. In phase 1, we compute the variables $n([a, b], x)$, $m([a, b], x)$ and $u([a, b], x)$. In phase 2, we compute $Credibility(o, s, K_{ad}^{min})$ where $o \in C_i$ for $i = 1, 2, \ldots, p$ by using the variables used in phase 1, namely $n([a, b], x)$, $m([a, b], x)$, and $u([a, b], x)$.

Let $|T|$ be the number of tuples in T. Algorithm 2 runs in polynomial time in $|T|$, p, n_s, and l. It is easy to verify that phase 1 takes $O(|T| + p^2 n_s)$. Consider phase 2. Computing one instance of A and computing one instance of B take $O(n_s^2)$ and $O(n_s)$, respectively. Since there are $O(pn_p)$ iterations, phase 2 takes $O(pn_p n_s^2)$. Since $n_p = O(n_s l)$, the complexity of phase 2 becomes $O(pn_s^3 l)$. Thus, the running time of Algorithm 2 is $O(|T| + p^2 n_s + pn_s^3 l)$.

THEOREM 6. *Algorithm 2 runs in $O(|T| + p^2 n_s + pn_s^3 l)$ time.*

The previous theorem means that computing the credibility of an individual only takes polynomial time in $|T|$, p, n_S, and l. In other words, this kind of attack is highly feasible.

B. PROOF OF LEMMAS/THEOREMS

PROOF OF THEOREM 1. We will prove that the credibility as computed by the formulae for credibility is exactly the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC by first considering a class Q in T^* where only two QID values in T^e, namely q_1 and q_2, are generalized to Q. Then, we relax the proof by considering a class Q where multiple QID values are generalized to Q.

Consider a QID value Q in T^*. Suppose q_1 and q_2 (in T^e) are generalized to Q in T^*. Let n_1 and n_2 be the number of tuples with value q_1 and q_2, respectively. Let x be the total number of sensitive tuples in Q.

Consider four cases. *Case* 1: $x \leq n_1$ and $x \leq n_2$. Without loss of generality, we consider $Credibility(o, s, K_{ad}^{min})$ where o has a QID value on q_1. We further consider a number of subcases. *Case* (a): $x = 1$. We have the sensitive

Algorithm. 2 Algorithm for Computing Credibility

// Phase 1(a): Initialization
obtain n_i for all i
for $a = 1$ to p **do**
  for $x = 0$ to n_s **do**
    initialize $n([a,a],x)$, $m([a,a],x)$ and $u([a,a],x)$
  end for
end for
// Phase 1(b): Recursion
// Compute $m([a,b],x)$ and $u([a,b],x)$
for $x = 0$ to n_s **do**
  for $a = p$ downto 1 **do**
    for $b = a+1$ to p **do**
      $m([a,b],x) \leftarrow 0$
      for $r = 0$ to n_a **do**
        if $x - r \geq 0$ **then**
          $m([a,b],x) \leftarrow m([a,b],x) + C_r^{n_a} \times m([a+1,b],x-r)$
        end if
      end for
      $u([a,b],x) \leftarrow 0$
      for $r = 0$ to $\lceil n_a/l \rceil$ **do**
        if $x - r \geq 0$ **then**
          $u([a,b],x) \leftarrow u([a,b],x) + C_r^{n_a} \times u([a+1,b],x-r)$
        end if
      end for
    end for
  end for
end for
// Compute $n([a,b],x)$
for $x = 0$ to n_s **do**
  for $a = p$ downto 1 **do**
    for $b = a+1$ to p **do**
      $n([a,b],x) \leftarrow 0$
      for $r = \lfloor n_a/l \rfloor + 1$ to n_a **do**
        if $x - r \geq 0$ **then**
          $n([a,b],x) \leftarrow n([a,b],x) + C_r^{n_a} \times m([a+1,b],x-r)$
        end if
      end for
      for $r = 0$ to $\lfloor n_a/l \rfloor$ **do**
        if $x - r \geq 0$ **then**
          $n([a,b],x) \leftarrow n([a,b],x) + C_r^{n_a} \times n([a+1,b],x-r)$
        end if
      end for
    end for
  end for
end for
// Phase 2: Computing credibility $Credibility(o, s, K_{ad}^{min})$
Let $cred_i$ be $Credibility(o, s, K_{ad}^{min})$ where $o \in C_i$
for $i = 1$ to p **do**
  $cred_i \leftarrow 0$
  for $j = 1$ to n_i **do**
    calculate A and B according to $n([a,b],x)$, $m([a,b],x)$ and $u([a,b],x)$
    $cred_i \leftarrow cred_i + \frac{A}{B} \times \frac{j}{n_i}$
  end for
end for

Table XXIII. Possible Combinations of Number of Sensitive Tuples when $x = 1$

	Number of sensitive tuples		Total number of cases
	$q1$	$q2$	
(a)	0	1	$C_0^{n_1} \times C_1^{n_2}$
(b)	1	0	$C_1^{n_1} \times C_0^{n_2}$

Table XXIV. Possible Combinations of Number of Sensitive Tuples when $x = 2$

	Number of sensitive tuples		Total number of cases
	$q1$	$q2$	
(a)	0	2	$C_0^{n_1} \times C_2^{n_2}$
(b)	1	1	$C_1^{n_1} \times C_1^{n_2}$
(c)	2	0	$C_2^{n_1} \times C_0^{n_2}$

tuple distribution table as shown in Table XXIII. It is easy to see that

$$\begin{aligned} \mathit{Credibility}(o, s, K_{ad}^{min}) &= \frac{\text{total number of cases for Scenario (b)}}{\text{total number of all possible cases}} \times \frac{1}{n_1} \\ &= \frac{C_1^{n_1} \times C_0^{n_2}}{C_0^{n_1} \times C_1^{n_2} + C_1^{n_1} \times C_0^{n_2}} \times \frac{1}{n_1} \\ &= \frac{n_1}{n_2 + n_1} \times \frac{1}{n_1} \\ &= \frac{1}{n_1 + n_2} \end{aligned}$$

which is equal to the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC Q.

Case (b): $x = 2$. Similarly, we have the sensitive tuple distribution table as shown in Table XXIV. We have

$$\begin{aligned} \mathit{Credibility}(o, s, K_{ad}^{min}) &= \frac{\text{total number of cases for Scenario (b)}}{\text{total number of all possible cases}} \times \frac{1}{n_1} \\ &\quad + \frac{\text{total number of cases for Scenario (c)}}{\text{total number of all possible cases}} \times \frac{2}{n_1} \\ &= \frac{C_1^{n_1} \times C_1^{n_2}}{C_0^{n_1} \times C_2^{n_2} + C_1^{n_1} \times C_1^{n_2} + C_2^{n_1} \times C_0^{n_2}} \times \frac{1}{n_1} \\ &\quad + \frac{C_2^{n_1} \times C_0^{n_2}}{C_0^{n_1} \times C_2^{n_2} + C_1^{n_1} \times C_1^{n_2} + C_2^{n_1} \times C_0^{n_2}} \times \frac{2}{n_1} \\ &= \frac{n_1 n_2}{\frac{n_2(n_2-1)}{2} + n_1 n_2 + \frac{n_1(n_1-1)}{2}} \times \frac{1}{n_1} \\ &\quad + \frac{\frac{n_1(n_1-1)}{2}}{\frac{n_2(n_2-1)}{2} + n_1 n_2 + \frac{n_1(n_1-1)}{2}} \times \frac{2}{n_1} \\ &= \frac{2(n_1 + n_2 - 1)}{n_1^2 + n_2^2 + 2n_1 n_2 - n_1 - n_2} \end{aligned}$$

$$= \frac{2(n_1 + n_2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}$$
$$= \frac{2}{n_1 + n_2}$$

which is equal to the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC Q.

Case (c): $x > 2$. Inductively, we can also derive that

$$Credibility(o, s, K_{ad}^{min}) = \frac{x}{n_1 + n_2}$$

which is equal to the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC Q.

We consider the other three cases. *Case* 2: $x \leq n_1$ and $x > n_2$, *Case* 3: $x > n_1$ and $x \leq n_2$, and *Case* 4: $x > n_1$ and $x > n_2$. With similar arguments, we also conclude that

$$Credibility(o, s, K_{ad}^{min}) = \frac{x}{n_1 + n_2}.$$

Now, we consider the class Q where multiple QID values are generalized to Q. Since the idea is similar and the key idea is no exclusion of any scenarios in the sensitive tuple distribution table, we obtain that the credibility is exactly the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC. □

PROOF OF LEMMA 2. To prove this lemma, we give an example where 2 QID's $q1$ and $q2$ are generalized to Q. There are 4 tuples of $q1$ and 2 tuples of $q2$. In total, there are 3 occurrences of the sensitive value set s in the 6 tuples. If 2-diversity is the goal, then we can exclude the case of 2 sensitive $q1$ tuple and 1 sensitive $q2$ tuple. After the exclusion, the credibility of any linkage between any individual to s still does not exceed 0.5. □

PROOF OF THEOREM 2. We shall transform the problem of Exact Cover by 3-Sets (X3C) [Holyer 1981] to the m-confidentiality anonymization problem. X3C is defined by: Given a set X with $|X| = 3q$ and a collection C of 3-element subsets of X. Does C contain an exact cover for X, namely a subcollection $C' \subseteq C$ such that every element of X occurs in exactly one member of C'?

Given an instance of X3C, we transform it to an instance of optimal m-confidentiality under global recoding as follows. Create a table T with two attributes $\mathcal{Q}$ and S, where $\mathcal{Q}$ is a QID attribute and S is a sensitive attribute that may contain sensitive values. For S, there is only one sensitive value s_v and one nonsensitive value s_n. We set $weight(\mathcal{Q}) = 1$. For each element x in X, create a tuple with $\mathcal{Q} = x$ and $S = s_v$. Hence, each value of x appears in exactly one tuple. Let the elements in C be $c_1, \ldots, c_N$. For each element $c_i = (x, y, z)$ in C, create a taxonomy $\mathcal{T}_i$. $\mathcal{T}_i$ contains ground elements of x, y, z, n_{i1}, n_{i2}, and n_{i3}, which are children of a root node r_i. Create 3 tuples with $\mathcal{Q} = n_{ij}$ and $S = s_n$, for $j = 1, 2, 3$.

The remaining of the proof is to show: C contains an exact cover for X if and only if there is a solution T^* for the 2-confidentiality problem with

$Dist(T, T^*) = e$ where $e = \frac{2q}{q+N}$. Firstly, we prove that if C contains an exact cover for X, then there is a solution T^* for the 2-confidentiality problem with $Dist(T, T^*) = \frac{2q}{q+N}$. Let C' be the exact cover for X. We know that every element of X occurs in exactly one member of C'. Then, for each $c_i = (x, y, z) \in C'$, the correspondence taxonomy $\mathcal{T}_i$ is used for the generalization of x, y, and z together with n_{i1}, n_{i2}, and n_{i3} because with global recoding all occurrences of an attribute value are recoded to the same value. Thus, for each generalization from $\mathcal{T}_i$, the information loss of these six tuples in T^* are 6. Since $|C'| = q$, the total information loss among all tuples (i.e., $\sum_{t^* \in T^*} \mathcal{IL}(t^*)$) is equal to $6q$. Since $Dist(T, T^*) = \frac{\sum_{t^* \in T^*} \mathcal{IL}(t^*)}{|T^*|}$ and the total number of tuples in T^* is equal to $3q + 3N$, we have $Dist(T, T^*) = \frac{6q}{3q+3N} = \frac{2q}{q+N}$. Besides, note that the adversary cannot launch a minimality attack since each QID value appears only in one tuple in the set of tuples. The adversary cannot exclude any possible combination of the table of sensitive tuple distribution. From Theorem 1, minimality attack is not possible. Besides, the frequency of each QID-EC with s_v in T^* is at most 0.5. Thus, there is a solution T^* for the 2-confidentiality problem with $Dist(T, T^*) = \frac{2q}{q+N}$.

Now, we prove that if there is a solution T^* for the 2-confidentiality problem with $Dist(T, T^*) = \frac{2q}{q+N}$, then C contains an exact cover for X. Similarly, since each QID value appears only in one tuple, it is impossible for the adversary to exclude any possible combination of the table of sensitive tuple distribution. From Theorem 1, minimality attack is not possible. Since the frequency of each QID-EC with s_v in T^* is at most 0.5 and $Dist(T, T^*) = \frac{2q}{q+N}$, T^* is a result of the generalizations by using exactly q taxonomies $\mathcal{T}_i$ containing disjoint ground values. Otherwise, either the frequency of some QID-EC's in T^* is greater than 0.5 or $Dist(T, T^*) > \frac{2q}{q+N}$, that is, each tuple with value s is generalized by exactly one generalization taxonomy. In other words, each element in X occurs in exactly one member of the set $C' \subseteq C$ such that $|C'| = q$ and each $c \in C'$ corresponds to $\mathcal{T}_i$ used for generalization. Thus, C contains an exact cover C' for X.

Besides, it is easy to see that the reduction runs in polynomial time. From Theorem 6 (in Section 4), we know that we can compute the credibility of each individual in polynomial time. Thus, we can verify problem optimal m-confidentiality in polynomial time. So, problem optimal m-confidentiality under global recoding is NP-complete. □

Proof of Theorem 3. We shall transform the problem of Partition into 4-Cliques [Holyer 1981] to the m-confidentiality anonymization problem. Partition into 4-Cliques is defined by: Given a simple graph $G = (V, E)$, with $|E| = 6k$ for some integer k, can the edges of G be partitioned into k edge-disjoint 4-cliques?

Given an instance of Edge Partition into 4-Cliques. Set $m = 6$. For each vertex $v \in V$, construct a QID attribute. For each edge $e \in E$, where $e = (v_1, v_2)$, create a record r_{v_1,v_2} in which the QID attribute values v_1 and v_2 are equal to 1 and all other QID attribute values equal to 0. Besides, we associate each record with a sensitive attribute S. We generate sensitive attribute values of all records as follows. If two edges share a common vertex, the sensitive attribute

values of their corresponding records are different. The aforesaid principle can be accomplished with the following steps. Firstly, the sensitive values of all records are set to 0. Then, we randomly find a record r where the corresponding edge is e. Let A be the set of all records where the corresponding vertices have a common vertex with e. Then, we can obtain a set A' containing the sensitive values of all records in A. Find the smallest positive value which does not occur in A'. Assign this value as the sensitive attribute value of record r. Repeat the previous steps for each of the remaining records with sensitive value $= 0$. It is noted that the preceding process resembles a process of edge coloring. However, since the aforesaid process does not require that the edges are colored with a limited (or optimal) number of colors, it can be done in polynomial time.

We define the cost in the 6-confidentiality problem to be the number of suppressions applied in the dataset. We show that this cost is at most $24k$ if and only if E can be partitioned into a collection of k edge-disjoint 4-cliques.

Suppose E can be partitioned into a collection of k disjoint 4-cliques. Consider a 4-clique Q with vertices v_1, v_2, v_3, and v_4. If we suppress the attributes v_1, v_2, v_3, and v_4 in the 6 records corresponding to the edges in Q, then a cluster of these 6 records are formed where each modified record has four *'s. Note that the the frequency of each sensitive value in this cluster is at most 1/6. Similar to Theorem 2, the adversary cannot launch a minimality attack since each QID value appears only in one tuple in the cluster. Thus, the dataset satisfies 6-confidentiality. The cost of the 6-confidentiality is equal to $6 \times 4 \times k = 24k$.

Suppose the cost for the 6-confidentiality problem is at most $24k$. As G is a simple graph, any six records should have at least four different attributes. So each record should have at least four *'s in the solution of 6-confidentiality. Then, the cost of 6-confidentiality is at least $6 \times 4 \times k = 24k$. Combining with the proposition that the cost is at most $24k$, we find that the cost is exactly equal to $24k$ and thus each record should have exactly four *'s in the solution. Each cluster should have exactly 6 records (with different sensitive values). Suppose the six modified records contain four *'s in attributes v_1, v_2, v_3, and v_4, the records contain 0's in all other nonsensitive attributes. This corresponds to a 4-clique with vertices v_1, v_2, v_3 and v_4. Thus, we conclude that the solution corresponds to a partition into a collection of k edge-disjoint 4-cliques.

Similar to Theorem 2, it is easy to see that the reduction runs in polynomial time. From Theorem 6 (in Section 4), we know that we can compute the credibility of each individual in polynomial time. Thus, we can verify problem optimal m-confidentiality in polynomial time. We conclude that optimal m-confidentiality under local recoding is NP-complete. □

PROOF OF THEOREM 4. In order to prove that T^* generated by algorithm MASK is m-confidential, we analyze how the adversary performs an attack, given that the adversary knowledge contains not only the knowledge described in Assumption 3 but also the mechanism of algorithm MASK. In the following, we show that the credibility computed is at most $1/m$. Let the privacy requirement considered be $\mathcal{R}$ (i.e., m-confidentiality). Let the privacy requirement for k-anonymity be $\mathcal{R}^k$. From T^*, the adversary knows that, in T^*, for each

QID-EC Q_i, the size of Q_i is at least k and the frequency of each sensitive value (in fraction) is at most $1/m$. We consider two cases.

Case 1. The published table T^* is equal to the minimal k-anonymous table T^k generated in step 1 of algorithm MASK; that is, $|\mathcal{V}|$ is equal to $\emptyset$. We know that algorithm MASK generates T^* such that the information loss of T^* is minimal with respect to the QID attributes. Note that T^* is a result of generalization for privacy requirement $\mathcal{R}^k$ (instead of privacy requirement $\mathcal{R}$). However, at the same time, T^* also satisfies m-diversity.

We prove that T^* also satisfies $\mathcal{R}$ in the following. Similar to Section 4, we can also compute the credibility by constructing the sensitive tuple distribution table with condition (1) and condition (2) (but not condition (3)) of Definition 10 accordingly in this case. It is noted that we do not need to consider condition (3) since the generalization step for generating T^* is caused by "unequal" QID values in the original table T (without the consideration of the sensitive attribute). In other words, the generalization is performed for $\mathcal{R}^k$ (instead of m-diversity). Since condition (3) is not considered, there is no exclusion of any combination of the number of sensitive tuples in the sensitive tuple distribution table in the adversary's analysis of this case. (However, the existence of the exclusion used in Section 4 is due to the fact that the generalization is caused by the consideration of both QID attributes and the sensitive attribute; that is, the generalization is performed for m-diversity instead of $\mathcal{R}^k$ where condition (3) is involved during the generation of the sensitive tuple distribution table.)

Since there is no exclusion of any combination in the sensitive tuple distribution table in this case, by Theorem 1, the credibility as computed by the formulae for credibility is exactly the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC. Besides, in T^*, for each QID-EC Q_i, the frequency of each sensitive value (in fraction) is at most $1/m$. We deduce that $Credibility(o, s, K_{ad}^{min})$ is at most $1/m$ for any individual o and any sensitive value set s. So, T^* satisfies $\mathcal{R}$.

Case 2. The published table T^* is not equal to T^k; that is, $|\mathcal{V}|$ is not equal to $\emptyset$. Then, the adversary knows that step 2(a) of algorithm MASK is performed. We consider two subcases.

Subcase (a). The total number of QID-EC's which satisfy m-diversity in T^k is smaller than $u(= (m-1) \times |V|)$. This case is impossible because T^* is already published.

Subcase (b). The total number of QID-EC's which satisfy m-diversity in T^k is equal to or greater than u. The analysis of the credibility in this case is different from Case 1. This analysis involves two major steps. The first step is that the adversary deduces that some of the nonsensitive values in T^* originally come from sensitive values in T. This is because some sensitive values are distorted or modified to become nonsensitive values in step 4 of algorithm MASK. The second step is similar to Section 4 and Case (1). Specifically, according to the sensitive values deduced in the first step, the adversary can compute the credibility by the sensitive tuple distribution table with condition (1) and condition (2) but not condition (3) of Definition 10 accordingly. Again, it is noted that we do not need to consider condition (3) since the generalization step for generating

T^* is caused by "unequal" QID values in the original table T (without the consideration of the sensitive attribute).

In the following, we will show that it is difficult for the adversary to achieve the first step. Then, with this result, we assume that we only need to consider the sensitive values in T^* to calculate the credibility in the second step.

For the first step, it is infeasible for the adversary to figure out what are the original sensitive values because the adversary does not have the knowledge about: (1) the size of $\mathcal{V}$, (2) the original frequency of the sensitive tuples in each QID-EC $\in \mathcal{V}$, and (3) which QID-EC's in T^* come from $\mathcal{V}$. It may be argued that the adversary can first consider all possible choices of the aforesaid knowledge, compute the credibility for each choice with the second step, and finally compute the final credibility with all choices. This approach does not work because without sufficient knowledge, we do not know the probability that each choice occurs. Assuming a random world assumption (i.e., all such probabilities have the same values) is also not reasonable. This is because, for example, the adversary cannot tell whether the probability that $|\mathcal{V}| = 1$ occurs is equal to the probability that $|\mathcal{V}| = 2$ occurs or not. With this reasoning, the deduction of the original sensitive values is impossible.

For the second step, we assume that the sensitive values considered come from T^*. By similar arguments as Case 1, since there is no exclusion of any combination, the credibility as computed by the formulae for credibility is exactly the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC. Besides, in T^*, for each QID-EC Q_i, the frequency of each sensitive value (in fraction) is at most $1/m$. $Credibilityx(o, s, K_{ad}^{min})$ is at most $1/m$ for any individual o and any sensitive value set s. So, T^* satisfies $\mathcal{R}$. □