使用展頻技術之數值型資料庫浮水印

楊英一
高雄第一科技大學
資訊管理系

吳大鈞
高雄第一科技大學
電腦與通訊工程系

蔡文祥
交通大學
資訊工程學系
亞洲大學
資訊工程學系

Speaker:吳大鈞 副教授

94.9.1

大綱

- 簡介
- 展頻技術
- 展頻技術應用於資料庫浮水印
  - 有原始資料庫可供參考之版權保護
  - 無原始資料庫可供參考之版權保護
- 結論

簡介

資料庫保護之迫切性

- 資料庫之資料易被有系統地盜拷
  - Rural Telephone Service Company vs. Feist Publication
  - 104 人力銀行 vs. 1111 人力銀行

- 資料庫市場之興起
  - 資料探勘技術之進步
  - 如 Walmart 出售過期之銷售資訊

簡介(續)

數位浮水印(Digital Watermarking)

- 嵌入與版權相關資訊於欲保護之物件中以保護版權或驗證真確性
- 可應用在資料庫之版權保護上

簡介(續)

資料庫之特性

- 資料庫由值組(tuple)組成
  - 值組之間沒有順序性
- 大多資料為文字與數字型態
  - 資料可改變空間較小
- 可能具有主鍵(prime key)或候選鍵

簡介(續)

資料庫之一般操作

- 新增資料
- 刪除資料
- 修改資料

展頻技術

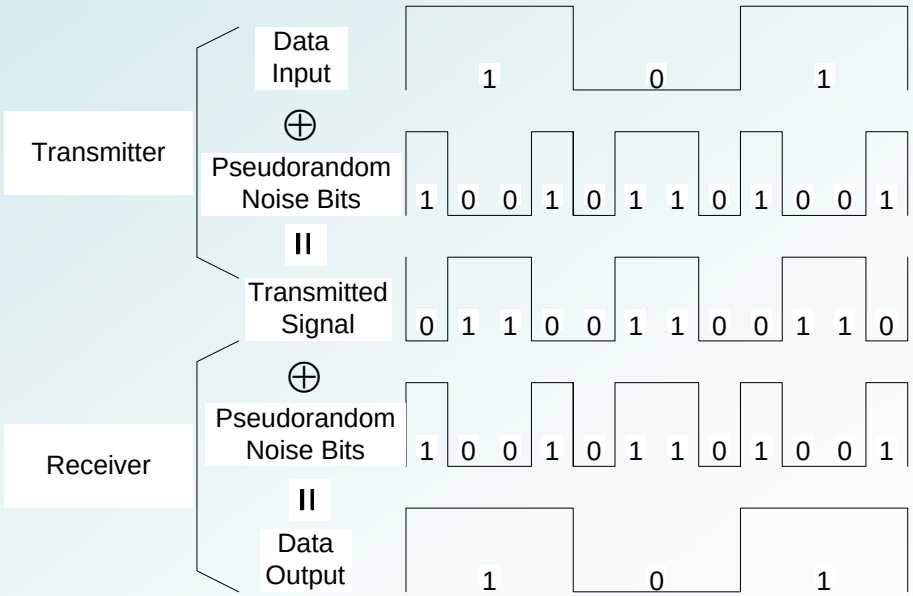
跳頻展頻(Frequency Hopping Spread Spectrum)

- 發送和接收雙方皆擁有頻道切換器
- 發送和接收雙方同步切換頻道

展頻技術(續)

直接序列展頻(Direct Sequence Spread Spectrum)

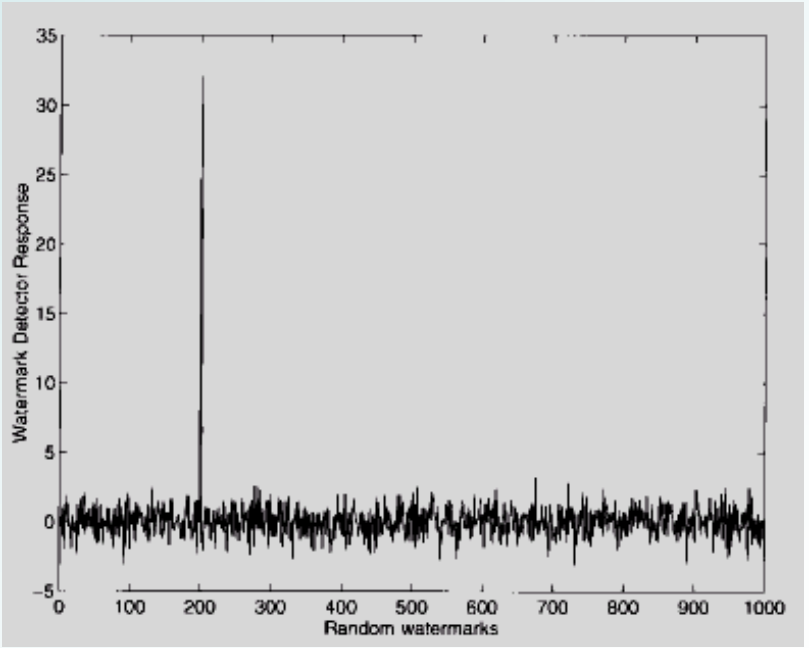
- 將資料位元延展為多個位元
- 利用雜訊產生器將資料擾亂



展頻技術於影像浮水印之應用

以Cox所提之方法為例

- 1000個不同key之浮水印偵測結果

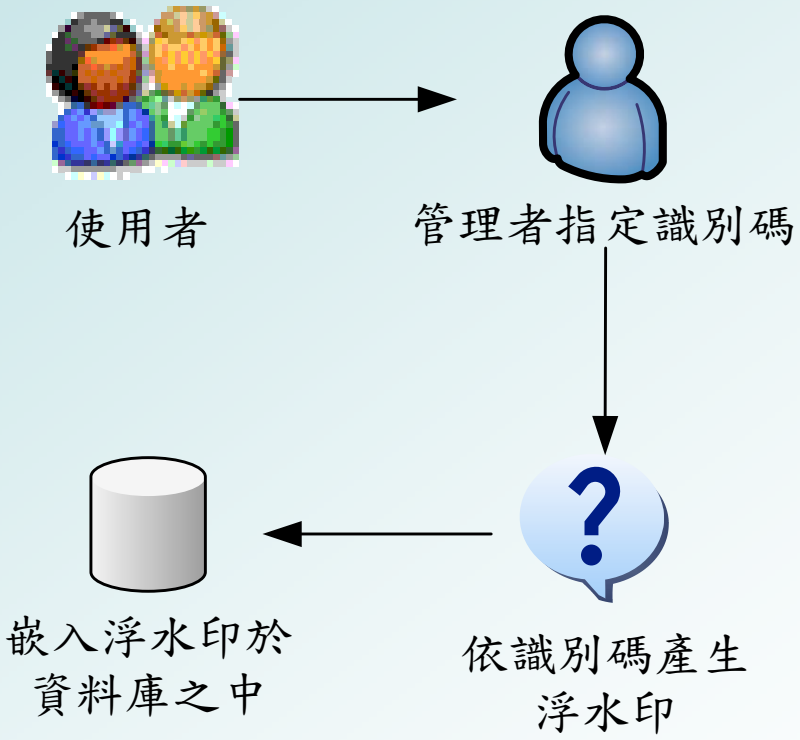


展頻技術應用於資料庫浮水印

- 基本概念
- 有原始資料庫可供參考之版權保護
- 無原始資料庫可供參考之版權保護

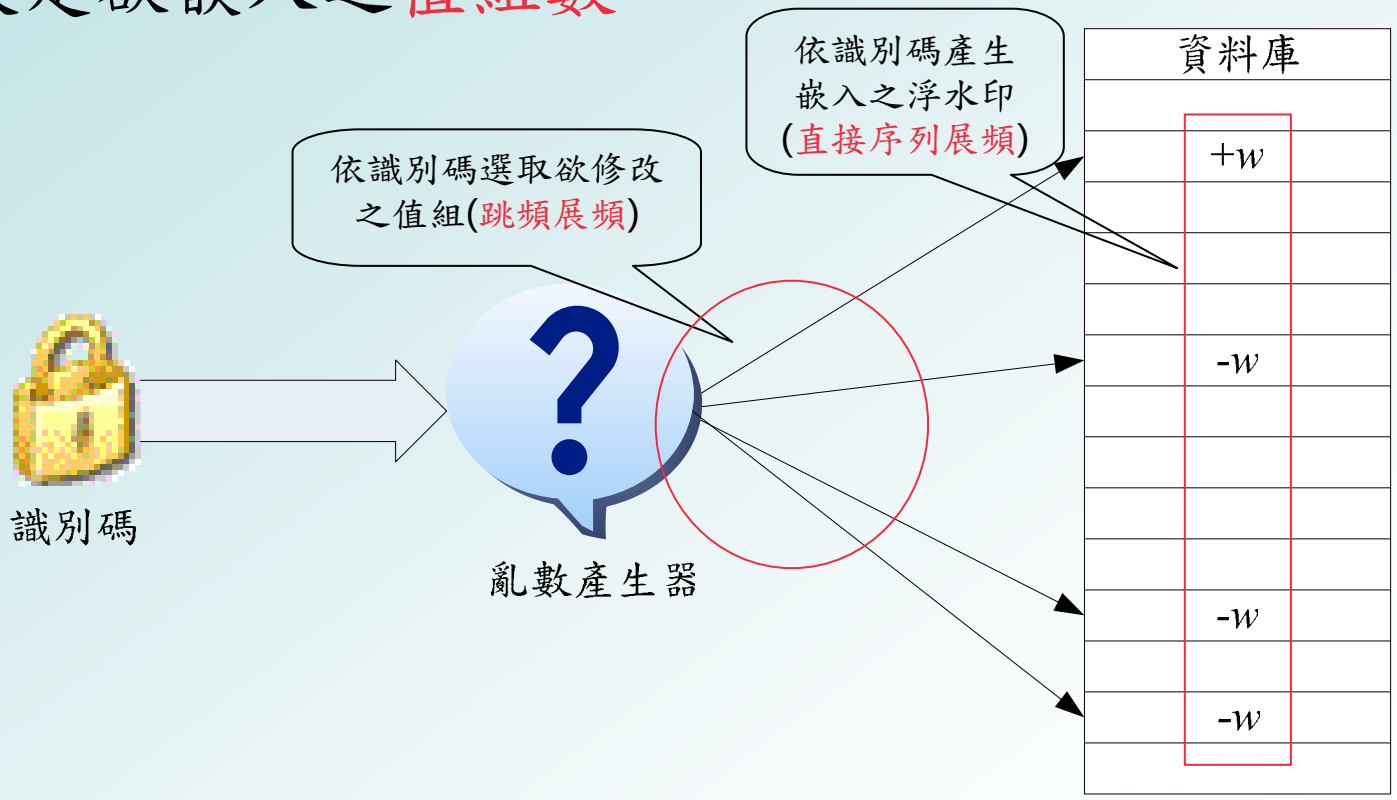
基本概念

- 嵌入使用者識別碼產生之浮水印訊號



基本概念(續)

- 選定識別碼
- 設定欲嵌入之值組數



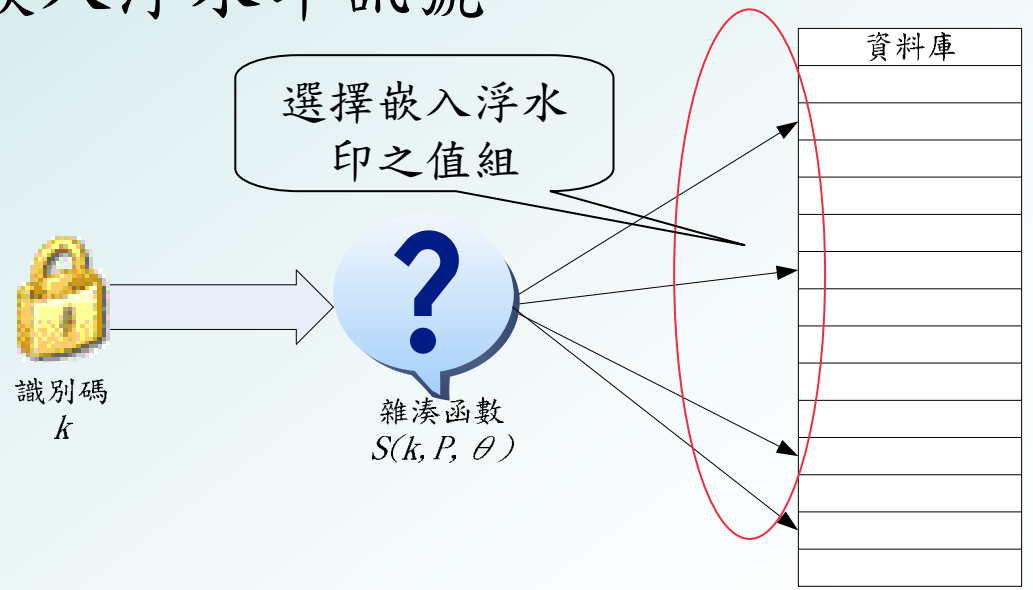
有原始資料庫可供參考之版權保護

- 浮水印嵌入
- 浮水印偵測
- 浮水印分析
- 實驗結果

浮水印嵌入

選擇嵌入值組

- 選定使用者識別碼k
- 選擇欲修改之值組比例θ
- 以各值組主鍵屬性之資料p與k及θ 決定此值組是否須嵌入浮水印訊號

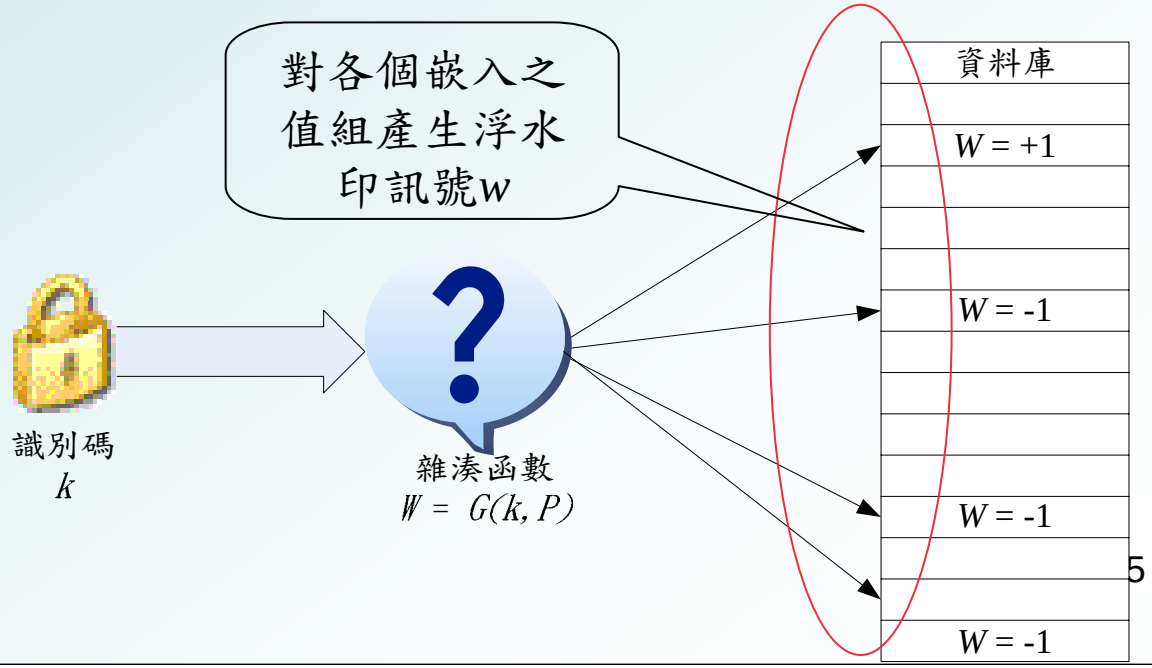


浮水印嵌入(續)

選擇嵌入值組之浮水印訊號

- 取識別碼k
- 主鍵屬性之資料p
- 產生嵌入該值組之浮水印訊號{+1, -1}

$$w = G(k, p)$$



浮水印嵌入(續)

嵌入浮水印訊號之方式

- 選擇可嵌入訊號之非主鍵屬性為目標屬性
- 設定其資料之可容忍誤差範圍 δ
- 依w對資料採取正向(+)或負向(-)之修改
- 若目標屬性屬性值為d，嵌入浮水印後之資料

$$\hat{d} = d + \alpha \times \delta \times w$$

資料之修改量

浮水印偵測

- 以原始資料庫與欲偵測資料庫之主鍵屬性進行交集(Intersection)之運算
- 以原始資料庫的主鍵屬性資料p與識別碼k及修改之值組比例θ，做為判斷值組是否有嵌入浮水印之依據

$$S(p,k,\theta)=\begin{cases} true & watermark\ embedded \\ flase & skip\ this\ record \end{cases}$$

浮水印偵測(續)

若值組T有嵌入之浮水印

- 利用原始資料庫之主鍵屬性資料p_T及原選用之識別碼k，計算出原先應嵌入T之浮水印訊號

$$w_T = G(k, p_T)$$

- 以被偵測資料屬性值與原資料屬性值相減擷取出實際嵌入之浮水印訊號

$$w'_T = \begin{cases} +1 & if\ \Delta' = d' - d \geq 0 \\ -1 & if\ \Delta' = d' - d < 0 \end{cases}$$

浮水印偵測 (續)

計算兩浮水印訊號之相關性

$$C_T = w_T \times w'_T$$

- 相關性計算結果

C_T	$w_T = +1$	$w_T = -1$
$w'_T = +1$	+1	-1
$w'_T = -1$	-1	+1

浮水印偵測 (續)

若兩浮水印訊號序列完全相同

- 累積m筆浮水印訊號之結果

$$C = \sum_{i=1}^{m} C_i = \sum_{i=1}^{m} w_i \times w_i' \approx m$$

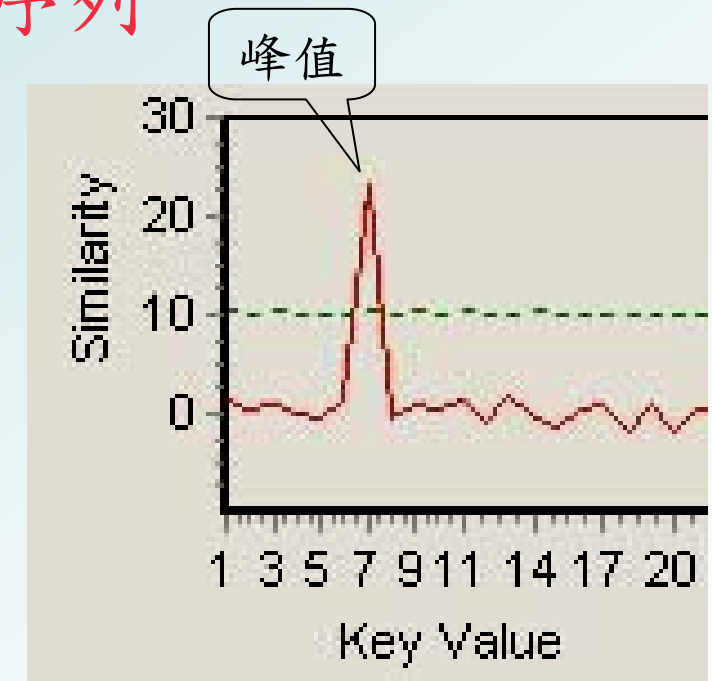
- 利用Cox等人所提出之相似度公式

$$Sim = \frac{\sum_{i=1}^{m} w_i \times w_i'}{\sqrt{\sum_{i=1}^{m} w_i' \times w_i'}} = \frac{C}{\sqrt{m}}$$

浮水印分析

比對不同識別碼*k*之相似度結果

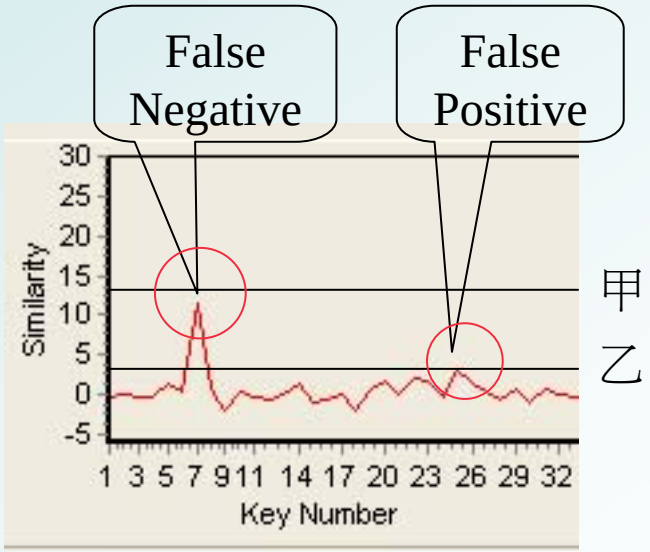
- 相似度若出現峰值，則視為含有某識別碼之浮水印訊號序列



浮水印分析 (續)

判定峰值之門檻值設定不易

- 設定太高，可能會產生false negative之問題
- 設定太低，可能會產生false positive之問題



浮水印分析 (續)

以Pearson相關係數進行偵測

- 比對原始資料修改之量 Δ
  與被偵測資料和原資料之差 Δ'

$$\Delta' = d' - d$$

- Pearson相關係數

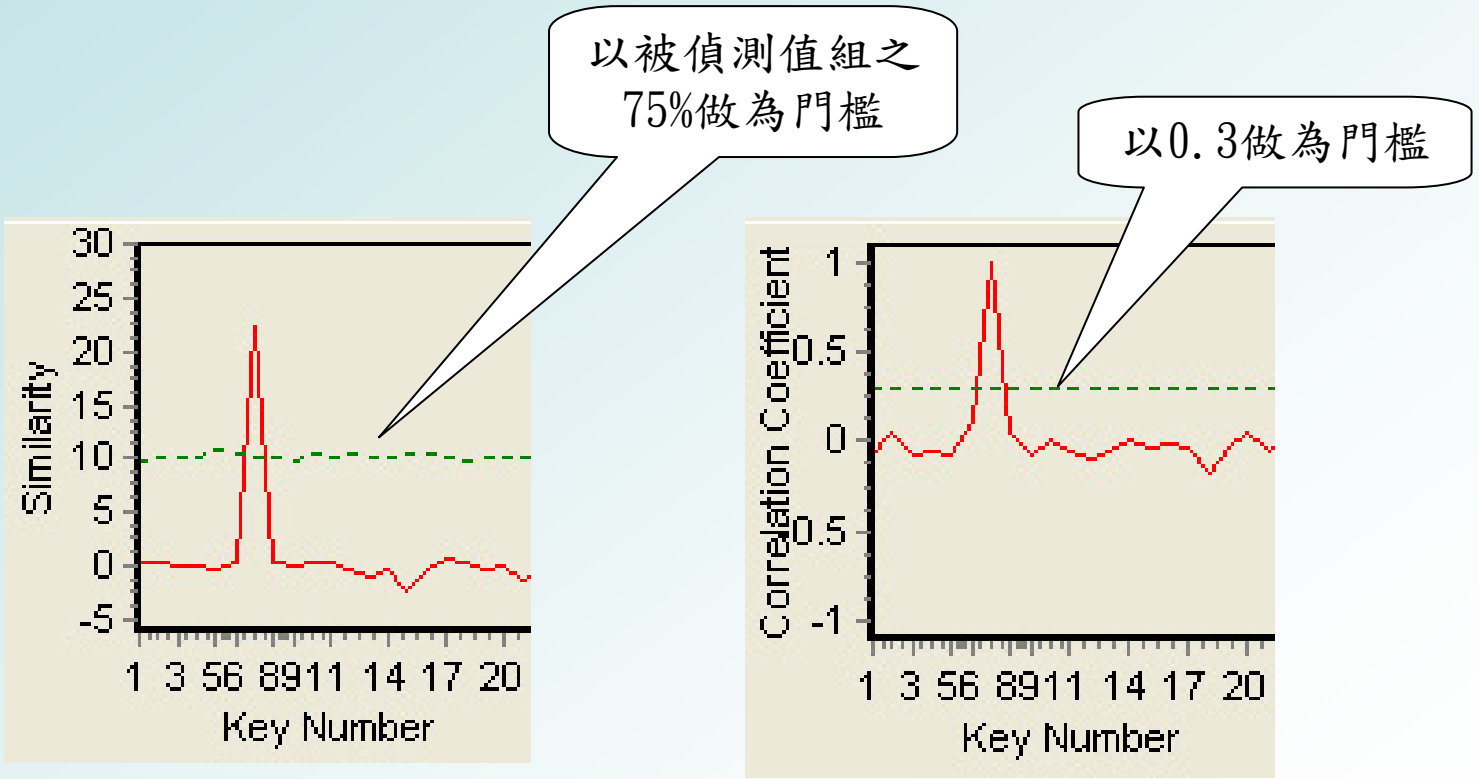
$$\rho = \frac{\sum_{i=1}^{m} \Delta_i \Delta_i' - \left(\sum_{i=1}^{m} \Delta_i \sum_{i=1}^{m} \Delta_i'\right) \Big/ m}{\sqrt{\sum_{i=1}^{m} \Delta_i^2 - \left(\sum_{i=1}^{m} \Delta_i\right)^2 \Big/ m} \sqrt{\sum_{i=1}^{m} \Delta_i'^2 - \left(\sum_{i=1}^{m} \Delta_i'\right)^2 \Big/ m}}$$

浮水印分析 (續)

- 在統計學中
  - $|\rho| \geq 0.7$ 兩者為高度相關
  - $0.7 > |\rho| \geq 0.3$ 兩者為低度相關
  - $0.3 > |\rho| \geq 0$ 兩者為不相關

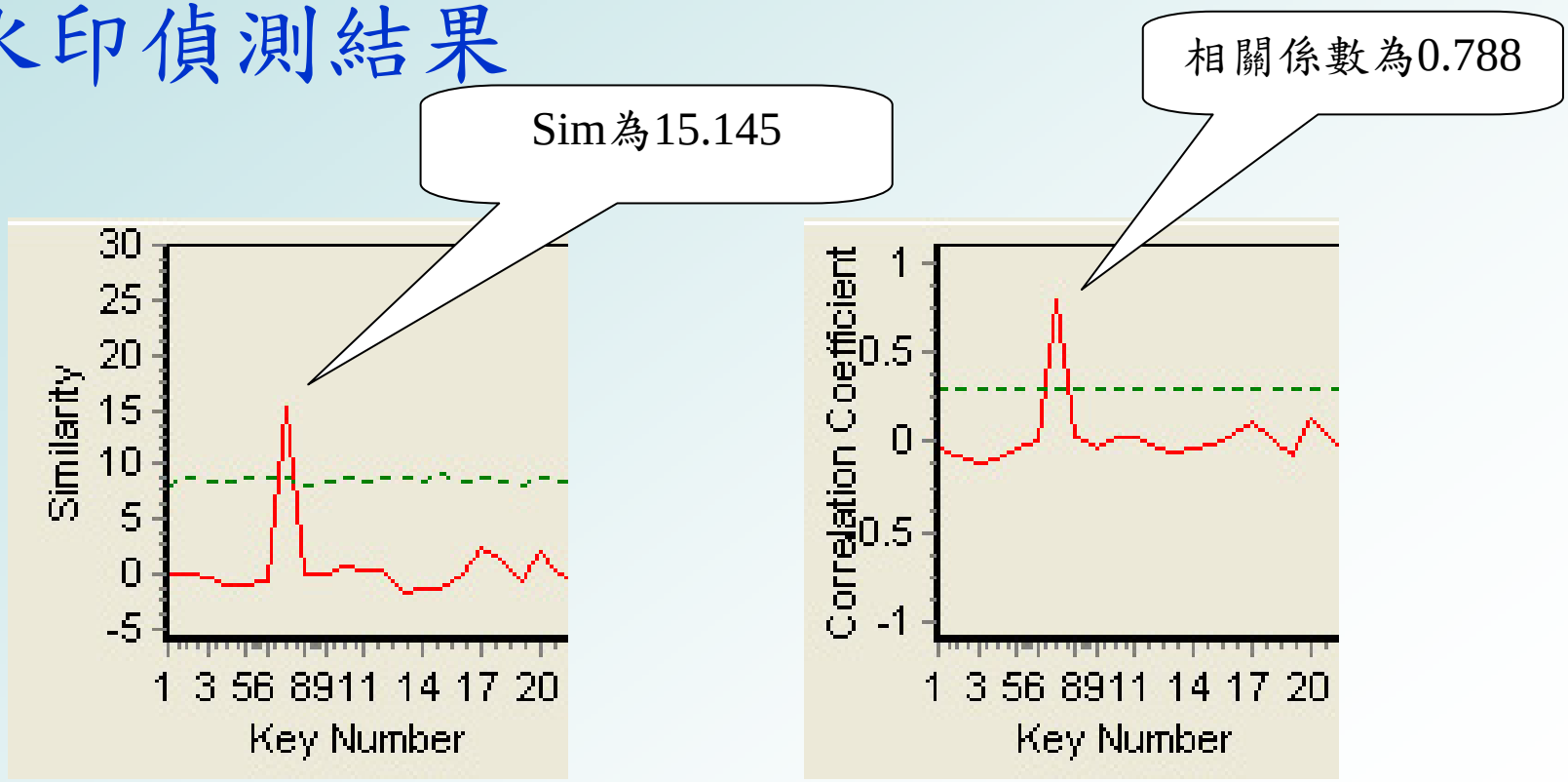
實驗結果

- 在10000筆值組之資料庫中嵌入浮水印後之浮水印偵測結果



實驗結果(續)

- 資料庫被刪除3000筆、新增3000筆、修改3000筆之連續攻擊後之浮水印偵測結果



無原始資料庫可供參考之版權保護

- 浮水印嵌入
- 浮水印偵測
- 浮水印分析
- 實驗結果

浮水印嵌入

部份步驟與有原始資料庫可供參考 相同

- 選擇嵌入值組
- 選擇嵌入值組之浮水印訊號

浮水印嵌入(續)

嵌入浮水印訊號之方式

- 選擇可嵌入訊號之非主鍵屬性
- 令 δ 為其資料之可忍受誤差範圍
- 假設原始資料為d，嵌入浮水印後之資料為 d'須符合

$$d' = \begin{cases} d + \delta/2 & w = 1 \; and \; (d \bmod \delta) < \delta/2 \\ d - \delta/2 & w = -1 \; and \; (d \bmod \delta) \geq \delta/2 \\ d & else. \end{cases}$$

之條件

浮水印偵測

- 以被偵測資料庫之主鍵屬性資料p^*及原選用之識別碼k及嵌入比例θ，做為判斷值組是否有嵌入浮水印之依據

$$S(p^*,k,\theta)=\begin{cases} true & watermark\ embedded \\ flase & skip\ this\ record \end{cases}$$

浮水印偵測(續)

若值組T有嵌入之浮水印

- 利用被偵測資料庫之主鍵屬性資料p_T*及原選用之識別碼k，計算出原先應嵌入T之浮水印訊號

$$w'_T = G(k, p^*_T) = \{+1, -1\}$$

- 由資料值擷取實際嵌入之浮水印訊號

$$w^*_T = \begin{cases} +1 & if\ d^*_T \bmod \delta \geq \delta/2 \\ -1 & if\ d^*_T \bmod \delta < \delta/2 \end{cases}$$

浮水印偵測(續)

- 計算兩浮水印訊號之相關性 $C_T = w_T^* \times w_T' = \{+1, -1\}$
- 累積m筆值組浮水印序列偵測之結果

$$C = \sum_{i=1}^{m} C_i$$

- 利用Cox等所提出之相似度公式

$$Sim = \frac{\sum_{i=1}^{m} C_i}{\sqrt{\sum_{i=1}^{m} (w_i^* \times w_i^*)}}$$

浮水印分析

- 在無嵌入識別碼之浮水印的情況下

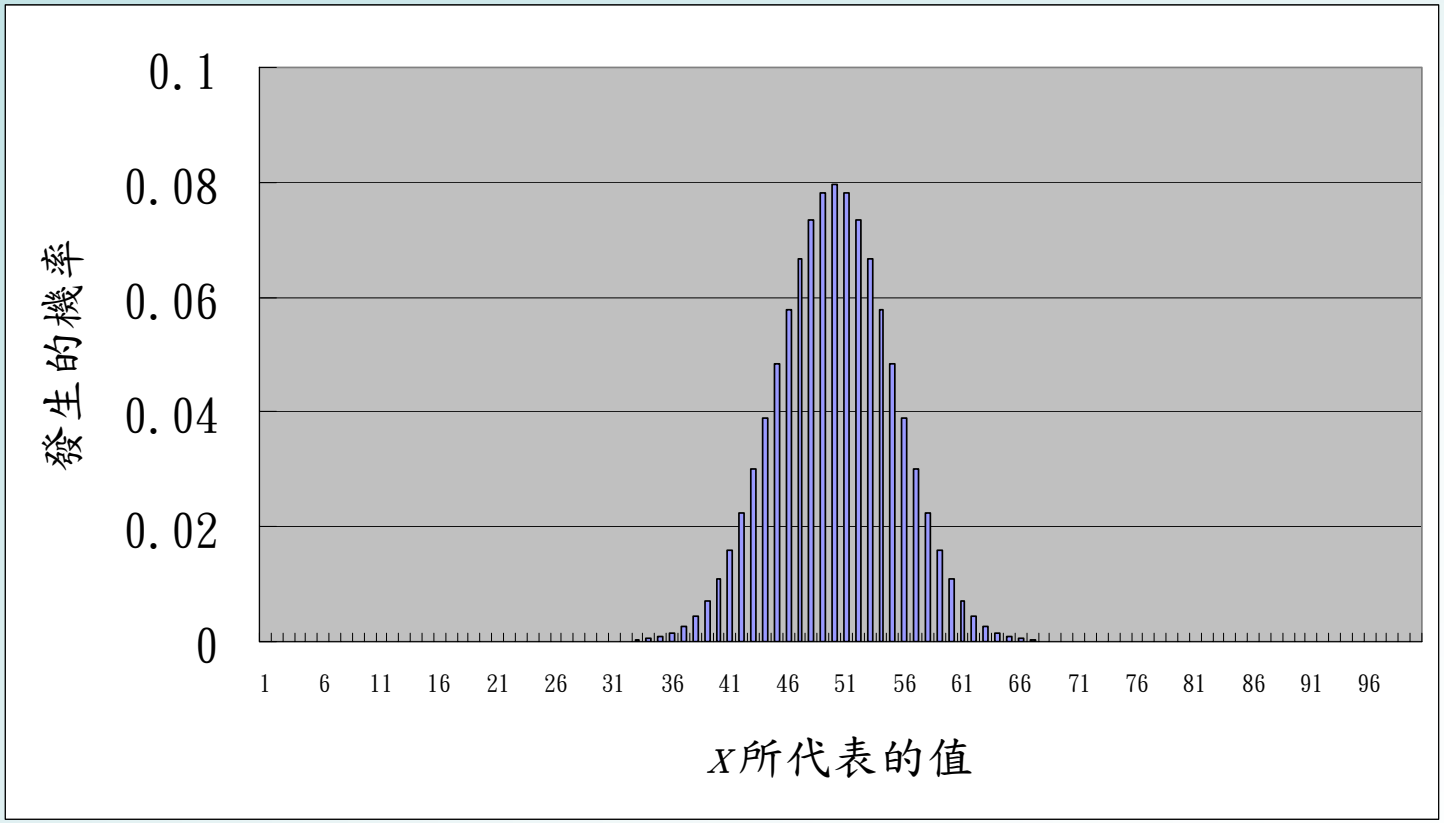
$$C_T = \{+1, -1\}$$

- C_T的結果為一次機率1/2的柏努利試行(Bernoulli Trail)
- m筆值組偵測結果的情況為二項分配(Binomial Distribution)，x為浮水印比對相符之值組數。

$$b(x; m, \frac{1}{2})$$

浮水印分析(續)

- 以m=100為例，$b(x;100,\frac{1}{2})$ 的機率分佈圖

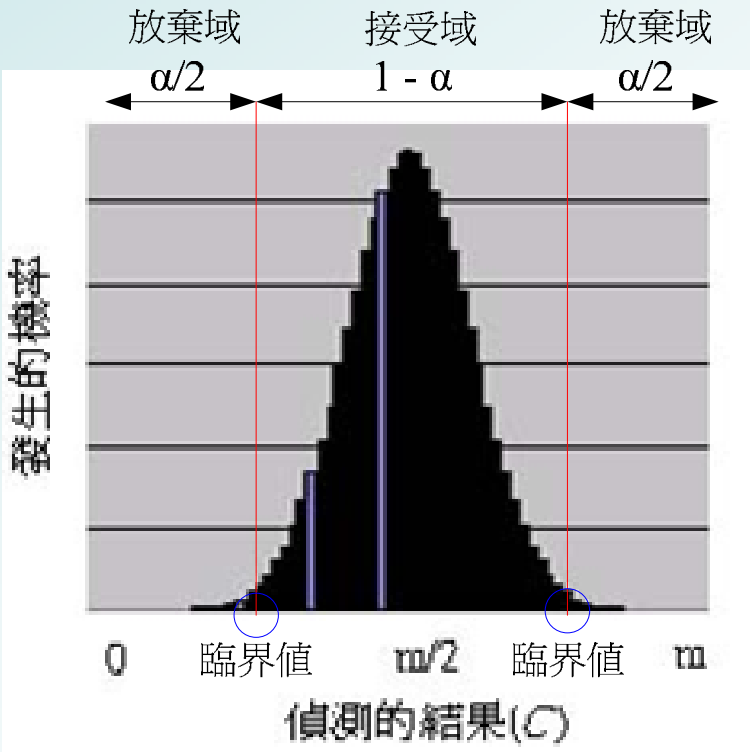


浮水印分析(續)

- 以假設檢定的方法判別浮水印訊號

$$\begin{cases} H_0 : \text{不是浮水印訊號} \\ H_1 : \text{是浮水印訊號} \end{cases}$$

- α 為信心水準



浮水印分析(續)

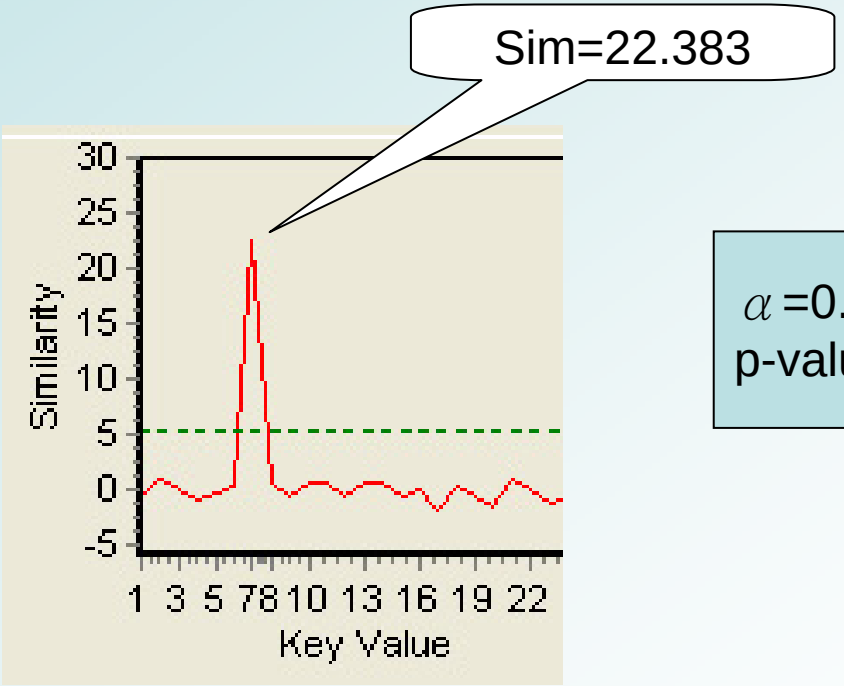
- 計算*p-value*以了解 false positive之機率

$$p-value=\sum_{x=n}^{m} b(x;m,\tfrac{1}{2})$$

$$=\sum_{x=n}^{m} C_x^m (\tfrac{1}{2})^x (1-\tfrac{1}{2})^{m-x} = \sum_{x=n}^{m} C_x^m (\tfrac{1}{2})^m$$

實驗結果

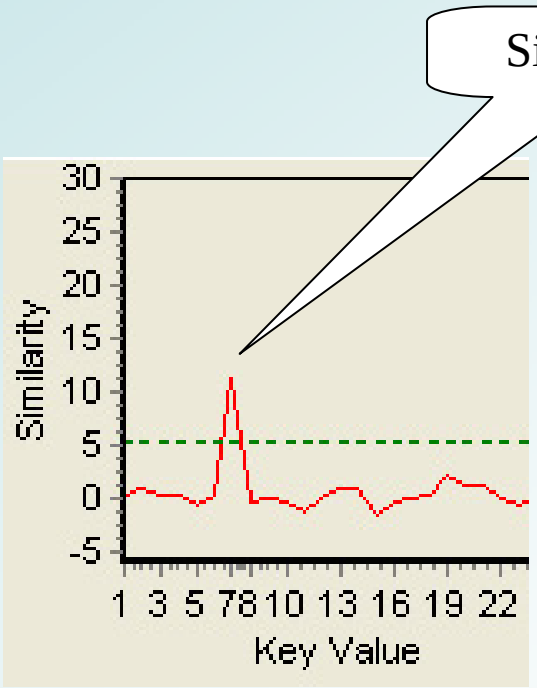
- 在10000筆值組之資料庫中嵌入浮水印後之浮水印偵測結果



α =0.0000001
p-value=3.05*10^{-151}

實驗結果(續)

- 資料庫被刪除3000筆、新增3000筆、修改3000筆之連續攻擊後之浮水印偵測結果



α =0.0000001
p-value=$2.42*10^{-30}$

結論

- 參考原始資料庫之版權保護方法
  - 優點: 可完全抵抗刪除和新增資料之攻擊
  - 缺點: 須保留嵌入浮水印時之原始資料
- 不須參考原始資料庫之版權保護方法
  - 優點: 不須保留原始資料
  - 缺點: 刪除和新增資料會產生輕微影響