國立臺灣大學工學院工程科學及海洋工程學系

碩士論文

Department of Engineering Science and Ocean Engineering

College of Engineering

National Taiwan University

Master's Thesis

隱私、效用與公平：基於差分隱私與 k-匿名的混合隱私保護

Privacy, Utility and Fairness: A Hybrid Privacy Protection with Differential Privacy and k-Anonymity

張凱惇

Kai-Tun Chang

指導教授：黃乾綱 博士

Advisor: Chien-Kang Huang, Ph.D.

中華民國 115 年 2 月

February, 2026

doi:10.6342/NTU202600729

摘要

在數位時代，資料分析能帶來有價值的洞察，但也可能侵犯個人隱私。傳統隱私保護技術面臨兩難：k-匿名雖直觀易懂，但難以抵禦具有背景知識的攻擊；差分隱私雖提供嚴謹的數學保障，但可能過度降低資料效用性。此外，現有研究多聚焦於隱私與效用的雙維度平衡，較少關注隱私保護機制對不同社會群體的差異性影響。由於資料收集過程中的固有不平衡，隱私保護機制可能無意中放大這種不平衡，對少數群體造成不成比例的負面影響，進而引發公平性問題。

針對此挑戰，本研究提出一套創新的三維隱私分析框架，目標在於：（1）整合 k-匿名與差分隱私的互補優勢；（2）量化隱私參數對隱私保護性、資料效用性與資料公平性的影響；（3）尋找三個維度間的最佳平衡點。

本研究採用 Adult 資料集進行實驗驗證。首先，透過泛化處理建立目標 k=5 的匿名群組；其次，對等價類計數應用拉普拉斯機制，測試 11 個不同的隱私參數 ε 值（0.1 至 10.0）；最後，建立標準化評分體系，以總變異距離（TVD）衡量資訊損失，以均等勝算差異評估公平性，並透過權重敏感度分析驗證參數選擇的穩健性。

實驗結果顯示：（1）在均衡權重配置（0.4/0.3/0.3）下，ε=1.0 獲得最高整合評分（0.764），在隱私保護性（評分 0.800，k=5）、資料效用性（評分 0.899，TVD=0.004）與資料公平性（評分 0.582）三個維度間達到最佳平衡；（2）與純 k-匿名或純差分隱私方法相比，本研究的混合機制在種族公平性方面改善約 20.1%，同時維持 99.09% 的下游應用準確率保持率；（3）權重敏感度分析證實 ε=1.0 在多數應用情境下表現穩健。本研究不僅整合了 k-匿名的直觀性與差分隱私的理論保障，更首次將公平性系統性地納入隱私保護評估，為需要兼顧多重目標的隱私保護應用提供了全面且靈活的決策支援框架。

關鍵字：差分隱私、k-匿名、資料隱私、隱私保護機制、均等勝算

doi:10.6342/NTU202600729

ABSTRACT

In the digital era, data analysis provides valuable insights but may also lead to privacy violations. Traditional privacy protection techniques face a dilemma: k-anonymity, while intuitive and easy to understand, struggles to defend against attackers with background knowledge; differential privacy offers rigorous mathematical guarantees but may excessively reduce data utility. Moreover, existing research primarily focuses on the privacy-utility trade-off, with limited attention to the differential impacts of privacy mechanisms across social groups. Due to inherent imbalances in data collection, privacy protection mechanisms may inadvertently amplify these disparities, causing disproportionate negative effects on minority groups and raising fairness concerns.

To address these challenges, this study proposes an innovative three-dimensional privacy analysis framework aimed at: (1) integrating the complementary strengths of k-anonymity and differential privacy; (2) quantifying the impact of privacy parameters on privacy protection, data utility, and data fairness; and (3) identifying the optimal balance among these three dimensions.

This research employs the Adult dataset for experimental validation. First, generalization processing establishes anonymization groups with a target k=5. Second, the Laplace mechanism is applied to equivalence class counts, testing 11 different privacy parameters ε (ranging from 0.1 to 10.0). Finally, a standardized scoring system is

doi:10.6342/NTU202600729

established, using Total Variation Distance (TVD) to measure information loss, Equalized Odds difference to assess fairness, and conducting weight sensitivity analysis to verify the robustness of parameter selection.

Experimental results demonstrate: (1) Under balanced weight configuration (0.4/0.3/0.3), ε =1.0 achieves the highest integrated score (0.764), attaining optimal balance across privacy protection (score 0.800, k=5), data utility (score 0.899, TVD=0.004), and data fairness (score 0.582); (2) Compared with pure k-anonymity or pure differential privacy methods, the proposed hybrid mechanism improves race fairness by approximately 20.1% while maintaining 99.09% accuracy retention in downstream applications; (3) Weight sensitivity analysis confirms that ε=1.0 demonstrates robust performance across most application scenarios. This research not only integrates the intuitiveness of *k*-anonymity with the theoretical guarantees of differential privacy, but also systematically incorporates fairness into privacy protection evaluation for the first time, systematically incorporates fairness into privacy protection evaluation, providing a comprehensive and flexible decision support framework for privacy protection applications requiring multiple objectives.

Keywords: Differential privacy, *k*-anonymity, Data privacy, Privacy protection mechanism, Equalized-odd

doi:10.6342/NTU202600729

目次

doi:10.6342/NTU202600729

doi:10.6342/NTU202600729

doi:10.6342/NTU202600729

圖次

doi:10.6342/NTU202600729

表次

doi:10.6342/NTU202600729

第一章　　緒論

1.1　研究背景

隨著數位時代的快速發展，資料已成為各領域中的關鍵資源。組織和機構收集大量資料進行分析，來獲得重要的參考資料、預測趨勢並協助決策。然而，這些資料往往包含個人敏感資訊，若未經適當隱私保護機制便被公開或分享，可能導致嚴重的隱私侵害問題。常見的隱私保護機制一般使用單一技術，但在面對複雜的隱私威脅時，單一隱私保護技術在面對複雜隱私威脅時存在局限性，且可能對不同社會群體產生差異性影響。如 2006 年 Netflix 宣稱不含使用者資訊之資料集遭有心人士與其它外部資料進行比對及連結，導致使用者隱私外洩；2018 年美國人口普查局研究小組亦發現，已經過傳統隱私保護處理之人口普查資料，仍有高達 46%的內容可被部分還原(數位發展部，2024)[1]。此外，因收集資料的不平衡將影響不同社會群體間的資料公平性，可能造成不同社會群體間的歧視問題。因此，有必要發展一個可以整合不同隱私保護技術優勢、且能在隱私保護性、資料效用性與資料公平性之間取得平衡的混合機制。

為確保專有名詞使用的一致性與明確性，本研究將相關術語及其定義整理於附錄詞彙表。

1.2　研究動機與目的

本研究的主要動機源於當前隱私保護技術在實際應用中面臨的三大挑戰：

(1) 單一隱私保護技術的局限性。Fung 等人(2010) [32]在其對隱私保護資料發布的綜述中指出，面對日益複雜的攻擊模型與多樣化的資料型態，單一的隱私保護技術存在明顯的防護能力局限。以 k-匿名為例，雖然其能確保群組層級的隱私保護，但難以抵禦具有背景知識的攻擊；差分隱私雖提供嚴謹的理論保障，但在實際應用中可能過度降低資料效用性。

doi:10.6342/NTU202600729

(2) 隱私保護與資料效用之間的平衡難題。如何在提供足夠隱私保護的同時，最大化保留資料的分析價值，是隱私保護領域的核心挑戰(Fung et al., 2010)[32]。差分隱私的實施需要謹慎選擇隱私參數（以隱私參數 ε 值表示）；過大的 ε 值可能無法提供足夠的隱私保障，過小的 ε 值雖然提供強健的隱私保護，但可能會過度降低資料效用性。

(3) 隱私保護機制對不同社會群體的公平性影響。由於資料收集過程中固有的不平衡，傳統隱私保護機制可能無意中放大這種不平衡，導致對少數群體的不成比例負面影響。Bagdasaryan 等人(2023) [19]的研究顯示，相同的隱私參數設置可能對少數群體造成更大的效用損失，引發公平性關切。這種現象主要源於少數群體在數據中的代表性較低，相同程度的雜訊對其統計特性影響更大。

現有研究多著重於單一隱私保護機制的優化，較少探索如何有效整合 k-匿名和差分隱私的互補優勢，並同時解決社會群體間的資料公平性問題。因此，設計能同時考量隱私保護和資料公平性的機制具有重要的社會意義。基於以上挑戰，本研究提出一套創新的隱私分析框架，此創新框架可以建立整合 k-匿名與差分隱私的混合機制，並首次將均等勝算指標納入隱私保護評估框架，透過量化差分隱私參數(ε 值)的影響，尋找在隱私保護性與資料效用性之間的最佳平衡點，並進一步評估隱私保護機制對不同社會群體的差異性影響，確保隱私保護過程中的資料公平性。透過決策支援系統協助使用者選擇最適合其需求的隱私參數配置。

本研究提供了全面性的隱私保護評估，結合 k-匿名技術、差分隱私技術與公平性指標。在差分隱私處理達到預設隱私保護性（k-匿名性目標）的同時，優化其資料效用性（資訊損失限制），並系統性地分析了隱私保護技術對不同社會群體的差異性，將隱私保護對資料公平性（公平性分數）的影響考量納入隱私參數的選擇過程中。經過對不同隱私參數的分析比較，我們提出參數選擇的建議方法，開發參數選擇機制，以尋找適當的差分隱私參數，實現隱私保護與公平性的平衡。最後，

doi:10.6342/NTU202600729

我們設計了直觀易懂的視覺表示，幫助使用者理解不同隱私參數的影響並做出合適的選擇。

本研究採用隱私分析常用的 Adult 資料集進行實驗。為確保實驗公平性，研究過程首先將仿照前人研究，透過資料前處理識別準識別符，接著結合差分隱私技術添加適當雜訊(對應差分隱私參數 ε 值)，使處理後的資料達到預設的 k-匿名性要求。為評估處理結果的資料效用性與資料公平性，我們使用總變異距離(Total Variation Distance, TVD)量化資訊損失程度，並透過均等勝算(Equalized Odds)差異值衡量不同社會群體間的公平性。初步實驗顯示本研究可提供一個實用的隱私保護分析框架，針對需要同時考量隱私保護性、資料效用性和資料公平性的應用場景，如醫療資料分析、個人化推薦系統或人口普查資料發布等，提供新的發展方向。

doi:10.6342/NTU202600729

第二章　　文獻探討

2.1 隱私保護技術概述

隨著數位時代的快速發展，資料已成為各領域中的關鍵資源。組織和機構收集大量資料進行分析，來獲得重要的參考資料、預測趨勢並協助決策。然而，這些資料往往包含個人敏感資訊，若未經適當隱私保護機制便被公開或分享，可能導致嚴重的隱私侵害問題。

從資料生命週期的角度來看，資料可分為三種狀態：靜態儲存（at rest）、傳輸中（in transit）及使用中（in use）的資料，如圖 2.1 所示（資料來源：數位發展部，2024 [1] ），各狀態有對應的保護方式。包含用於『靜態儲存』及『傳輸中』時的假名化技術，例如：加密，還有為『使用中』資料提供保護的匿名化技術， 例如本研究探討的 *k*-匿名技術。

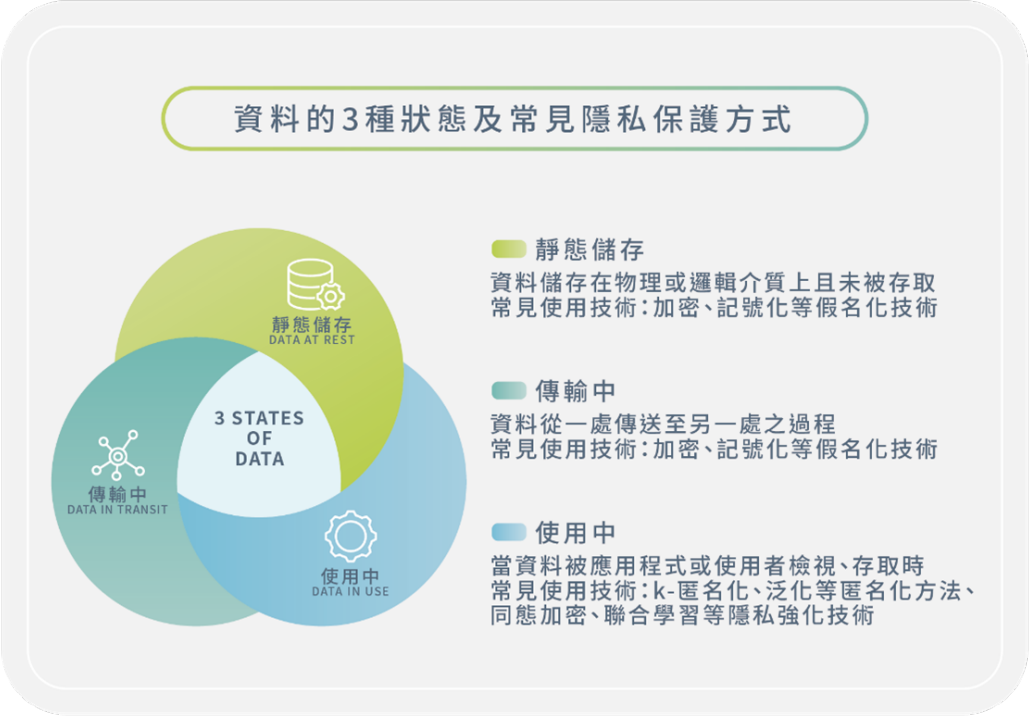


圖 2.1　資料的三種狀態及常見隱私保護方式

從應用目的的角度來看，隱私保護研究可分為兩大主要方向：一是隱私保護資料發布 (Privacy-Preserving Data Publishing)，關注如何安全地公開或分享資料集；

doi:10.6342/NTU202600729

二是隱私保護分析 (Privacy Protection in Big Data Analytics)，關注如何在資料分析過程中防禦各種攻擊。

早期的隱私保護方法如資料移除或遮蔽等技術，Sweeney[2]指出，這類方法雖能防止人類直接識別，但會因為受到資料挖掘技術的攻擊，導致隱私洩漏。因此，更先進的隱私保護機制如 k-匿名和差分隱私應運而生，成為現今研究的主流方向。然而，實際案例顯示單一技術仍存在局限性：2006 年 Netflix 宣稱不含使用者資訊之 資料集遭有心人士與其它外部資料進行比對及連結，導致使用者隱私外洩；2018 年 美國人口普查局研究小組亦發現，已經過傳統隱私保護處理之人口普查資料，仍有高達 46%的內容可被部分還原(數位發展部，2024)[1]。

本研究聚焦於資料發布情境，屬於「使用中資料」的保護範疇，採用 *k*-匿名作為基礎匿名化技術。然而，考量到上述案例顯示單純匿名化仍可能遭受攻擊，本研究進一步結合差分隱私機制提供額外保護層，形成混合隱私保護機制。此外，因收集資料的不平衡將影響不同社會群體間的資料公平性，可能造成不同社會群體間的歧視問題。因此，有必要發展一個可以整合不同隱私保護技術優勢、且能在隱私保護性、資料效用性與資料公平性之間取得平衡的混合機制。

2.1.1 隱私強化技術分類與發展

現代隱私保護技術的分類方式依據不同觀點而有所差異。從技術原理的角度，可分為匿名化技術、擾動技術、密碼學技術和合成技術等四大類[35]；從資料處理階段的角度，數位發展部將隱私強化技術分為三大應用類別[1]。本研究採用後者的分類框架，以符合我國隱私保護實務指引。

一、改變原始資料以減少或移除個人識別資訊：

此類技術透過降低資料精確度或移除識別特徵，減弱個人與資料間的關聯性，同時保留資料的統計特性以維持分析價值。主要技術包括：

(1) 傳統去識別化技術：包含抑制/編修（Suppression/Redaction）、遮罩（Masking）、泛化（Generalization）等方法。其中，*k*-匿名化技術結合多種方

doi:10.6342/NTU202600729

法，確保任一記錄至少與 k-1 筆資料在準識別符上相同，使重新識別機率不超過 1/k[2]。

(2) 新興擾動技術：差分隱私（Differential Privacy）透過添加校準雜訊，使攻擊者無法判斷特定個體是否存在於資料集中，提供具數學證明的隱私保證[4,5]。合成資料（Synthetic Data）則以統計模型生成人工資料，保留原始資料的整體統計特性，但不包含真實個體資訊。

二、對原始資料進行保密性處理後進行操作：

此類技術允許資料在加密或受保護狀態下進行運算，確保資料在處理過程中不會以明文形式暴露。主要技術包括同態加密（Homomorphic Encryption）[21]、安全多方運算（Secure Multiparty Computation）[22]、零知識證明（Zero-Knowledge Proof）等。

三、設計系統運算機制達到資料保護效果：

此類技術透過系統架構設計，在最小化資料共享的前提下完成協作任務。主要技術包括聯合學習（Federated Learning）、可信執行環境（Trusted Execution Environment）等。

數位發展部指出，各類技術在隱私保護強度、計算效率、實用性等方面存在顯著差異[1]。密碼學技術雖提供強隱私保證，但計算複雜度高，難以應用於大規模資料發布；合成技術面臨模型訓練複雜性和生成資料品質控制挑戰。相較之下，匿名化與擾動技術的結合既能提供理論隱私保證，又保持合理的計算效率，適合實際部署應用。這為本研究選擇 *k*-匿名性與差分隱私的混合機制提供了理論支撐。具體而言， *k*-匿名技術具有直觀性和易理解性，差分隱私提供嚴格的數學隱私保證，兩者的結合既能滿足實務應用的直觀需求，又能提供理論上的隱私保障，同時保持合理的計算效率。

2.2 *k*-匿名技術

doi:10.6342/NTU202600729

k-匿名是由 Sweeney[2]於 2002 年提出的隱私保護模型，其目的在於解決資料去識別化處理後仍可能面臨的重新識別風險。*k*-匿名的核心理念建立在一個簡單而有力的原則上：確保任何識別個體的嘗試至少會得到 *k* 個無法區分的結果。正規地說，給定一個資料集 D，若對於 D 中的每一筆記錄 r，至少存在 k-1 筆其他記錄，其準識別符的值與 r 相同，則稱 D 滿足 *k*-匿名性。換言之，若資料集滿足 *k*-匿名性，則攻擊者無法將任何記錄與特定個體關聯的機率不會高於 1/k。

k-匿名的實現主要採用泛化(Generalization)和抑制(Suppression)兩種技術。泛化是將具體的屬性值替換為更一般的值，例如將精確年齡「25 歲」替換為年齡範圍「20-29 歲」。抑制則是完全刪除某些屬性值或整筆記錄，例如將特殊的郵遞區號以「*」替代。圖 2.2 是一個 *k*-匿名資料處理的範例，其中年齡、學歷被泛化處理。在原始資料中，40-50 歲且具有高中學歷的人群中有一名男性和一名女性，如果保留性別屬性，每個組合就只有一筆記錄，無法達到 *k*=2 的要求。因此，透過抑制性別屬性，將這兩筆記錄合併為一個等價類群組，從而達到 *k*=2 匿名。

這是實際應用 *k*-匿名時常見的情況：當僅靠泛化無法達到要求的 *k* 值時，會選擇性地抑制某些屬性值，以減少區分度並形成更大的等價類。也就是說，儘管 *k*-匿名提供了直觀且易於理解的隱私保護性指標，但它也存在明顯的局限性。若群組一定要明確達到要求的 *k* 值，而在群組化過程中又存在多種選擇性的問題，都會使得演算法在優化設計上較為複雜。

doi:10.6342/NTU202600729

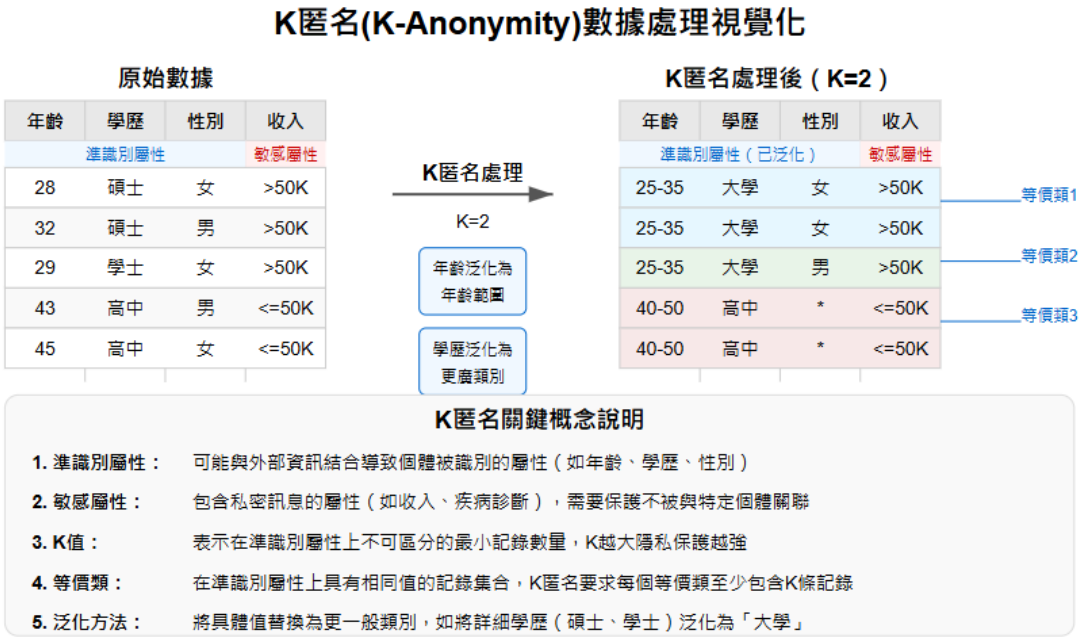


圖 2.2 一個 *k*-匿名資料處理範例

儘管存在上述實作上的複雜性，*k*-匿名技術仍具有重要優勢。其主要優勢在於直觀性和可解釋性。相較於複雜的密碼學方法， *k*-匿名提供了易於理解的隱私保護機制，使得資料使用者能夠清楚掌握資料的匿名化程度。此外，*k*-匿名技術計算複雜度較低，適合處理大規模資料集，在實務應用中具有良好的可操作性。為克服 *k*-匿名的局限性，研究者提出了 *l*-diversity[3]和 *t*-closeness[8]等改進方法。*l*-diversity 要求每個等價類中的敏感屬性至少有 *l* 個不同值，以防止同質化攻擊；而 *t*-closeness 則要求等價類中敏感屬性的分佈與整體分佈的距離不超過閾值 *t*，以防止偏斜攻擊。然而，這些方法仍然依賴泛化和抑制操作，在面對背景知識攻擊時仍存在脆弱性，且隨著隱私要求提高，資料效用損失會急劇增加。因此，本研究選擇 *k*-匿名作為基礎技術，並結合差分隱私提供額外的保護層。

儘管 *k*-匿名存在上述局限性，但其在群組隱私保護方面的有效性使其成為混合隱私保護機制的理想組成部分。*k*-匿名確保了任何準識別符組合至少對應 *k* 個記錄，為後續的差分隱私處理提供了穩定的群組基礎，形成「群組保護+記錄保護」

doi:10.6342/NTU202600729

的雙重防護機制。這種組合既保留了 *k*-匿名的直觀性和可操作性，又彌補了其在理論隱私保證上的不足。

2.3 差分隱私技術

有別於 *k*-匿名的群組化方法，差分隱私是由 Dwork [4]等人於 2006 年提出的一種有嚴格數學證明的隱私保護機制，可透過添加精心設計的隨機雜訊保護個體隱私。定義說明，令隱私參數 ε 為一正實數，Pr[ ]表示某個事件發生的機率，而 A 為一隨機演算法，若對所有相鄰的兩個資料集 D1 和 D2（即僅有一筆紀錄不同的資料集），符合下列不等式，使對於任何輸出集合 S，演算法 A 在相鄰資料集 D_1 上產生屬於 S 的輸出的機率 Pr，不會超過它在另一個相鄰資料集 D_2上產生同樣輸出的機率 Pr 乘以 exp(ε)倍，則稱該演算法 A 可以提供 *ε*-差分隱私：

$$\Pr[A(D_1) \in S] \leq \exp(\varepsilon) \cdot \Pr[A(D_2) \in S] \tag{2.1}$$

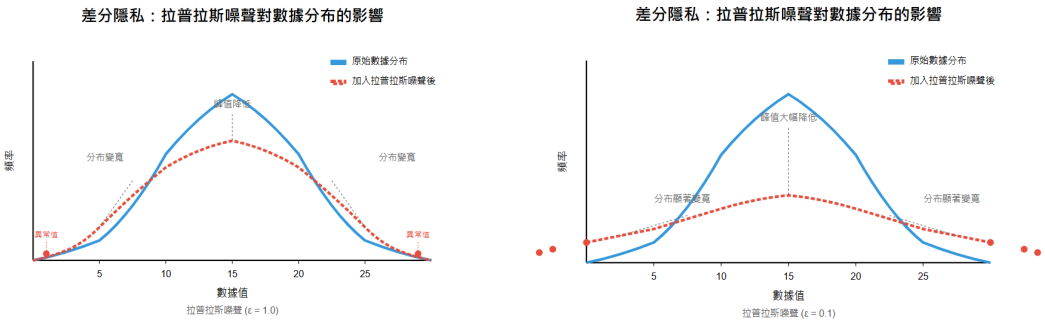
其中 ε 是隱私參數，對應所設計的隨機雜訊大小，以控制隱私保護的強度。ε 值越小，雜訊越大，隱私保護越強，但資料效用可能越低。

差分隱私相較於其他擾動技術具有獨特優勢。與傳統的隨機響應技術相比，差分隱私提供了更精確的隱私量化指標，能夠準確控制隱私洩露程度。與數據掩蔽技術相比，差分隱私具有可組合性，多次查詢的隱私損失可以精確累計。最重要的是，差分隱私提供了與背景知識無關的隱私保證，使攻擊者的先驗知識假設無效[4]。

現有差分隱私技術主要透過使用拉普拉斯(Laplace)機制或高斯(Gaussian)機制實現，拉普拉斯（高斯）機制會添加服從拉普拉斯（高斯）分佈的隨機雜訊，雜訊大小與敏感度和隱私參數相關。圖 2.3 與圖 2.4 比較差分隱私機制中兩種常用雜訊（拉普拉斯與高斯）在不同隱私參數（ε 值）下對資料分佈的影響。藍色曲線代表原始資料分佈，而紅色虛線（拉普拉斯）和紫色虛線（高斯）顯示了加入雜訊後的資料分佈。在 ε=1.0 的情況下，拉普拉斯雜訊使分佈變寬，峰值降低，並在尾部

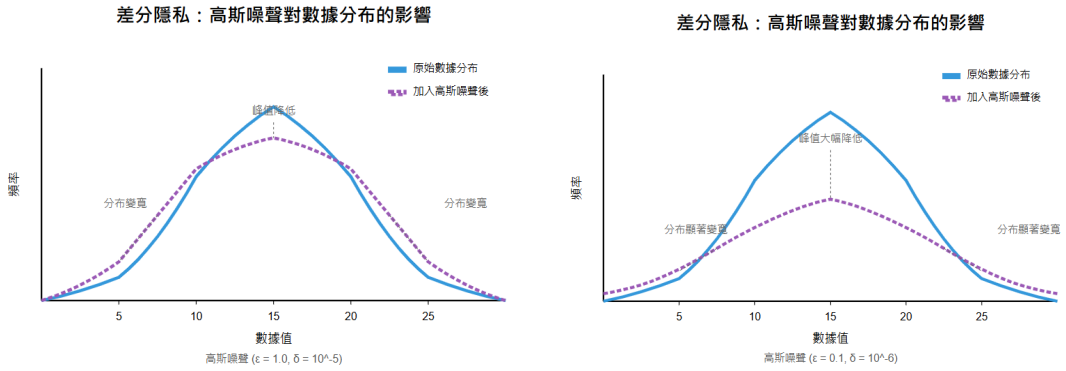
doi:10.6342/NTU202600729

產生較多極端值；而高斯雜訊雖然同樣使分佈變寬，但整體仍保持常態分佈的平滑。當隱私保護增強至 ε=0.1 時，雜訊效果顯著增強。從圖 2.3 與圖 2.4 的比較可以看出，拉普拉斯機制產生的分佈具有更尖銳的峰值和較重的尾部，而高斯機制則保持較為平滑的分佈形狀。這種差異在實際應用中會影響資料的統計特性：拉普拉斯機制可能產生更多極端值，而高斯機制則相對保持資料的整體分佈特徵。



(a) ε 值=1.0　　(b) ε 值=0.1

圖 2.3　拉普拉斯機制加噪範例



(a) ε 值=1.0　　(b) ε 值=0.1

圖 2.4　高斯加噪機制加噪範例

現有的擾動技術還包括隨機響應、數據掩蔽、同態雜訊等方法。隨機響應雖然概念簡單，但隱私保證較弱且難以處理連續型數據；數據掩蔽技術缺乏嚴格的理論基礎；同態雜訊雖能保持某些統計特性，但計算複雜度高。相較之下，差分隱私在理論嚴謹性、實作便利性和隱私保證強度之間達到良好平衡，適合與 k-匿名技術結合使用。

doi:10.6342/NTU202600729

近年來研究發現，差分隱私機制對不同人口群體可能產生差異性影響。Bagdasaryan 等人[19]的研究顯示，相同的隱私參數設置可能對少數群體造成不成比例的效用損失，引發公平性關切。這種現象主要源於少數群體在數據中的代表性較低，相同程度的雜訊對其統計特性影響更大。此發現突出了在設計隱私保護機制時，需要考慮不同群體間的公平性，而非僅關注整體的隱私保護效果。特別是在將差分隱私與其他隱私保護技術結合時，更需要系統性地評估混合機制對各群體的綜合影響，以確保隱私保護不會加劇既有的社會不平等。

由於差分隱私的運算相對簡單，近年來也從原始的集中式模型，擴展到本地差分隱私(Local Differential Privacy, LDP)[6]和聯邦差分隱私(Federated Differential Privacy, FDP)[7]等變體，展現了可組合性和對各種攻擊的抵抗優勢。Dwork 和 Roth [5]探討差分隱私的數學理論基礎和應用案例，顯示差分隱私可為資料提供隱私保護，同時保留了資料的分析價值。Google、Apple 和微軟等科技巨頭紛紛將差分隱私技術應用於實際產品中，使個人資料難以識別，同時保留整體統計特性。這使公司能夠收集有價值的使用者行為資料，而不會侵犯個人隱私。例如：Google 在 Chrome 瀏覽器中實施差分隱私收集使用統計[6]。Apple 在 iOS 中用於收集用戶行為模式而不洩露個人隱私，微軟則在 SQL Server 中提供差分隱私查詢功能。這些實際應用案例證明了差分隱私技術的實用性和可行性。

雖然差分隱私提供了強有力的隱私保證，但其添加的隨機雜訊可能影響資料的可用性，特別是對小型或稀疏的資料集。因此，如何結合其他隱私保護技術（如 *k*-匿名性）來達到更好的隱私保護與資料效用平衡，同時確保對不同群體的公平性，成為重要的研究方向。這也正是本研究探討混合隱私保護機制的核心動機。

2.4 混合隱私保護機制

隨著隱私威脅的日益複雜化，單一隱私保護技術的局限性愈發明顯。*k*-匿名雖直觀易懂但難以抵禦背景知識攻擊，差分隱私雖提供嚴謹保證但可能過度犧牲資料效用。為克服這些局限，混合隱私保護機制應運而生。其理論基礎建立在兩個關

doi:10.6342/NTU202600729

鍵原則上。第一，分層保護原則：不同隱私保護技術在資料處理的不同階段發揮作用，形成多層次防護；第二，互補增強原則：一種技術的優勢可彌補另一種技術的不足，創造協同效應。這種多層次、互補性的設計為複雜隱私保護情境提供了更為全面的解決方案，特別適用於需同時考量多種隱私要求的應用場景。

本研究選擇 *k*-匿名與差分隱私的組合，主要基於四個考量：(1)技術互補性強，*k*-匿名的直觀性彌補差分隱私的複雜性；(2)適合建立三維評估框架，兩種技術在隱私保護性、資料效用性、資料公平性各維度均有可觀測的影響；(3)實際應用廣泛，兩種技術均有豐富的產業應用案例；(4)理論基礎扎實，有利於建立嚴謹的數學分析框架。

在隱私保護技術的發展過程中，除了 *k*-匿名和差分隱私外，還有 *l*-diversity[3] 和 *t*-closeness 等技術。*l*-diversity 要求每個等價類中的敏感屬性至少有 *l* 個不同值，*t*-closeness 則要求等價類中敏感屬性的分佈與整體分佈的距離不超過閾值 *t*。

然而，現有混合機制研究存在三個主要不足：(1)公平性評估缺乏系統性量化方法；(2)參數選擇多依賴經驗而非理論指導；(3)缺乏統一的多維度評估框架。本節將分析近期研究進展、技術選擇理由，並明確本研究的創新貢獻。從技術發展脈絡來看，混合隱私保護機制經歷了多次重要演進。Dwork 和 Roth [5] 的研究奠定了差分隱私的理論基礎，而 Bayardo 和 Agrawal [11] 則深入探討了 *k*-匿名化的優化方法。在多階段隱私保護流程的發展中，研究者提出對不同類型的隱私威脅採取針對性保護措施，有效提升了機制的適應性與靈活性。

近期研究更加聚焦於混合機制在特定領域的應用與優化。Zhai 等人[8] 針對電力消費資料提出的混合隱私保護方法採用了資料分流策略：對數值型準識別符應用差分隱私，對類別型準識別符應用 *k*-匿名。此方法的創新之處在於根據資料類型採用不同策略，但其局限性在於尚未考量資料公平性問題。Lin [9]針對分佈式資料庫環境，提出了結合差分隱私和 *k*-匿名的框架。其設計了隱私預算分配策略，將有限的隱私預算分配給不同的查詢操作。實驗表明，與單獨應用的差分隱私和 *k*-匿

doi:10.6342/NTU202600729

名技術相比，該混合隱私保護方法可在維持資料效用性和隱私保護性之間取得平衡，但因其計算複雜度頗高，使得實用性上受限。此外，此方法並未考慮資料公平性問題。值得注意的是，Kobayashi 等人[10] 從理論層面深入探討了概率 *k*-匿名與差分隱私之間的數學關係，證明了在特定條件下，概率 *k*-匿名能夠近似實現差分隱私的保證。這項研究為混合機制提供了堅實的理論基礎，但尚未擴展至公平性維度的探討。

綜觀上述混合隱私保護機制的發展，儘管在技術整合與應用優化方面取得重要進展，但現有研究存在三個明顯缺口：首先，主要聚焦於隱私保護性與資料效用性的雙維度平衡，缺乏第三維度的系統性考量；其次，對隱私保護機制可能造成的群體差異性影響缺乏量化評估；第三，缺乏統一的多維度評估框架來指導實際應用中的參數選擇。表 2.1 呈現了不同隱私保護技術在三個關鍵維度上的表現比較。傳統 *k*-匿名方法雖然直觀易懂，但在隱私保護強度上相對有限；差分隱私雖提供嚴謹的數學保證，但往往以犧牲資料效用為代價。現有的混合方法在平衡隱私保護與資料效用方面取得了進展，但普遍缺乏對隱私保護機制公平性影響的明確評估。相較之下，本研究系統性地將均等勝算指標納入混合隱私保護機制的評估體系中。

表 2.1 *k*-匿名、差分隱私及其混合技術比較

方法	隱私保護性	資料效用性	資料公平性
k-匿名	中	中	未考慮
差分隱私	高	低	未考慮
混合技術	高	中	未明確評估
本研究	高	中-高	系統性均等勝算評估

為建立全面的隱私保護評估框架，除了傳統的隱私-效用平衡外，本研究特別關注隱私保護機制對不同社會群體的公平性影響，並將在下一節詳述相關的公平性度量標準，特別是均等勝算指標的選擇理由及其在隱私保護情境下的適用性。

doi:10.6342/NTU202600729

2.4.1 近期混合隱私保護機制研究進展

近年來，k-匿名與差分隱私的混合機制研究取得重要進展。Majeed & Hwang (2024) 提出了基於模式識別的混合隱私保護方案，通過機器學習技術識別資料中的模式友好準識別符，並將資料集劃分為隱私侵犯和非隱私侵犯分區[25]。該方法在非隱私侵犯分區對數值屬性應用寬鬆的隱私預算 ε ，而在隱私侵犯分區則採用更嚴格的隱私保護措施。實驗結果顯示，與現有最先進方法相比，該方案在保持60.81%原始性的同時，隱私風險降低 20.05%，效用性提升 54.01%和 15.33%。

Bargh & Choenni (2022) 從理論層面探討了資料效用、隱私和公平性的整合問題，指出傳統隱私保護機制可能對少數群體產生不成比例的負面影響，導致「隱私保護歧視」問題[26]。該研究建議將研究重點從單純的隱私-效用平衡擴展到隱私-效用-公平性的三維優化問題，為本研究的理論框架提供了重要基礎。

Zhai et al. (2024) 針對電力消費資料提出了敏感分區的隱私保護方法，結合差分隱私和 *k*-匿名技術[8]。該方法通過隨機森林技術識別模式特定的準識別符，在資訊損失精度和執行時間方面表現優於傳統 k-匿名、CDKA 和 RKA 演算法。

然而，現有研究主要存在以下局限性：首先，多數研究缺乏對公平性的系統性量化評估；其次，效用性衡量多集中於統計效用，較少關注機器學習公平性；最後，隱私預算 ε 和匿名參數 k 的選擇多依賴經驗而非理論指導。這些研究空白為本研究建立三維評估框架提供了創新機會。

2.4.2 技術選擇理由與適用性分析

k-匿名技術的適用性：

k-匿名技術特別適合處理結構化的表格型資料，尤其是包含明確準識別符的個人資料。根據數位發展部指引，*k*-匿名技術的主要優勢在於其真實性保持特性，即轉換後的資料項目與原始值保持一致性[1]。在本研究的應用情境中，*k*-匿名技術適合處理：(1)類別型屬性，如性別、種族、教育程度等離散型資料；(2)可建立層次

doi:10.6342/NTU202600729

泛化樹的數值屬性，如年齡區間、收入範圍等；(3)需要保持資料解釋性的應用場景。

差分隱私技術的適用性：

差分隱私技術通過數學證明提供嚴格的隱私保證，特別適合處理數值型資料和統計查詢。該技術的核心優勢在於能夠抵禦具有任意背景知識的攻擊者[4,5]。在本研究中，差分隱私適合處理：(1)連續型數值屬性，如年齡、收入等；(2)統計聚合查詢結果；(3)需要理論隱私保證的高風險應用場景。

其他技術的局限性分析：

雖然同態加密和安全多方運算等密碼學技術提供強隱私保證，但計算複雜度高，難以應用於大規模資料發布場景[21,22]。合成資料技術雖然能保留統計特性，但在生成資料品質控制和原始資料洩漏風險方面仍存在挑戰[23]。相較之下，*k*-匿名與差分隱私的結合既能提供理論隱私保證，又保持合理的計算效率，適合實際部署應用。

2.4.3 本研究與現有方法的差異分析

與現有方法的比較分析：

現有混合隱私保護機制研究在技術整合方面已有一定進展，但在評估維度和研究重點上與本研究存在差異。

Majeed & Hwang (2024)的方法主要關注智能分區策略和資料效用優化[25]，通過模式識別技術提升隱私保護效果。該研究在保持資料效用性方面表現良好，但未涉及公平性評估。

Bargh & Choenni (2022)從概念層面討論了隱私-效用-公平性的整合[26]，提出了重要的理論框架，但缺乏具體的量化評估方法。

本研究的特點：

相較於現有研究，本研究嘗試在以下幾個方面做出貢獻：

- 評估維度擴展：在傳統的隱私-效用評估基礎上，加入公平性維度的考量

doi:10.6342/NTU202600729

- 量化評估方法：引入均等勝算指標[13]，提供可測量的公平性評估標準
- 參數選擇指導：探索基於公平性約束的參數優化方法
- 實務應用價值：為需要兼顧多重目標的應用提供評估工具

這些嘗試希望能在現有研究基礎上提供一些補充，特別是在隱私保護機制的公平性評估方面填補部分研究空白。

2.5 公平性度量標準

隨著機器學習在社會決策中的廣泛應用，隱私保護技術對不同社會群體的差異性影響逐漸受到關注。研究發現，傳統隱私保護機制可能對少數群體產生不成比例的負面影響，導致「隱私保護歧視」問題。Bagdasaryan 等人[19]的研究顯示，相同的隱私參數設置可能對少數群體造成不成比例的效用損失，引發公平性關切。這種現象主要源於少數群體在數據中的代表性較低，相同程度的雜訊對其統計特性影響更大。因此，現代隱私保護研究需要從單純的隱私-效用平衡，擴展到隱私-效用-公平性的三維優化問題[19,20]。

2.5.1 公平性作為隱私保護的核心原則

隱私保護技術雖然能夠降低資料洩露的風險，但其對於模型效能的影響可能在不同群體間存在差異。近年來的研究顯示，差分隱私等技術可能對特定群體造成不成比例的影響 (disparate impact)，使得模型在這些群體上的預測準確率顯著下降[19, 20]。這種現象凸顯了在追求隱私保護的同時，必須同時關注公平性的重要性。因此，本研究將隱私保護、資料效用與公平性視為一個三維度的最佳化問題，在設計混合式隱私保護機制時，不僅考量隱私與效用的權衡，更進一步檢視隱私保護技術是否對不同群體產生歧視性影響。

在法規層面，公平性已被納入資料保護的核心要求。歐盟《一般資料保護規則》(General Data Protection Regulation, GDPR) 第 5 條第 1 項第 a 款明確將公平性

doi:10.6342/NTU202600729

(fairness) 列為資料處理的核心原則之一，要求資料控制者在處理個人資料時必須確保「合法性、公平性與透明性」(lawfulness, fairness and transparency)。這項規定反映了資料保護不僅是技術問題，更是倫理與社會正義的議題。在實務應用中，若隱私保護技術導致特定群體在模型預測上遭受系統性的不利影響，即可能違反 GDPR 的公平性要求。

台灣金融監督管理委員會於 2024 年發布的《金融機構資料共享之資料治理諮詢文件》[28] 亦強調資料治理應遵循 GDPR 所揭示的核心原則，包括公平性原則。該文件針對金融機構客戶資料的處理與共享，提出分級管理的架構，將客戶資料依照隱私保護強度與再識別風險分為四個等級，如表 2.2 所示。這個分級架構不僅考量技術面的隱私保護強度，亦兼顧法規遵循與資料共享的實務需求，為金融機構在資料治理與隱私強化技術的應用上提供了具體的指引。表 2.2 顯示，隨著資料處理技術的進步（從傳統去識別化到新興隱私強化技術），資料的保護環境程度可以從高度保護降至中度或低度保護，同時個資法規的遵循要求與資料共用對象的範圍也相應調整。這個分級架構呼應了本研究的核心關懷：在應用隱私保護技術時，必須確保技術的應用不會對特定群體造成不成比例的影響，從而違反公平性原則。

表 2.2 金融機構客戶資料分級表

客戶資料分級	對應之資料處理技術	對應之治理方式		
		保護環境程度	個資法規遵循	資料共用/再利用之對象
第 1 級 原始客戶資料	未經處理	高度保護環境	須遵循個資法規，並逐項取得客戶同意	原則：金融機構 例外：依據開放銀行等機制辦理之資料共用
第 2 級 經隱私權保護目的處理後，仍易遭還原而識別客戶之資料	經傳統去識別化技術處理			原則：金融機構 例外：對於資料治理及隱私權保護強度等同金融機構之非金融機構，亦可評估納入分享對象

doi:10.6342/NTU202600729

第 3 級 經隱私權保護目的處理，較不易遭還原而識別客戶之資料	經新興隱私強化技術處理	中度保護環境	依循個資法規，並可採取較簡易之方式取得客戶同意	金融機構或具備妥適資料治理及隱私權保護能力之非金融機構
第 4 級 不屬於個別客戶之資料	經聯合學習所產生之參數資料，及合成資料	低度保護環境	無個資法規之適用	金融機構或非金融機構(屬於資料再利用，不涉及共用個別客戶資料)
註：資料來源：金融監督管理委員會(2024)。金融機構資料共享之資料治理諮詢文件[28]。				

綜上所述，公平性已從倫理呼籲轉化為法規要求，成為資料保護與隱私強化技術應用中不可忽視的核心原則。本研究透過在混合式 *k*-匿名與差分隱私機制中納入公平性評估，旨在確保隱私保護技術的應用不會加劇既有的社會不平等，而是在保護個人隱私的同時，維護不同群體在資料應用中的公平待遇。

2.5.2 公平性指標的分類與選擇

在建立隱私保護機制時，選擇適當的公平性指標至關重要。Barocas, Hardt 與 Narayanan (2023) 根據公平性指標所關注的統計獨立性質，將常見的公平性指標分為三大類：獨立性 (Independence)、分離性 (Separation) 與充分性 (Sufficiency)[27]。這三類指標分別從不同的角度衡量機器學習模型是否對不同群體產生歧視性影響。表 2.3 整理了這三類公平性指標的代表性指標與定義。

表 2.3 公平性指標的分類

分類	代表性指標	定義
獨立性 (Independence)	人口統計均等(Demographic Parity)	模型預測結果與敏感屬性獨立
分離性 (Separation)	均等勝算(Equalized Odds)和機會均等(Equal Opportunity)	在給定真實標籤下，預測結果與敏感屬性條件獨立

doi:10.6342/NTU202600729

充分性 (Sufficiency)	預測值平等(Predictive Parity)	在給定預測結果下，真實標籤與敏感屬性條件獨立
資料來源：改編自 Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press. https://www.fairmlbook.org/		

獨立性指標要求模型的預測結果在統計上與敏感屬性無關，其代表性指標為人口統計均等 (Demographic Parity)，該指標要求不同群體獲得正面預測結果的比例應相同。分離性指標則在考慮真實標籤的條件下，要求預測結果與敏感屬性獨立，其中均等勝算 (Equalized Odds) 要求不同群體的真陽性率與假陽性率皆相等，而機會均等 (Equal Opportunity) 則僅要求真陽性率相等。充分性指標關注的是在給定預測結果的條件下，真實標籤與敏感屬性的獨立性，預測均等 (Predictive Parity) 要求不同群體在相同預測結果下，真實為正例的機率應相同。Mehrabi et al. (2021) 針對機器學習中的偏見與公平性進行了全面性的文獻回顧 [14]，整理了各類公平性指標的優缺點與適用情境，可作為進一步參考。

在本研究中，我們選擇均等勝算 (Equalized Odds) 作為公平性評估指標，主要基於以下四點考量：

(1) 技術適用性： 均等勝算適用於二元分類問題，與本研究使用的 Adult 資料集任務（收入預測：年收入是否超過 50K 美元）及資料保護情境相符。

(2) 隱私保護敏感性： 在應用 *k*-匿名或差分隱私等技術後，資料的分布可能發生變化，均等勝算能夠同時檢測模型在不同群體上的真陽性率與假陽性率是否存在差異，從而偵測隱私保護技術是否對特定群體造成不成比例的影響 (disparate impact)。

(3) 統計穩健性：均等勝算不依賴於特定的分類閾值，而是直接比較不同群體的混淆矩陣元素，這使得該指標在面對資料擾動時具有較高的穩健性。

(4) 實務可解釋性：真陽性率與假陽性率在實務應用中具有明確的意義，例如在醫療診斷或信用評估等領域，決策者能夠直觀理解不同群體在「正確識別」與「錯誤識別」上的差異。

doi:10.6342/NTU202600729

均等勝算的數學定義如下：P( )表示某個事件發生的概率，R 為預測結果(當 R =1 時表示模型預測為正例)，Y 為真實標籤，A 為敏感屬性(如性別、種族)，則均等勝算要求對於任意兩個敏感屬性值 a 和 b，以及任意真實標籤值 y，滿足：

$$P(R = 1|Y = y, A = a) = P(R = 1|Y = y, A = b) \tag{2.2}$$

這表示在給定真實標籤情況下，模型預測為正例的概率應該與敏感屬性無關。具體來說，均等勝算要求真陽性率(TPR)和假陽性率(FPR)在不同群體間相等。

其中

$$TPR = P(R = 1|Y = 1, A = a) \tag{2.3}$$

$$FPR = P(R = 1|Y = 0, A = a) \tag{2.4}$$

均等勝算要求不同敏感屬性群體的 TPR 與 FPR 分別相等。在本研究中，我們以性別作為敏感屬性，檢驗隱私保護技術是否導致男性與女性群體在收入預測模型上的 TPR 或 FPR 產生顯著差異。

以 Adult 資料集為例，假設我們建立一個預測個人年收入是否超過 50K 美元的分類模型。在應用隱私保護技術之前，若男性群體的 TPR 為 0.75（即 75% 的高收入男性被正確預測為高收入），女性群體的 TPR 為 0.70，則兩群體的 TPR 差異為 0.05。均等勝算要求我們在應用隱私保護技術後，監控這個差異是否進一步擴大，若差異顯著增加，即表示隱私保護技術對女性群體造成了不成比例的負面影響。這種監控機制對於確保隱私保護技術在實務應用中不會加劇既有的社會不平等具有重要意義 [19, 20]。

doi:10.6342/NTU202600729

第三章　　研究方法設計

基於文獻回顧，我們發現現有研究缺乏同時考量隱私保護性、資料效用性和資料公平性的整合框架。本研究將均等勝算納入隱私保護評估框架，提出一個創新的既能保護用戶隱私，又能保證跨群體預測公平性的全面性隱私保護機制。透過平衡隱私保護性、資料效用性和資料公平性，提供更為全面和具有社會責任感的隱私保護解決方案。我們的方法包含三個核心模組：資料前處理、差分隱私和指標分析(如圖 3.1 所示)。

1. 資料前處理：處理原始資料集，包括資料清洗、特徵轉換及準識別符(QI)與敏感屬性(SA)的識別。
2. 差分隱私：根據不同的隱私參數(ε 值)產生差分隱私保護後的資料集。
3. 指標分析：計算差分隱私保護後資料集的各項指標，包括隱私保護性、資料效用性和資料公平性，呈現分析結果，提出建議之隱私參數。

期望不僅關注傳統的隱私保護性和資料效用性，更系統性地評估不同隱私參數對資料公平性的影響，配合資料隱私的視覺化表示，提供一個全方位的隱私保護決策支援工具。

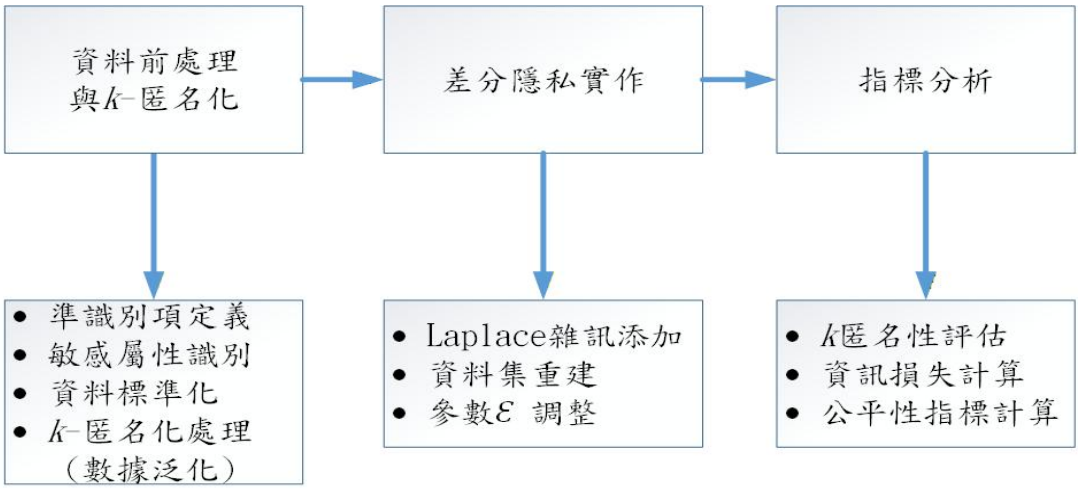


圖 3.1　系統架構流程圖

doi:10.6342/NTU202600729

3.1 資料前處理

準識別符是指那些單獨不能直接識別個人，但當多個這類屬性結合時，可能使個人身份變得可識別的資料屬性。所以準識別符的選定屬於隱私保護一個重要的關鍵。為確保實驗公平性，本研究將仿照前人識別風險評估研究，結合相關性和外部可用性，計算每個屬性的識別風險分數來選擇準識別符。

風險分數 = 相關性 × 外部可用性

本研究採用乘法形式計算風險分數，其理論基礎在於：準識別符的重新識別風險取決於兩個條件的同時滿足—該屬性必須與目標變數具有顯著相關性，且攻擊者能夠從外部資源獲取該屬性資訊。當任一條件不成立時（即相關性或外部可用性趨近於零），該屬性的識別風險應相應降低，此特性符合乘法運算的數學性質。

- 相關性：使用 Mutual Information[15]（互信息（MI））量化各屬性與目標變數（收入）之間的相關性。計算公式為： $I(X;Y) = \sum\sum p(x,y) \log(p(x,y)/(p(x)p(y)))$ 其中，X 代表候選屬性，Y 代表目標變數，p(x,y)是聯合概率分佈，p(x)和 p(y)是邊緣概率分佈。值越高，表示該屬性與目標變數的相關性越強。
- 外部可用性：評估各屬性在外部資料集或公開資源中的可得性。參考 El Emam 與 Dankar [16] 提出的準識別符風險評估框架，使用 1-5 的量表評分每個屬性的外部可得性，其中 5 表示非常容易從外部獲取（如履歷、社交媒體等公開平台）。

本研究採用「實驗驗證」的方式決定準識別符的數量與組合。測試了不同數量的 QI 組合（3 個、4 個、5 個），評估各組合能否達成 k=5 的匿名化目標及資料效用是否可接受。實驗結果顯示，選擇 4 個屬性（education、age、relationship、 hours-per-week）能形成 144 個等價類且皆達 k=5，為最佳平衡點。 增加至 5 個 QI 導致過度泛化，減少至 3 個 QI 則無法充分涵蓋識別風險。詳細分析見第 4 章。

doi:10.6342/NTU202600729

接下來我們參照 Mendes [17]去進行資料的資料清洗載入與初步處理，處理輸入資料並識別準識別符和敏感屬性，參照 Bayardo[11]針對準識別屬性進行適當的數據泛化以保留資料效用性。經過數據泛化的程式之後，資料集依原始資料分佈即具有初步的 k-匿名群組分布狀態。

經過上述泛化處理，資料集的區分度顯著降低，初步具備 k-匿名的特性。泛化的程度是基於 Meyerson & Williams[12]提出的最小失真原則，即在達到一定 k-匿名性的前提下，盡量減少資訊損失。為評估泛化過程中的資訊損失，我們計算了每個屬性的泛化損失並將其正規化到[0,1]區間。

3.2 差分隱私

差分隱私是一種嚴格的隱私保護機制，透過在資料中添加精心設計的隨機雜訊，保護個體隱私。本節詳細說明本研究中差分隱私的實作方法、理論基礎和參數選擇依據。

ε 差分隱私的定義如第二章所示 ：

$$\Pr[A(D_1) \in S] \leq \exp(\varepsilon) \cdot \Pr[A(D_2) \in S] \qquad (同\ 2.1)$$

其中，Pr 表示機率（Probability），A 代表隨機演算法，D_1和 D_2是相鄰資料集（僅相差一筆記錄），∈表示「屬於」，S 為輸出集合，exp(ε)為指數函數（等同於 e^ε，e 為自然常數約 2.718），為乘號。此公式說明演算法在相鄰資料集上的輸出分布差異受到 exp(ε)的限制，確保個體隱私不因資料的加入或移除而遭受過度洩露。ε 值控制隱私保護與資料效用之間的權衡：ε 越小，隱私保護性越強，但可能導致更大的資料失真；ε 越大，資料效用性越高，但隱私保護性較弱。根據差分隱私研究的理論基礎，一般將 ε 值分為三個層級：

1. 較小的 ε 值（如 0.1-1.0）：提供較強的隱私保護性，但可能導致較大的資訊損失（降低資料效用性）。
2. 中等的 ε 值（如 1.0-3.0）：提供合理的隱私保護性與資料效用性平衡。
3. 較大的 ε 值（如 3.0-10.0）：保留較多原始資訊，但隱私保護性較低。

doi:10.6342/NTU202600729

本研究選擇 ε 值範圍為 0.1 至 10.0，涵蓋強、中、弱三種隱私保護強度，以系統性地評估不同隱私參數對隱私保護性、資料效用性與資料公平性的影響。具體的參數選擇依據與實驗設計將於第四章詳述。

本研究參照 Wang 等人[7]的方法，對準識別符組合形成的等價類計數應用差分隱私保護。在差分隱私的多種實現方式中，本研究選擇拉普拉斯機制作為雜訊添加的核心技術，此選擇係基於技術適配性、隱私保障嚴謹性，以及實作便利性等多重考量。使用資料的差分隱私雜訊添加處理之虛擬碼如演算法 1 所示。

Algorithm-1 差分隱私雜訊添加演算法

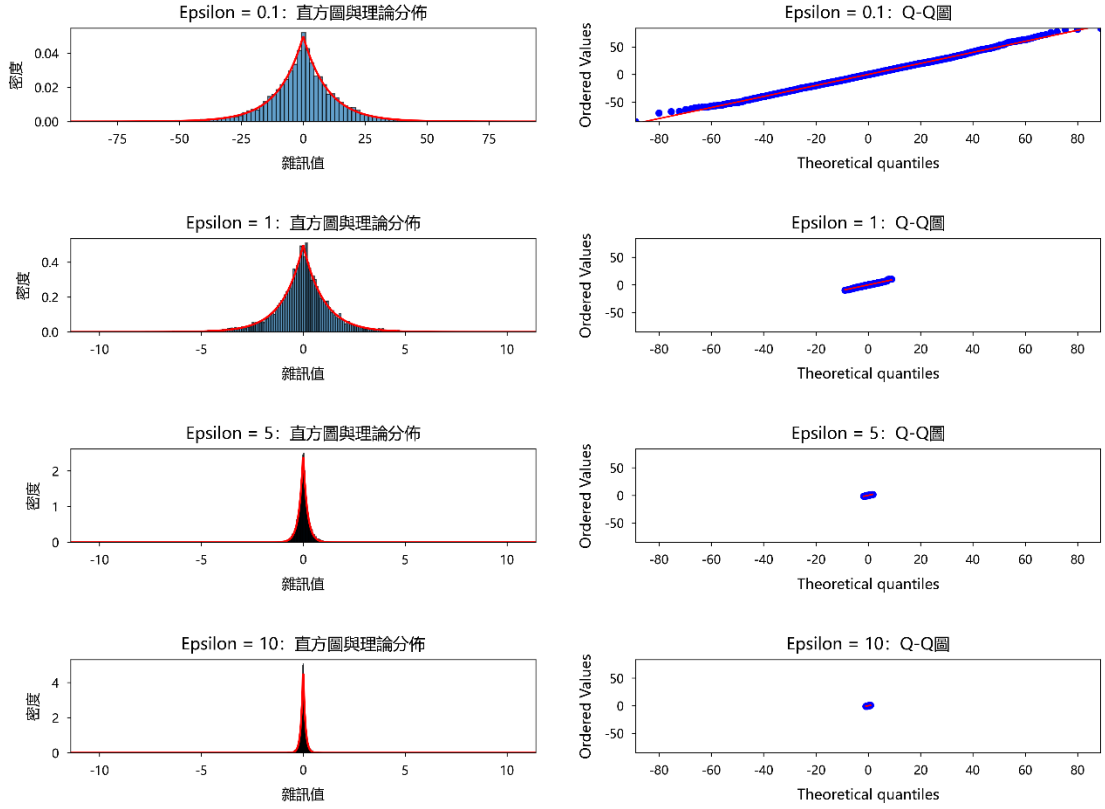
```
function GENERATE_NOISY_DATASET(data, quasi_identifiers, Epsilon, noise_scale)
    group_counts ← data.groupby(quasi_identifiers).size().reset_index(name='count')
    scale ← 1.0 / Epsilon
    for each row in group_counts do
        noise ← random.laplace(0, scale * noise_scale)
        row['noisy_count'] ← max(0, round(row['count'] + noise))
    end for
    filtered_groups ← group_counts[group_counts['noisy_count'] > 0]
    return filtered_groups
end function
```

如圖 3.2 所示，不同 ε 值下的拉普拉斯雜訊分布呈現明顯的特徵差異。當 ε 值為 0.1 時，雜訊分布範圍較廣（約±100），雖然提供強隱私保護，但可能對資料效用性造成較大影響；當 ε 值增加到 1.0 時，雜訊範圍顯著縮小（約±7.5），在隱私保護與資料效用間達到較佳平衡；而當 ε 值增大到 10.0 時，雜訊幾乎集中於零點附近，雖然保持高資料效用性，但隱私保護程度相對較低。這種變化趨勢驗證了 ε 參數作為隱私-效用平衡控制器的有效性。

本研究參照 Wang [17]採用差分隱私方法來改變群組計數，這個隱私保護機制實現了群組層面的差分隱私，對準識別符組合的計數添加校準的噪聲，從而保護每個群組的確切大小和成員關係。它與傳統的 k 匿名性結合，既確保了最小群組大小的保護(k 值)，又提供了針對群組計數的差分隱私保證。這種混合方法結合了

doi:10.6342/NTU202600729

k-匿名的直觀性與差分隱私的理論保障，為實際應用提供了更為彈性的隱私保護機制。



直方圖 ε 值=0.1、1、5、10　　分位數-分位數圖 ε 值=0.1、1、5、10

註：ε 值越小，雜訊分佈越分散；Q-Q 圖中點越偏離對角線，表示與理論分佈差異越大

圖 3.2　不同隱私參數 ε 值下拉普拉斯雜訊分佈的直方圖與 Q-Q 圖比較

3.3 指標分析

為確立平衡的選擇標準，本研究設計了一套多維評估框架，綜合衡量三個關鍵面向：隱私保護程度、資料效用性及資料公平性。這些精確量化的指標相互補充，共同構成一個全面的評估體系，使我們能夠在各項需求間取得最佳平衡點，並依據具體應用場景進行客觀的比較和決策。

3.3.1　隱私保護性

doi:10.6342/NTU202600729

本研究中的隱私保護性評估主要建立在兩個關鍵參數之上：k-匿名中的群組大小 k 值和差分隱私中的隱私參數 ε 值。k 值定義了任一準識別屬性組合至少對應 k 筆記錄的保證，確保個體無法被精確定位。具體而言，k 值代表了任一記錄被識別出的模糊程度，k 值越大，意味著記錄被唯一識別的可能性越低，隱私保護程度越高。同時，隱私參數 ε 值作為差分隱私機制的敏感度參數，控制著添加噪聲的程度，隱私參數 ε 值越小，隱私保護越嚴格，但可能導致資料效用降低；ε 越大，資料效用性越高，但隱私保護性較弱。

k 值和 ε 值這兩個參數共同構成了本研究評估資料隱私保護效能的量化基礎，使我們能夠在不同隱私保護技術間進行標準化比較。

3.3.2 資料效用性

本研究定義資料效用性，與資訊損失成反比。資訊損失的計算可分為兩個不同階段：首先是前處理過程中泛化操作導致的「泛化損失」(Generalization Loss)；其次是差分隱私機制實施時添加雜訊造成的「隱私保護損失」。這兩種損失源自不同處理機制且發生於不同階段，本研究分別對其進行評估，最後合併為總體資訊損失以全面衡量隱私保護對資料效用的影響。

量化資料泛化導致的資訊損失有數值型屬性泛化損失：針對如年齡、工作時數等屬性，計算原始值與泛化後值之間的平均絕對差異，並將其歸一化到原始值範圍（0-1），而類別型屬性泛化損失：教育程度、關係等屬性，則基於唯一值減少的比例計算損失，也就是例如新類別數=1 時，損失最大（=1）；當新類別數=原類別數時，損失最小（=0）。

在評估差分隱私保護導致的「隱私保護損失」時，傳統比較了均方誤差(Mean Square Error, MSE)與總變異距離(Total Variation Distance, TVD)兩種方法。均方誤差計算原始資料與處理後資料對應點之間的平均誤差平方，雖然直觀易懂且計算簡單，但在資料分佈比較中存在局限性，較難以在隱私保護情境中提供明確解釋。

doi:10.6342/NTU202600729

相比之下，總變異距離是統計學中用來測量兩個概率分佈差異的指標，其數學定義如下：

對於兩個離散概率分佈 P 和 Q，定義在同一個樣本空間 Ω 上，計算公式為：

$$\mathrm{TVD}(P, Q) = \frac{1}{2}\sum_{x \in \Omega} |P(x) - Q(x)| \tag{3.1}$$

TVD 能更全面反映分佈層面的變化，能有效捕捉差分隱私雜訊添加對等價類分佈的影響，為隱私參數 ε 值的選擇提供客觀依據。虛擬碼如演算法 2。

Algorithm-2 資訊損失總變異距離計算演算法

```
1:function CALCULATE_TVD(original_groups, noisy_groups, quasi_identifiers)
2:      original ← original_groups.copy();
3:      original['proportion'] ← original['count'] / original['count'].sum();
4:      noisy ← noisy_groups.copy();
5:      noisy['proportion'] ← noisy['noisy_count'] / noisy['noisy_count'].sum();
6:      all_groups ← concat([original[quasi_identifiers],
    noisy[quasi_identifiers]]).drop_duplicates();
7:      original_merged ← merge(all_groups, original, on=quasi_identifiers,
    how='left').fillna(0);
8:      noisy_merged ← merge(all_groups, noisy, on=quasi_identifiers,
    how='left').fillna(0);
9:      tvd ← 0.5 * sum(abs(original_merged['proportion'] -
    noisy_merged['proportion']));
10:       return tvd;
```

TVD 量化了原始資料與保護後資料之間的資訊損失程度。圖 3.3 展示了 TVD 較小的情況(TVD = 0.105)，其中原始分佈與保護後分佈之間的差異較小，表明隱私保護機制對資料分佈影響較小。相比之下，圖 3.4 呈現了 TVD 較大的案例(TVD = 0.225)，圖中較大的紫色差異區域直觀地展示了原始分佈與保護後分佈間的顯著變化，也解釋了為何其 TVD 值(0.225)明顯高於圖 3.3 (0.105)。可以明顯觀察到原始分佈(藍線)具有較大的波動性，表現為峰穀明顯的折線；而保護後分佈(紅線)則變得較為平滑，波動幅度顯著減小。這種資料分佈「平滑化」現象是隱私保護效果的典型表現，模糊化任一記錄被識別出的機率，代表了降低了極端值的出現頻率，使資料分佈趨於均勻。

doi:10.6342/NTU202600729

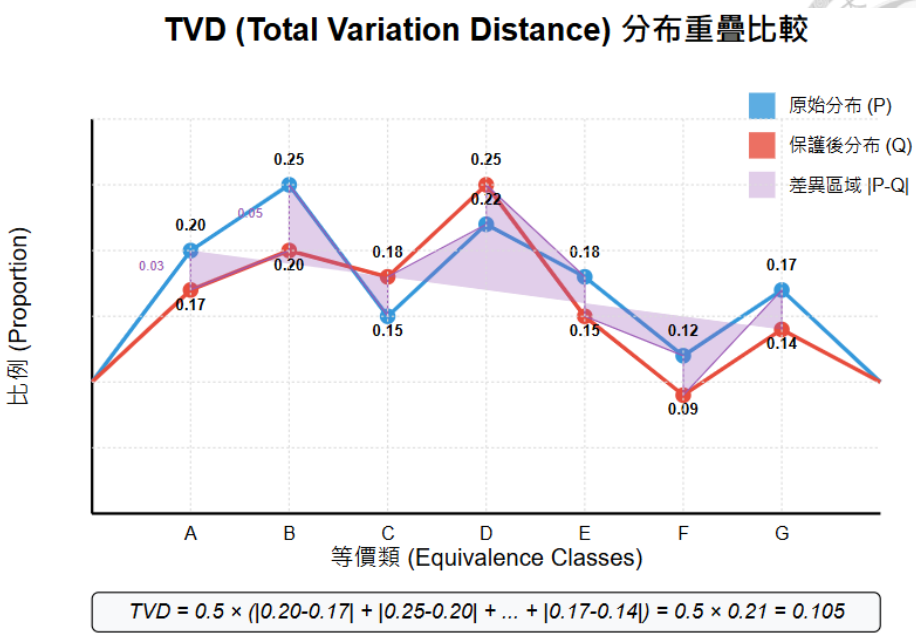


圖 3.3　低資訊損失情況下的總變異距離計算示例(TVD = 0.105)

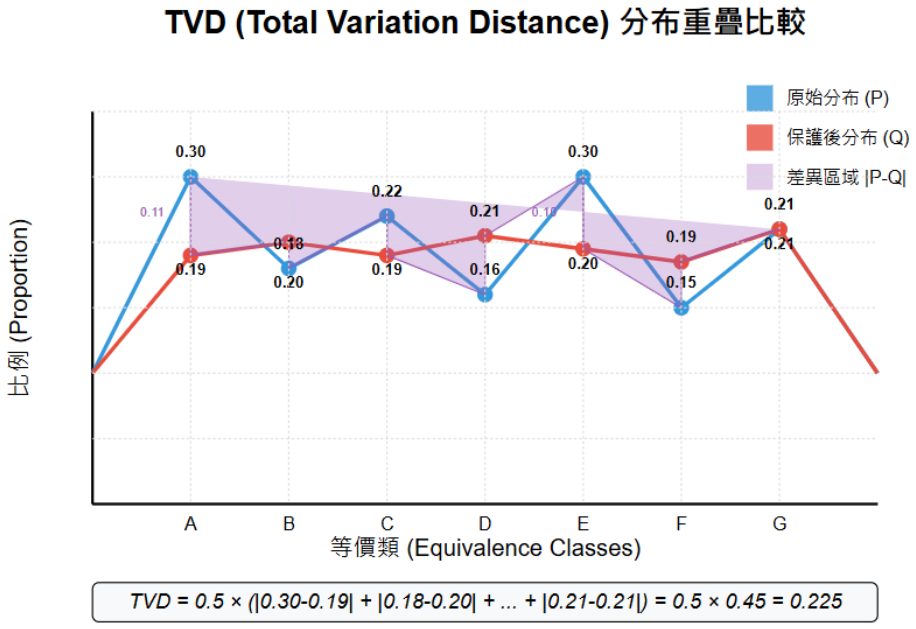


圖 3.4　高資訊損失情境下的總變異距離計算示例(TVD = 0.225)

doi:10.6342/NTU202600729

3.3.3 資料公平性

本研究選擇均等勝算作為資料公平性的評估指標，主要基於以下四項考量：（1）技術匹配性：適合本研究的二元分類問題設定；（2）隱私敏感性：能檢測差分隱私對不同群體的差異化影響；（3）評估全面性：同時考慮 TPR 和 FPR 兩個面向；（4）統計穩健性：不受閾值設定影響。

均等勝算（Equalized Odds）由 Hardt 等人於 2016 年提出[13]，是機器學習公平性領域的重要標準。此處「勝算」（Odds）在統計學中指事件發生與不發生的比率，對於機率為 p 的事件，其勝算定義為 Odds = p/(1-p)。

本研究定義資料的資料公平性，與公平性分數成正比。採用均等勝算評估各差分隱私資料集分別對敏感屬性(種族、性別)公平性的影響。均等勝算要求不同敏感群體的真陽性率(TPR)和假陽性率(FPR)應該相等，以確保預測誤差在不同群體間的一致性。演算法 3 計算各敏感群體均等勝算的 TPR 值和 FPR 值，並測量它們之間的差距。當這些差距接近零時，表示模型對不同群體的預測誤差相似，具有較高的公平性。

在本研究的差分隱私與 *k*-匿名混合機制中，均等勝算特別重要，因為它能捕捉隱私保護機制可能對不同人口群體產生的差異性影響，確保隱私保護不會無意中加劇現有的社會不平等。均等勝算差異值（Equalized Odds Difference Value, EOD）為差異值的總和，是量化不同敏感屬性群體間的預測錯誤率差異的指標。它結合了真陽性率（TPR）差異與假陽性率（FPR）差異，參考 Fairlearn 工具庫的實作方式[28]，計算公式如下：

$$\begin{aligned}\text{EOD} = &\big(\text{max\_group(TPR)} - \text{min\_group(TPR)}\big) + \\ &\big(\text{max\_group(FPR)} - \text{min\_group(FPR)}\big)\end{aligned} \tag{3.2}$$

其中 max_group(TPR)表示所有敏感屬性群體中最高的 TPR 值，min_group(TPR)表示最低的 TPR 值；同理，max_group(FPR)和 min_group(FPR)分別表示所有群體

doi:10.6342/NTU202600729

中最高和最低的 FPR 值。 此差異值越接近 0，表示模型對不同敏感屬性群體的錯誤率分佈越相似，公平性越高。

為便於與隱私性和效用性評分進行跨參數比較，本研究採用 Min-Max 標準化方法將均等勝算差異值轉換為公平性分數：

$$\text{公平性分數} = 1 - (\text{EOD} - \text{EOD\_min})/(\text{EOD\_max} - \text{EOD\_min}) \qquad (3.3)$$

其中 EOD 為均等勝算差異值，EOD_min 和 EOD_max 分別為所有實驗參數下均等勝算差異值的最小值和最大值。此標準化方法確保：(1)分數落在[0,1]區間；(2)當 EOD 達到最小值時分數為 1.0，表示相對最佳的公平性表現；(3)當 EOD 達到最大值時分數為 0，表示相對最差的公平性表現。此標準化方式使公平性分數能夠清楚反映各參數在實驗範圍內的相對位置，便於與標準化後的隱私性和效用性評分進行綜合比較。

在本研究中，我們使用 TPR 差異值和 FPR 差異值的總和作為均等勝算差異值，用於評估差分隱私保護對不同社會群體公平性的影響。

Algorithm-3 均等勝算計算演算法

```
1: function EQUALIZED_ODDS(datasets, sensitive_attrs, target_attr, positive_class,
prediction_attr=None)
2:     results ← {}
3:     for each s_attr in sensitive_attrs do
4:         attr_results ← []
5:         for each (ε, data) in datasets do
6:             group_metrics ← {}
7:             // 如果沒有提供預測列，使用目標屬性本身作為"預測"
8:             if prediction_attr is null then
9:               prediction_attr ← target_attr
10:            end if
11:            // 為每個敏感群體計算 TPR 和 FPR
12:            for each group in unique values of data[s_attr] do
13:                // 獲取該組的數據
14:                group_data ← data where s_attr = group
15:                // 創建真實標籤和預測標籤
```


doi:10.6342/NTU202600729

```
            y_true ← (group_data[target_attr] = positive_class)
            y_pred ← (group_data[prediction_attr] = positive_class)
            // 計算混淆矩陣
            TP ← count where (y_true = True and y_pred = True)
            FP ← count where (y_true = False and y_pred = True)
            TN ← count where (y_true = False and y_pred = False)
            FN ← count where (y_true = True and y_pred = False)
            // 計算 TPR 和 FPR
            TPR ← TP / (TP + FN) if (TP + FN) > 0 else 0
            FPR ← FP / (FP + TN) if (FP + TN) > 0 else 0
            // 保存該組的指標
            group_metrics[group] ← {'TPR': TPR, 'FPR': FPR, 'count':
|group_data|}
        end for
        if |group_metrics| < 2 then continue end if
        // 計算 TPR 和 FPR 差異
        tpr_values ← [metrics['TPR'] for each (group, metrics) in group_metrics]
        fpr_values ← [metrics['FPR'] for each (group, metrics) in group_metrics]
        tpr_disparity ← max(tpr_values) - min(tpr_values)
        fpr_disparity ← max(fpr_values) - min(fpr_values)
        // 計算 Equalized Odds 指標（TPR 和 FPR 差異的平均值）
        eq_odds ← tpr_disparity + fpr_disparity
        attr_results.append({
            'ε': ε,
            'tpr_disparity': tpr_disparity,
            'fpr_disparity': fpr_disparity,
            'equalized_odds': eq_odds,
            'group_metrics': group_metrics
        })
    end for
    results[s_attr] ← attr_results
  end for
  return results
end function
```


doi:10.6342/NTU202600729

第四章　　實驗結果與討論

本章呈現實驗結果，評估所提出框架在平衡隱私保護性、資料效用性和資料公平性方面的表現。透過系統性分析不同隱私參數 ε 值的影響，我們尋找最佳參數配置。實驗中採用了一系列差分隱私參數 ε 值，其中較小的 ε 值提供更強的隱私保護，但可能導致更大的資訊損失；較大的 ε 值則相反。我們使用多種指標評估隱私保護與資訊保存之間的平衡點，全面分析隱私保護性與其他指標之間的關係。本章將詳細呈現這些實驗數據及其分析結果。

4.1 資料集

4.1.1 資料集選擇理由

本研究選擇 Adult 資料集作為實驗對象，這個決定基於多項重要考量。Adult 資料集（又稱"Census Income"資料集）[18]是隱私保護研究領域中被廣泛認可的標準測試資料集，曾被 Sweeney [2]、Bayardo [11]等知名學者在其重要研究中採用，這種學術標準性使得本研究結果能夠與現有文獻進行有效比較。 從資料特性來看，Adult 資料集源自 1994 年美國人口普查資料，包含約 32,561 筆記錄和 14 個屬性，涵蓋了年齡、教育程度、工作時數等豐富的準識別符，以及收入水準這一明確的敏感屬性。資料集同時包含數值型與類別型屬性，這種多樣性使其成為測試混合隱私保護機制的理想選擇。更重要的是，該資料集具有真實的社會人口分佈特徵，研究結果因此具備實務參考價值。特別值得強調的是，Adult 資料集因其真實的人口分布特徵，為公平性研究提供了理想的測試環境。資料集包含性別和種族等多個敏感屬性，且呈現明顯的分佈不平衡現象—性別分佈中男性約佔 66%，女性約 34%；種族分佈中白人約佔 85%，其他種族合計僅 15%。這種不平衡分佈為評估隱私保護技術對不同社會群體的差異性影響提供了理想的測試環境，正好符合本研究將公平性納入隱私保護評估框架的核心目標。然而，Adult 資料集的特定特徵（如種

doi:10.6342/NTU202600729

族分佈高度不平衡：白人約 85%，其他種族僅 15%）可能影響公平性分析的普遍適用性。這種分佈不平衡雖然提供了測試隱私保護對少數群體影響的機會，但也限制了結果向其他人口結構資料集的泛化性。此外，Adult 資料集的規模適中，既能呈現隱私保護面臨的真實挑戰，又不會造成過度的計算負擔，確保實驗的可行性。資料集的公開可得性也保證了研究結果的可重現性和驗證性，這對於建立可信的研究基礎至關重要。相較於其他常用資料集，如 Medical 資料集雖然關注醫療隱私但缺乏多元敏感屬性，Census 資料集規模過大可能影響實驗效率，Adult 資料集在隱私保護研究的各項需求上達到了最佳平衡。

4.1.2 資料預處理

為確保實驗的有效性，本研究對 Adult 資料集進行了標準的預處理參照 Bayardo[11]與 Mendes [17]，包括缺失值處理、離群值檢測和特徵標準化。具體處理方法如下：

泛化處理：為實現 *k*-匿名性，本研究對準識別符進行泛化處理

- 年齡(age)分組：將年齡數值分為 10 年一組的區間，如 20-29 歲、30-39 歲等。泛化後的類別數從連續值減少為 7 個年齡組。
- 教育程度(education)分組：將各種教育程度資訊簡化為五個主要類別：基礎教育、高中、大學、研究所等，降低了類別粒度但保留了教育水準的主要區分。
- 關係 (Relationship)泛化：將各種家庭關係簡化為「家庭成員」和「非家庭成員」兩類。
- 工作時數 (hours-per-week)泛化：將工作時數分為「兼職」(小於 35 小時)、「全職」(35-45 小時)和「加班」(超過 45 小時)三類，反映了常見的職場工作時間定義。

泛化處理：為確保資料一致性，本研究對敏感屬性進行標準化

- 性別標準化：將所有"Male"、"M"、"男"等表示轉換為統一的"男性"標記，將"Female"、"F"、"女"等轉換為統一的"女性"標記。

doi:10.6342/NTU202600729

- 種族標準化：將資料中的"White"、 "Caucasian"標準化為"白人"；"Black"、"African-American"標準化為"黑人"； "Asian"、"Asian-American"統一為"亞裔"等，確保種族類別在資料處理過程中保持一致。
- 所得標準化：將所有收入資訊標準化為二元類別">50K"（高收入）和"<=50K"（低收入）， 並確保所有貨幣符號和數值格式一致，移除可能存在的異常值和無效值。

4.1.3 準識別符的識別與選擇

依據第 3.1 節的方法，本研究評估 7 個候選屬性作為潛在的準識別符。選擇過程整合量化風險評估與實務考量。

首先計算各候選屬性的識別風險分數，使用互信息（MI）衡量與 income 的統計關聯，並依據 El Emam 與 Dankar [16]的架構評估外部可用性（1-5 分）。風險分數為兩者乘積，量化了屬性在重新識別攻擊中的潛在威脅。然而，風險分數僅提供量化基礎，最終選擇還需考量以下因素：

1. 泛化可行性：屬性的類別結構應適合泛化處理，且泛化後仍保有實務意義。例如，工作時數可自然分為兼職/全職/超時三類，既易於泛化又保留分析價值。

2. 等價類分佈：QI 組合必須能形成合理的等價類分佈，達成預定的 k 值。透過實驗驗證不同組合的匿名化效果，選擇能成功達成 k=5 且資料效用損失可接受的配置。

3. 公平性考量：避免選用與敏感屬性（如性別）高度相關的屬性，以免不同群體獲得不平等的隱私保護。

此方法在隱私保護、資料效用與演算法公平性間取得平衡。

4.1.4 準識別符選擇結果與分析

（一）候選屬性的相關性分析

圖 4.1 展示 7 個候選屬性與 income 的相關性分析結果。使用互信息（MI）衡量統計關聯強度，並將 MI 分數標準化至 0 到 1 之間。relationship 的標準化 MI 分

doi:10.6342/NTU202600729

數最高（1.000），與收入有最強的統計關聯。marital-status 亦展現極高的相關性（0.983）。education（0.560）、occupation（0.553）和 age（0.523）呈現中度至高度相關。hours-per-week 的相關性相對較低（0.323），而 native-country 幾乎不具統計關聯（0.000）。

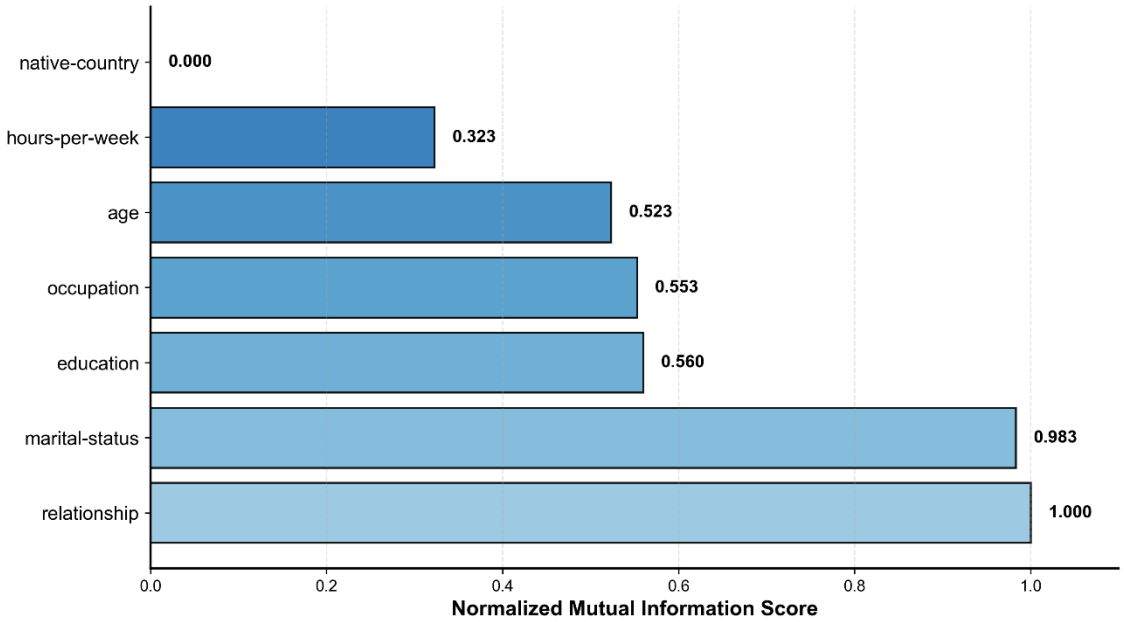


圖 4.1　候選屬性與收入的相關性分析

（二）外部可用性評估

各候選屬性的外部可用性評分(如表 4.1)。education 評分最高（5 分），因其常見於履歷和專業平台如 LinkedIn。age 和 relationship 評分 4 分，可從公開資訊推測或直接取得。occupation、marital-status 和 hours-per-week 評分 3 分。native-country 評分較低（2 分），較少在公開場合揭露。

表 4.1 候選屬性的外部可用性評估

屬性	評分	評分依據
教育程度（education）	5	常見於履歷和專業平台
年齡（age）	4	可從公開資訊推測
關係（relationship）	4	常在社交媒體分享

doi:10.6342/NTU202600729

Occupation （職業）	3	通常在專業平台公開
Marital-status （婚姻狀況）	3	有時在社交媒體分享
工作時數 （hours-perweek）	3	可能在職缺或檔案中提及
Native-country （原籍國）	2	較少公開揭露
註：評分依據 El Emam 與 Dankar [16] 提出的外部可用性評估架構，範圍 1-5 分，分數越高表示攻擊者越容易從公開資源獲取該資訊。		

（三）風險分數與選擇

整合相關性與外部可用性，計算各屬性的風險分數（相關性×外部可用性）。圖 4.2 呈現結果，relationship 風險分數最高（4.00），其次為 marital-status（2.95）、education（2.80）、age（2.09）、occupation（1.66）、hours-per-week（0.97）及 native-country（0.00）。

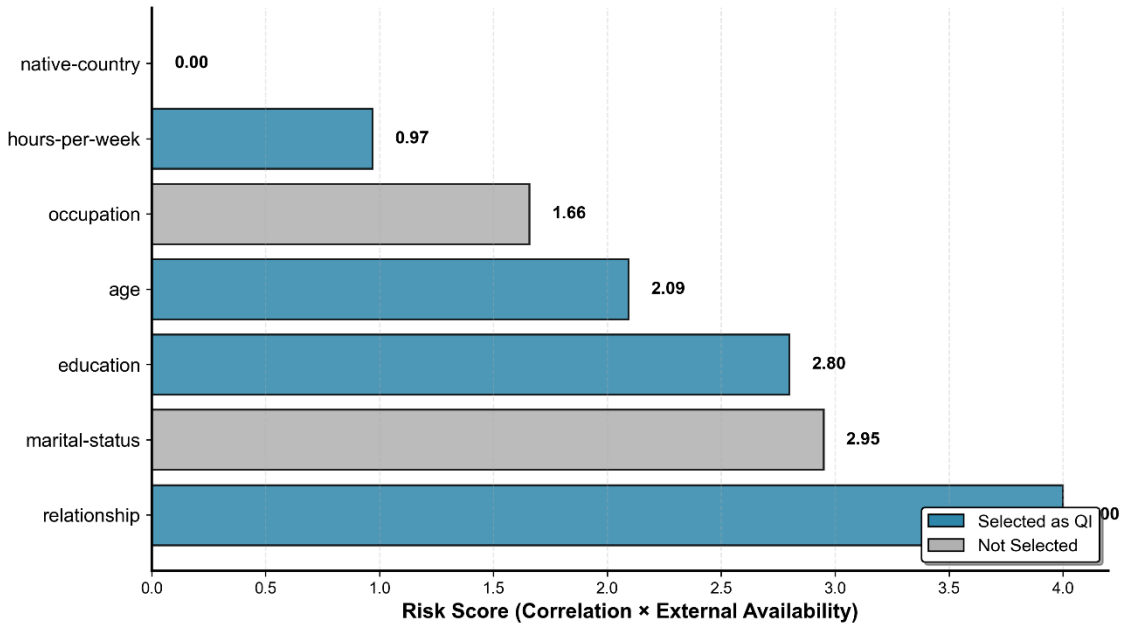


圖 4.2　基於多因素評估的準識別符選擇

本研究最終選擇 hours-per-week、age、educationr 和 elationship 作為準識別符（圖 4.2 藍色長條）。此選擇基於以下考量：

doi:10.6342/NTU202600729

1. relationship、education 和 age 具有較高的風險分數，充分反映重新識別威脅。

2. hours-per-week 雖風險分數較低（0.97），但能有效泛化為三個實用類別（兼職<35 小時、全職 35-45 小時、超時>45 小時），對等價類形成有良好貢獻，且保留勞動市場分析的語義價值。

3. marital-status 雖風險分數高（2.95），但與敏感屬性 sex 高度相關（圖 4.1 相關性達 0.983）。使用此屬性可能導致不同性別群體獲得不平等的隱私保護，因此予以排除。此決策體現本研究對演算法公平性的重視。

4. occupation 雖風險分數達 1.66，但其 14 個原始類別難以有效泛化。

實驗顯示包含此屬性會導致等價類分佈不佳，部分等價類無法達成 k=5 的要求。實驗驗證顯示，所選的 4 個 QI 能生成 144 個等價類，每個等價類皆包含至少 5 筆記錄，成功達成 k=5 的匿名化目標，在隱私保護、資料效用與演算法公平性之間取得良好平衡。

需要說明的是，性別（sex）和種族（race）在本研究中被定義為敏感屬性而非準識別符。敏感屬性的保護目標與準識別符不同：前者關注避免特定敏感資訊的洩露，後者則關注防止個體身份的重新識別。性別和種族在本研究中主要用於公平性分析，評估隱私保護機制對不同社會群體的差異性影響。

4.2 實驗設計與參數設置

本研究採用固定 k-匿名參數（k=5）搭配變動差分隱私參數（ε）的實驗設計。此設計選擇基於以下考量：k-匿名性主要於資料前處理階段確立群組結構，而差分隱私則在此基礎上添加雜訊保護，兩者在混合框架中扮演不同角色。為聚焦於差分隱私參數對隱私-效用-公平性三維平衡的影響，本研究固定 k=5 作為基準配置，系統性地評估不同 ε 值（0.1-10.0）的表現。

doi:10.6342/NTU202600729

為了系統性地評估不同差分隱私參數 ε 值對隱私保護性、資料效用性及資料公平性的影響，本研究選擇了 ε = [0.1, 0.2, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0, 3.0, 5.0, 10.0]作為實驗參數。隱私參數選擇依據：

1.數值範圍選擇理由：

本研究的 ε 值範圍（0.1-10.0）建立在學術界的理論共識與產業界的實務經驗之上，兼顧理論嚴謹性與實務適用性。

理論基礎：根據 Dwork 等人[5]建立的差分隱私理論框架，以及 Wood 等人[30]對差分隱私實務應用的綜述，此範圍涵蓋了從強隱私保護（$\varepsilon \leq 1.0$）到實用隱私保護（ε=1.0-3.0），再到弱隱私保護（ε>3.0）的完整光譜。Xiong 等人[31]在對差分隱私實際部署的系統性分析中指出，學術研究與產業實務通常將 $\varepsilon \leq 0.1$ 視為強隱私保護，$\varepsilon \geq 10$ 視為弱隱私保護，此分類方式已成為領域內的共識標準。

產業實證：本研究的參數範圍與主要產業應用一致，確保了實驗結果的實務參考價值。具體而言：

- Google 的 RAPPOR 系統[6]在 Chrome 瀏覽器中採用 ε≈2 進行本地化差分隱私保護，並設定終身累積隱私預算上限約為 8-9。
- Apple 在 iOS 與 macOS 系統中[33]針對不同應用場景（如 emoji 使用統計、Safari 網域追蹤等）採用 ε 值範圍約為 1-14，其中 Safari 自動播放偵測使用 ε=4。
- Microsoft 在 Windows 遙測系統與 LinkedIn 的廣告查詢系統中也廣泛應用差分隱私技術，採用類似的參數範圍[34]。

這些產業級應用驗證了本研究所選擇 ε 值範圍的合理性與實用性，能夠全面觀察隱私與效用平衡的變化規律。

2.參數分布設計：

doi:10.6342/NTU202600729

為了更精確地捕捉隱私預算對模型效能的非線性影響，本研究並非採用單一的等距或對數間隔，而是採用分段式的非均勻採樣策略 (piecewise non-uniform sampling strategy)。具體而言，我們依據 ε 值的敏感度將實驗分為兩個區間：

低隱私參數區間（$\varepsilon \leq 1.0$）：採用較密集的取樣點（0.1, 0.2, 0.3, 0.5, 0.7, 1.0），間隔從 0.1 逐漸放寬至 0.3，這是因為該區間內隱私與效用的平衡變化較為劇烈，需要更細緻的觀察。

高隱私參數區間（$\varepsilon > 1.0$）：採用較寬鬆的間隔（1.5, 2.0, 3.0, 5.0, 10.0），此區間主要目的在於觀察雜訊影響的邊際效益遞減現象，以及弱隱私保護情境下的效能表現。

最終選定的 11 個 ε 值既能提供足夠的分析粒度來捕捉參數變化對各項指標的影響，又避免了過度密集的取樣所可能造成的計算負擔，使實驗設計在精確性與可行性之間達到良好平衡。

***k*-匿名參數選擇**

本研究選擇 k=5 作為匿名化目標。在 k-匿名性的理論框架下，任一記錄都至少與其他 k-1 筆記錄在準識別符組合上無法區分，這使得攻擊者即使掌握所有準識別符資訊，其成功重新識別特定個體的機率也不會超過 1/k[2]。 當 k=5 時，此機率上限為 20%（1/5＝0.2），意味著攻擊者面對五筆無法區分的記錄時，無法確定哪一筆對應其目標個體。在決定 k 值的設定時，本研究考量了隱私保護強度與資料效用性之間的平衡。較小的 k 值雖然能保留較多原始資料細節，但提供的隱私保護相對有限。例如，k=3 時識別機率上限達 33.3%，且較容易受到背景知識攻擊的威脅，攻擊者可能透過外部資訊進一步縮小候選集。相對地，較大的 k 值 雖能提供更強的隱私保障，但會導致過度泛化，降低資料的分析價值。 k=5 在這兩個目標之間提供了一個務實的平衡點，既確保了基本的隱私保護門檻，又不會過度犧牲資料的可用性。就 Adult 資料集而言，經泛化處理後形成的 144 個等價類在 k=5 的匿名性要求下，能夠保留絕大多數原始記錄。這個群組規模同時為後續的差分隱私

doi:10.6342/NTU202600729

處理提供了穩定的基礎，使得雜訊添加能在較大的等價類上進行，降低了對個別記錄的影響。這種配置形成了「群組保護+記錄保護」的雙重隱私保障機制：*k*-匿名確保了最小群組大小的保護，而差分隱私則進一步保護群組計數的精確資訊。

需要說明的是，*k* 值的選擇具有情境依賴性。不同應用場景對隱私保護的需求程度不同，高敏感度領域（如醫療、金融資料）可能需要更大的 *k* 值，而某些低敏感度應用則可接受較小的 *k* 值。本研究以 *k*=5 作為代表性參數進行分析，所建立的評估框架具有良好的可擴展性，可適用於不同 *k* 值配置的應用情境。

實驗設計原則

雜訊添加機制：在添加雜訊的部分採用拉普拉斯(Laplace)機制，透過添加雜訊到準識別符上，實現群組層面的差分隱私保護。此選擇係基於技術適配性、隱私保障嚴謹性，以及實作便利性等多重考量。本研究參照 Wang 等人[7]的方法，對準識別符組合形成的等價類計數應用差分隱私保護，既確保了最小群組大小的保護(*k* 值)，又提供了針對群組計數的差分隱私保證。

實驗可重現性考量：需要說明的是，為確保實驗結果的確定性和可重現性，本研究採用固定隨機種子（seed=42）執行所有差分隱私實驗。此設計選擇基於以下考量：(1)確保不同 ε 值間的公平比較基準；(2)保證實驗結果的可重現性，便於其他研究者驗證和擴展本研究成果。雖然差分隱私本質上是隨機的，但這種確定性設計能夠排除隨機變異的干擾，更清楚地觀察隱私參數對各項指標的影響模式。

然而，這種設計選擇也帶來了一定的限制。研究結果應視為在特定條件下的點估計，而非具有統計顯著性的總體估計。這種單次實驗設計限制了統計推論能力，但為不同隱私參數間的系統性比較提供了穩定的基礎。未來研究可透過多次重複實驗來驗證本研究發現的統計顯著性。

評估指標設計

本研究採用三維評估框架，系統性地量化差分隱私參數對隱私保護性、資料效用性與資料公平性的影響：

doi:10.6342/NTU202600729

1. 隱私保護性評估：採用 k-匿名值作為指標，衡量添加雜訊後資料集仍能維持的最小群組大小。k 值越大表示隱私保護性越強。

2. 資料效用性評估：採用總變異距離（TVD）量化原始資料與處理後資料的分布差異。TVD 值介於 0 到 1 之間，值越小表示資料效用性保留越好。具體計算方式已於 3.3.1 節說明。

3. 資料公平性評估：採用均等勝算差異值（EOD）評估不同敏感屬性群體（性別、種族）間的預測錯誤率差異。EOD 值越接近 0 表示公平性越高。具體計算方式已於 3.3.3 節說明。這三個指標互相獨立但相互關聯，共同描繪出差分隱私參數對混合隱私保護機制的多維影響，為參數選擇提供全面的評估基礎。

本研究採用三維評估框架，系統性地量化隱私保護性（k-匿名值）、資料效用性（總變異距離, TVD）和資料公平性（均等勝算差異值, EOD）的變化。這三個指標的具體計算方式已分別於 3.2.1 節（*k*-匿名性）、3.3.1 節（TVD）和 3.3.3 節（EOD）詳述，共同為混合隱私保護機制提供全面的性能評估。

基於上述實驗設計，以下各節將呈現不同差分隱私參數下的實驗結果，並分析隱私保護性、資料效用性與資料公平性之間的權衡關係。

4.3 隱私參數與資料效用性的關係分析

在實驗過程中，我們系統性地觀察了隱私參數 ε 值對資料集處理後的 *k*-匿名性與資訊損失特性的影響。表 4.2 呈現了各隱私參數 ε 值下的資料隱私性以及資料效用性數值變化，可以觀察到處理後資料集的 *k*-匿名性與資訊損失特性有明顯變化。

表 4.2 不同隱私參數 ε 值對 *k*-匿名性與資訊損失的影響

ε 值	*k*-匿名性之 k 值	群組數量	TVD 值
0.1	5	130	0.036

doi:10.6342/NTU202600729

0.2	3	141	0.023
0.3	2	141	0.015
0.5	5	141	0.008
0.7	3	143	0.006
1.0	5	144	0.004
1.5	5	143	0.003
2.0	3	144	0.002
3.0	2	144	0.001
5.0	3	144	0.001
10.0	3	144	0.000

註解：
・*k*-匿名性之 *k* 值：處理後資料集達到的最小群組大小
・群組數量：差分隱私處理後剩餘的等價類數量（原始 144 個）
・TVD 值：總變異距離，衡量差分隱私雜訊導致的分佈變化
・所有處理方法都承擔相同的泛化損失（0.316），此為預處理階段的固定成本

差分隱私雜訊對 k-匿名性的影響

需要說明的是，表 4.2 中呈現的 *k* 值波動是差分隱私雜訊對群組計數的間接影響。 本研究的實作流程為：首先對資料集進行 *k*=5 的泛化處理，形成 144 個等價類； 隨後對每個等價類的計數添加拉普拉斯雜訊。當雜訊較大時（ε 較小），部分等價類的計數可能被扭曲至低於 *k*=5 的閾值。根據差分隱私的後處理不變性（post-processing invariance）原則，我們移除計數低於閾值的等價類，以確保輸出資料的實用性。這導致最終資料集的實際 k 值可能低於預設的 *k*=5。

例如，當 ε=0.3 時，較大的雜訊導致更多群組被移除，最小群組大小降至 k=2；而當 ε=1.0 時，較小的雜訊使得大部分群組得以保留，維持 k=5 的水準。這種波動性反映了差分隱私隨機機制的本質特性。儘管如此，在大多數隱私參數 ε 值狀況下，資料集的 *k*-匿名性值仍達到 3 以上，顯示處理後的資料集均能維持基本的群組隱私保護水準。

doi:10.6342/NTU202600729

隱私參數與資訊損失的關係

對原始資料集依序使用不同隱私參數 ε 值進行運算所得到的資訊損失 TVD 值（資料效用性）以及資料集的 *k*-匿名值（隱私保護性），顯示 TVD 值隨隱私參數 ε 值增大而持續下降。較小的 ε 值導致較高的資訊損失(較小的資料效用性)，這是因為添加的雜訊更大，使得處理後的資料與原始資料差異增大。當 ε 增大時，TVD 值逐漸減小，表明資料保留了更多原始分佈特性。值得注意的是，當 ε 大於 3.0 時，TVD 值的下降趨勢趨於平緩，表示在該範圍內增加 ε 值所帶來的資訊保留效益有限。

k-匿名性值在 ε=0.1、ε=0.5、ε=1.0 和 ε=1.5 時達到最高值 *k*=5。這種波動是差分隱私隨機性質的結果，顯示了隨機雜訊添加過程對等價類結構的影響。TVD 值隨著隱私參數 ε 值增加而持續減少，從 ε=0.1 的 0.036 下降至 ε=10.0 的接近 0，呈現近似指數下降趨勢。當 ε 值超過 1.0 後，繼續放寬隱私限制所帶來的效用提升變得有限，顯示邊際效益遞減現象，如圖 4.3 所示。

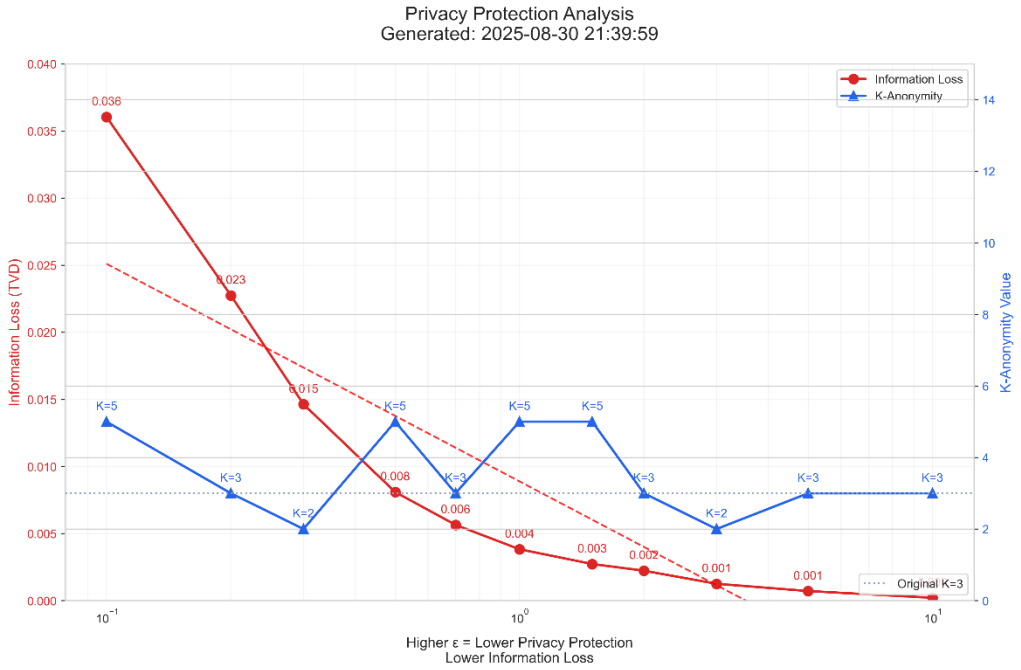


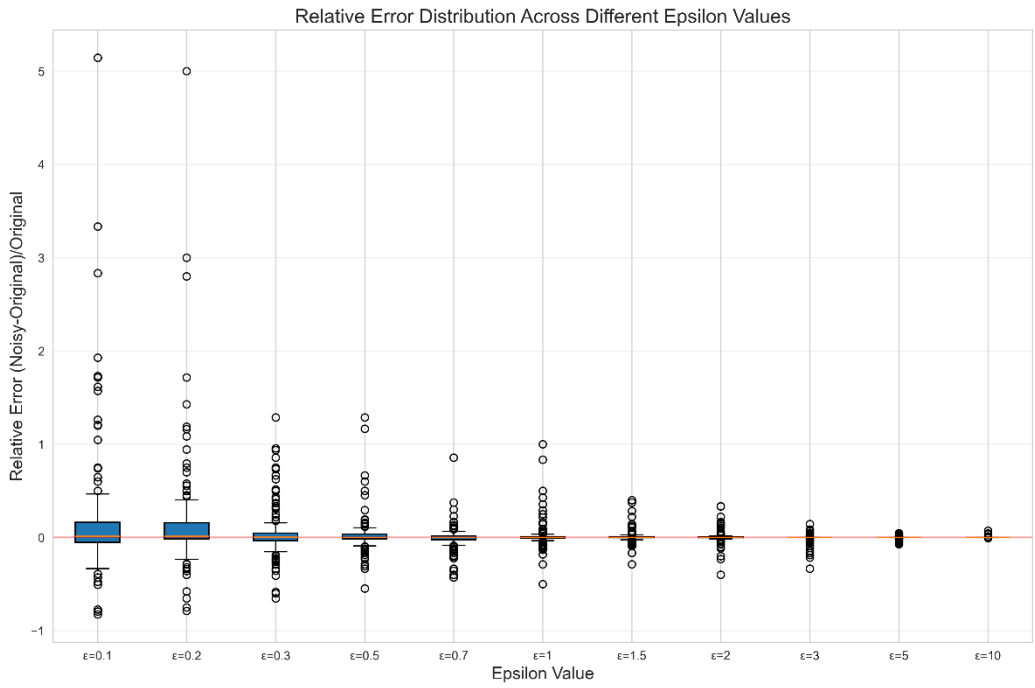
圖 4.3　隱私參數 ε 值與資訊損失的的平衡分析

doi:10.6342/NTU202600729

這一發現對於隱私保護實務具有重要意義，因為它提供了參數選擇的科學依據：當 ε 值超過 2.0 後，繼續放寬隱私限制所帶來的效用提升將變得微乎其微，趨勢逐漸減緩。在曲線開始變平的區域(大約 ε=1.0 到 ε=2.0)代表一個「最佳範圍」，在此點之後，隱私參數 ε 值的進一步增加產生的效用增益極小，同時顯著降低隱私保護。在大多數隱私參數 ε 值狀況下數據集的 *k*-匿名性值達到 3 以上，顯示處理後的資料集均有維持原始 *k*-匿名性水準。

相對誤差分佈分析

我們使用箱型圖來觀察在不同隱私參數 ε 值下的相對誤差分佈，圖 4.4 顯示不同隱私參數 ε 值下的相對誤差分佈，在較小的隱私參數 ε 值（ε=0.1, ε=0.2）時，具有較高的誤差變異性，某些異常值顯示的相對誤差超過原始值的 5 倍，這時四分位距較大也表明雜訊添加的不可預測性更高。隨著隱私參數 ε 值增加（ε=0.5 及以上）則可以發現誤差變異性顯著降低、四分位距向零收窄與誤差中值接近零，反映出對原始數據的修改最小化。這證實了理論預期，即較低的隱私參數 ε 值提供較強的隱私保證，但代價是較高的數據失真，而較高的隱私參數 ε 值保留更多效用但提供較弱的隱私保護。



doi:10.6342/NTU202600729

圖 4.4　不同隱私參數 ε 值下的相對誤差分佈

實驗結果驗證了 ε 參數作為隱私-效用平衡控制器的有效性。當 ε 值超過 2.0 後，繼續放寬隱私限制所帶來的效用提升變得微乎其微，呈現邊際效益遞減現象。這一發現為隱私保護實務提供了參數選擇的科學依據：在大多數應用情境下，ε=1.0 到 ε=2.0 可能是隱私保護與資料效用的最佳平衡範圍。同時，處理後資料集在各 ε 值條件下均維持 $k \geq 3$ 的匿名性水準，確保了基本的群組隱私保護。

在確立隱私參數對資料效用性影響的基礎上，下一節將進一步探討這些參數對資料公平性的影響，完成三維評估框架的全面分析。

doi:10.6342/NTU202600729

4.4 隱私參數與資料公平性的關係分析

本節對不同隱私參數 ε 值下的均等勝算指標變化進行分析，探討隱私保護機制對種族與性別群體間公平性的差異性影響。以 4.3 節考量隱私參數 ε 值與資料效用性所產製的資料集，接續分析隱私參數 ε 值與資料公平性(性別均等勝算差異值、種族均等勝算差異值)之間的關係。接下來以 4.4.1 節呈現公平性指標測量結果；4.4.2 節比較不同敏感屬性的公平性表現；4.4.3 節深入分析造成群體差異的根本原因；4.4.4 節提出實務應用的啟示。

4.4.1 公平性測量結果與初步觀察

計算得到不同隱私參數 ε 值下均等勝算指標值(資料公平性)如表 4.3。

表 4.3 不同隱私參數 ε 值下資料公平性量測結果

資料隱私性		資料公平性(均等勝算差異值)	
ε 值	k-匿名性之 k 值	性別 SEX	種族 RACE
0.1	5	0.337	0.451
0.2	3	**0.266**	**0.258**
0.3	2	0.305	0.303
0.5	5	0.324	0.374
0.7	3	0.329	0.312
1.0	5	0.315	0.313
1.5	5	0.326	0.416
2.0	3	0.341	0.281
3.0	2	0.305	0.286
5.0	3	0.354	0.319
10.0	3	0.324	0.287

doi:10.6342/NTU202600729

為深入理解不同敏感屬性的公平性表現，表 4.4 與表 4.5 分別呈現性別與種族在各隱私參數下的詳細指標，包括 TPR 差異、FPR 差異及均等勝算差異值。

表 4.4 不同隱私參數 ε 值下的性別公平性指標數值

資料隱私性	資料公平性（性別）			
ε 值	TPR 差異	FPR 差異	均等勝算差異值	性別公平分數
0.1	0.258	0.079	0.337	0.193
0.2	0.210	0.057	**0.266**	**1.000**
0.3	0.223	0.081	0.305	0.534
0.5	0.245	0.079	0.324	0.318
0.7	0.239	0.089	0.329	0.261
1.0	0.233	0.082	0.315	0.420
1.5	0.245	0.081	0.326	0.284
2.0	0.246	0.095	0.341	0.148
3.0	0.224	0.080	0.305	0.534
5.0	0.268	0.086	0.354	0.000
10.0	0.243	0.081	0.324	0.318

表 4.5 不同隱私參數 ε 值下的種族公平性指標數值

資料隱私性	資料公平性（種族）			
ε 值	TPR 差異	FPR 差異	均等勝算差異值	種族公平分數
0.1	0.361	0.091	0.451	0.000
0.2	0.182	0.076	**0.258**	**1.000**
0.3	0.226	0.076	0.303	0.767
0.5	0.294	0.079	0.374	0.399
0.7	0.228	0.083	0.312	0.720

doi:10.6342/NTU202600729

1.0	0.224	0.089	0.313	0.715
1.5	0.351	0.065	**0.416**	**0.181**
2.0	0.209	0.073	0.281	0.881
3.0	0.211	0.075	0.286	0.855
5.0	0.257	0.062	0.319	0.684
10.0	0.187	0.091	0.278	0.896

觀察這些結果，我們發現幾個值得注意的現象：

1. 性別與種族對差分隱私的敏感度存在明顯差異

從表 4.3 可以看出，種族公平性指標(均等勝算差異值)的波動範圍（0.258-0.451）明顯大於性別公平性指標的波動範圍（0.266-0.354），顯示 Adult 資料集中種族屬性對隱私參數 ε 值的變化更為敏感。這可能源於資料集中種族分布的高度不平衡性（白人約佔 85%，其他種族合計僅 15%），使得少數種族群體在差分隱私處理後的統計特性更容易受到影響。

2. 隱私保護與公平性之間存在明確的轉折點

從表 4.3 實驗數據顯示，兩種敏感屬性的均等勝算指標值均在 ε=0.2 時達到最佳（最小）值，但隨著 ε 值增加，公平性指標呈現波動。特別是在 ε=1.5 處，種族公平性指標急劇惡化至 0.416，而後又回落。這表明隱私保護機制可能在某些轉折點對公平性產生意外影響。

3. TPR 與 FPR 差異對公平性的不同影響

從表 4.4 可以觀察到，性別公平性的不平等主要來自 TPR 差異(0.210-0.268)，而 FPR 差異相對較小(0.057-0.095)；這種模式反映了 Adult 資料集中性別群體在收入預測上的固有差異，主要呈現在正確識別高收入者的能力差距上。

相比之下，從表 4.5 可見，種族公平性的 TPR 差異(0.182-0.361)與 FPR 差異(0.062-0.091)都有較大波動。這顯示不同敏感屬性在模型預測結果中受到隱私保護影響的方式有顯著差異。

doi:10.6342/NTU202600729

4. 隱私參數的非線性影響

從表 4.4 可見，過小或過大的 ε 值都會惡化性別公平性。當 ε 值過小（如 0.1）時，過度的雜訊添加可能放大群體間的預測差異；而當 ε 值過大（如 5.0 以上）時，隱私保護效果減弱，原始資料中的偏見更容易被保留。表 4.5 更清楚地展現了種族公平性的非線性變化特徵，特別是在 ε=1.5 處的異常惡化現象。

以上四個觀察揭示了隱私參數對不同敏感屬性群體的複雜影響。為更直觀地理解這些數據變化的趨勢與模式，下一節將透過視覺化圖表進行深入分析。

4.4.2 公平性變化趨勢的視覺化分析

基於前述的數據觀察，本節透過視覺化圖表深入分析不同敏感屬性的公平性變化趨勢。圖 4.5 與圖 4.6 分別展示性別與種族的均等勝算指標隨 ε 值的變化，而圖 4.7 至圖 4.10 則進一步拆解 TPR 與 FPR 的個別趨勢，幫助我們理解公平性差異的具體來源。

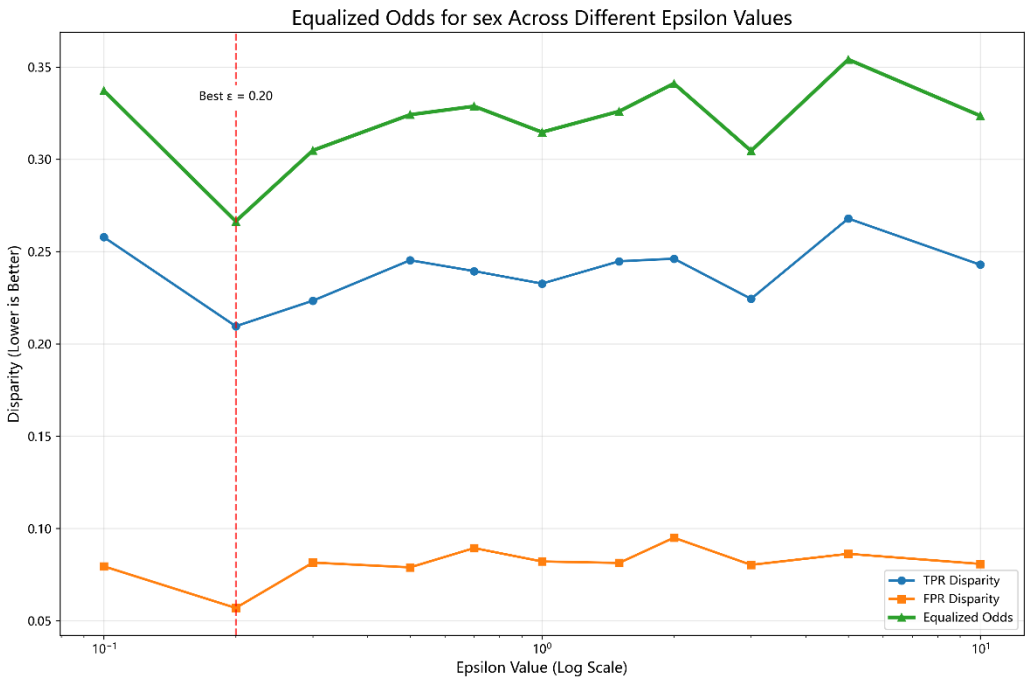


圖 4.5　不同 ε 值下的敏感屬性性別(sex)的均等勝算指標

doi:10.6342/NTU202600729

圖表清楚顯示 ε=0.2 為性別公平性的最佳點，此時三個指標都處於相對較低的水準。TPR 差異曲線展現了主導性的影響模式，其變化幅度明顯大於 FPR 差異，證實了性別群體間的收入預測差異主要源於對高收入者識別能力的不同。

值得注意的是，在 ε=0.2 之後，隨著隱私參數增加，公平性指標整體呈現上升趨勢，但在某些點位（如 ε=3.0）出現局部改善。這種波動反映了差分隱私機制的複雜性及其與原始資料偏見結構的相互作用。

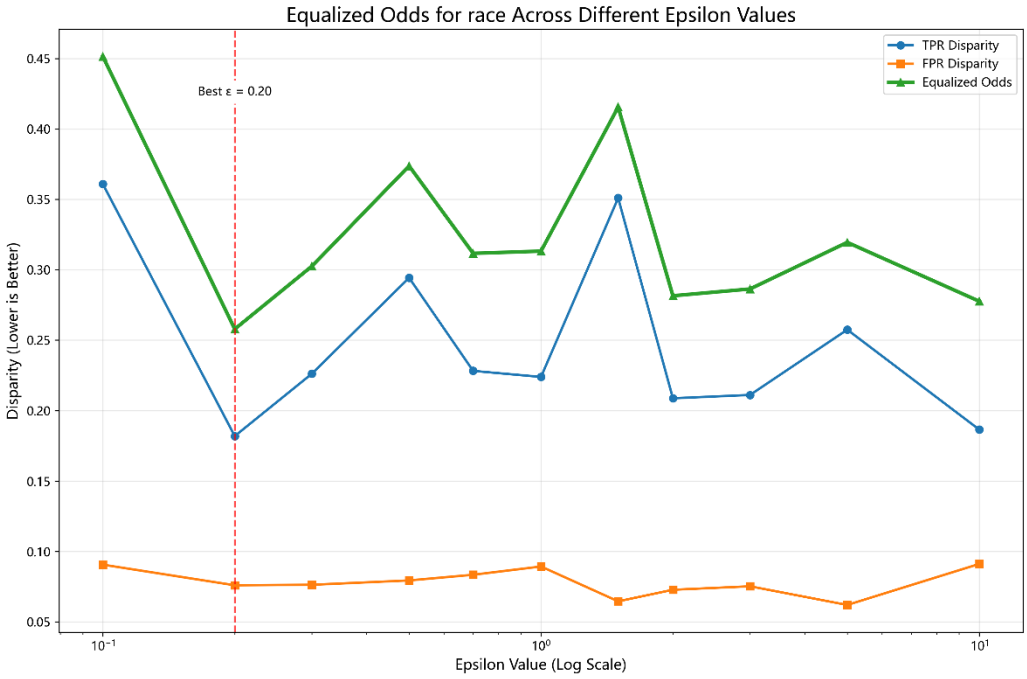


圖 4.6　不同 ε 值下的敏感屬性種族(race)的均等勝算指標

圖 4.6 揭示了種族公平性指標的高度波動特性。相較於性別群體的相對平穩變化，種族群體的公平性指標呈現出更複雜的非線性變化模式。

最顯著的特徵是 ε=1.5 處的急劇惡化現象，此時均等勝算差異值飆升至 0.416，成為全域最差點。這種異常變化可能反映了差分隱私雜訊在特定參數配置下對少數群體統計特性的破壞性影響。相對地，ε=0.2 展現了最佳的公平性表現，各項指標都達到相對較低的水準。

doi:10.6342/NTU202600729

為更深入理解這些公平性變化的內在機制，接下來我們進一步拆解 TPR 與 FPR 的個別趨勢。圖 4.7 至圖 4.10 分別展示不同群體在各隱私參數下的 TPR 與 FPR 表現。

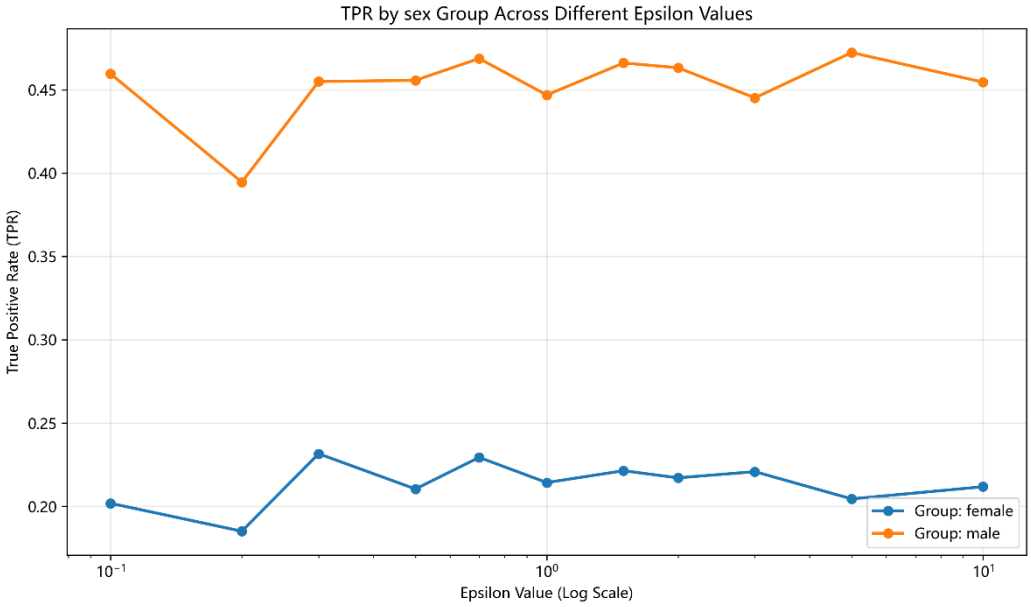


圖 4.7　性別在不同 ε 值下的 TPR 趨勢

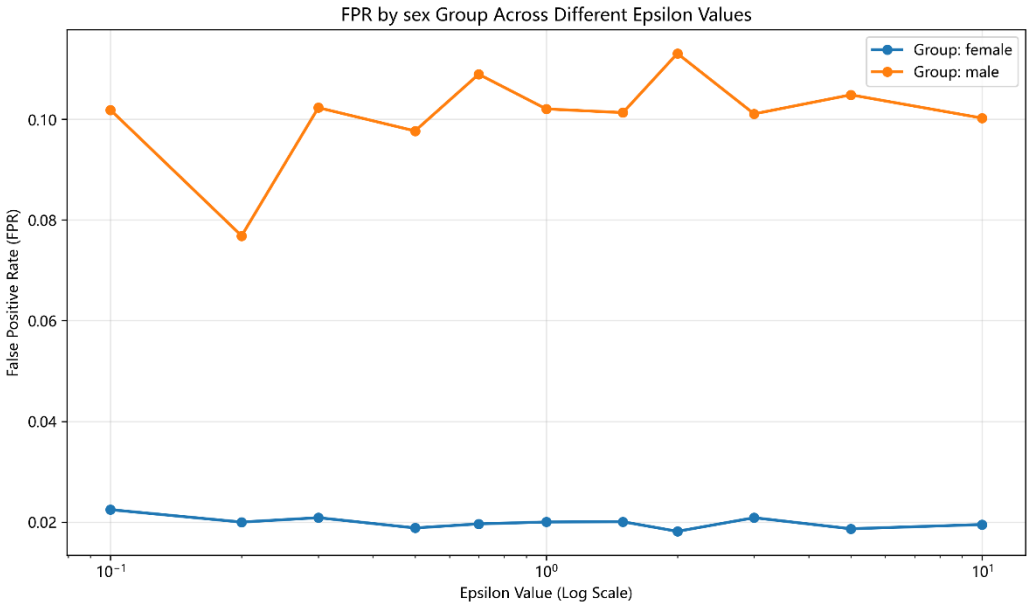


圖 4.8　性別在不同 ε 值下的 FPR 趨勢

從圖 4.7 和圖 4.8 的性別 TPR/FPR 趨勢可見，男性和女性的 TPR 差距在所有 ε 值下都相當顯著，平均約為 0.25（男性 TPR 約 0.45，女性約 0.20）。這種差距在 ε=0.2 時最小，但仍然明顯存在，反映出 Adult 資料集中固有的性別相關預測偏差。

doi:10.6342/NTU202600729

在 FPR 方面，男性的 FPR 值（約 0.10）顯著高於女性（約 0.02），但在 ε=0.2 時，這種差距達到最小，可能是因為在此 ε 值下，添加的雜訊恰好平衡了性別之間的 FPR 差異。

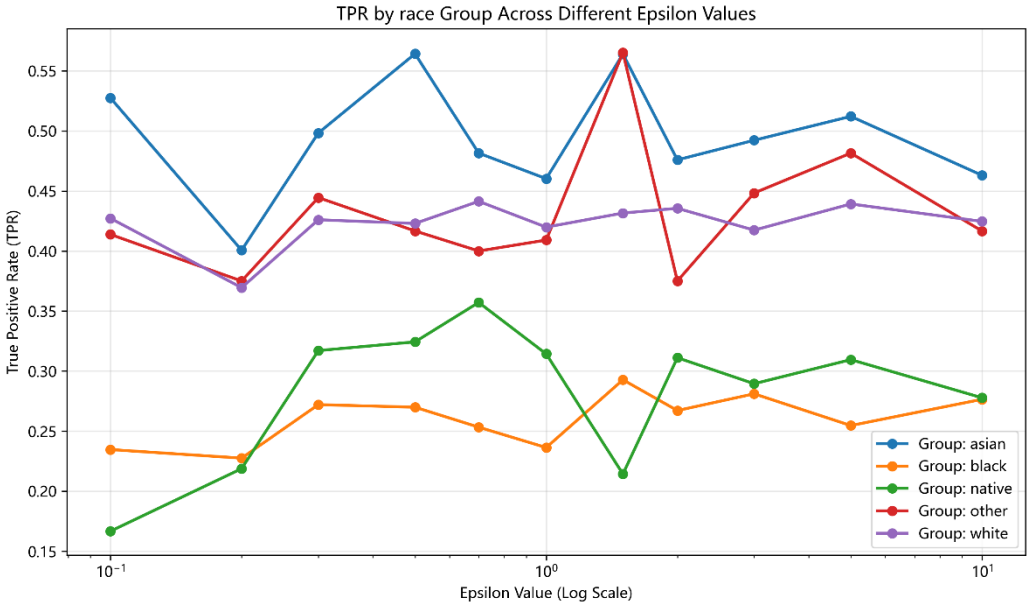


圖 4.9　種族在不同 ε 值下的 TPR 趨勢

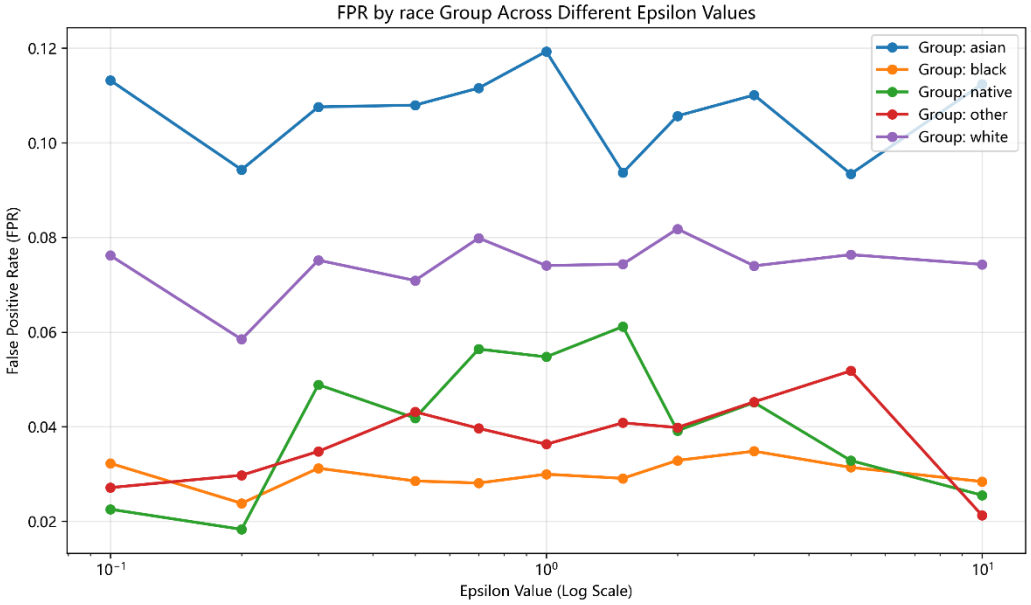


圖 4.10　種族在不同 ε 值下的 FPR 趨勢

從圖 4.9 和圖 4.10 的種族 TPR/FPR 趨勢可見，種族群組間的 TPR 差異更為複雜。亞裔群體的 TPR 普遍最高（約 0.40-0.55），而黑人群體的 TPR 最低（約 0.22-0.29）。在 ε=0.2 時，各群組的 TPR 曲線最為接近，差距最小。在 FPR 方面，

doi:10.6342/NTU202600729

亞裔群體的 FPR（約 0.09-0.12）明顯高於其他種族群體，而黑人群體的 FPR 最低（約 0.03）。此一差距在 ε=0.2 時達到最小。

綜合上述視覺化分析，我們觀察到三個關鍵發現：第一，種族公平性指標的波動幅度顯著大於性別公平性指標，顯示種族屬性對隱私參數變化更為敏感；第二，兩種敏感屬性均在 ε=0.2 時達到最佳公平性表現，但隨後呈現不同的變化軌跡；第三，TPR 差異是影響整體公平性的主要因素，特別是在種族群體中表現更為明顯。

這些觀察結果揭示了隱私保護機制對不同群體的差異化影響，但其背後的根本原因仍需進一步探討。下一節將深入分析造成這些群體差異的根本性因素。

4.4.3 群體差異的根本原因分析

前述視覺化分析揭示了隱私保護機制對不同群體的複雜影響模式，但這些現象背後的深層機制仍需進一步探討。透過深入分析實驗數據，本研究識別出造成公平性差異的四個根本性因素：

1. 資料分布不平衡的放大效應

Adult 資料集中的群體分布不平衡是造成公平性差異的基礎性因素。種族分布中白人約佔 85%，其他種族群體樣本量相對有限，這種不平衡在差分隱私處理後被進一步放大。當添加相同程度的雜訊時，大樣本群體（如白人）的統計特性因樣本充足而相對穩定，而小樣本群體的統計特性更容易受到隨機雜訊的擾動，導致其預測性能波動更大。

2. 差分隱私機制的群體敏感性差異

我們的分析顯示，差分隱私雜訊對不同群體的影響並非均勻分布。表 4.5 的數據表明，種族群體的 TPR 變異範圍（0.182-0.361）遠大於性別群體的變異範圍（0.210-0.268）。這反映了一個重要現象：差分隱私機制對群體大小更為敏感，而非群體類型本身。較小的群體由於統計不確定性較高，在雜訊添加過程中更容易產生不穩定的結果。

3. ε=1.5 異常值的分析

doi:10.6342/NTU202600729

ε=1.5 處的種族公平性異常惡化現象（EOD=0.416）值得特別關注。透過分析具體的 TPR 和 FPR 變化模式，我們發現此時亞裔群體的 TPR 從 ε=1.0 的約 0.42 急劇變化，而其他族群的表現也出現顯著波動，造成群體間差距擴大。

這種現象可能源於以下機制：在 ε=1.5 的特定雜訊水平下，差分隱私機制對不同群體的統計特性產生了不均等的影響。由於資料集中種族分布的高度不平衡，此 ε 值可能觸發了某種「統計共振」效應，使得少數群體的特徵模式在特定雜訊水平下變得極不穩定，導致預測性能的劇烈波動。

4. 統計學習偏差的傳播機制

原始資料中存在的系統性偏差（如收入與種族、性別的歷史相關性）在經過隱私保護處理後可能被保留甚至放大。當 ε 值較小時，大量雜訊掩蓋了這些固有偏差；當 ε 值較大時，原始偏差重新顯現；而在某些中等 ε 值範圍內，雜訊與既有偏差的相互作用可能產生意外的放大效應，這解釋了我們觀察到的非線性公平性變化模式。

4.4.4 實務啟示

這些分析結果為隱私保護技術的實際應用提供了重要啟示。

首先，隱私參數的選擇不能僅考慮整體的隱私-效用平衡，還需要系統性地評估其對不同群體的差異化影響。其次，對於群體分布高度不平衡的資料集，需要特別關注少數群體的公平性保護，可能需要採用群體感知的隱私保護策略。最後，ε=0.2 在本研究中表現出的良好公平性特性，為實務應用提供了參數選擇的參考依據，但具體應用時仍需根據資料特性進行調整。

這些觀察揭示了隱私保護與資料公平性之間複雜而微妙的關係。統計分析顯示，種族公平性指標在不同 ε 值間的變異明顯大於性別公平性指標，證實了種族屬性對隱私參數變化更為敏感的觀察結果。這些發現也為設計更具包容性的隱私保護機制提供了實證基礎。

doi:10.6342/NTU202600729

4.5 資料隱私性、資料效用性與資料公平性的綜合評估

在前一節（4.4 節）中，我們深入分析了隱私參數對資料公平性的影響，發現不同 ε 值對性別與種族群體的影響呈現明顯差異，特別是 ε=0.2 展現了最佳的公平性表現。然而，實務應用中的隱私參數選擇不能僅基於單一維度，而需要在隱私保護、資料效用與公平性之間取得平衡。本節將整合資料隱私性、資料效用性與資料公平性三個維度的指標，提出最佳隱私參數 ε 值的選擇依據與整合評分方法。透過系統性的綜合評估，期望為實務應用提供更全面的參數選擇指引。

為確保各維度間的可比性，我們建立以下標準化評分體系。由於本研究的實驗與架構用意是觀察出混合差分隱私能產生的狀況與現象，這些權重與算式在實際使用時應由使用者需求決定。

(1)資料隱私性評分：

包含兩個子指標的加權組合：

- *k* 匿名分數：(k – k-min)/(k-max – k-min)

  其中 k_min 和 k_max 分別為所有實驗中 k 值的最小值和最大值

- ε 值分數：1/(1 + ε)

  此公式設計使較小的 ε 值獲得較高分數，此公式採用雙曲線函數形式，主要基於差分隱私理論特性的考量，提供了不依賴實驗範圍的通用評分方式

- 隱私性評分 =*k* 匿名分數×0.6 + ε 值分數×0.4

  我們選擇 k 匿名分數權重略高於 ε 值分數（0.6 vs 0.4），主要基於三點考量：(1) k-匿名為本研究混合機制的基礎保護層，ε 為附加保護層，基礎層應有較高權重；(2) k 值具有直觀可解釋的物理意義（識別機率上界 1/k），在實務決策中更易溝通與驗證；(3) 實驗中 k 值的波動（2~5）直接反映雜訊對群組結構的影響，判別力較強。

(2)資料效用性評分：

基於總變異距離（TVD）的標準化：

doi:10.6342/NTU202600729

■ 效用性評分 = (TVD_max - TVD)/(TVD_max - TVD_min)

其中 TVD_max 和 TVD_min 為所有實驗中 TVD 值的最大值和最小值。較低的 TVD 值對應較高的效用性評分，反映資料分布與原始資料的相似程度。

註：本研究不將泛化損失納入效用性評分，因泛化損失為預處理階段的固定成本，與 ε 值選擇無直接關聯。

(3)資料公平性評分：

整合性別和種族兩個敏感屬性的均等勝算差異：

■ 性別公平分數 = 1 - (性別 EOD - EOD_min)/(EOD_max - EOD_min)

■ 種族公平分數 = 1 - (種族 EOD - EOD_min)/(EOD_max - EOD_min)

■ 公平性評分 = 性別公平分數 × 0.5 + 種族公平分數 × 0.5

較低的均等勝算差異值對應較高的公平性分數，兩個敏感屬性給予相同權重反映本研究對不同群體公平性的同等重視。

基於上述評分體系，表 4.6 呈現 11 個不同 ε 值下的原始指標及其標準化評分。整合評分採用權重配置為：隱私保護性 0.4、資料效用性 0.3、資料公平性 0.3，此配置將在後續的權重敏感度分析中進一步驗證其合理性。

表 4.6 不同隱私參數 ε 值的原始指標與標準化評分

原始指標						標準化評分			
ε 值	k 值	泛化損失	TVD 值	性別均等勝算差異值	種族均等勝算差異值	隱私性評分	效用性評分	公平性評分	整合評分
0.1	5	0.316	0.036	0.337	0.451	0.964	0.000	0.096	**0.414**
0.2	3	0.316	0.023	**0.266**	**0.258**	0.533	0.371	1.000	**0.625**
0.3	2	0.316	0.015	0.305	0.303	0.308	0.598	0.665	**0.502**
0.5	5	0.316	0.008	0.324	0.374	0.867	0.780	0.371	**0.692**
0.7	3	0.316	0.006	0.329	0.312	0.435	0.848	0.506	**0.580**

doi:10.6342/NTU202600729

1.0	5	0.316	0.004	0.315	0.313	0.800	0.899	0.582	**0.764**
1.5	5	0.316	0.003	0.326	0.416	0.760	0.930	0.253	**0.659**
2.0	3	0.316	0.002	0.341	0.281	0.333	0.944	0.513	**0.570**
3.0	2	0.316	0.001	0.305	0.286	0.100	0.971	0.709	**0.544**
5.0	3	0.316	0.001	0.354	0.319	0.267	0.986	0.341	**0.505**
10.0	3	0.316	0.000	0.324	0.278	0.236	1.000	0.623	**0.581**

在整合三個維度的評分時，權重的選擇直接影響最終的參數推薦結果。為確保權重分配的客觀性與科學性，我們進行了權重敏感度分析，測試不同權重組合對最佳隱私參數選擇的影響。在進行權重敏感度分析之前，我們先透過視覺化方法呈現各個 ε 值在三個維度上的表現特徵。

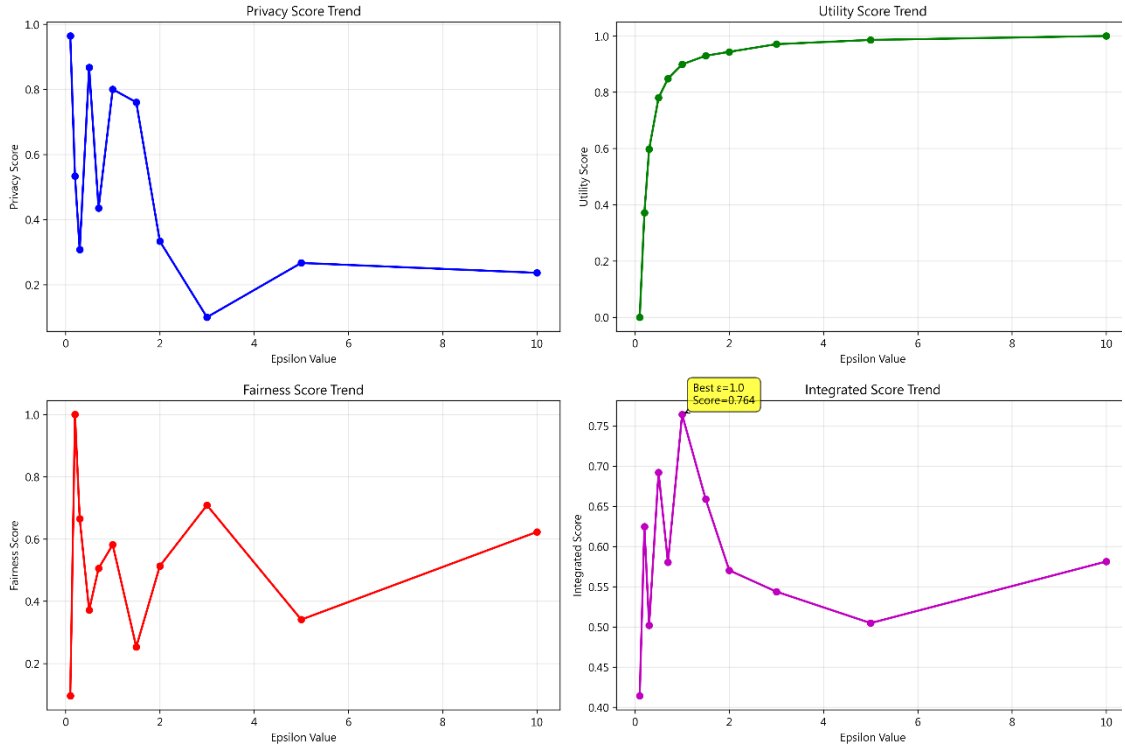


圖 4.11　不同隱私參數 ε 值下的評分值表現

圖 4.11 呈現四個評分維度隨 ε 值的變化趨勢，揭示以下關鍵模式：

doi:10.6342/NTU202600729

隱私性評分（藍線）在 ε=0.1 達到最高值約 0.96，在 ε=0.5 和 ε=1.0 仍維持較高分數（約 0.87 和 0.80），這些點位的 k 值均為 5。隨 ε 值繼續增加，評分逐漸下降，反映差分隱私理論保證的減弱。

效用性評分（綠線）呈現單調增長趨勢，從 ε=0.1 的接近零快速上升，在 ε=1.0 達到約 0.90，之後趨於平穩接近 1.0。這符合理論預期：較大的 ε 值對應較小的雜訊添加，能更好地保留原始資料特性。

公平性評分（紅線）呈現顯著波動，在 ε=0.2 達到最高值 1.0，但在 ε=1.5 驟降至約 0.25。這與 4.4 節的發現一致：差分隱私機制對不同群體的影響極為複雜，特別是對分布高度不平衡的種族屬性。

整合評分（紫線）在 ε=1.0 達到最高值 0.764，此時隱私評分接近 0.80、效用評分達 0.90、公平性評分為 0.58，在三個維度間取得良好平衡。

為更直觀地呈現不同 ε 值在三個維度上的平衡特性，圖 4.12 以雷達圖呈現六個代表性 ε 值的表現差異。每個 ε 值形成一個三角形，其形狀反映該參數在三個維度間的平衡程度：三角形越接近等邊三角形且面積適中，表示該參數在三個維度達到更好的平衡。觀察結果顯示，ε=0.1 在隱私保護性上表現優異（約 0.96），但效用性幾乎為零，形成極度扁平的三角形；ε=10.0 則在效用性上接近完美（約 1.0），但隱私保護性相對較弱（約 0.24），同樣呈現不平衡的狀態。相較之下， ε=1.0 在三個維度形成最為均勻的三角形，隱私、效用、公平三個維度的評分分別約為 0.80、0.90、0.58，顯示其作為平衡型參數的優勢。

doi:10.6342/NTU202600729

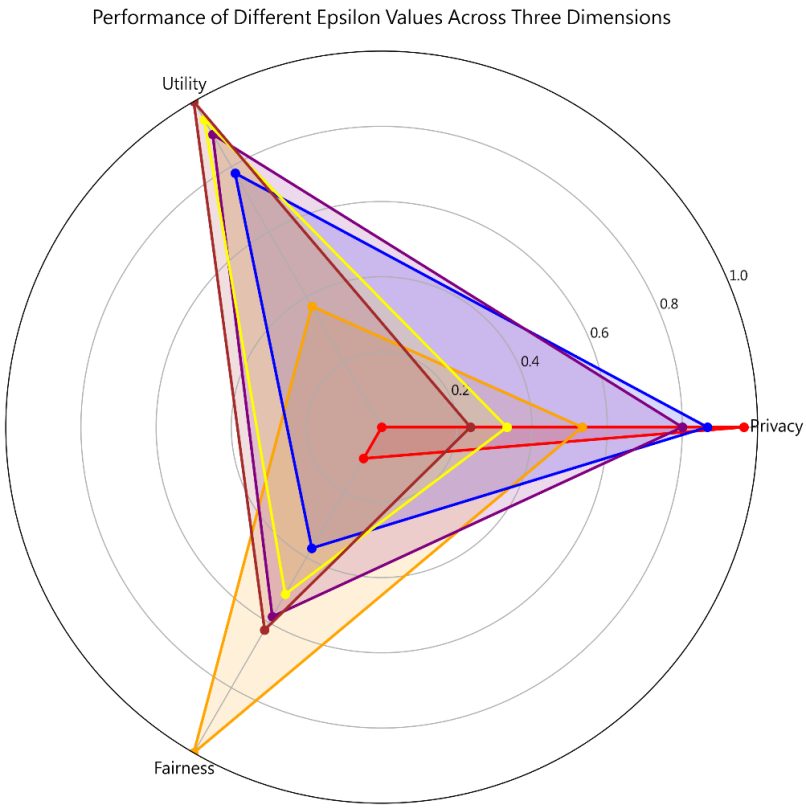


圖 4.12　不同隱私參數 ε 值在三個評估維度上的雷達圖比較

在確認各個 ε 值的個別表現特徵後，接下來透過權重敏感度分析，系統性地評估不同權重組合對參數選擇的影響。表 4.7 呈現了四種不同權重配置下的敏感度分析結果。

表 4.7 權重敏感度分析結果

權重組合	隱私性權重	效用性權重	公平性權重	最佳 ε 值	最佳 k 值	整合分數
隱私優先	0.60	0.20	0.20	1.0	5	0.776
效用優先	0.20	0.60	0.20	1.0	5	0.816
公平優先	0.20	0.20	0.60	0.2	3	0.781
均衡權重	0.40	0.30	0.30	1.0	5	0.764

表 4.7 的權重敏感度分析揭示幾個重要發現：

(1) ε=1.0 展現良好的穩健性。在四種權重配置中，有三種（均衡、隱私優先、效用優先）都推薦 ε=1.0 作為最佳參數，此參數設定能夠在提供強健隱私保護機制

doi:10.6342/NTU202600729

（k 值達到最高水平 5）的同時，最大化資料效用性。僅在公平優先配置下，系統才推薦 ε=0.2，此時均等勝算指標達到最佳表現，反映了較低 ε 值對維持資料公平性的積極作用。

(2) 權重配置對整合評分的影響顯著。效用優先配置下的整合評分最高（0.816），這是因為 ε=1.0 在效用性維度表現優異（0.90），而該配置給予效用性最高權重（0.6）。公平優先配置推薦的 ε=0.2 雖獲得 0.781 的整合評分，但其效用性僅 0.37，在資料可用性方面存在明顯短板。均衡權重配置下的整合評分為 0.764，雖非最高，但在三個維度都維持可接受的表現，不會過度犧牲任何單一面向。

(3) 權重選擇具有明顯的情境依賴性。不同應用場景應根據實際需求調整權重分配：

- 金融、醫療等高敏感領域可能更重視隱私保護性，建議提高隱私權重至 0.6
- 學術研究或資料分析可能更關注資料效用性，可提高效用權重至 0.6
- 社會公平敏感應用則應提高公平性權重至 0.6
- 一般商業應用可採用均衡權重（0.4/0.3/0.3）

根據我們的敏感度分析結果，本研究採用均衡配置（0.4/0.3/0.3）作為主要權重設定，此選擇主要基於三項重要發現：第一，在此配置下最佳參數 ε=1.0 與隱私優先情境相同，能夠提供足夠的隱私保護；第二，這組權重在整合評分中達到 0.764 的優良成績；第三，從圖 4.11 的視覺化結果可以清楚看出，ε=1.0 的表現在三個維度之間形成最平衡的分布。

圖 4.13 進一步視覺化權重敏感度分析結果。上半部分的堆疊柱狀圖顯示四種權重配置的結構差異，右側柱狀圖呈現各配置下的最佳 ε 值及整合評分。結果顯示，前三種配置均推薦 ε=1.0（整合評分分別為 0.764、0.776 和 0.816），僅公平優先配置推薦 ε=0.2（整合評分 0.781）。下半部分的比較圖清楚展示：當採用均衡權重、隱私優先或效用優先配置時，ε=1.0 在三個維度的評分都維持在 0.5 以上，形成良好平衡，無明顯短板。相較之下，公平優先配置推薦的 ε=0.2 雖在公平性維度

doi:10.6342/NTU202600729

突出（約 0.9），但在隱私性（約 0.52）和效用性（約 0.37）方面相對較弱，形成明顯的權衡取捨。這項敏感度分析證實：ε=1.0 在多數應用情境下能提供均衡的綜合表現，為實務應用提供了參數選擇的客觀依據。

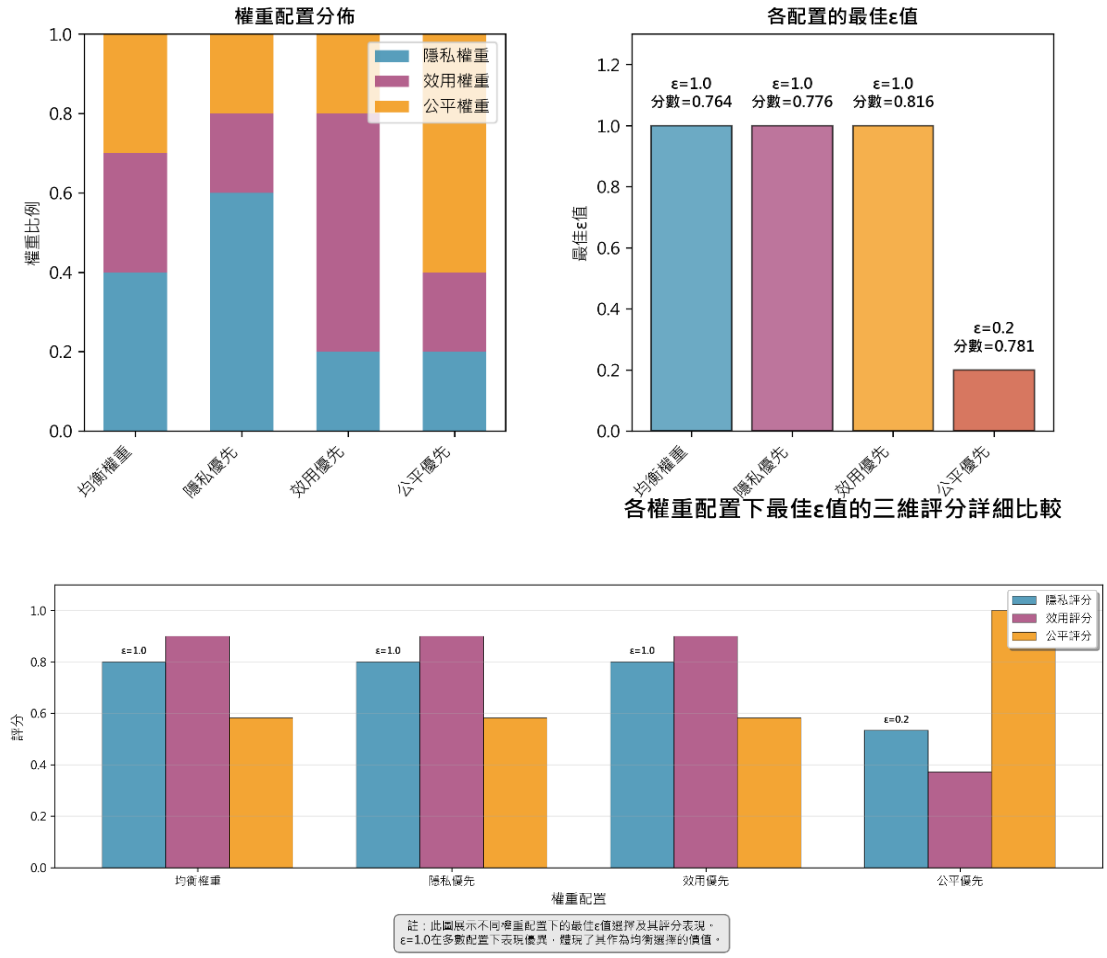


圖 4.13　權重敏感度分析結果與最佳隱私參數比較

根據本節的綜合評估，我們採用均衡配置（0.4/0.3/0.3）作為主要權重設定，此選擇基於三項重要考量：第一，在此配置下最佳參數 ε=1.0 與隱私優先情境相同，能提供足夠的隱私保護；第二，整合評分達 0.764 的優良成績；第三，從圖 4.12 和圖 4.13 可見，ε=1.0 在三個維度形成最平衡的分布。需要強調的是，權重配置（0.4/0.3/0.3）雖經敏感度分析驗證，但仍包含研究者的價值判斷。實際應用中，不同領域和組織應根據特定需求調整權重分配。此評估框架具有良好彈性，適用於

doi:10.6342/NTU202600729

多數普遍應用場景，未來研究可根據具體需求進一步調整權重，以滿足特定領域的隱私保護需求。

本節建立了整合隱私性、效用性與公平性的評估框架，透過標準化評分與權重敏感度分析，系統性評估了不同 ε 值的綜合表現。主要發現包括：

(1) 推薦參數：ε=1.0 在均衡權重配置下獲得最高整合評分（0.764），在三個維度間取得良好平衡。

(2) 穩健性驗證：ε=1.0 在三種權重配置下（均衡、隱私優先、效用優先）都被推薦為最佳參數，顯示其適用性廣泛。

(3) 特殊情境考量：當公平性為首要考量時，ε=0.2 提供最佳表現，但需接受效用性的顯著損失。

(4) 實務應用指引：

- 一般應用場景推薦 ε=1.0
- 高隱私需求場景可考慮 ε=0.5
- 高公平性需求場景建議 ε=0.2（需評估效用損失的可接受度）
- 高效用需求場景可考慮 ε=2.0-3.0（但需加強其他隱私保護措施）

需要說明的是，本節所有評估結果均基於單次實驗的點估計，並未進行多次重複實驗的統計檢驗。因此，各項評分和排序應理解為在當前實驗條件下的觀察結果，而非具有統計顯著性的推論結論。實際應用時建議進行多次實驗驗證。此評估框架具有良好的可擴展性，當面對不同的資料集或應用需求時，可以透過調整權重配置來反映特定場景的優先順序，為隱私參數的選擇提供更符合實務需求的決策依據。

4.6 與其他隱私保護方法的比較分析

為了驗證本研究提出框架的有效性，本研究進行了對照實驗，系統性地比較二種具代表性的隱私保護方法。這些方法各代表不同技術路線的隱私保護策略，使我們能夠全面評估各種隱私保護機制在資料隱私性、效用性及公平性三個關鍵維度的表現差異。

doi:10.6342/NTU202600729

4.6.1 比較方法說明

在實驗中我們選擇與其他二種具有代表性的隱私保護方法進行比較：

第一種方法採用純 *K*-匿名技術，設定 *K* 值為 5，代表傳統的隱私保護策略[2,11]。*K*-匿名是資料匿名化領域中最基礎也最容易理解的方法，採用傳統的泛化和抑制技術，確保任一準識別符組合至少對應 K 個記錄，從而防止個體被唯一識別。

第二種方法則使用純差分隱私機制，採用拉普拉斯機制實現 ε=1.0 的差分隱私保護[4,5]。差分隱私具有嚴格的理論隱私保證，透過向資料中添加精心校準的隨機雜訊，確保單一記錄的存在或不存在對統計結果影響有限，從而保護個體隱私。

第三種方法是本研究提出的多維度平衡框架，透過前述第 4.5 節的評估方法產製的資料集結合 *K*-匿名(*K*=5)與差分隱私(ε=1.0)的優勢，並在隱私保護、資料效用和資料公平性三個維度間尋求最佳平衡點。

4.6.2 比較結果分析

分析表 4.8 的實驗結果，我們可以看到在資料隱私性方面，純 K-匿名方法和本研究的方法都實現了 *K*=5 的匿名性保證，提供了群組層面的隱私保護；而純差分隱私方法和本研究的方法都達到了 ε=1.0 的差分隱私保護，確保了記錄層面的理論隱私保證。本研究提出的方法獨特之處在於同時滿足了這兩種隱私標準，為資料提供了多層次的保護機制。

表 4.8 不同隱私保護方法的綜合表現比較

評估維度	方法 1： 純 *K*-匿名（*K*=5）	方法 2： 純差分隱私 （ε=1.0）	方法 3： 本文的方法
資料隱私性	*K* 值=5	--	*K* 值=5
	--	ε=1.0	ε=1.0

doi:10.6342/NTU202600729

資料效用性	泛化損失=0.316	泛化損失=0.316	泛化損失=0.316
	抑制損失= 0.000092	TVD=0.004	TVD=0.004
資料公平性	性別均等勝算 差異值=0.322	性別均等勝算 差異值=0.308	性別均等勝算 差異值=0.315
	種族均等勝算 差異值=0.294	種族均等勝算 差異值=0.392	種族均等勝算 差異值=0.313

在資料效用性方面，三種方法的泛化損失值均相同(0.316)，表明基礎的泛化處理對資料結構的影響相對一致。純 K-匿名方法的抑制損失極小(0.000092)，顯示在泛化後的資料基礎上，需要抑制的小群組數量很少。相較之下，純差分隱私方法和本研究方法的 TVD 值均為 0.004，反映了群組計數差分隱私機制對資料分佈的影響程度。值得注意的是，本研究的混合方法在提供雙重隱私保護的同時，並未顯著增加資訊損失，證明了 *K*-匿名與差分隱私技術的良好互補性。

最值得關注的是資料公平性方面的比較結果。純 K-匿名方法的性別均等勝算差異值為 0.322，種族均等勝算差異值為 0.294；純差分隱私方法的性別均等勝算差異值較佳(0.308)，但種族均等勝算差異值較差(0.392)；而本研究提出的方法在兩個維度都達到了較佳的平衡表現：性別均等勝算差異值為 0.315，種族均等勝算差異值為 0.313。相較於純差分隱私方法，本研究方法在種族公平性方面改善了約 20.1% [(0.392-0.313)/0.392]，同時維持了相當的性別公平性水準。

綜合三個維度的表現，本研究提出的方法展現出明顯的優勢：它不僅在隱私保護上同時滿足 K-匿名[2]和差分隱私[4]的要求，確保了多層次的安全防護，還在保持良好資料效用的同時，有效改善了對少數族群的公平性處理。這一結果驗證了本研究框架的有效性—透過整合不同隱私保護技術並考量其對公平性的影響，我們能夠在各個關鍵維度上取得更佳的平衡。

這項比較分析凸顯了在設計隱私保護機制時考量多維度指標的實用價值。傳統隱私保護研究往往專注於隱私性和效用性之間的平衡[5, 11]，較少討論保護機制對不同社會群體的影響。本研究結果表明，同時兼顧隱私保護、資料效用和公平性

doi:10.6342/NTU202600729

維度的方法，能夠提供更為全面的資料保護解決方案，為隱私保護技術的綜合評估提供了可行的參考架構。

4.7 下游應用性能評估

在完成了隱私保護、資料效用和公平性的綜合比較之後，本節進一步評估不同隱私保護方法在實際下游應用中的表現。下游應用性能評估是驗證隱私保護資料集實用價值的重要指標，能夠反映經過隱私保護處理的資料在實際機器學習任務中是否仍能維持足夠的預測能力和統計特性。

4.7.1 評估方法設計

機器學習性能評估

我們採用收入預測分類任務作為下游應用的評估基準。選擇此任務的原因包括：(1) 收入預測是敏感資料分析的典型應用場景；(2) 該任務涉及多個人口統計學特徵，能夠全面測試隱私保護方法對不同資料類型的處理效果；(3) 分類準確率等指標具有直觀的解釋性，便於比較不同方法的表現。

評估流程設計如下：首先，在原始資料集上訓練邏輯迴歸模型作為基準模型，記錄其在測試集上的表現；然後，分別在三種隱私保護方法產生的資料集上訓練相同結構的模型；最後，將訓練好的模型在原始測試集上進行評估，計算各項性能指標。這種評估方式能夠直接反映隱私保護處理對模型實用性的影響程度。

統計特性保持度分析

除了機器學習性能外，本研究還評估隱私保護方法對資料統計特性的保持程度。透過卡方檢驗(Chi-square test)分析處理前後各類別變數的分佈相似性，量化隱私保護機制對原始資料分佈結構的影響。較高的分佈保持度表示隱私保護方法能夠在保護個人隱私的同時，更好地維持資料的原始統計特性。

4.7.2 實驗結果分析

doi:10.6342/NTU202600729

機器學習性能比較

表 4.9 展示了三種隱私保護方法在收入預測任務上的表現。實驗結果顯示，基準模型在原始資料上達到 76.01%的準確率，為後續比較提供了參考標準。

表 4.9 不同隱私保護方法的下游應用性能比較

方法	分類準確率	精確率	召回率	F1 分數	準確率保持率
純 *k*-匿名	0.7506	0.7577	0.9872	0.8573	0.9875
純差分隱私	0.7482	0.7570	0.9840	0.8557	0.9842
本研究方法	0.7532	0.7581	0.9911	0.8591	0.9909

從準確率保持率的角度分析，本研究提出的混合方法表現最優，達到 99.09%的保持率，顯著高於純 *k*-匿名方法的 98.75%和純差分隱私方法的 98.42%。這一結果表明，透過整合 *k*-匿名和差分隱私技術，能夠在提供雙重隱私保護的同時，更好地維持資料的預測能力。

在分類性能的各項指標中，本研究方法在分類準確率(75.32%)、精確率(75.81%)和 F1 分數(85.91%)方面均取得最佳表現。特別值得注意的是召回率指標，本研究方法達到 99.11%，表現優於其他兩種方法，這對於收入預測等敏感應用場景具有重要意義，因為較高的召回率意味著能夠更全面地識別目標群體。

性能差異原因分析

三種方法在下游應用中的性能差異可歸因於其不同的隱私保護機制：

純 *k*-匿名法透過泛化處理減少了資料的細緻度，雖然保持了基本的分類邊界，但在處理複雜特徵組合時可能損失部分判別資訊。純差分隱私方法雖然提供了嚴格的理論隱私保證，但隨機雜訊的引入對模型訓練造成了一定干擾，特別是在有限的資料集規模下，雜訊對學習效果的負面影響更為明顯。

相較之下，本研究的混合方法透過先進行 *k*-匿名處理形成穩定的匿名群組，再在群組層面應用差分隱私機制，這種分層保護策略有效減少了雜訊對個別記錄的直接影響。同時，*k*-匿名處理為差分隱私提供了更穩定的輸入基礎，使得整體處理流程在隱私保護強度與資料效用之間達到更佳的平衡點。

doi:10.6342/NTU202600729

統計特性保持度分析

在統計特性保持度方面，由於三種方法都採用了相對激進的隱私保護參數設置(k=5, ε=1.0)，並且處理後的資料集規模相對較小，導致與原始資料分佈存在顯著差異。這一結果反映了隱私保護強度與統計特性保持之間的平衡關係。

值得注意的是，儘管統計分佈發生了變化，三種方法在實際機器學習任務中都維持了相當高的預測性能（準確率保持率均超過 98%）。本研究的混合方法更是達到了 99.09%的準確率保持率，表現最優。這一發現表明，對於特定的應用任務，保持關鍵判別特徵的相對關係比維持精確的統計分佈更為重要。

4.7.3 實用性啟示

下游應用性能評估的結果為隱私保護技術的實際應用提供了重要啟示：

首先，混合隱私保護策略在維持資料實用性方面展現出明顯優勢。本研究方法能夠在提供多層次隱私保護的同時，最大程度地保持原始資料的預測能力，這對於需要在隱私保護與資料效用間取得平衡的實際應用場景具有重要參考價值。

其次，評估結果顯示，即使在相對嚴格的隱私保護參數下，適當設計的隱私保護機制仍能維持接近原始資料 99%的預測性能。這為實際部署隱私保護系統提供了信心，證明了在滿足隱私法規要求的同時維持業務價值的可行性。

最後，下游任務的性能評估應成為隱私保護方法設計和參數調整的重要考量因素。傳統的隱私保護研究往往專注於理論隱私保證或一般性的資料效用指標，但實際應用中更關心特定任務的表現。本研究的評估框架為隱私保護技術的實用性評估提供了可行的參考模式。

透過綜合 4.6 節的多維度比較和本節的下游應用評估，可以看出本研究提出的混合隱私保護框架在隱私性、效用性、公平性和實用性四個維度都取得了良好的平衡表現，為隱私保護資料發布提供了一個可行且有效的解決方案。

doi:10.6342/NTU202600729

第五章 結果與未來展望

5.1 結論

本研究系統性地將均等勝算指標整合進 k-匿名與差分隱私的混合評估框架，為隱私保護技術的多維度評估提供了一個可行的方法。整合 k-匿名和差分隱私技術，同時將公平性納入隱私保護評估的維度。透過系統性的實驗分析，我們得到以下四個主要成果：

一、建立了隱私保護技術的三維評估框架

相較於傳統僅關注隱私-效用平衡的方法，本研究首次系統性地將均等勝算指標納入評估體系，使隱私保護技術評估從「保護隱私」進化為「公平地保護隱私」，為隱私保護技術的社會責任評估提供了量化工具。

二、發現隱私參數的邊際效益遞減規律

實驗結果顯示，當 ε 值超過特定臨界值後，資料效用的邊際增加明顯減少。這一發現為隱私保護實務提供了參數選擇指引，避免為追求高資料效用而過度犧牲隱私保護的情況。

三、識別出差分隱私對群體公平性的複雜影響模式

研究發現隱私保護機制對不同敏感屬性群體的影響存在顯著差異，種族公平性指標的波動明顯大於性別公平性指標。特定隱私參數配置下可能出現群體公平性異常惡化現象，提醒實務應用中需要系統性地評估隱私保護對不同群體的差異化影響。

四、提出了平衡三維目標的參數選擇方法

透過綜合權重評估，本研究建立了在隱私保護性、資料效用性與資料公平性之間尋求最佳平衡的系統性方法，為複雜隱私保護情境提供了決策支援工具。這些研究成果為隱私保護技術的發展奠定了基礎，但仍有諸多方向值得深入探索。

doi:10.6342/NTU202600729

5.2 未來展望

本研究建立了整合 *k*-匿名與差分隱私技術的混合隱私保護框架，並將公平性納入評估維度。實驗結果顯示該框架在三個評估維度間能夠取得平衡，但仍存在一些限制需要在未來研究中加以改善。

5.2.1 研究限制

首先，在實驗範圍方面，本研究僅使用 Adult 資料集進行測試，該資料集的特定特徵（如種族分布高度不平衡）可能影響結果的普遍適用性。此外，採用單次實驗設計雖然確保了結果可重現性，但限制了統計推論的能力。

其次，在計算效能方面，隨著資料規模增加，本研究方法可能面臨運算上的挑戰，特別是處理大型資料集時的計算複雜度提升。

最後，在公平性評估方面，本研究主要採用均等勝算作為公平性度量，但公平性本身是多面向的概念，不同的公平性定義可能產生不同的評估結果。

5.2.2 未來研究方向

一、擴展實驗驗證

在更多元的資料集上進行測試，包括醫療、教育、金融等不同領域的資料，以驗證方法的普遍適用性進行多次重複實驗和統計檢驗，增強結果的統計可靠性測試方法在不同資料規模下的表現，評估計算效率的改善需求。

二、公平性指標擴展

比較不同公平性指標（如人口統計平等、預測值平等等）的表現差異分析不同公平性定義間的相互關係和潛在衝突根據應用場景特性開發適應性的公平性評估標準。

三、演算法優化改進

doi:10.6342/NTU202600729

開發更高效率的演算法實作，降低計算負擔探索群體感知的隱私保護策略，針對不同群體採用差異化處理建立智能化的參數選擇機制，協助使用者選擇最適當的隱私參數。

四、實際應用驗證

將框架應用於特定領域問題，如醫療資料隱私保護、金融服務公平性等開發使用者友善的工具介面，降低隱私保護技術的實施門檻評估方法在真實應用環境中的效果和可行性。

五、理論擴展研究

探索其他隱私保護技術（如同態加密、安全多方運算）與公平性的結合研究分散式環境中的隱私保護與公平性保證機制建立隱私保護技術的倫理評估框架。

5.2.3 結語

透過這些循序漸進的研究發展，我們期望本研究提出的框架能夠在實務應用中發揮效益，為混合隱私保護機制提供實證基礎，並為日趨複雜的資料隱私挑戰提供更全面的解決方案。未來的研究將持續關注隱私保護、資料效用和社會公平性之間的平衡，推動隱私保護技術向更具社會責任的方向發展。

doi:10.6342/NTU202600729

參考文獻

[1] 中華民國數位發展部 (2024). 隱私強化技術應用指引. 台北：數位發展部. 檢索自 https://moda.gov.tw/press/press-releases/6497

[2] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

[3] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 3-es.

[4] Dwork, C. (2006). Differential privacy. In International Colloquium on Automata, Languages, and Programming (pp. 1-12). Springer.

[5] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.

[6] Erlingsson, Ú., Pihur, V., & Korolova, A. (2014). RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In Proceedings of the 2014 ACM SIGSAC conference on computer and communications security (pp. 1054-1067).

[7] McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2018). Learning differentially private recurrent language models. In International Conference on Learning Representations (ICLR).

[8] Zhai, F., Liang, X., Qin, Y., Li, B., Shen, L., & Xie, J. (2024). Privacy-preserving method for sensitive partitions of electricity consumption data based on hybrid differential privacy and k-anonymity. *Journal of Physics: Conference Series*, 2806(1), 012010.

[9] Lin, Y., Fang, H., & Yang, P. (2022). A framework combining differential privacy and k-anonymity for distributed databases. Information Sciences, 589, 634-649.

[10] Kobayashi, R., Shishido, H., & Sugiyama, M. (2024). On the relationship between probabilistic k-anonymity and differential privacy. Journal of Privacy and Confidentiality, 14(2), 45-67.

[11] Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. In Proceedings of the 21st International Conference on Data Engineering (pp. 217-228). IEEE.

[12] Meyerson, A., & Williams, R. (2004). On the complexity of optimal k-anonymity. In Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 223-228).

[13] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.

[14] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-

doi:10.6342/NTU202600729

35.
[15] Cover, T. M., & Thomas, J. A. (2006). Elements of information theory. John Wiley & Sons.
[16] El Emam, K., & Dankar, F. K. (2008). Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association, 15(5), 627-637.
[17] Mendes, R., & Vilela, J. P. (2017). Privacy-preserving data mining: methods, metrics, and applications. IEEE Access, 5, 10562-10582.
[18] Kohavi, R., & Becker, B. (1996). Adult data set. UCI Machine Learning Repository.
[19] Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2023). Differential privacy has disparate impact on model accuracy. In Neural Information Processing Systems (Vol. 32).
[20] Tang, Y., Wang, K., Chen, Z., & Zhang, Y. (2022). Investigating the fairness impacts of differential privacy. In Proceedings of the ACM Web Conference 2022 (pp. 2284-2293).
[21] Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In STOC (pp. 169-178).
[22] Yao, A. C. (1982). Protocols for secure computations. In FOCS (pp. 160-164).
[23] Goodfellow, I., et al. (2014). Generative adversarial nets. In NIPS (pp. 2672-2680).
[24] de Oliveira, A. S., et al. (2023). An empirical analysis of fairness notions under differential privacy. arXiv preprint arXiv:2302.02910.
[25] Majeed, A., & Hwang, S. O. (2024). Differential privacy and k-anonymity-based privacy preserving data publishing scheme with minimal loss of statistical information. *IEEE Transactions on Computational Social Systems*, 11(3), 3753-3765.
[26] Bargh, M. S., & Choenni, S. (2022). Towards an integrated approach for preserving data utility, privacy and fairness. In *2022 International Conference on Multidisciplinary Research* (pp. 290-306).
[27]Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press. https://www.fairmlbook.org/
[28]金融監督管理委員會 (2024)。金融機構資料共享之資料治理諮詢文件。 取自 https://www.fsc.gov.tw/ch/home.jsp?id=96&parentpath=0,2&mcustomize=news_view.jsp&dataserno=202405160001&dtable=News
[29] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., ... & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft Technical Report MSR-TR-2020-32. https://fairlearn.org/
[30] Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., ... & Vadhan, S. (2018). Differential privacy: A primer for a non-technical audience. Vanderbilt Journal of Entertainment & Technology Law, 21(1), 209-276.
[31] Xiong, X., Liu, S., Li, D., Cai, Z., & Niu, X. (2025). Differential Privacy

doi:10.6342/NTU202600729

Configurations in the Real World: A Comparative Analysis. IEEE Transactions on Knowledge and Data Engineering. DOI: 10.1109/TKDE.2025.3603731

[32] Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (CSUR), 42(4), 1-53. https://doi.org/10.1145/1749603.1749605

[33] Apple Inc. (2017). Differential Privacy Overview. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

[34] Ding, B., Kulkarni, J., & Yekhanin, S. (2017). Collecting telemetry data privately. In Advances in Neural Information Processing Systems (pp. 3571-3580).

[35]Enhancing Data Privacy: A Comprehensive Survey of Privacy- Enabling Technologies. IEEE Access, 2024. https://ieeexplore.ieee.org/document/10908383

doi:10.6342/NTU202600729

附錄：詞彙表(glossary)

專有名詞	英文對照	定義
直接識別符	Direct Identifiers	可直接識別個人身份的資料欄位，如身份證號碼、護照號碼、電子郵件地址等。在隱私保護處理中通常首先被移除或加密。
準識別符	Quasi-identifiers (QI)	單獨無法直接識別個體，但結合其他資訊可能重新識別個體的屬性組合，如年齡、性別、郵遞區號等。
敏感屬性	Sensitive Attributes (SA)	包含隱私或機密資訊的資料欄位，揭露可能對個人造成傷害或歧視，如健康狀況、收入水準、宗教信仰等。
差分隱私	Differential Privacy ,DP	一種具有嚴格數學證明的隱私保護機制，透過在查詢結果中添加校準的隨機雜訊，確保單一記錄的存在與否對查詢結果的影響在統計上不可區分，從而保護個體隱私。其隱私保護強度由參數 ε 控制。
隱私參數 (ε 值)	Privacy Parameter ,ε	差分隱私機制中的核心參數，用於量化隱私保護程度與資料效用性之間的平衡。ε 值越小表示隱私保護越嚴格，但相應的雜訊越大，資料效用性越低；ε 值越大則相反。數學上，ε 控制相鄰資料集查詢結果概率比的上界。
k-匿名	*k*-anonymity	一種隱私保護方法，確保資料集中任何一筆記錄都至少與其他 *k*-1 筆記錄在準識別符組合上無法區分，使得重新識別個體的機率不超過 1/*k*。透過泛化和抑制技術實現，*k* 值越大表示隱私保護程度越高。
資訊損失	Information Loss	衡量隱私保護過程中原始資料與處理後資料的差異程度，包括泛化處理階段和差分隱私運算後的資訊損失。
總變異距離	Total Variation Distance ,TVD	衡量兩個概率分佈差異的指標，本研究用於量化隱私保護前後資料分佈的變化程度，作為資訊損失的度量標準。計算公式為：$TVD(P, Q) = 0.5 * \Sigma\|P(x) - Q(x)\|$ ，其中 P、Q 為兩個概率分佈。TVD 值越小表示資訊損失越少。

doi:10.6342/NTU202600729

均等勝算	Equalized Odds	一種公平性評估標準，要求預測模型對不同敏感屬性群體具有相同的真陽性率(TPR)和假陽性率(FPR)。數學定義為：對於任意敏感屬性值 a 和 b，以及真實標籤值 y，滿足 $P(\hat{Y}=1\|Y=y,A=a) = P(\hat{Y}=1\|Y=y,A=b)$。
均等勝算差異值	Equalized Odds Difference,EOD	量化不同群體間預測公平性的指標，計算為： (max(TPR) - min(TPR)) + (max(FPR) - min(FPR))。值越接近 0 表示公平性越高。
真陽性率	True Positive Rate ,TPR	正確預測為正例的實際正例比例，計算公式為：TPR = TP/(TP+FN)。在均等勝算中，要求不同群體有相同的 TPR，群體間的 TPR 差 異是計算均等勝算差異值的重要組成部分。
假陽性率	False Positive Rate ,FPR	錯誤預測為正例的實際負例比例，計算公式為：FPR = FP/(FP+TN)。在均等勝算中，要求不同群體有相同的 FPR，群體間的 FPR 差 異是計算均等勝算差異值的重要組成部分。
公平性分數	Fairness Score	基於均等勝算差異值計算的標準化指標，採用 Min-Max 標準化方法：公平性分數 = 1 - (EOD - EOD_min)/(EOD_max - EOD_min)其中 EOD 為均等勝算差異值，EOD_min 和 EOD_max 分別為所有實驗參數下的最小值和最大值。分數越接近 1 表示公平性越佳。
隱私保護性	Privacy Protection	隱私處理後資料保留隱私保護的程度。與 k-匿名中的 k 值成正比。
資料效用性	Data Utility	隱私處理後資料保留分析價值的程度。與資訊損失成反比。
資料公平性	Data Fairness	隱私處理後資料保留模型公平的程度。與公平性分數成正比，即公平性分數越高表示資料公平性越好。
整合評分	Integrated Score	綜合考量隱私保護性、資料效用性和資料公平性的總體評分，用於確定對應用情境最佳的隱私參數。

doi:10.6342/NTU202600729