隱私強化技術與自主權

王大為　　中央研究院資訊科學研究所

本文將從技術面解釋與健保資料相關之隱私強化機制，並說明技術現況，以供參考。

I.　背景

為達到隱私保障之目的，利用技術的方法降低資料集裡面的資料洩露個人隱私的風險，稱為隱私強化技術。筆者研究隱私強化技術二十幾年，曾經執行行政院衛生署八十八年下半年及八十九年度委託研究計畫「健保資料庫個人隱私保護機制研究」(DOH89-NH-032)以及正在執行衛福部「健保署一一零年度民眾資料自主權框架與醫療資源共享模式委辦案」(案號 C1100039499)。也曾經擔任醫學資訊學會理事長(2011-2015)以及中央研究院資訊服務處長(2010-2017)。對於資訊隱私強化技術、健康資料大數據的潛力以及資訊系統開發與維護所需的資源，皆有深刻的理解。同時擔任「衛生福利部衛生福利資料統計應用管理審議會」委員(1030702~1110701)因此對於行政管理方面也有足夠的經驗。以下將介紹隱私強化技術中的去識別化技術，說明去識別化強度與資料可用性之間的權衡。

II.　去識別化技術介紹

我們簡略的將隱私強化技術分為三種類型：第一種類型以資料集為標的，探詢的議題為如何處理資料集使得資料集中的個人資料無法「識別出個人」或「無法推論出某個人的某項私密資料」。K-匿名化(K-anonymity)就是此類型的例子。第二種類型是以演算模式為標的，探詢的是如何讓握有機密資料的雙方或多方可以共同完成演算得到結果，但過程與結果不會洩漏計算結果以外的資訊。這個領域稱為多方私密計算(multiparty private computation)。第三種類型是同時考慮資料集以及所要進行的運算或查詢，以 differential privacy 為這方面的代表。在本文中我們將聚焦於第一種類型的現況，原因是此為國內目前討論的重點。當我們討論去識別化(de-identification)也往往是以這個類型為基準。

K-匿名化是一個資料集的性質，一個資料集裡面每一個人的資料都可以發現有 K-1 個人與他的資料一模一樣。這是 K-匿名化的最簡單的定義。而所謂用 K-匿名化來做去識別化，意思是將一個資料集經過處理後可以產生一個 K-匿名化的資料集。最常見的處理方式為刪除資料或將資料模糊化。例如將 90 歲以上的人的資料刪除或是將生日取代為年齡。很明顯地如果刪去的資料越多或是模糊的程度越大，那麼所產出的滿足 K-匿名化的資料集的用處可能就越小。這就是隱私保護程度與資料效用(utility)之間的權衡(tradeoff)。

健保資料庫包含了許多個人醫療相關的資料，也正因為這樣詳細的資料可以透過資料分析產生重要洞見(insight)。但這樣的資料要完成產生出符合 K-匿名化的資料集且要有不錯的效用，這幾乎是不可能的任務。因此才改成資料中心的模式，在分析的過程中所使用的資料集並不符合 K-匿名化的要求，而是在資料攜出時候對於攜出資料做了限制。這樣的做法的確增加了隱私的保護，但是是否與 K-匿名化有相同的效果，似乎不是簡單的論證。

III. 去識別化技術在隱私保護的角色

首先引用美國總統科技顧問(president' s council of advisors on science and technology, PCAST)遞交的一份有關巨量資料與隱私的報告[1](Big Data and Privacy: A Technological Perspective, May 2014)中有關去識別化技術的描述：面對巨量高維度的資料，去識別化後資料被再識別(re-identification)的風險增加.且其面對未來新產生的分析技術與新出現的資料集會有始料未及的困難。因此「傳統的」去識別化技術要成為隱私保護的基石會有困難，但仍然可以提供多一層的保護。此處傳統的去識別化技術包含了K-匿名化。其原文如下：

"Anonymization remains somewhat useful as an added safeguard, but it is not robust against near-term future reidentification methods. PCAST does not see it as being a useful basis for policy. Unfortunately, anonymization is already rooted in the law, sometimes giving a false expectation of privacy where data

[1] REPORT TO THE PRESIDENT. BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE. Executive Office of the President. President's Council of Advisors on. Science and Technology. May 2014

lacking certain identifiers are deemed not to be personally identifiable information and therefore not covered by such laws as the Family Educational Rights and Privacy Act (FERPA)."

值得順帶一提的上述引文的最後一句話，認為很不幸地有些法律誤以為匿名化即可讓資料不再是個人資料而不在該法之範疇內。我們的個資法是否也落入此窠臼?值得探究。

Ohm 一篇頗受矚目的論文 "Broken Promises of Privacy: Responding to The Surprising Failure of Anonymization"[2]引發了一連串對於去識別化技術的討論與辯論。我以 2014 年的筆戰作為此論辯的縮影，在 Cavoukia 的文章，"Setting the Record Straight: De-Identification Does Work"[3]中認為大多數的 re-identification 事件都是因為 de-identification 做得不好或沒有依照標準指引來做而造成的，並提出許多媒體的報導多誇大或誤解了學術論文中的敘述。並引用 Narayanan 等人所寫的 "Robust de-anonymization of large sparse datasets"[4]對 Netflix 公開資料所做的 re-identification 作為例子。同時認為大多數人沒有進行複雜再識別演算的能力 Narayanan 等人則寫了 "No silver bullet: De-identification still doesn't work"[5]一文回應，其主要論述為進行再識別者所擁有的輔助資料很難在進行去識別化時候做一個合理的界定，且認為再識別風險的評估需要許多假設頂多只能算是經驗式的論證(heuristic)而非正規方法的證明(formal method)。另外也指出擁有再識別所需的軟體能力者可能有百萬之譜。文中提到了 Heritage Health Prize 的公開資料集當初請他

[2] Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization (August 13, 2009). UCLA Law Review, Vol. 57, p. 1701, 2010; U of Colorado Law Legal Studies Research Paper No. 9-12. Available at SSRN: https://ssrn.com/abstract=1450006

[3] Ann Cavoukian and Daniel Castro, Big Data and Innovation, Setting the Record Straight: De-identification Does Work June, 16 2014
http://www2.itif.org/2014-big-data-deidentification.pdf

[4] Arvind Narayanan and Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets," Proceedings of the 2008 IEEE Symposium on Security and Privacy, 2008, http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

[5] Arvind Narayanan and Edward W. Felten, No silver bullet: De-identification still doesn't work July, 9 2014

進行再識別評估作為例子。這引起了負責 Heritage Health Prize 去識別化的 Emam 為文回應[6]。

我舉這個例子想說明幾件事：1.去識別化的效果在學術界仍有不同的看法；2.去識別化還沒有完整的數學模式；3.若以去識別化為基礎認為不再是個人的資料故無隱私問題需相當謹慎；4.去識別化仍有其價值但須嚴謹且明確說明所做的假設。

從 2014 年至今，去識別與再識別的技術有許多的進展，但概括來說前述的看法仍然是正確的。另外近來再識別的研究指出即使是抽樣的資料，只要資料夠多人數夠多仍然會有不可忽略的風險能夠再識別個人[7]。

IV 個資風險分析範例

對於個資經過去識別化後的風險分析是許誤解的來源。健保署也可能做了類似的分析，但因為筆者並未得到相關資訊，因此利用下面的例子來說明之[8]。

2011 年四月，加州的醫療照護組織 Heritage Provider Network(NPH)舉辦了一個健康資料分析競賽，目的是建構模型來預測病人明年住院的天數。所提供的資料為去前年去年與今年的申報資料(claim data)。資料集包括了十三萬三千名病人。

作者提出了三種可能的再識別手段分別是愛八卦鄰居、與選舉人登記資料(voter' s registration data)比對，以及與 state in patient data 比對。這三種模式的重點是限制攻擊者的背景資訊，也就是除了公開的資料集以外可能蒐集到的資訊。在分析去識別化資料是否已經足以保護個人，如何處理背景資訊（有些文獻稱為輔助資料，auxiliary data）。最後的分析結果發現約 0.84%的人遇到八卦的鄰居可能會被識別出來，而其他兩個攻擊的成功率都極低。

---

[6] Khaled El Emam and Luk Arbuckle Why de-identification is a key solution for sharing data responsibly July, 24 2014

[7] Rocher, L., Hendrickx, J.M. & de Montjoye, YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communication 10, 3069 (2019). https://doi.org/10.1038/s41467-019-10933-3

[8] El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Rose S, Howard J, Gluck J De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset J Med Internet Res 2012;14(1):e33

接下來我們概述一下另外一位接受邀請擔任攻擊腳色嘗試再識別的分析報告[9]。在他的分析中得到如下的結論：

"I performed my own re-identification analysis with a slightly different set of assumptions, one that seems at least as realistic to me as the original. In particular, I assumed that the adversary knows the year but not the month or day of each visit. I was then able to show that one would derive dramatically different re-identification probabilities — up to 12.5% of members"

12.5% 和 084% 有著相當大的差距，而得到天壤之別的結論的原因為 "with slightly different set of assumptions"。因此爭辯這兩個結論的對錯實際上要爭辯的是雙方誰所做的假設比較「合理」。

從以上的資訊，我們可以發現去識別化過程與其風險分析並沒有一致的標準，而且不同的假設可能可以導致相當不同的結果。以保密為由而使去識別化方法與風險分析無法由第三方檢視，是相當不合適的作法。此外利用風險認定都有爭議的方法來論述這些資料無從識別個人因此不再視為個資，並不恰當。

V 尊重自主的技術

在過去的討論中，常聽到一個論述為尊重個人自主權的機制可能行政成本太高。在此提出利用提供選擇退出(option-out)的機制，可以達到尊重個人自主；又能讓社會在以資料為核心的競爭中保有競爭力。簡單的說選擇退出機制就是一種意願表達的機制，目前疫苗注射登記系統也有相同的功能。因此建構這樣的機制來成就台灣成為更尊重個人的社會，實在是相當划算的事情。而最需要心力與資源的是與人民溝通資料的重要性，以及個人資料將如何受到保護。如此一來去識別化技術可以做為增加的保障，而不需要勉強做為已經無法識別個人因此非個資的論述基礎的技術基礎。

VI 結論

[9] Arvind Narayanan An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset http://randomwalker.info/publications/heritage-health-re-identifiability.pdf

去識別化與資料效用有權衡關係，大量且細緻的資料要做到可接受的去識別化程度並同時保有資料效用，幾乎不可能。全民健保為強制保險，要讓去識別化成為論證非個人資料的基礎，如何進行與效用的權衡，定非能由一單位在不透明的情形下單獨為之。去識別的風險評估可因些許假設的差異而得到不同的結果，完成可受公評的評估所需的資源與人力，恐遠大於提供選擇退出機制來表達對個人自主權的尊重。以上為筆者的意見，也是建構民眾資料自主權框架與醫療資源共享模式思考的起點。

附錄：

個資風險分析細節資訊
資料集中包含了下列欄位：

MemberID, Age, Sex Patient's sex, DaysInHospital Y2, DaysInHospital Y3

每一筆的申報資料則包含了：

MemberID, ProviderID, PCP: Unique identifier for the primary care provider

Year: Indicator of claim year (year 1, year 2, or year 3)

Specialty: Specialty of provider

PlaceOfService: Place of service

CPTCode: Current Procedural Terminology these codes provide a means to accurately describe medical, surgical, and diagnostic services, are used for processing claims and for medical review, and are the national coding standard under HIPAA

LOS: Length of stay in hospital

DSFC: Number of days since first claim computed from the first claim for that patient for each year

PayDelay: Number of days of delay between date of service and date of payment of the claim

Diagnosis: ICD-9-CM code

進行隱私強化的步驟如下

首先進行一些基礎的步驟：

1.產生代碼 (create Psuedonyms)

The MemberID, ProviderID, Vendor, and PCP fields 這幾個欄位都轉換成不可逆的代碼 pseudonyms。

2.Top-coding 將少數太高或太低的值轉換成一個區間，例如 95 歲以上。雖然常見的經驗法則是切在 99.5%但為了保險將 PayDelay, DaysInHospital 這兩個變數切在 99%，也就是超過 99%的值就會是「X 以上」。

3.刪除一些申報資料

若某個人的申報資料太多，則很容易被識別出來.這裡是以 95%為閥值。也就是有百分之五的人的申報資料會被刪除一些，作者也提供了一個計分方式刪除分數最高的那幾筆申報資料。

4.移除有高機敏資料的個人，例如 human immunodeficiency virus infection 的病人、墮胎病人，以及罕見且容易看出來的病人。

5.Suppression of Provider, and PCP Identifiers, 有些醫院很容易識別出來，有些醫院很容易猜到去那裏的患者是什麼毛病。

經過去識別化後有些欄位的值被模糊化了(generalized)。例如：年齡以十年為級距，超過八十歲的紀錄成 80+，每年住院總天數紀錄為兩周以內及兩周以上。Sepciality, PlaceOfService, CTPCode,Diagnoisis 都被以群組的方式記錄。

在做再識別攻擊演練之前，作者做了如下的假設

Fact: The dataset that was being released for the HHP consisted of a small sample of all HPN patients.

Fact: All entrants in the competition had to sign (or click through) an agreement saying that they would not attempt to re-

identify patients in the dataset, contact patients, or link the HHP data with other datasets that would add demographic, socioeconomic, or clinical data about the patients (where such data could make the risk of re-identification much higher).

Assumption: It would not be possible for an adversary to know whether the record for a particular patient was in the HHP dataset. If an adversary made a guess, it would be equal to the sampling fraction. Most patients would themselves not know whether they were members of HPN, and therefore the most realistic sampling fraction to use would be from the population of counties in California covered by HPN. However, to err on the conservative side, we assumed that an adversary would know whether a patient was a member of HPN in our calculations of re-identification risk.

Assumption: An adversary would have background information about only a subset of the claims of a patient in the dataset. For example, if a patient had 100 claims, we did not deem it plausible for the adversary to know the exact information in all of those 100 claims and to use that information for re-identification purposes. Rather, we assumed the adversary would have information about only a subset of these claims. This has previously been referred to as the power of the adversary, and various methods have been used to account for power when de-identifying transactional data