PRIVACY PRESERVING DATA PUBLISHING - GR

Digital security and privacy

Georgios Spathoulas

Msc in ”Informatics and computational bio-medicine”

University of Thessaly

DATA FORMAT

Στην βασική μορφή του PPPDP ο εκδότης του πίνακα έχει έναν πίνακα της μορφής:

D(Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes), όπου :

- **Explicit Identifier** Το σύνολο των πεδίων, όπως όνομα ή αριθμός κοινωνικής ασφάλισης που περιέχουν πληροφορίες μέσω των οποίων μπορεί να αναγνωριστεί το άτομο που σχετίζεται με την εγγραφή του πίνακα
- **Quasi Identifier (QID)** Το σύνολο των πεδίων που ενδεχομένως μπορεί να αποκαλύψουν το άτομο που σχετίζεται με τα δεδομένα
- **Sensitive Attributes** Τα πεδία που περιέχουν ευαίσθητα προσωπικά δεδομένα, όπως ασθένεια, μισθός ή αναπηρία
- και **Non-Sensitive Attributes** Όλα τα υπόλοιπα πεδία που δεν σχετίζονται με τις παραπάνω κατηγορίες

ΑΠΟΤΡΟΠΗ LINKING ATTACKS

Για να αποτραπούν πιθανά linking attacks, ο εκδότης εκδίδει ένα ανωνυμοποιημένο πίνακα

T (QID' , Sensitive Attributes, Non-Sensitive Attributes)

- **QID'** πρόκειται για μία ανωνυμοποιημένη μορφή του αρχικού **QID** που παράγεται εκτελώντας εργασίες ανωνυμοποιήσης στα πεδία του **QID** στον αρχικό πίνακα D
- Οι εργασίες αυτές έχουν ως στόχο να αποκρύψουν μέρος της πληροφορίας, ώστε αρκετές εγγραφές να μην είναι δυνατό να διαχωριστούν μεταξύ τους
- Συνεπώς εάν κάποιος συνδεθεί με μία εγγραφή του πίνακα, συνδέεται επίσης και με όλες τις άλλες εγγραφές που έχουν ίδιες τιμές στο **QID**

EXAMPLES

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

(b) External table

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

(c) 3-anonymous patient table

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

(d) 4-anonymous external table

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Bob	Professional	Male	[35-40)
Cathy	Artist	Female	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Professional	Male	[35-40)
Gladys	Artist	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

Παραδείγματα επιθέσεων

EXAMPLE

- Υποτίθεται ότι ένα νοσοκομείο πρόκειται να δημοσιοποιήσει τα αρχεία των ασθενών του **Table (a)** σε ένα ερευνητικό κέντρο
- Υποτίθεται ότι το ερευνητικό κέντρο έχει πρόσβαση και σε έναν άλλο πίνακα **Table (b)** και γνωρίζει ότι για κάθε άνθρωπο για τον οποίο υπάρχει εγγραφή στο Table (b), υπάρχει εγγραφή και στο Table (a)
- **Ενώνοντας** τα δύο tables **στα κοινά πεδία** Job, Sex, και Age μπορεί κανείς να συνδέσει το όνομα κάποιου με την ασθένειά του
- Για παράδειγμα ο Doug, male, lawyer και 38 ετών, αναγνωρίζεται ως ασθενής HIV από το qid = Lawyer, Male, 38

K-ANONYMITY

Για την προστασία από το record linkage μέσω QID, χρησιμοποιούμε την έννοια **k-anonymity**:

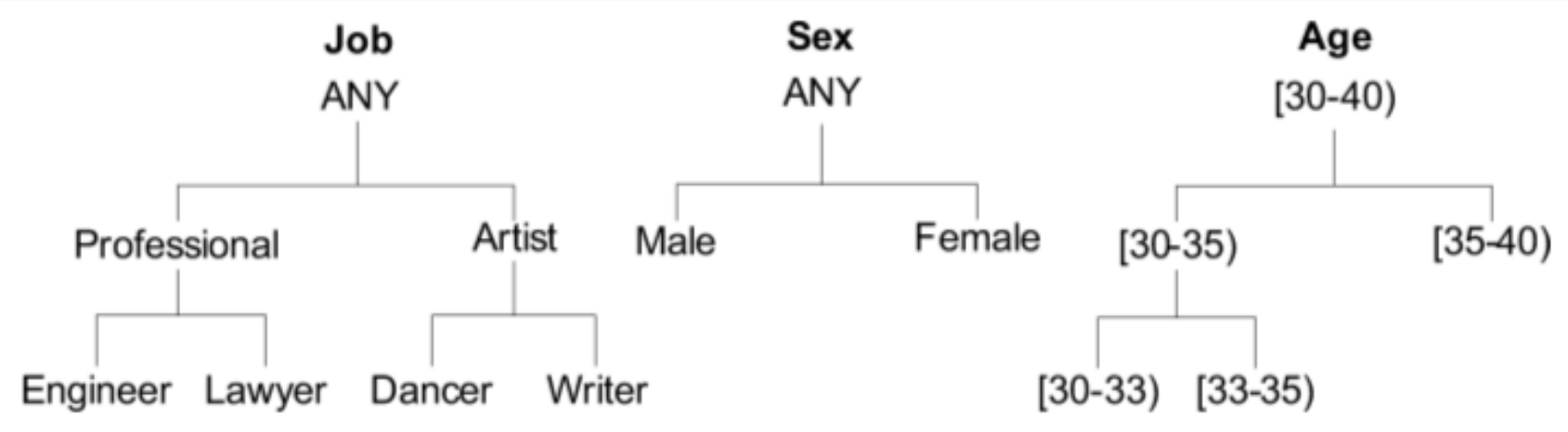
k-anonymity

για κάθε εγγραφή του πίνακα υπάρχουν επιπλέον k-1 εγγραφές με τον ίδιο συνδυασμό τιμών στα qid πεδία

- Με άλλα λόγια το ελάχιστο μέγεθος ενός QID group είναι k
- Ενας πίνακας που ικανοποιεί το παραπάνω κριτήριο λέγεται k-anonymous
- Σε έναν k-anonymous πίνακα, κάθε εγγραφή είναι μη διαχωρίσιμη από τουλάχιστον k – 1 άλλες εγγραφές, σε σχέση με το QID
- Συνεπώς η πιθανότητα να συνδεθεί το θύμα με μία εγγραφή είναι μέσω του QID είναι το πολύ 1/k

ΠΑΡΑΔΕΙΓΜΑ

- Το Table (c) δείχνει έναν **3-anonymous πίνακα** γενικεύοντας το QID = Job, Sex, Age του Table (a) χρησιμοποιώντας τα δένδρα της επόμενης διαφάνειας
- Παρατηρούμε δύο διαφορετικά group βάσει του QID:
  - Professional, Male, [35–40)
  - and Artist, Female, [30–35)
- Αφού κάθε group περιέχει τουλάχιστον 3 εγγραφές, **ο πίνακας είναι 3-anonymous**
- Εάν συνδεθούν το Table (b) με το Table (c) μέσω του QID, κάθε εγγραφή του b συνδέεται με καμία ή με τουλάχιστον 3 εγγραφές του c



Taxonomy trees

ATTRIBUTE LINKAGE

- Σε μία **attribute linkage** επιθεση, ο επιτιθέμενος μπορεί να μην εντοπίσει την εγγραφή του στόχου αλλά να συμπεράνει την τιμή των πεδίων Sensitive Attributes του πίνακα T, βασιζόμενος στο σετ των τιμών αυτών των πεδίων που προκύπτουν από τις εγγραφές του QID group στο οποίο ανήκει
- Σε πολλές περιπτώσεις μία τιμή στα Sensitive Attributes είναι κυρίαρχη σε ένα group, **ακόμα και αν ικανοποιείται το κριτήριο του k-anonymity**
- Για να αντιμετωπίσουμε αυτές τις επιθέσεις προσπαθούμε να ελαττώσουμε την συσχέτιση των τιμών στο QID με τα Sensitive Attributes

ΠΑΡΑΔΕΙΓΜΑ ATTRIBUTE LINKAGE

- Από το **Table (a)**, ο επιτιθέμενος μπορεί να συμπεράνει ότι όλες οι γυναίκες dancers ετών 30 πάσχουν από HIV
- **Dancer, Female, 30 → HIV with 100% confidence**
- Χρησιμοποιώντας αυτό το συμπέρασμα στο **Table (b)**, βρίσκουμε ότι η Emily πάσχει από HIV με 100% confidence, υπό την προϋπόθεση ότι υπάρχει μία εγγραφή για αυτήν στο **Table (a)**

ℓ-DIVERSITY

- Η αρχή του **ℓ-diversity** χρησιμοποιείται για την αποφυγή του attribute linkage
- Απαιτεί κάθε qid group να περιέχει τουλάχιστον ℓ διαφορετικές τιμές στα sensitive attributes
- Το μοντέλο ℓ-diversity αυτομάτως ικανοποιεί και το k-anonymity, με k = ℓ, αφού κάθε qid group περιέχει τουλάχιστον ℓ εγγραφές
- Κάποιες τιμές στα sensitive attributes είναι συχνότερες από τις άλλες σε ένα group, επιτρέποντας έτσι στον επιτιθέμενο να υποθέσει με μεγάλο ποσοστό επιτυχίας ότι μία εγγραφή που ταιριάζει με αυτό το group έχει και την επικρατούσα τιμή στα sensitive attributes
- Για αυτό τον λόγο προτείνεται μία ισχυρότερη έννοια το **entropy ℓ-diversity**

ENTROPY ℓ-DIVERSITY

Ένας πίνακας είναι entropy ℓ-diverse αν για κάθε qid group :

$$-\sum_{s} P(qid, s)\log(P(qid, s)) \geq \log(\ell) \tag{1}$$

- όπου **S** είναι **sensitive attribute**, και **P(qid, s)** είναι **το ποσοστό των εγγραφών του qid group που έχουν μία συγκεκριμένη τιμή s**
- Το αριστερό μέρος της ανίσωσης, ονομάζεται entropy του sensitive attribute, και έχει μεγαλύτερες τιμές για πεδία των οποίων οι τιμές είναι ομοιόμορφα κατανεμημένες
- Οπότε μία μεγαλύτερη οριακή τιμή ℓ σημαίνει μικρότερη βεβαιότητα του επιτιθέμενου για την επιλογή μίας τιμής για το θύμα

ΠΑΡΑΔΕΙΓΜΑ ENTROPY ℓ-DIVERSITY

- Για το Table (c)
- Για το πρώτο group Professional, Male, [35 -40) ,
  $-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = \log(1.9)$
- και για το δεύτερο group Artist, Female, [30-35) ,
  $-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = \log(1.8)$
- Οπότε ο πίνακας ικανοποεί το entropy ℓ-diversity με $\ell \leq 1.8$

ΜΕΤΡΩΝΤΑΣ INFORMATION LOSS

ILoss είναι ένα metric για την μέτρηση του information loss όταν γενικοποιούμε μία τιμή σε μία γενικότερη

Εάν για ένα **πεδίο A**, μία ειδική τιμή **v** γενικοποιείται σε μία πιο γενική τιμή $\mathbf{v_g}$, τότε το information loss υπολογίζεται ως ο παρακάτω λόγος

$$\text{ILoss}(v_g) = \frac{|v_g| - 1}{|D_A|} \tag{2}$$

Το $|v_g|$ είναι το πλήθος των τιμώς που γενικοποιούνται σε v_g και $|D_A|$ είναι το πλήθος όλων των δυνατών τιμών του πεδίου A

ΜΕΤΡΩΝΤΑΣ ΤΟ INFORMATION LOSS

Για να υπολογισθεί το information loss **για μία εγγραφή** τα losses των πεδίων της προσθέτονται (πιθανώς με την χρήση βαρών w)

$$\text{ILoss}(r) = \sum_{u_g \in r} w_i * \text{ILoss}(v_g) \tag{3}$$

Για να υπολογιστεί το information loss για **όλο τον πίνακα** τα losses των εγγραφών του προστίθενται

$$\text{ILoss}(T) = \sum_{r \in T} \text{ILoss}(r) \tag{4}$$