Πανεπιστήμιο Πειραιώς

Τμήμα Ψηφιακών Συστημάτων

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Τεχνοοικονομική Διοίκηση & Ασφάλεια Ψηφιακών Συστημάτων»

Θέματα ασφάλειας και προστασίας της ιδιωτικότητας σε πληροφοριακά συστήματα Υγείας.

Φοιτητής	Τζαχρήστας Στέφανος (MTE1133)
Καθηγητές	Λαμπρινουδάκης Κωνσταντίνος
Ημερομηνία	22 Δεκεμβρίου 2014

Περιεχόμενα

Εισαγωγή

Τα πληροφοριακά συστήματα υγείας ανήκουν ίσως στις πιο ευαίσθητες ομάδες πληροφοριακών συστημάτων σε θέματα ασφάλειας και προστασίας της ιδιωτικότητας.

Θέματα μεγάλου ενδιαφέροντος τα οποία θα μελετηθούν με στόχο τόσο να εντοπιστούν πιθανές επιθέσεις όσο και για να ελεγχθεί η αποτελεσματικότητα συγκεκριμένων μέτρων ασφάλειας είναι: Αυθεντικοποίηση χρηστών από διαφορετικά security domains (π.χ. απομακρυσμένη πρόσβαση ιατρών σε Νοσοκομεία, ανταλλαγή δεδομένων μεταξύ ανεξάρτητων μονάδων υγείας κλπ.), τεχνικές ανωνυμοποίησης για παραγωγή δεδομένων που μπορούν να αξιοποιηθούν σε ιατρική έρευνα, θέματα προστασίας της ιδιωτικότητας που ανακύπτουν κατά τη χρήση αισθητήρων για την καταγραφή ιατρικών σημάτων. Ψηφιακές υπογραφές κλπ..

Η ανωνυμοποίηση έχει νόημα όταν τα δεδομένα πρόκειται να χρησιμοποιηθούν για δευτερογενή σκοπό και όχι για την παροχή ή βελτίωση της φροντίδας του ασθενούς. Τα ευαίσθητα αυτά δεδομένα μπορούν να λάβουν μέρος σε ιατρικές ή και άλλες έρευνες σχετικές με την δημόσια υγεία, πιστοποιήσεις ή ακόμη και για διαφημιστικούς σκοπούς. Η ορθή χρήση αυτών μπορεί να συμβάλει στην πρόληψη, την θεραπεία και τον έλεγχο ασθενειών, με αποτέλεσμα την βελτίωση της ποιότητας ζωής των πολιτών. Ο Ηλεκτρονικός Φάκελος (EHR) ενός ασθενούς αποτελείται από ιατρικά δεδομένα, εξετάσεις και αποτελέσματα. Τα στοιχεία αυτά αποτελούν ευαίσθητα προσωπικά δεδομένα και χρήζουν ιδιαίτερης προσοχής καθώς η δημοσιοποίησή τους έχει νομικές κυρώσεις.

Η συναίνεση ενός ασθενούς για χρήση των προσωπικών δεδομένων του είναι μια λύση, η οποία όμως είναι δύσκολα εφαρμόσιμη. Επιπλέον υπάρχει ήδη ένα τεράστιο πλήθος από ιατρικά δεδομένα, όπου για να ληφθεί η συγκατάθεση των ατόμων θα χρειαστεί μεγάλο χρονικό διάστημα και σε μερικές περιπτώσεις θα είναι αδύνατο. Ένα μοντέλο το οποίο εξασφαλίζει την ανωνυμοποίηση των δεδομένων, που δεν θα προσβάλει την ιδιωτικότητα και την ακεραιότητα των δεδομένων σε ένα βαθμό, θα είναι ιδανικό για την περίπτωση αυτή. Στόχος της παρούσας εργασίας θα είναι η μελέτη ενός τέτοιου μοντέλου που να διασφαλίζει την ανωνυμοποίηση των δεδομένων, αλλά παράλληλα θα τα καθιστά ικανά για στατιστική ανάλυση.

Τα ιατρικά δεδομένα χωρίζονται σε δυο βασικές κατηγορίες οι οποίες χρήζουν ξεχωριστής ανάλυσης και αντιμετώπισης. Τα cross-sectional δεδομένα αφορούν ασθενείς όπου επισκέπτονται ένα νοσοκομείο μόνο μια φορά. Αυτό σημαίνει πως εάν το δείγμα μας αφορά γεννήσεις βρεφών και μια μητέρα γέννησε το 2009 και το 20014, δεν υπάρχει τρόπος να ταχτοποιηθεί ότι πρόκειται για το ίδιο άτομο. Στο δείγμα μας εισάγεται η έννοια του ρίσκου και υπολογίσαμε ότι το μέγιστο επιτρεπτό ρίσκο θα είναι Pr(επανα-προσδ. |αναγνώρισης)≤0,115. Βάση αυτού του ρίσκου έγινε ανωνυμοποίηση του δείγματος με Εντροπία 58.26% - 64.24%. Τα longitudinal δεδομένα αφορούν ασθενείς που επισκέπτονται πάνω από μια φορά ένα νοσοκομείο. Τεχνικές όπως η κ-ανωνυμοποίηση και η απόκρυψη δεν δουλεύουν το ίδιο σωστά όπως στα Cross-sectional δεδομένα. Υπολογίσαμε το όριο του ρίσκου μας Pr(επαν.-προσδιορισμού)≤0.09 και βάση αυτού πήραμε ένα ανώνυμο δείγμα με Εντροπία 44.7%. Τέλος έγινε ανάλυση ενός πραγματικού cross-sectional δείγματος με χρήση όλων των γνωστών μεθόδων ανωνυμοποίησης και μελετήσαμε την απώλεια πληροφορίας και τι επιπτώσεις έχουν οι εκάστοτε αλλαγές στις παραμέτρους που εισάγονται για την ανωνυμοποίησή του.

Σε αυτό το σημείο είναι απαραίτητο να δοθεί μία συνοπτική παρουσίαση των κεφαλαίων που ακολουθούν. Στο Κεφάλαιο 1 γίνεται αναλυση του προβλήματος της ανωνυμοποιησης. Εισάγεται η εννοια του ρίσκου, εξετάζονται οι πιθανες επιθεσεις και μελεταται ο ποσοτικος προσδιορισμος του ρισκου αυτου. Στη συνέχεια στο Κεφάλαιο 2 αναλύονται οι βασικές μέθοδοι ανωνυμοποίησης (k-anonymity, l-diversity, t-closeness) και εξετάζονται οι ευπάθειες αυτών. Τέλος στο Κεφάλαιο 3 εξετάζονται δυο βασικά δείγματα (Cross-Sectional και Longitudinal) ως προς την ανωνυμοποίησή τους εφαρμόζοντας την έννοια του ρίσκου. Επιπλέον εξετάζονται σε ένα Αληθινό Cross-Sectional δείγμα όλες οι μέθοδοι ανωνυμοποίησης και γίνεται σύγκριση και ανάλυση των αποτελεσμάτων αυτών.

Θεωρητικό υπόβαθρο

Για την ευκολότερη κατανόηση των εννοιών με τις οποίες πραγματεύεται η εργασία στα επόμενα κεφάλαια, κρίνεται απαραίτητη μία συνοπτική παρουσίαση του θεωρητικού υπόβαθρου, η οποία θα θέσει τα θεμέλια για την μελέτη και υλοποίηση

του εργαλείου ανωνυμοποίησης. Στην συνέχεια δίνονται οι ορισμοί οι οποίοι βρίσκονται στον πυρήνα του θεωρητικού μοντέλου της ιδιωτικότητας δεδομένων.

- **Γνωρίσματα (attributes)**

Έστω ένας πίνακας $B(A_{1,\dots} A_n)$ με πεπερασμένο αριθμό πλειάδων. Το πεπερασμένο σύνολο των γνωρισμάτων του πίνακα B είναι το $\{ A_{1,\dots} A_n \}$.

- **Πίνακας οντοτήτων (entity-specific table)**

Έστω πίνακας $B(A_{1,\dots} A_n)$ με σύνολο γνωρισμάτων $\{ A_{1,\dots} A_n \}$. Θα λέμε ότι ο πίνακας B είναι πίνακας οντοτήτων αν κάθε εγγραφή του πίνακα αντιστοιχεί σε ένα άτομο.

- **Identifiers**

Identifiers (ή ευαίσθητα χαρακτηριστικά) είναι χαρακτηριστικά που προσδιορίζουν μονοσήμαντα ένα άτομο (π.χ., ΑΜΚΑ, Αρ. Ταυτότητας)

- **Confidential**

Confidential είναι τα χαρακτηριστικά που περιέχουν ευαίσθητες πληροφορίες (π.χ., είδος ασθένειας)

- **Non Confidential**

Non Confidential είναι χαρακτηριστικά που δεν θεωρούνται ευαίσθητα και ως εκ τούτου η δημοσιοποίηση τους δεν δημιουργεί πρόβλημα.

- **Κλάση ισοδυναμίας (equivalence class)**

Ως κλάση ισοδυναμίας, ορίζουμε κάθε σύνολο εγγραφών που έχουν ίδιες τιμές ψευδο-αναγνωριστικών. Οι κλάσεις ισοδυναμίας είναι ξένες μεταξύ τους.

- **Ψευδο-αναγνωριστικά (quasi-identifiers)**

Έστω ένας πληθυσμός οντοτήτων U, ένας πίνακας οντοτήτων , $T(A_{1,\dots} A_n)$, μία συνάρτηση $f_c : U \rightarrow T$ και μία συνάρτηση $f_g : U \rightarrow T'$ με $U' \subseteq U$. Ένα ψευδο-αναγνωριστικό του T, συμβολίζεται με Q_t και είναι ένα σύνολο γνωρισμάτων $\{ A_{1,\dots} A_j \} \subseteq \{ A_{1,\dots} A_n \}$ για το οποίο ισχύει ότι: $\exists\, p_i \in U : f_g(f_c(p_i)[Q_t]) = p_i$

- **Release Table (RT)**

To Release Table είναι το δείγμα που πρόκειται να δημοσιευτεί, αφού έχει υποστεί την επεξεργασία.

- **Tuples**

Tuples είναι το σύνολο των attributes του Release Table (RT). Τα attributes του RT διαφέρουν από αυτά του PT, μετά την αφαίρεση όλων των «ευαίσθητων» attributes, όπως τα Identifiers και τα Confidential. Τα Quasi-identifier μπορούν να αφαιρεθούν επίσης, αλλά με συγκεκριμένο τρόπο που παρέχει η τεχνική ανωνυμίας που χρησιμοποιούμε, όπως θα δούμε παρακάτω.

Non Confidential	Quasi-Identifier			Confidential	Identifiers
Όνομα	DoB	Φύλο	Τ.Κ.	Ασθένεια	Α.Τ.
Andre	21/1/76	Male	53715	H. Disease	125487
Beth	13/4/86	Female	53715	Hepatitis	565468
Carol	28/2/76	Male	53703	Bronchitis	584431
Dan	21/1/76	Male	53703	Broken Arm	879965
Ellen	13/4/86	Female	53706	Flu	230568
Eric	28/2/76	Female	53706	Hang Nail	365982

- **Ευαίσθητα γνωρίσματα (sensitive attributes)**

Ένα γνώρισμα κάποιου πίνακα οντοτήτων θα χαρακτηρίζεται ως ευαίσθητο αν δεν πρέπει να επιτραπεί σε έναν επιτιθέμενο να ανακαλύψει την τιμή του για οποιοδήποτε άτομο στο σύνολο των δεδομένων.

- **Αποκάλυψη ταυτότητας (identity disclosure)**

Κατηγορία επίθεσης, κατά την οποία ο επιτιθέμενος αποσκοπεί στην αναγνώριση της ταυτότητας ενός ατόμου χρησιμοποιώντας έναν ή περισσότερους δημοσιευμένους (και πιθανώς ανωνυμοποιημένους) πίνακες οντοτήτων.

- **Αποκάλυψη (ευαίσθητου) γνωρίσματος (attribute disclosure)**

Κατηγορία επίθεσης, κατά την οποία ο επιτιθέμενος αποσκοπεί στον καθορισμό της τιμής ενός ή περισσότερων ευαίσθητων γνωρισμάτων μίας εγγραφής (ή ισοδύναμα ενός ατόμου) με μεγάλη πιθανότητα σε ένα πίνακα οντοτήτων.

- **Θετική αποκάλυψη γνωρίσματος (positive attribute disclosure)**

Αποκάλυψη γνωρίσματος κατά την οποία η δημοσίευση ενός ανωνυμοποιημένου πίνακα οδηγεί στον ορθό εντοπισμό από έναν επιτιθέμενο της τιμής ενός ευαίσθητου γνωρίσματος με μεγάλη πιθανότητα.

- **Αρνητική αποκάλυψη γνωρίσματος (negative attribute disclosure)**

Αποκάλυψη γνωρίσματος κατά την οποία η δημοσίευση ενός ανωνυμοποιημένου πίνακα οδηγεί στον ορθό αποκλεισμό από τον επιτιθέμενο κάποιας τιμής του ευαίσθητου γνωρίσματος.

- **Γενίκευση δεδομένων (generalization)**

Η γενίκευση δεδομένων αναφέρεται στην διαδικασία απεικόνισης τιμών από ένα αρχικό πεδίο, έτσι ώστε διαφορετικές τιμές του αρχικού πεδίου να απεικονίζονται σε μία τιμή στο πεδίο προορισμού. Αυτό στην γενική περίπτωση επιτυγχάνεται με χρήση της ιεραρχίας γενίκευσης (generalization hierarchy) όπου κάθε τιμή του αρχικού πεδίου μπορεί να απεικονιστεί στο αμέσως γενικότερο επίπεδο και αυτή με την σειρά της στο αμέσως γενικότερο. Μπορούμε λοιπόν να σκεφτούμε την ιεραρχία γενίκευσης σαν ένα δέντρο που τα φύλλα απεικονίζονται στην τιμή του γονέα, αυτή στου δικού του γονέα πηγαίνοντας μέχρι την ρίζα του δέντρου (*) που σημασιολογικά αντιστοιχεί σε όλες τις τιμές. Γενίκευση μπορούμε να έχουμε τόσο σε αριθμητικές τιμές όσο και σε κατηγορικές. Οι αριθμητικές τιμές κατά την γενίκευσή τους αντιστοιχίζονται σε ένα διάστημα τιμών. Για παράδειγμα οι τιμές 32, 37 και 39 θα μπορούσαν να γενικευθούν στην τιμή "3*" που σημαίνει οποιαδήποτε τιμή από 30 ως 39 ή στην τιμή [31-40] κλπ. Η τιμή [31- 40] θα μπορούσε να γενικευθεί κι αυτή σε μια άλλη τιμή όπως "≤50". Για τις κατηγορικές τιμές μπορούμε να θεωρήσουμε ως παράδειγμα μια ιεραρχία όπου οι Η.Π.Α και ο Καναδάς γενικεύονται στην τιμή "Βόρεια Αμερική", η Βραζιλία και η Αργεντινή γενικεύονται στην τιμή "Νότια Αμερική" και με την σειρά τους οι δύο αυτές τιμές γενικεύονται στην τιμή "Αμερική".

Σε ότι αφορά την ανωνυμοποίηση, το πόσα επίπεδα πρέπει να "ανέβει" μια τιμή κατά την γενίκευσή της εξαρτάται κυρίως από την συχνότητα εμφάνισης των διάφορων τιμών. Είναι σαφές ότι η ιεραρχία δημιουργείται ομαδοποιώντας τιμές με κάποιο ή κάποια κοινά χαρακτηριστικά.

- **Αρχή της ιδιωτικότητας (privacy principle):**

Η επιτυχία του επιτιθέμενου είναι δυνατόν να μετρηθεί με την διαφορά της αρχικής του πεποίθησης ότι η ζητούμενη τιμή είναι s και της τελικής του πεποίθησης, η οποία διαμορφώνεται μετά τον εντοπισμό ενός συνόλου εγγραφών που μπορεί να αντιστοιχούν στο αναζητούμενο άτομο στον ανωνυμοποιημένο πίνακα δεδομένων. Η διαφορά αυτή πρέπει να είναι μικρή.

- **Κανονικοποιημένη Ποινή Βεβαιότητας (Normalized Certainty Penalty)**

Έστω κλάση ισοδυναμίας G και γνώρισμα A_N. Για αριθμητικά δεδομένα η Κανονικοποιημένη Ποινή Βεβαιότητας μίας κλάσης ισοδυναμίας ορίζεται ως:

$$NCP_{A_N}(G) = \frac{max_{A_N}^G - min_{A_N}^G}{max_{A_N} - min_{A_N}}$$

Φυσικά, είναι προφανές πως ενώ η παραπάνω μετρική προσφέρει χρήσιμη πληροφορία για την απώλεια πληροφορίας μέσα σε μία κλάση ισοδυναμίας, χρειαζόμαστε μία μετρική για να την μετράμε καθολικά κατά μήκος όλων των κλάσεων. Για αυτό το σκοπό εισάγεται η έννοια της Συνολικής Ποινής Βεβαιότητας.

- **Συνολική Ποινή Βεβαιότητας (Global Certainty Penalty)**

Έστω P το σύνολο όλων των κλάσεων ισοδυναμίας στον ανωνυμοποιημένο πίνακα, N ο αριθμός των εγγραφών στον αρχικό πίνακα και d η διάσταση των κλάσεων ισοδυναμίας. Η Συνολική Ποινή Βεβαιότητας ορίζεται ως :

$$GCP_{(P)} = \frac{\sum_{G \in P} |G| NCP(G)}{dN}$$

Το πλεονέκτημα αυτού το ορισμού είναι κυρίως το εύρος από 0 έως 1 που δίνει, με 0 να είναι η ιδανική περίπτωση μη απώλειας δεδομένων.

1. Συναίνεση ή Ανωνυμοποίηση

Οι περισσότεροι νόμοι αναφορικά με την ιδιωτικότητα των προσωπικών δεδομένων βασίζονται στη συναίνεση του καθενός, αναφορικά με την χρήση ή μη των προσωπικών του δεδομένων. Εάν τα δεδομένα είναι ανώνυμα, προφανώς δεν τίθεται θέμα συγκατάθεσης από κάποιον. Είναι λογικό λοιπόν να εξετάσουμε πρώτιστα την περίπτωση της συναίνεσης. Στην περίπτωση των ιατρικών δεδομένων το να πείσεις κάποιον ασθενή να συναινέσει για την χρήση των δεδομένων του, ειδικά την ώρα που ασχολείται με άλλα πιο σοβαρά θέματα είναι ιδιαίτερα δύσκολο. Επιπλέον έχει παρατηρηθεί ότι οι νέοι όλο και περισσότερο γίνονται καχύποπτοι αναφορικά με την χρήση των προσωπικών τους δεδομένων. Είναι αρκετά μεγάλο το κόστος ώστε να επενδύσει κάποιος σε ενημέρωση των πολιτών σχετικά με την συναίνεση και αποτελεί μια χρονοβόρα διαδικασία.

Πολλοί οργανισμοί διαθέτουν ήδη αρκετά μεγάλες βάσεις δεδομένων με ιατρικούς φακέλους ασθενών. Στην προκείμενη περίπτωση θα πρέπει να ενημερωθούν ένας ένας οι χιλιάδες ασθενείς, ζητώντας την συγκατάθεση τους, κάτι το οποίο φαντάζει αδύνατο. Τίθενται πρακτικά προβλήματα όπως π.χ. πολλοί από αυτούς μπορεί να μην ζουν ή ακόμη τα στοιχεία επικοινωνίας τους να έχουν αλλάξει. Επιπλέον ακόμη και εάν ξεπεραστούν τα προβλήματα αυτά, υπάρχουν αποδείξεις ότι τα άτομα που δίνουν συγκατάθεση με αυτά που αρνούνται, διαφέρουν σε σημαντικά χαρακτηριστικά [1] και το τελικό αποτέλεσμα του δείγματος δεν θα είναι το αναμενόμενο.

Κόστος

Το κόστος που καλείται να πληρώσει ένας οργανισμός στην περίπτωση που αποκαλυφθούν προσωπικά δεδομένα τα όποια έχουν δημοσιοποιηθεί ως ανώνυμα ανέρχεται και είναι της τάξης των $200 περίπου για κάθε άτομο[2]. Αυτό περιλαμβάνει και τις περιπτώσεις που τα δεδομένα αυτά βρίσκονται σε ένα usb δίσκο που χάθηκε ή σε κάποιο υπολογιστή που εκλάπη ή ακόμη και σε μια βάση δεδομένων που έχει προσβληθεί από κάποιον κακόβουλο. Σε κάθε τέτοια περίπτωση πρέπει να ενημερωθούν τα άτομα που έχουν προσβληθεί και οι οργανισμοί που είναι υπεύθυνοι για την ασφάλεια.

Μια ανωνυμοποίηση δεδομένων που δεν γίνει με το σωστό τρόπο, μπορεί να αποβεί μοιραία για έναν οργανισμό εάν τα άτομα μπορούν εκ νέου να προσδιοριστούν. Χαρακτηριστικά αναφέρουμε την περίπτωση της AOL που δημοσιοποίησε για ερευνητικούς σκοπούς 0.5εκατ δεδομένα χρηστών της με ερωτήματα που έκαναν στην μηχανή αναζήτησής της. Λίγο αργότερα αποκαλύφθηκε ότι κάθε χρήστης μπορούσε να προσδιοριστεί με βάση τα ερωτήματα που έκανε. Έτσι η υπόθεση οδηγήθηκε στα δικαστήρια με την εταιρία να πληρώνει 5εκατ. στους χρηστές της και από 1εκατ στους δικηγόρους [3].

Οργανισμοί πιστοποίησης ελέγχουν επιπλέον εάν τηρούνται οι προδιαγραφές ανωνυμοποίησης. Σε μερικές περιπτώσεις ο οργανισμός Health Insurance Portability and Accountability Act (HIPAA) στις ΗΠΑ επιβάλει κυρώσεις, όταν διαπιστώνει αδυναμίες στις πρακτικές ανωνυμοποίησης που εφαρμόζονται [4].

Τεχνικές Ανωνυμοποίησης

Με τον ορό ανωνυμοποίηση αναφερόμαστε στις ενέργειες που ακολουθούμε για να προστατεύουμε μια οντότητα σε μια βάση δεδομένων. Πιο συγκεκριμένα το ISO/TS 25237 ορίζει την ανωνυμοποίηση, ως την διαδικασία που αφαιρεί την σχέση μεταξύ των δεδομένων και του θέματος. Υπάρχουν δυο τεχνικές ανωνυμοποίησης, η προσθήκη μάσκας (mask) και ο από-προσδιορισμός (de-identification).

Οι δυο παραπάνω τεχνικές αφορούν διαφορετικού τύπου δεδομένα σε μια βάση δεδομένων. Σε μερικά πεδία θα προστεθεί μάσκα (ονόματα, ΑΜΚΑ κτλ) ενώ κάποια άλλα θα από-προσδιοριστούν (ηλικία, Τ.Κ., αρ. παιδιών κτλ.).

Μάσκα

Αναφορά στην τεχνική αυτή παρουσιάζεται στο ISO 25237 χωρίς όμως να προσδιορίζονται συγκεκριμένες τεχνικές. Μια προφανής τεχνική είναι η διαγραφή, αφαιρώντας ένα ολόκληρο πεδίο. Επίσης η μάσκα περιλαμβάνει και την τεχνική αντικατάστασης πεδίων με τυχαίες τιμές από μια άλλη μεγάλη βάση δεδομένων [2]

Από-προσδιορισμός

Γενικά υπάρχουν 3 μέθοδοι από-προσδιορισμού ιατρικών δεδομένων που έχουν αναπτυχτεί τα τελευταία χρονιά:

- Από-προσδιορισμός με χρήση λιστών.
- Από-προσδιορισμός με χρήση λογικών βημάτων.
- Από-προσδιορισμός με χρήση της εννοίας του ρίσκου.

Λίστες

Ένα παράδειγμα είναι το πρότυπο Safe Harbor που εφαρμόζει ο HIPAA. Αυτό το πρότυπο προσδιορίζει 18 αντικείμενα που πρέπει να αφαιρεθούν ή αν γενικευτούν. Εάν γίνουν αυτές οι ενέργειες, τα δεδομένα θεωρούνται ότι είναι ανώνυμα συμφώνα με τον HIPAA.

Η μέθοδος αύτη του από-προσδιορισμού έχει σχολιαστεί έντονα καθώς δεν διασφαλίζει ότι τα δεδομένα είναι ανώνυμα. Αντίθετα μπορεί κάποιος να δημιουργήσει δεδομένα βασισμένα στο πρότυπο του Safe Harbor τα όποια μπορούν να προσδιοριστούν πολύ εύκολα.

Λογικά Βήματα

Αυτή η προσέγγιση άφορα λίστες που έχουν δημιουργηθεί με σκοπό τον από-προσδιορισμό των δεδομένων. Αυτές οι λίστες περιέχουν κανόνες και συνθήκες πάνω στις οποίες γίνεται ο απο-προσδιορισμός. Επειδή σαν τεχνική είναι δύσκολη και δυσνόητη, τα λάθη είναι συχνά και πολλές φορές χρησιμοποιείται χωρίς να υπάρχει η επαρκής γνώση. Έχει παρατηρηθεί για παράδειγμα με την χρήση λιστών να επιτρέπεται η δημοσιοποίηση «σπανίων ασθενών» που προσδιορίζουν μοναδικά κάποιο record.

Εμείς θα ασχοληθούμε με αυτή την πρακτική η όποια είναι σύμφωνη με τους κανόνες προστασίας της ιδιωτικότητας του HIPAA καθώς και με άλλους κυβερνητικούς οδηγούς προστασίας που έχουν δημοσιοποιηθεί:

- "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," από τον οργανισμό Υγείας την Αμερικής
- "Anonymization: Managing Data Protection Risk Code of Practice," από τον οργανισμό πληροφοριών του Ηνωμένου Βασιλείου
- ""Best Practice' Guidelines for Managing the Disclosure of De-Identified Health Information," Από τον οργανισμό για την Υγεία του Καναδά
- "Statistical Policy Working Paper 22, Report on Statistical Disclosure Limitation Methodology," από την Στατιστική Υπηρεσία της Αμερικής

Από τους παραπάνω οδηγούς συνάγονται 12 χαρακτηριστικά που απαιτούνται για τον από-προσδιορισμό. Έτσι εάν κάποιος πληροί αυτές τις προδιαγραφές, τότε είναι συμβατός με τα στάνταρ που έχουν θέσει οι οργανισμοί.

Περιπτώσεις

Ανάλογα με την χρήση και το είδος των δεδομένων χρησιμοποιούνται διαφορετικές τεχνικές ανωνυμοποίησης. Οι περιπτώσεις που έχουμε είναι οι έξης:

- Έρευνα: Πρόκειται για την πιο απλή περίπτωση, όπου το περιεχόμενο της βάσης μπορεί να προσδιοριστεί πλήρως. Ο αναλυτής λοιπόν επιλέγει πως θα ανωνυμοποίησει τα δεδομένα, δημιουργεί το αρχείο και το στέλνει στους ερευνητές.
- Δημοσιοποίηση: Υπάρχει μεγάλη ανάγκη για την δημοσιοποίηση των ιατρικών δεδομένων. Αυτό περιλαμβάνει κλινική ερεύνα ή ακόμη και κυβερνητικά προγράμματα.

- Βάσεις Δεδομένων: Πολλοί οργανισμοί διατηρούν εξωτερικά, σε βάσεις δεδομένων τα στοιχειά τους, ακόμη και για στατιστικούς λογούς. Τα δεδομένα πρέπει να ανωνυμοποιηθούν προτού σταλούν.
- Δημόσια Υγεία: Για την παρακολούθηση της δημόσιας υγείας, για ερευνητικούς σκοπούς άλλα και στατιστικούς λογούς, τα δεδομένα πρέπει να συλλέγονται κάπου κεντρικά. Προτού σταλούν από τον κάθε οργανισμό θα πρέπει να ανωνυμοποιηθούν.
- Ιατρικά Μηχανήματα : Τα ιατρικά μηχανήματα που είναι εγκατεστημένα σε διάφορα κέντρα και παρακολουθούν ασθενείς. Αυτά στέλνουν τις ηλεκτρονικές αυτές πληροφορίες σε κάποιο κεντρικό σταθμό, άλλα πρώτου πρέπει να διασφαλίζεται η ανωνυμοποίηση.
- Συναγερμοί: Αυτή η περίπτωση διαφέρει από τις υπόλοιπες. Μπορεί να φαρμακευτική εταιρία να ενημερώνεται όταν κάποιο φάρμακο χορηγείται σε κάποιον ασθενή. Η πληροφορία αυτή θα πρέπει να απο-προσδιοριστεί πρώτα και να ανωνυμοποιηθούν οι πληροφορίες της. Αυτό άφορα συγκεκριμένους ασθενείς και μόνο και όχι μια ολόκληρη βάση δεδομένων.
- Ανάπτυξη Λογισμικού: Για λογούς δοκίμων και εξέλιξης των προγραμμάτων, δίνεται πρόσβαση σε συστήματα σε εταιρίες λογισμικού ακόμη σε περιβάλλοντα παράγωγης. Αυτή η πρόσβαση θα πρέπει αν γίνεται με τρόπο που να διαφυλάσσονται αυτά τα προσωπικά δεδομένα.

2. Η μέθοδος του Ρίσκου στον απο-προσδιορισμό

Πριν αρχίσουμε την μελέτη για το πώς μπορούμε να από-προσδιορίσουμε ιατρικά δεδομένα, χρειάζεται να περιγράψουμε την μεθοδολογία που θα χρησιμοποιήσουμε.

Εισαγωγή

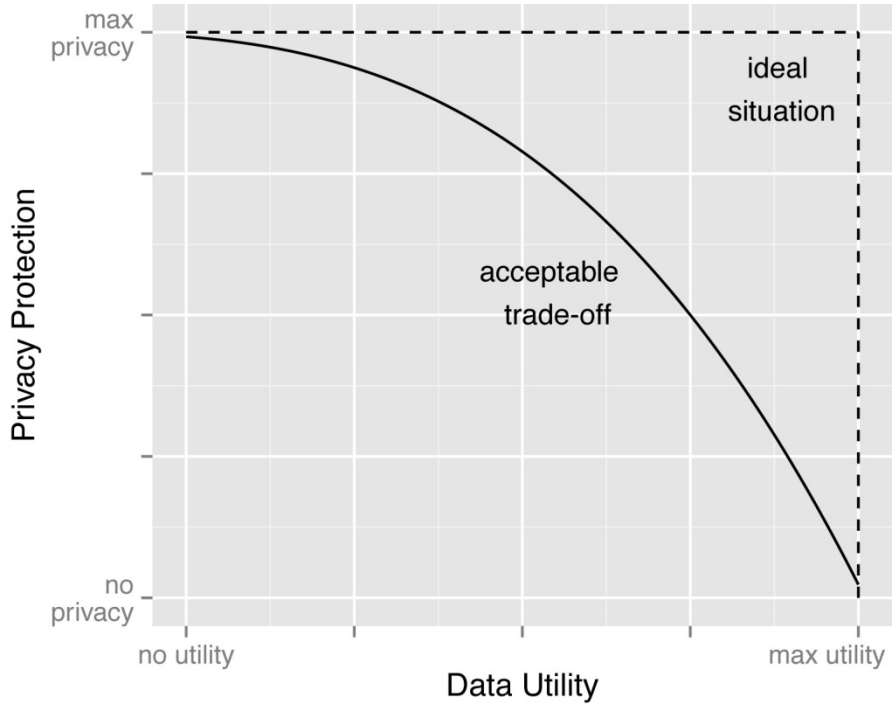
Υπάρχουν κάποιοι βασικοί κανόνες αναφορικά με τον από-προσδιορισμό

Το ρίσκο του επανα-προσδιορισμού

Το ρίσκο του επανα-προσδιορισμού μπορεί να μετρηθεί. Έχοντας λοιπόν ένα μετρό για τον επανα-προσδιορισμό μπορούμε να ορίσουμε πόσο μεγάλος ή μικρός θα είναι ο βαθμός του από-προσδιορισμού που θα εισάγουμε στην βάση μας. Η μέτρηση αύτη γίνεται με βάση κάποιες παραδοχές αναφορικά με τα δεδομένα που έχουμε ή ακόμη και τις τεχνικές (μέσα) της ενδεχόμενης επίθεσης.

Ιδιωτικότητα και δεδομένα

Είναι σημαντικό τα RT (βάσεις) που θα παράγουμε να περιέχουν χρήσιμα και αξιοποιήσιμα δεδομένα. Όπως παρατηρούμε στο Σχήμα 2.1 η ιδιωτικότητα και τα δεδομένα (ακεραιότητα) είναι αντιστρόφως ανάλογη. Δεν μπορούμε λοιπόν να έχουμε όλα τα δεδομένα μας ακεραία και ταυτόχρονα την μεγίστη δυνατή προστασία της ιδιωτικότητας.



Σχήμα 2.1

Ρίσκο επανα-προσδιορισμού

Δεν μπορούμε να εγγυηθούμε μηδενικό ρίσκο επαναπροσδιορισμού. Αυτό εξαρτάται πάντα από το γενικότερο πλαίσιο στο οποίο τοποθετούνται τα δεδομένα αυτά. Εάν πρόκειται να δημοσιοποιηθούν το ρίσκο είναι μεγαλύτερο από το να σταλούν σε ένα ερευνητικό κέντρο το οποίο πληροί με τη σειρά του μηχανισμούς προστασίας και διατηρεί ένα επίπεδο ασφαλείας.

Απο-προσδιορισμός

Πολλές διαφορετικές τεχνικές έχουν αναπτυχτεί για να διασφαλίσουν ότι το ρίσκο του επανα-προσδιορισμού είναι μικρό. Μερικές από αυτές βασίζονται στην παρατήρηση και όποιες άλλες είναι πιο τεχνικές με επέμβαση στα δεδομένα. Στην πραγματικότητα χρησιμοποιείται συνδυασμός και των δυο αυτών μεθόδων.

Στάδια από-προσδιορισμού

Υπάρχουν κάποια σταδία που πρέπει να γίνουν έτσι ώστε να φτάσουμε στο επιθυμητό βαθμό ανωνυμοποίησης των δεδομένων μας.

1. Επιλογή άμεσων και έμμεσων προσδιοριστών.

Άμεσοι προσδιοριστές (Direct Identifiers) ονομάζονται εκείνα τα πεδία σε ένα δείγμα τα όποια μπορούν να προσδιορίσουν μοναδικά μια οντότητα π.χ. ΑΜΚΑ, ΑΦΜ κτλ. Οι άμεσοι προσδιοριστές είναι οι ιδιότητες, οι οποίες συνδέονται άμεσα με μια και μόνο οντότητα. Έμμεσοι προσδιοριστές (Indirect Identifiers ή Quasi-Identifiers) ονομάζονται εκείνα τα πεδία σε ένα PT (δείγμα) τα όποια προσδιορίζουν πιθανόν μια οντότητα, αλλά συνήθως προσδιορίζουν μια ομάδα ατόμων π.χ. Τ.Κ., περιοχή, Ημ. Γέννησης. Ο διαχωρισμός αυτών των δυο προσδιοριστών είναι πολύ σημαντικός παράγοντας στην ανωνυμοποίηση ενός δείγματος. Χρησιμοποιούμε τεχνικές απόκρυψης (masking) για να ανωνυμοποίησης τους αμέσους προσδιοριστές, ενώ για τους εμέσους προσδιοριστές χρησιμοποιούμε τεχνικές γενίκευσης (generalization). Εάν δεν είμαστε σίγουροι για τον τύπο του προσδιοριστή και τα δεδομένα μας πρόκειται να χρησιμοποιηθούν για στατιστική ανάλυση, επιλέγουμε τον έμμεσο, διότι έτσι διατηρούμε τα δεδομένα που θα να είναι απαραίτητα στην ανάλυση μας. Στην περίπτωση της απόκρυψης, χάνονται τελείως αυτά τα δεδομένα, αφού διαγράφονται εντελώς από το RT (δείγμα) μας.

2. Επιλογή ορίου κινδύνου

Το όριο του κίνδυνου παριστάνει το μέγιστο αποδεκτό ρίσκο που θα έχουμε κατά τον διαμοιρασμό των δεδομένων μας. Αυτό το μέγεθος πρέπει να είναι μετρήσιμο και υπάρχουν δύο βασικοί παράγοντες που το επηρεάζουν:

- Εάν τα δεδομένα μας θα είναι δημοσιά ορατά.
- Ο βαθμός των προσωπικών δεδομένων που περιέχει το δείγμα.

3. Επιθέσεις

Υπάρχουν 4 πιθανά είδη επιθέσεων σε ένα δείγμα. Τα τρία πρώτα αφορούν τον παραλήπτη του δείγματος όπου είναι κάποια γνωστή οντότητα ενώ στο τελευταίο ο επιτιθέμενος δεν είναι γνωστός και άφορα δείγματα που είναι εκτίθενται δημόσια:

- Ο παραλήπτης εσκεμμένα προσπαθεί να επανα-προσδιορίσει τα δεδομένα.
- Ο παραλήπτης άσκοπα επανα-προσδιορίζει τα δεδομένα.
- Γίνεται παραβίαση και υποκλέπτονται τα δεδομένα από τον παραλήπτη.
- Κάποιος κακόβουλος κάνει επίθεση και προσπαθεί να επανα-προσδιορίσει τα δεδομένα.

4. Από-προσδιορισμός δεδομένων

Η διαδικασία του από-προσδιορισμού για την Ανωνυμοποίηση μιας βάσης περιλαμβάνει μια ή περισσότερες από τις παρακάτω τεχνικές:

Γενίκευση (Generalization)

Με αυτή την μέθοδο γενικεύουμε κάποιο πεδίο, οπότε ελαττώνουμε την ακρίβεια στον προσδιορισμό που μπορεί να παρέχει. Για παράδειγμα η Ημ. Γέννησης μπορεί να γενικευτεί μόνο στο έτος.

Απόκρυψη (Suppression)

Με αυτή την μέθοδο διαγράφουμε κάποιο πεδίο το οποίο μπορεί να προσδιορίσει μοναδικά μια οντότητα. Για παράδειγμα σε ένα αρχείο από Μαιευτήριο (που περιέχει μητέρες) η Ημ. Γέννησης μιας 54-χρονης μητέρας είναι αρκετά πιθανό να προσδιορίζει μοναδικά το άτομο αυτό.

Υποσύνολο

Μπορούμε να δημοσιοποιήσουμε ένα τυχαίο υποσύνολο της βάσης δεδομένων μας παρά το σύνολο της.

Αυτές είναι οι βασικές μέθοδοι που χρησιμοποιούνται και είναι αποδεκτές στην ανωνυμοποίηση βάσεων στα συστήματα υγείας. Έχουν δημιουργηθεί και άλλες μέθοδοι οι οποίες έχουν σοβαρά μειονεκτήματα. Η προσθήκη θορύβου στα δεδομένα μπορεί εύκολα να αντιστραφεί χρησιμοποιώντας μεθόδους φιλτραρίσματος. Η διαφορική ιδιωτικότητα (Differential Privacy), η όποια έχει κάποιους περιορισμούς, καθίσταται ακατάλληλη για εφαρμογή στα συστήματα υγείας [9].

5. Καταγραφή

Είναι απαραίτητη η καταγραφή της μεθόδου και των βημάτων που έχουν γίνει για να επιτευχτεί η ανωνυμοποίηση. Επίσης στην περίπτωση επανα-προσδιορισμού είναι σημαντικό να υπάρχει ένα τέτοιο αρχείο που να βοηθά στον προσδιορισμό των σφαλμάτων που έχουν γίνει [10].

Ποσοτικός Προσδιορισμός Ρίσκου

Όπως προαναφέραμε πρέπει να καταφέρουμε να μετρήσουμε ποσοτικά το ρίσκο επανα-προσδιορισμού ενός δείγματος και να πάρουμε τα κατάλληλα μέτρα ανάλογα με το αποτέλεσμα.

Τ1. Εσκεμμένη προσπάθεια επανα-προσδιορισμού

Οι περισσότερες επιθέσεις αυτού του τύπου λαμβάνουν μέρος σε ένα σχετικά ασφαλές περιβάλλον όπου έχουμε παραδώσει τα δεδομένα μας. Είναι εφικτό όμως το σενάριο να παραχωρήσουμε τα δεδομένα αυτά σε έναν ερευνητή και εκείνος με την σειρά του (αφού έχουν υπογράφει τα απαραίτητα δικαιολογητικά προστασίας της ιδιωτικότητας) να τα προωθήσει σε ένα Πανεπιστήμιο. Η πιθανότητα κάποιος από τον χώρο αυτό να επανα-προσδιορίσει τα δεδομένα είναι:

$$Pr(\text{επανα-προσδ.}, \text{προσπάθεια}) = Pr(\text{προσπάθεια}) \times Pr(\text{επανα-προσδ.} \mid \text{προσπάθεια})$$

Η *Pr(προσπάθεια)* είναι η πιθανότητα κάποιος από το σύνολο των ατόμων που έχουν πρόσβαση σε αυτή την βάση να προσπαθήσει να κάνει επανα-προσδιορισμό. Εάν για παράδειγμα στο Πανεπιστήμιο που προωθήθηκε είχαν πρόσβαση 100 άτομα και μόνο ένας ήταν κακόβουλος, τότε η πιθανότητα αυτή θα είναι *Pr(προσπάθεια)*=1/100=0,01. Εάν υποθέσουμε ότι στο παράδειγμα μας δεν έχει υπογραφεί κάποια σύμβαση προστασίας της ιδιωτικότητας, τότε είναι λογικό ότι η *Pr(προσπάθεια)* θα είναι μεγάλη.

Η πιθανότητα *PR(επανα-προσδ. | προσπάθεια)* υπολογίζεται απευθείας στην Βάση Δεδομένων και θα το εξετάσουμε στη συνεχεία.

Τ2. Ακούσια προσπάθεια επανα-προσδιορισμού

Το συγκεκριμένο παράδειγμα βασίζεται στην τυχαία αναγνώριση κάποιου μέσα από μια Βάση Δεδομένων. Εάν για παράδειγμα στο δείγμα ανακαλύψει κάποιος άσκοπα ένα γνωστό του ή κάποιον συγγενή του ή έστω ένα δημόσιο πρόσωπο, αυτό αποτελεί παράδειγμα άσκοπης προσπάθειας επανα-προσδιορισμού. Η πιθανότητα είναι:

Pr (επανα-προσδ., αναγνώρισης) = Pr (αναγνώρισης) × Pr (επανα - προσδ. | αναγνώρισης)

Η *Pr(αναγνώρισης)* είναι η πιθανότητα να αναγνωριστεί κάποιος μέσα από την βάση δεδομένων. Για παράδειγμα εάν το δείγμα μας περιέχει στοιχειά με ασθενείς με καρκίνο, η πιθανότητα αυτή είναι, να γνωρίζει κάποιος κάποιον άλλο με καρκίνο. Η πιθανότητα αυτή μπορεί να υπολογιστεί βασιζόμενοι στο γεγονός ότι κατά μέσο ορό ο άνθρωπος έχει 150φίλους (Αριθμός Dunbar [7]). Γνωρίζοντας επίσης την συχνότητα εμφάνισης στο πληθυσμό p βάση μετρήσεων, έχουμε την πιθανότητα *PR(αναγνώρισης)* $= 1 - (1 - p)^{150/2}$ (Διαιρούμε τους φίλους στο μισό, αφού η συχνότητα είναι μεγαλύτερη στις Γυναίκες).

Τ3. Παραβίαση δεδομένων

Έρευνες δείχνουν ότι οι περισσότερες επιθέσεις παραβίασης αφορούν απώλεια ή κλοπή κινητών συσκευών. Τα τελευταία στοιχειά δείχνουν ότι το 27% των συμβεβλημένων με τους κανόνες ασφάλειας του HIPAA αναφέρει από μια παραβίαση αυτού του τύπου κάθε χρόνο.

$$Pr(\text{επανα-προσδ., απώλειας}) = Pr(\text{απώλειας}) \times Pr(\text{επανα-προσδ. | απώλειας})$$

Άρα συμφώνα με όσα αναφέρθηκαν παραπάνω $Pr(\text{απώλειας})=0,27$.

Τ4. Δημόσια δεδομένα

Σε αυτή την κατηγορία τα δεδομένα μας είναι δημοσιευμένα και κάποιος επιτιθέμενος διαθέτει τα μέσα και προσπαθεί να επανα-προσδιορίσει κάποιες οντότητες. Η πιθανότητα λοιπόν είναι:

$Pr(\text{επανα-προσδ.})$ και εξαρτάται από τον τύπο των Δεδομένων που διαθέτουμε.

Μέτρηση του ρίσκου επανα-προσδιορισμού

Υπάρχον διάφοροι τρόποι για να μετρηθεί το ρίσκο του επανα-προσδιορισμού. Ποιο είναι όμως το κόστος αναγνώρισης κάποιας εγγραφής σε ένα δείγμα; Το να αναγνωρίσεις μια μεμονωμένη εγγραφή δεν σημαίνει κάτι. Το κόστος αναγνώρισης πλήρως μιας οντότητας κυμαίνεται από $1 -$15 [8]. Στην περίπτωση που κάποιος κακόβουλος υποκλέψει φαρμακευτικές συνταγές το κόστος κυμαίνεται περίπου σε $1.20 για τον κάθε ασθενή [9].

Πιθανότητες

Όταν δημοσιοποιούμε μια βάση, μπορούμε να βρούμε την πιθανότητα επανα-προσδιορισμού μεμονωμένα για κάθε εγγραφή της. Το ζητούμενο όμως δεν είναι

αυτό, θέλουμε μια συνολική πιθανότητα για να συμπεράνουμε εάν το δείγμα μας πληροί τις προϋποθέσεις ασφάλειας της ιδιωτικότητας για να δημοσιοποιηθεί. Έτσι έχουμε δυο προσεγγίσεις για αυτό τον σκοπό, το Μέγιστο ρίσκο και το Μέσο ρίσκο.

Μέγιστο ρίσκο

Στο Μέγιστο ρίσκο θέτουμε στο δείγμα μας ως συνολικό ρίσκο επανα-προσδιορισμού, την πιθανότητα της εγγραφής εκείνης με την μεγίστη πιθανότητα προσδιορισμού. Βασιστήκαμε στο γεγονός ότι κάποιος που θα προσπαθήσει να επαναπροσδιορίσει ένα δείγμα θα ελέγξει πρωταρχικά έγγραφες που έχουν την μεγαλύτερη πιθανότητα επανα-προσδιορισμού.

Μέσο ρίσκο

Εάν κάποιος προσπαθεί να επανα-προσδιορίσει κάποιον σε ένα δείγμα με αναγνώριση τότε είναι εύλογο να χρησιμοποιήσουμε το Μέσο ρίσκο. Το πρόβλημα με αυτό είναι ότι επιτρέπει σε έγγραφες οι οποίες είναι μοναδικές να εκτεθούν. Για παράδειγμα εάν κάποιος ηλικιωμένος (100 ετών) ζει σε ένα συγκεκριμένο Τ.Κ. είναι πολύ εύκολο για κάποιον που τον γνωρίζει, να βρει τις έγγραφες του στο δείγμα. Χρειαζόμαστε λοιπόν έναν πιο αυστηρό ορισμό του Μέσου ρίσκου.

- Πρέπει να διασφαλίσουμε ότι το Μέγιστο ρίσκο είναι <0.5 ή ακόμη καλυτέρα <0.33. Έτσι εξασφαλίζουμε ότι δεν θα έχουμε προβλήματα αναγνώρισης όπως με τον ηλικιωμένο.
- Εάν το Μέγιστο ρίσκο είναι πάνω από το όριο που έχουμε θέσει >0.5 ή >0.33 τότε αυτό θα είναι το ρίσκο του δείγματος μας. Αλλιώς εάν το Μέγιστο ρίσκο είναι κάτω από το όριο μας, τότε το ρίσκο του δείγματος μας θα είναι το κανονικό Μέσο ρίσκο.

Απώλεια Πληροφορίας

Όταν από-προσδιορίζουμε ένα δείγμα, χάνουμε κομμάτια χρήσιμης πληροφορίας (αφού χρησιμοποιούμε τεχνικές γενίκευσης και απόκρυψης). Η απώλεια πληροφορίας

μας δίνει μια μέτρηση του πόσο παραποιήθηκαν τα δεδομένα μας. Ιδανικά θα έπρεπε να υπολογίσουμε ένα μέγεθος πριν και μετά τον απο-προσδιορισμό άλλα κάτι τέτοιο είναι δύσκολο, καθώς απαιτεί γνώση όλης της διαδικασίας του απο-προσδιορισμού. Δυο από αυτά που μας δίνουν τις πιο χρήσιμες πληροφορίες είναι τα ακόλουθα

Εντροπία

Η εντροπία αποτελεί μια μέτρηση της αβεβαιότητας. Στο δείγμα μας είναι το μέγεθος της ακριβείας που χάνεται και ευθύνεται για αλλαγές γενίκευσης και απόκρυψης. Όσο μεγαλύτερη είναι η εντόπια τόσο περισσότερη πληροφορία χάνεται. Ακόμη και εάν η εντροπία είναι ποσοτικό μέγεθος, μπορεί να μετατραπεί σε ποσοστό, διαιρώντας την με την μεγίστη εντροπία του δείγματος. Έτσι μπορεί κάποιος εύκολα να συγκρίνει την εντροπία πριν και μετά τον από-προσδιορισμό.

Απώλεια

Μια μέτρηση των τιμών ή των εγγράφων που λείπουν σε ένα δείγμα. Μεγάλη απώλεια εγγραφών, μειώνουν την στατιστική ισχύ ενός δείγματος. Εάν πάρουμε για παράδειγμα έναν έμμεσο-προσδιοριστή(quasi identifier), σπάνιες ή ακραίες τιμές θα υποκρυφθούν από το δείγμα για να γίνει ο απο-προσδιορισμός. Έτσι η απόκρυψη δεν θα είναι τελείως τυχαία.

Αυτά τα δυο μεγέθη μας δίνουν τις διαφορετικές προσεγγίσεις στο κομμάτι της απώλειας της πληροφορίας. Η εντροπία επηρεάζεται μόνο από τις γενικεύσεις που εφαρμόζουμε. Οπότε μας δίνει μια μέτρηση, του πόση πληροφορία χάνεται λόγο της γενίκευσης και του απο-προσδιορισμού. Η απώλεια επηρεάζεται μόνο από την απόκρυψη. Ο συνδυασμός αυτών των δυο μεγεθών μας δίνει μια συνολική εικόνα της απώλειας της πληροφορίας. Σκοπός μας είναι να περιορίσουμε όσο το δυνατόν την απώλεια στο ελάχιστο.

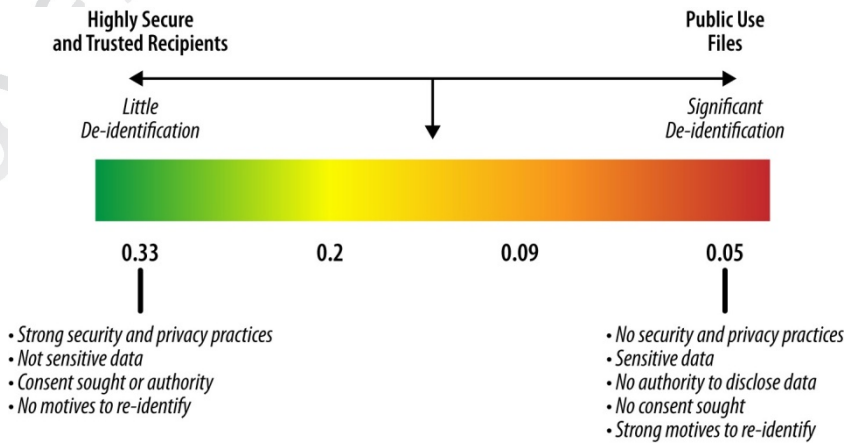
Πρέπει να τονίσουμε ότι η απώλεια έχει διαφορετικές ερμηνείες. Είναι το ποσοστό των γραμμών που έχουν υποστεί απόκρυψη σε ένα έμμεσο-προσδιοριστή ή το ποσοστό των κελίων (γραμμή και στήλη) στους έμμεσους-προσδιοριστές που έχουν υποστεί απόκρυψη.

Ας πάρουμε για παράδειγμα την περίπτωση που έχουμε 100 έγγραφες και 2 έμμεσους-προσδιοριστές. Ας υποθέσουμε ότι έχουμε 3 εγγραφές με λάθος πληροφορίες στο δείγμα μας (γραμμές 3,5 και 7), συνεπώς έχουμε 3% απώλεια πληροφορίας. Ας υποθέσουμε ότι κατά τον απο-προσδιορισμό κάνουμε απόκρυψη στον δεύτερο έμμεσο-προσδιοριστή στις εγγραφές 5, 10 και 18. Στην περίπτωση αυτή μόνο σε 2 νέες έγγραφες γίνεται απόκρυψη κατά τον από-προσδιορισμό (γραμμές 10 και 18). Οπότε έχουμε 5 έγγραφες στις οποίες λείπουν τιμές, άρα 5% απώλεια τιμών. Συνεπώς η απώλεια της πληροφορίας είναι 2% και αφορά απώλεια εγγραφών.

Ας θεωρήσουμε έναν άλλο τύπο απώλειας, βασιζόμενοι στο ποσοστό απώλειας κελιών που αφορούν έμμεσους-προσδιοριστές. Στο προηγούμενο παράδειγμα είχαμε απώλεια 3 κελιά από 200 (100 για κάθε έμμεσο-προσδιοριστή) ή 1.5% απώλεια κελιών. Μετά τον απο-προσδιορισμό έχουμε απώλεια 6 κελία ή 3% απώλεια κελιών, συνεπώς 1.5% απώλεια πληροφορίας.

Όρια ρίσκου

Ποια είναι η αποδέκτη πιθανότητα ώστε να δημοσιοποιήσουμε προσωπικά δεδομένα; Στην πραγματικότητα υπάρχουν οργανισμοί που καθορίζουν αυτές τις τιμές για το μέγιστο αποδεκτό ρίσκο που βλέπουμε στο παρακάτω σχήμα.



Σχήμα 2.2

Όπως φαίνεται όταν ένα δείγμα δημοσιοποιείται το όριο είναι μεταξύ 0,09 και 0,05. Συνεπώς μπορούμε να προσδιορίσουμε το όριο του ρίσκο για την επίθεση Τ4 που εξετάσαμε και *Pr(επανα-προσδ.)<0,05.*

Για να αποφανθούμε ποιο όριο μεταξύ του 0,09 και 0,05 πρέπει να επιλέξουμε, πρέπει πρώτα να εξετάσουμε την ευαισθησία των δεδομένων μας. Εάν το δείγμα μας περιέχει εξαιρετικά ευαίσθητα δεδομένα, τότε θα πρέπει να επιλέξουμε χαμηλότερο όριο. Για δεδομένα τα όποια δεν είναι δημόσια, το όριο μας είναι μεταξύ του 0,1 και 0,05. Στο παράδειγμα της επίθεσης Τ1 εάν η *Pr(επανα-προσδ.)=0,3* και το συνολικό όριο είναι στο 0,1 τότε η πιθανότητα *PR(επανα-προσδ. | προσπάθεια)= 0,1/0,3=0,33.*

Επίθεση	**Ρίσκο**	**Επιτρεπτό όριο**	**Τύπος ρίσκου**
Τ1	*Pr(επανα-προσδ., προσπάθεια)*	0.1 έως 0.05	Μέσο ρίσκο
Τ2	*Pr(επανα-προσδ., αναγνώρισης)*	0.1 έως 0.05	Μέσο ρίσκο
Τ3	*Pr(επανα-προσδ., απώλειας)*	0.1 έως 0.05	Μέσο ρίσκο
Τ4	*Pr(επανα-προσδ.)*	0.09 έως 0.05	Μέγιστο ρίσκο

Σχήμα 2.3

Πως χρησιμοποιούνται οι παραπάνω τεχνικές για να οριστεί το όριο του ρίσκου;

Κλάση Ισοδυναμίας

Όλες οι έγγραφες που έχουν τις ίδιες τιμές στους έμμεσους-προσδιοριστές π.χ. έγγραφες με ιδία Ημ. Γέννησης.

Μέγεθος Κλάσης Ισοδυναμίας

Το πλήθος μιας ισοδύναμης κλάσης. Οι ισοδύναμες κλάσεις αλλάζουν κατά τον από-προσδιορισμό. Εάν έχουμε 3 έγγραφες με ιδία Ημ. Γέννησης, μετά την γενίκευση μπορεί να προκύψουν 12 εγγραφές με Ημ. Γέννησης στο διάστημα 17-19.

K-ανωνυμία

Η πιο κοινή μέθοδος προστασίας ενάντια στον επανα-προσδιορισμό. Αυτή θέτει ένα όριο στο δείγμα κάθε κλάση να περιέχει κ ισοδύναμες κλάσεις [6]. Πολλοί αλγόριθμοι κ-ανωνυμία χρησιμοποιούν γενίκευση και απόκρυψη.

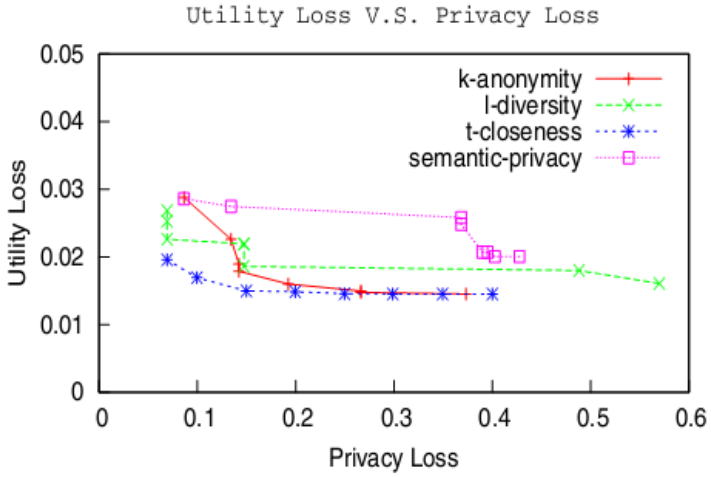
3. Τεχνικές Ανωνυμοποίησης

Στην ενότητα αυτή θα εξετάσουμε τις τεχνικές ανωνυμοποίησης που έχουν αναπτυχθεί, τις δυνατότητες και τις ευπάθειες που προκύπτουν σε κάθε μια από αυτές. Η εισαγωγή της ανωνυμίας ξεκίνησε με τον ορισμό της κ-ανωνυμίας από την L. Sweeny [12]. Το μοντέλο l-diversity [23] αποτελεί μια επέκταση του μοντέλου της k-ανωνυμίας που μειώνει την υποδιαίρεση της αναπαράστασης των δεδομένων με τη χρήση τεχνικών generalization και suppression. Η t-closeness [24] είναι μια περαιτέρω βελτίωση της l-diversity, που χρησιμοποιείται για την διατήρηση της ιδιωτικότητας σε σύνολα δεδομένων, με τη μείωση της λεπτομέρειας της αναπαράστασης αυτών. Παρακάτω θα εξετάσουμε λεπτομερώς την χρήση και την εφαρμογή των τεχνικών αυτών.

κ-ανωνυμία (k-anonymity)

Εισαγωγή

Η ραγδαία συλλογή ιατρικών δεδομένων και η επεξεργασία τους (προσωπικών ή μη) έχει δημιουργήσει δυο κύριες ανησυχίες σχετικά με την προστασία της ιδιωτικότητας και την χρησιμότητα των δεδομένων αυτών. Εάν επικεντρωθούμε στην προστασία της ιδιωτικότητας των δεδομένων αυτών, το ποσό της χρήσιμης πληροφορίας που θα εξαχθεί θα είναι μικρό. Αντίστοιχα, όσο πιο ακέραια είναι τα δεδομένα μας, τόσο η ιδιωτικότητα μειώνεται. Πιο συγκεκριμένα η εργασία των Tiancheng Li and Ninghui Li [11] μας δείχνει ότι η ιδιωτικότητα και η βαρύτητα των δεδομένων βαίνουν σε μια υπερβολή.



Σχήμα 3.1

Για την επεξεργασία των δεδομένων, οι μεγάλες εταιρίες στην Αμερική, αποθηκεύουν και επεξεργάζονται δεδομένα χρησιμοποιώντας εργαλεία **Εξόρυξης(Data Mining Tools)** με σκοπό να ανακαλύψουν νέα πρότυπα σε μεγάλες βάσεις θυσιάζοντας την ιδιωτικότητα. Ο σκοπός της εξόρυξης δεδομένων είναι να εξαρθούν συμπεράσματα από αυτά τα δεδομένα, μετατρέποντας τα σε μια κατανοητή για τον άνθρωπο δομή. Το πρόβλημα που προκύπτει είναι παγκόσμιο, καθώς τα ευαίσθητα ιατρικά δεδομένα χρησιμοποιούνται με τέτοιο τρόπο από τις μεγάλες εταιρίες (π.χ. Ασφαλιστικές) ή άλλους οργανισμούς.

1. Χρησιμότητα Δεδομένων

Με τον όρο χρησιμότητα των δεδομένων εννοούμε τη μετατροπή των δεδομένων σε μια άνθρωπο-κατανοητή δομή για περαιτέρω επεξεργασία. Η εξόρυξη δεδομένων είναι η βασική τεχνολογία με την οποία συσχετίζονται τεράστια σύνολα δεδομένων που παράγονται χρήσιμες πληροφορίες. Η εξόρυξη δεδομένων περιλαμβάνει μια σειρά από επιστημονικούς τομείς, όπως η τεχνητή νοημοσύνη, μηχανική μάθηση, τις στατιστικές, τα συστήματα βάσεων δεδομένων.

Η αξιοποίηση των ιατρικών δεδομένων χρησιμοποιείται κυρίως για εμπορικούς σκοπούς, όπως η διαφήμιση, ασφάλιση κ.λπ. Ωστόσο, η χρήση τους είναι επίσης απαραίτητη για επιστημονικούς σκοπούς. Για παράδειγμα, τα νοσοκομεία μπορεί να

επιθυμούν να μοιραστούν πληροφορίες μεταξύ τους για την επίλυση των διαφόρων ασθενειών ή άλλων περιστατικών. Το πρόβλημα είναι, όταν ο διαμοιρασμός των ιατρικών αρχείων δεδομένων μιας ΒΔ γίνεται στο σύνολό της και έτσι παραβιάζεται το δικαίωμα της ιδιωτικότητας μεταξύ γιατρού και ασθενούς. Η γενική ιδέα, είναι ότι τα δεδομένα πρέπει να μοιράζονται με τέτοιο τρόπο, που να μην παραβιάζει το απόρρητο των μεμονωμένων εγγραφών.

2. Προστασία Προσωπικών Δεδομένων

Η ιδιωτικότητα των προσωπικών δεδομένων ορίζει ότι πρέπει να δημοσιοποιούνται οι λιγότερες δυνατές πληροφορίες σχετικά με τα άτομα. Αυτό είναι ένα πολύ σοβαρό ζήτημα στις μέρες μας, καθώς οι εταιρείες θα συνεχίσουν να συλλέγουν, να διαχειρίζονται και να πουλάνε τεράστιες ποσότητες δεδομένων των χρηστών, όπου παραβιάζονται βασικά δικαιώματα. Πώς μπορούμε να δημοσιοποιήσουμε μια βάση δεδομένων χωρίς να διακυβεύεται η προστασία της ιδιωτικής ζωής;

Το πρόβλημα της ανωνυμίας

Μια απλοϊκή προσέγγιση είναι να αφαιρεθούν τα αναγνωριστικά στοιχεία, όπως το όνομα, ο αριθμός κοινωνικής ασφάλισης (ΑΜΚΑ), αριθμοί τηλεφώνου ή ΑΔΤ που μπορεί να συνδέσει κάποιος άμεσα μια οντότητα. Ωστόσο, αυτή η τεχνική εξακολουθεί να μην εξασφαλίζει την προστασία της ιδιωτικότητας. Κάποιος μπορεί να συλλέγει δεδομένα που έχουν δημοσιοποιηθεί και κατόπιν να ξανασυνδέει αυτά με άλλα δεδομένα και να προσδιορίζει μοναδικά μια οντότητα. Τα δεδομένα συχνά περιέχουν αναγνωριστικά στοιχεία, όπως η φυλή, η ημερομηνία γέννησης, το φύλο, και ταχυδρομικός κώδικας, τα οποία μπορεί να συνδεθούν με άλλα διαθέσιμα στο κοινό δεδομένα και να προσδιορίσουν έτσι συγκεκριμένες οντότητες. Η Sweeney παρατήρησε ότι για το 87% του πληθυσμού στις Ηνωμένες Πολιτείες, ο συνδυασμός της Ημερομηνίας Γέννησης, Φύλο και Τ.Κ αντιστοιχεί σε ένα και μοναδικό άτομο!

Re-identification by linking

Hospital Patient Data

DoB	Sex	Zipcode	Disease
1/21/76	Male	53715	Heart Disease
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Flu
2/28/76	Female	53706	Hang Nail

Vote Registration Data

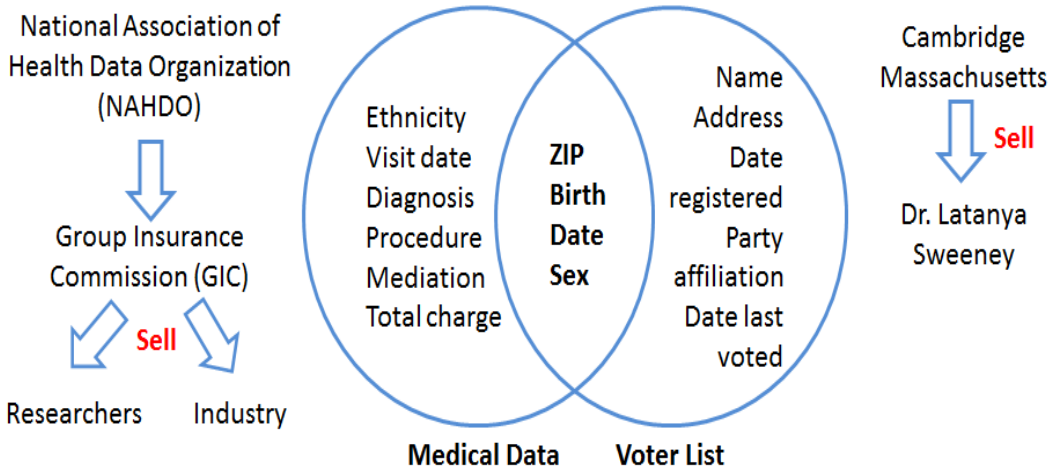
Name	DoB	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237

Andre had 87% to have heart disease!

Σχήμα 3.2

Για παράδειγμα στο Σχήμα-3.2 με τη συνένωση των βάσεων δεδομένων, παρόμοια με τη λίστα δεδομένων των ασθενών στο νοσοκομείο, με χαρακτηριστικά όπως η Ημ. Γέννησης, Φύλο και Τ.Κ. μπορούν να χρησιμοποιηθούν για την αναγνώριση ατόμων. Στην πραγματικότητα, μετά τη συσχέτιση των δεδομένων βρήκαμε ότι το άτομο με το έτος γεννήσεως = 01/21/76, Φύλο = Άνδρας και Τ.Κ. = 53715 σε ποσοστό της τάξης του 87% είναι ο Andre που έχει καρδιακές παθήσεις.

Ένα δεύτερο παράδειγμα είχε αναφερθεί από την έρευνα της Δρ L. Sweeney [12]. Όπως μπορούμε να δούμε από το Σχήμα-3.3, η επιτροπή National Association of Health Data Organization (NAHDO) στις ΗΠΑ έχει θεσπίσει νομοθετικές εντολές για τη συλλογή των δεδομένων του νοσοκομείου. Τα δεδομένα αυτά (αριστερά κύκλος στο σχήμα-3) περιέχουν ένα υποσύνολο των πεδίων, Τ.Κ., Ημερομηνία Γέννησης, Φύλο και εθνότητα [13].



Σχήμα 3.3

Στην πολιτεία της Μασαχουσέτης, η Επιτροπή για Ομαδικές Ασφαλίσεις (GIC), ένας οργανισμός που είναι υπεύθυνος για τα στοιχεία ασφάλισης και υγείας για τους κρατικούς υπαλλήλους, προχώρησε στην αγορά συγκεκριμένων ιατρικών δεδομένων ασθενών από την NAHDO. Επειδή τα δεδομένα θεωρηθήκαν ότι ήταν ανώνυμα, η GIC έδωσε ένα αντίγραφο των δεδομένων σε ερευνητές και προχώρησε στην πώληση ενός άλλου αντιγράφου στη βιομηχανία [14].

Η Δρ L. Sweeney αγόρασε τον εκλογικό κατάλογο από το Πανεπιστήμιο Gabriel της Μασαχουσέτης [15]. Ο εκλογικός κατάλογος περιείχε ένα υποσύνολο των χαρακτηριστικών (δεξιά κύκλος στο Σχήμα 3.3), όπως όνομα, διεύθυνση, ταχυδρομικός κώδικας, Ημερομηνία γέννησης, και το φύλο. Όπως έχουμε πει αυτό είναι το πρόβλημα της ανωνυμίας. Εκτός και αν οι πίνακες αυτοί φαίνεται να είναι ανώνυμοι, μπορούμε να τους επασυνδέσουμε με άλλους και να προχωρήσουμε στον εντοπισμό ατόμων. Η έρευνα της Δρ. L. Sweeney δείχνει ότι ο κ. William, κυβερνήτης της Μασαχουσέτης εκείνη τη στιγμή, πάσχει από μια συγκεκριμένη ασθένεια (η οποία προφανώς δεν ανακοινώθηκε). Η διαδικασία της συσχέτισης των δεδομένων, δείχνει ότι έξι άνθρωποι είχαν συγκεκριμένη ημερομηνία γέννησής, μόνο τρεις από αυτούς ήταν άνδρες, και αυτός ήταν ο μόνος με το συγκεκριμένο 5-ψήφιο ταχυδρομικό κώδικα.

Τα δύο αυτά παραδείγματα δείχνουν ότι το πρόβλημα της εκ νέου σύνδεσης και του μοναδικού προσδιορισμού είναι γεγονός και εγείρει σοβαρά ζητήματα.

κ-ανωνυμία

Για να αποφευχθεί η επίθεση με την συνδεσιμότητα δύο RT, η Samarati και η Sweeney[14,15] πρότειναν, την k-anonymity. Η κ-ανωνυμία, σε συνδυασμό και με άλλες τεχνικές (generalization (γενίκευση), suppression (απόκρυψη)), έχει προταθεί ως μια προσέγγιση για την προστασία της ταυτότητας των οντοτήτων, ενώ απελευθερώνει αληθείς πληροφορίες.

Non Confidentia	Quasi-Identifier			Confidential	Identifiers
Name	**DoB**	**Gender**	**Zipcode**	**Disease**	**SSN**
Andre	1/21/76	Male	53715	Heart Disease	125487
Beth	4/13/86	Female	53715	Hepatitis	565468
Carol	2/28/76	Male	53703	Brochitis	584431
Dan	1/21/76	Male	53703	Broken Arm	879965
Ellen	4/13/86	Female	53706	Flu	230568
Eric	2/28/76	Female	53706	Hang Nail	365982

Σχήμα 3.4

Ορισμός 1 για Quasi-identifier.

Έστω T (A1, ..., An) είναι ένας πίνακας. Ένα Quasi-identifier του πίνακα T, είναι ένα σύνολο από attributes {Ai, ..., Aj}⊆{A1, ..., An} του οποίου η δημοσιοποίηση του θα πρέπει να ελέγχεται πρώτα.

Για παράδειγμα, στο παραπάνω Σχήμα 3.4 φαίνεται ότι τα Quasi-identifier είναι DoB, Gender, Zipcode} υποσύνολο του {Name, DoB, Gender, Zipcode, Disease, SSN}

Τεχνικές k-anonymity

Το μοντέλο κ-ανωνυμία προτείνει δύο τεχνικές, την απόκρυψη και την γενίκευση των δεδομένων. Ας υποθέσουμε ότι έχουμε έναν πίνακα με n tuples και m attributes και έστω k > 1 ένας ακέραιος. Θέλουμε να δημοσιοποιήσουμε την τροποποιημένη έκδοση αυτού του πίνακα, όπου μπορούμε να κάνουμε suppress τις τιμές ορισμένων κελιών στον πίνακα. Ο στόχος είναι να ελαχιστοποιηθεί ο αριθμός των κελιών που είναι suppress, ενώ πρέπει να εξασφαλίσουμε ότι για κάθε tuple στον τροποποιημένο πίνακα, υπάρχουν τουλάχιστον k-1 άλλα tuples στον τροποποιημένο πίνακα ίδια με αυτό.

1. Απόκρυψη

Η απόκρυψη είναι η τεχνική για την προστασία των ευαίσθητων πληροφοριών με την αφαίρεση αυτών. Η απόκρυψη μπορεί να μειώσει το ποσό της απαιτουμένης γενίκευσης για να ικανοποιήσει τις προϋποθέσεις της κ-ανωνυμίας.

2. Γενίκευση

Ο όγκος της πληροφορίας που χάνεται γενικεύοντας τα δεδομένα σε κάποιο επίπεδο γενίκευσης (όσο πιο ψηλά στην ιεραρχία γενίκευσης τόσο μεγαλύτερη απώλεια πληροφορίας)

Παραδείγματα:

- ταχυδρομικός κώδικας: μπορούν να γενικευτούν με τη αποκοπή των λιγότερο σημαντικών ψηφίων. Ένα βήμα της γενίκευσης των αριθμών 20222 και 20223 είναι ο 2022*
- Φύλο: μπορεί να γενικευθεί σε μια αφηρημένη λογική οντότητα. Ένα παράδειγμα της γενίκευσης των φύλων Ασιάτης, Μαύρος, Λευκός είναι το άτομο.
- Ταχυδρομική διεύθυνση: μπορεί να γενικευθεί για το δρόμο (αποκοπή του αριθμού), στη συνέχεια για την πόλη, το νομό, το κράτος, και ούτω καθεξής.

Πίνακες Γενίκευσης

Η διαδικασία γενίκευσης έχει συγκεκριμένα βήματα για την αφαίρεση που καθορίζονται στο Ιεραρχικό Χάρτη Γενίκευσης (GHM). Ο GHM αποτελείται από Κλειδιά και Τιμές με τα παρακάτω χαρακτηριστικά.

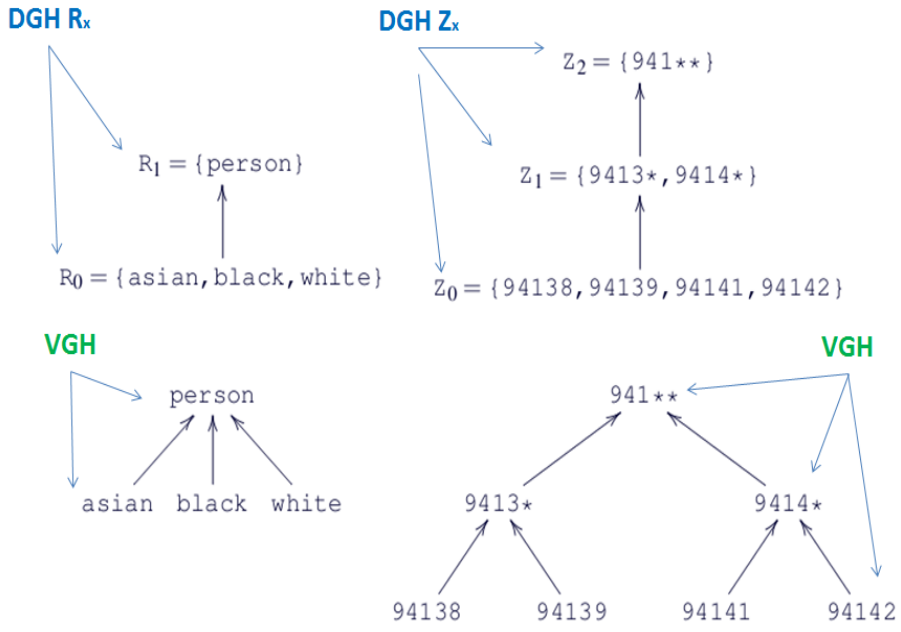
1. Ιεραρχικός Χάρτης Γενίκευσης (DGHD) ή Dom:

• Κάθε Di έχει το πολύ μία άμεση περιοχή γενίκευση Dj. Εξασφαλίζει την αιτιοκρατία στη διαδικασία γενίκευσης.

• Για κάθε τομέα Di έχει το πολύ μία άμεση περιοχή γενίκευσης Dj.

2. Σχέση Γενίκευσης VGH:

• Για κάθε τομέα Di (συμβολίζεται V) υπάρχει μια μοναδική αξία στον τομέα Dj, ως άμεση γενίκευση του Di.



Σχήμα 3.5

Όπως φαίνεται από το Σχήμα 3.5, μπορούμε να δούμε τις ιεραρχίες γενίκευσης για τα χαρακτηριστικά Race (R0) και Φύλο (S0). Η Γενίκευση του Race (R0) είναι ο τομέας (R1) με τιμές από το asian, black, white που γενικεύονται στο person.

Παράδειγμα Απόκρυψης

Όπως μπορούμε να δούμε από το Σχήμα 3.6, εάν PT είναι ο πίνακας μας θα πρέπει να κάνουμε suppress κάποιες τιμές για να επιτύχουμε το πρώτο στάδιο της ανωνυμίας. Όπως μπορούμε να δούμε, τα tuples 1 και 6 καθορίζονται μοναδικά από τον ΤΚ. Έτσι δημοσιοποιήθηκε ο RT_1 χωρίς αυτά τα records.

	Race	Zipcode
1	Asian	94142
2	Asian	94141
3	Asian	94139
4	Asian	94139
5	Asian	94139
6	Black	94138
7	Black	94139
8	White	94139
9	White	94141

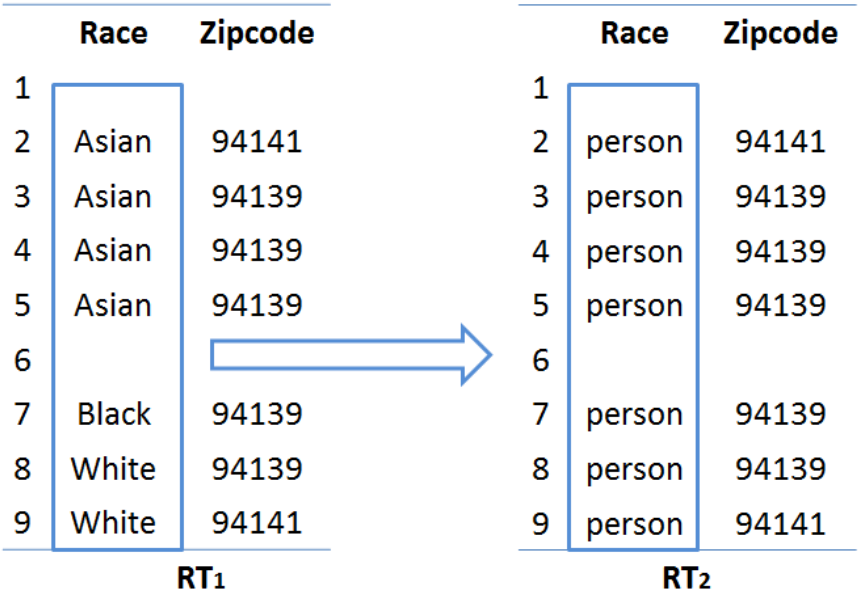
PT

	Race	Zipcode
1	Asian	94141
2	Asian	94139
3	Asian	94139
4	Asian	94139
5		
6	Black	94139
7	White	94139
8	White	94141

RT_1

Σχήμα 3.6

Παράδειγμα Γενίκευσης - Race



Σχήμα 3.7

Η διαδικασία της Γενίκευσης είναι πιο ιδιαίτερη καθώς δεν αφαιρει πληροφορίες από ένα tuple, άλλα από συγκεκριμένες τιμές. Το Σχήμα 3.7 δείχνει ότι γενικεύουμε το attribute Race από το Dom R0 στο R1 το οποίο καθορίζει τις τιμές στο Γενίκευση hierarchy map.

	Race	Zipcode
1		
2	Asian	94141
3	Asian	94139
4	Asian	94139
5	Asian	94139
6		
7	Black	94139
8	White	94139
9	White	94141

RT_1

	Race	Zipcode
1		
2	Asian	9414*
3	Asian	9413*
4	Asian	9413*
5	Asian	9413*
7	Black	9413*
8	White	9413*
9	White	9414*

RT_3

Σχήμα 3.8

Το Σχήμα 3.8 δείχνει πως μπορούμε να γενικεύσουμε το zipcode από το Dom Z0 στο Z1 με βάση τις τιμές του, το οποίο καθορίζεται από τον Χάρτη Ιεραρχίας Γενίκευσης (Σχήμα 3.5).

2-ανωνυμία σε δείγμα

Παράδειγμα 1

Αν θέλουμε για παράδειγμα να δημοσιοποιήσουμε έναν πίνακα με 2-Anonymity μπορούμε αν κάνουμε Suppress και Generalize την τιμή Race. Όπως απεικονίζεται στο Σχήμα 3.9 επιλέγουμε να περάσουμε από το R0 στον τομέα R1. Ο Πίνακας RT4 έχει (με κόκκινο) τουλάχιστον 2 ίδια tuples.

	Race: R_0	Zipcode: Z_0
1	Asian	94142
2	Asian	94141
3	Asian	94139
4	Asian	94139
5	Asian	94139
6	Black	94138
7	Black	94139
8	White	94139
9	White	94141

PT

suppress

	Race: R_0	Zipcode: Z_0
1		
2	Asian	94141
3	Asian	94139
4	Asian	94139
5	Asian	94139
6		
7	Black	94139
8	White	94139
9	White	94141

RT_1

generalize

	Race: R_1	Zipcode: Z_0
1		
2	person	94141
3	person	94139
4	person	94139
5	person	94139
6		
7	person	94139
8	person	94139
9	person	94141

RT_4

Σχήμα 3.9

Παράδειγμα 2

Έχουμε κάνει suppress και generalize τον πίνακα PT, όπως είδαμε στο προηγούμενο παράδειγμα. Η διαφορά είναι ότι τώρα έχουμε επιλέξει να κάνουμε Generalize το χαρακτηριστικό ΤΚ σε συνδυασμό με τον Χάρτη Ιεραρχίας Γενίκευσης (Σχήμα 3.5). Όπως απεικονίζεται στο Σχήμα 3.10 επιλέγουμε να περάσουμε από το Z0 στο Z1. Ο Πίνακας RT2 (με κόκκινο) έχει τουλάχιστον 2 ίδια tuples.

Race: R_0	Zipcode: Z_0
Asian	94142
Asian	94141
Asian	94139
Asian	94139
Asian	94139
Black	94138
Black	94139
White	94139
White	94141

PT

suppress

Race: R_0	Zipcode: Z_0
Asian	94141
Asian	94139
Asian	94139
Asian	94139
Black	94139
White	94139
White	94141

RT_1

generalize

Race: R_0	Zipcode: Z_1
Asian	9414*
Asian	9413*
Asian	9413*
Asian	9413*
Black	9413*
White	9413*
White	9414*

RT_2

Σχήμα 3.10

Περιορισμοί Γενίκευσης και Απόκρυψης

Όπως είδαμε στο προηγούμενο παράδειγμα, υπάρχουν δύο διαφορετικοί τρόποι υπολογισμού του κ-ανωνυμία σε συνδυασμό με το GHM. Δεν είναι όλες οι γενικεύσεις όμως εξίσου ικανοποιητικές. Το ερώτημα που προκύπτει είναι, "είναι αποδοτικό να κάνουμε Γενίκευση και Απόκρυψη συνέχεια;"

Ελάχιστες απαιτήσεις της κ-ανωνυμίας

Η απάντηση σε αυτό είναι αρκετά απλή και προφανής. Το αποτέλεσμα με εφαρμογές γενίκευση και απόκρυψη δεν πρέπει να ξεπέρνα το όριο του k. Η δυσκολία είναι να υπολογίσει κανείς το κατάλληλο k ανάλογα με το πρόβλημα που έχουμε να αντιμετωπίσουμε.

Έχουν προταθεί διάφορα κριτήρια για να φτάσει κάποιος στην ελάχιστη κ-ανωνυμία:

- **Ελάχιστη απόλυτη απόσταση**: Προτιμά αυτή τη γενίκευση με τον μικρότερο συνολικά αριθμό βημάτων.
- **Ελάχιστη σχετική απόσταση**: Προτιμά τη γενίκευση με τον μικρότερο συνολικά αριθμό των σχετικών βημάτων (ένα βήμα γίνεται σχετικό, διαιρώντας το με το ύψος της ιεραρχίας του τομέα στον οποίο αναφέρεται)
- **Τη μέγιστη διανομή**: προτιμά τα γενίκευση με τον μεγαλύτερο αριθμό διαφορετικών tuples.
- **Ελάχιστη απόκρυψη**: προτιμά τη γενίκευση που κάνουν suppress τα λιγότερα tuples, το οποίο είναι αυτό με το μεγαλύτερο πλήθος.

Ταξινόμηση των τεχνικών κ-ανωνυμίας

Το μοντέλο κ-ανωνυμία μπορεί να εφαρμόσει τεχνικές generalization και suppress σε διαφορετικά επίπεδα λεπτομέρειας.

Γενίκευση μπορεί να εφαρμοστεί στο επίπεδο των:

• *Attribute* (AG) – Single Column: γενικεύει όλες τις τιμές στη στήλη

• *Single cell* (CG): για μια συγκεκριμένη στήλη, ο πίνακας μπορεί να περιέχει τιμές σε διαφορετικά επίπεδα γενίκευσης.

Απόκρυψη μπορεί να εφαρμοστεί στο επίπεδο των:

• *Tuple* (TS) - Row: ένα απόκρυψη αφαιρεί μια ολόκληρη tuple

• *Attribute* (AS) - Column: ένα απόκρυψη επισκιάζει όλες τις τιμές μιας στήλης

• *Single Cell* (CS): Ο k-anonymized πίνακας μπορεί να εξαλείψει μόνο ορισμένα κελιά ενός συγκεκριμένου tuple / attribute

	Suppression			
Generalization	*Tuple*	*Attribute*	*Cell*	*None*
Attribute	**AG_TS**	**AG_AS** ≡ AG_	**AG_CS**	**AG_** ≡ AG_AS
Cell	**CG_TS** not applicable	**CG_AS** not applicable	**CG_CS** ≡ CG_	**CG_** ≡ CG_CS
None	**_TS**	**_AS**	**_CS**	**_** not interesting

Σχήμα 3.11

Το Σχήμα 3.11 παρουσιάζει τους πιθανούς συνδυασμούς της Γενίκευσης και Απόκρυψης που αφορούν tuples, cells, attributes ή τίποτα.

Αλγόριθμοι για τον υπολογισμό ενός k-anonymous πίνακα

Για τον υπολογισμό του ελάχιστου k-anonymous πίνακα, έχουν προταθεί πολλοί αλγόριθμοι ανάλογα με τη Γενίκευση και Απόκρυψη των attributes, κελίων ή σειρών.

Παρακάτω συγκεντρώνουμε μερικούς από τους αλγορίθμους, που έχουμε αναφέρει παραπάνω.

- Αλγόριθμοι για Γενίκευση και Απόκρυψη (AG_TS και AG)
- Αλγόριθμος του Bayardo-Agrawal (k-Optimize)
- Αλγόριθμος Samarati
- Incognito από LeFevre, DeWitt και Ramakrishnan
- Αλγόριθμοι Heuristic
- Αλγόριθμοι για Μοντέλα CS και CG.

Ευπάθειες

Ακόμα και όταν έχει ληφθεί επαρκής μέριμνα για την αναγνώριση του QI, η k-ανωνυμία είναι ευάλωτη σε επιθέσεις. Τέτοιου είδους επιθέσεις αφορούν την αναγνώριση τιμής ευαίσθητων δεδομένων και είναι δύσκολο να αντιμετωπιστούν με τη χρήση της k-ανωνυμίας.

1. Επίθεση εναντία στην k-ανωνυμία (unsorted matching attack)

Η επίθεση αυτή βασίζεται στην ίδια διαλογή των πλειάδων που απελευθερώνονται σε διαφορετικά RTs. Όπως μπορούμε να δούμε από το Σχήμα 3.12 τα RT1 και RT2 αντιστοιχούν σε δύο πίνακες μετά την γενίκευση της Φυλής και του T.K. αντίστοιχα. Επειδή οι πλειάδες των πινάκων RT1 και RT2 δημοσιεύτηκαν με την ίδια ταξινόμηση, είναι δυνατόν ένας αντίπαλος να συγχωνεύσει RT1 και RT2 σε έναν νέο πίνακα και να συνδέσει τις αντίστοιχες εγγραφές. Λύση σε αυτό το πρόβλημα θα μπορούσε να είναι η τυχαιοποίηση των πλειάδων θέσης μέσα στο κάθε δείγμα.

	Race	Zipcode
1		
2	Asian	94141
3	Asian	94139
4	Asian	94139
5	Asian	94139
6		
7	Black	94139
8	White	94139
9	White	94141

PT

	Race	Zipcode
1		
2	person	94141
3	person	94139
4	person	94139
5	person	94139
6		
7	person	94139
8	person	94139
9	Person	94141

RT_1

	Race	Zipcode
1		
2	Asian	9414*
3	Asian	9413*
4	Asian	9413*
5	Asian	9413*
6		
7	Black	9413*
8	White	9413*
9	White	9414*

RT_2

Σχήμα 3.12

2. Συμπληρωματικές επίθεσεις-απελευθέρωση της k-ανωνυμίας

Αυτή η επίθεση βασίζεται στην απελευθέρωση δείγματος από το ίδιο PT. Μεταξύ 2 ή περισσότερων k-anonymous πίνακων που έχουν κυκλοφορήσει δημόσια, μπορούμε να δημιουργήσουμε έναν τρίτο πίνακα, που δεν είναι k-anonymous. Για παράδειγμα, έχουμε έστω RT1 και RT2 δύο δείγματα μετά την γενίκευση της Φυλής και Τ.Κ. Επειδή, οι πλειάδες της RT1 και RT2 κυκλοφόρησαν από το ίδιο PT είναι δυνατόν ένας αντίπαλος να συγχωνεύσει πλειάδες σε και να τις επανα-συνδέσει δημιουργώντας έναν νέο όχι ανώνυμο πίνακα. Λύση σε αυτό το πρόβλημα θα μπορούσε να είναι η απελευθέρωση του πίνακα RT2 βάση του RT1 και όχι του PT.

3. Χρονική επίθεση

Το τρίτο πρόβλημα της k-Ανωνυμίας αφορά την δυναμική αλλαγή των στοιχείων των πινάκων. Ανά πάσα στιγμή η πρόσθεση, η αφαίρεση και η αλλαγή πλειάδων σε ένα σύνολο εγγραφών μπορεί να εκθέσει τη βάση σε κίνδυνο.

Έστω ότι, για t=0 υπάρχει ένας πίνακας T_0 και από αυτόν προκύπτει ένας ανωνυμοποιημένος πίνακας A_0 ο οποίος ικανοποιεί την k-Ανωνυμία. Αν σε χρόνο t, προστεθούν στον αρχικό πίνακα κάποιες εγγραφές τότε προκύπτει ο πίνακας T_t.

Με γενίκευση του T_t προκύπτει ο πίνακας A_t, ο οποίος επίσης ικανοποιεί την k-ανωνυμία. Λόγω του ότι δεν υπάρχει καμία εγγύηση η οποία να εξασφαλίζει ότι ο πίνακας A_t έχει σαν βάση τον A_0, τότε όπως και στο προηγούμενο παράδειγμα η σύνδεση των δύο πινάκων μπορεί να μην ακολουθεί την k-Ανωνυμία.

λ-Διαφορετικότητα (l-Diversity)

Σκοπός της προστασίας ιδιωτικότητας σε ένα σύνολο εγγραφών, δεν είναι μόνο η ασφάλεια της ταυτότητας μιας εγγραφής (identity disclosure). Είναι και ταυτόχρονα η διασφάλιση ότι ο επιτιθέμενος δεν θα μπορέσει εύκολα να βρει από αυτό το σύνολο εγγραφών, προσωπικά στοιχεία για ένα άτομο (attribute disclosure).

Εισαγωγή

Στην προηγούμενη ενότητα είδαμε ότι η k-ανωνυμία σε ένα δείγμα μπορεί να αποκαλύψει ευαίσθητες πληροφορίες. Εάν ένα δείγμα ικανοποιεί την k-ανωνυμία τότε οποιοσδήποτε ο οποίος γνωρίζει μόνο Άμεσους προσδιοριστές για κάποια οντότητα, δεν μπορεί να την προσδιορίσει με πιθανότητα μεγαλύτερη του 1/κ. Η k-ανωνυμία μας προστατεύει ως προς την αποκάλυψη της ταυτότητας κάποιας οντότητας στο δείγμα, αλλά δεν παρέχει προστασία σε μεμονωμένες τιμές της οντότητας. Το μοντέλο ℓ-διαφορετικότητα παρέχει προστασία της ιδιωτικότητας, ακόμη και όταν κάποιος προσπαθεί να συνδέσει τα δεδομένα με αλλά. Η κύρια ιδέα πίσω από την ℓ-διαφορετικότητα είναι ότι οι τιμές των ευαίσθητων χαρακτηριστικών πρέπει να αναπαρίστανται σε κάθε ομάδα.

Εάν υπάρχουν ℓ "αναπαραστάσεις" ευαίσθητων τιμών σε ένα σύνολο από πλειάδες (block) του RT του οποίου τα μη-ευαίσθητα χαρακτηριστικά είναι γενικευμένα τότε κάποιος χρειάζεται ℓ - 1 κατεστραμμένα κομμάτια της πληροφορίας για την εξάλειψη ℓ - 1 πιθανών ευαίσθητων τιμών για να εξαχθεί ένα θετικό συμπέρασμα. Έτσι, θέτοντας την παράμετρο ℓ, ο κάτοχος των δεδομένων μπορεί να καθορίσει πόση προστασία παρέχεται κατά γνωστικό υπόβαθρο, ακόμα και αν το υπόβαθρο της γνώσης είναι άγνωστο στον ιδιοκτήτη. Συνοψίζοντας τα παραπάνω, φτάνουμε στα ακόλουθα συμπεράσματα.

Ένα μπλοκ είναι ℓ-διαφορετικό αν περιέχει τουλάχιστον ℓ «καλές-εκπροσωπούμενες» τιμές για ένα συγκεκριμένο ευαίσθητο χαρακτηριστικό. Ένα RT είναι ℓ-διαφορετικό αν κάθε μπλοκ του είναι ℓ-διαφορετικό.

Σαν παράδειγμα, παρουσιάζουμε ένα 3-διαφορετικό RT, βάση του PT στο σχήμα 3.13

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

PT

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Σχήμα 3.13

Τα πλεονεκτήματα του μοντέλου ℓ-ποικιλομορφία συνοψίζονται παρακάτω:

- Η ℓ-Διαφορετικότητα δεν απαιτεί τη γνώση της πλήρους κατανομής των ευαίσθητων και μη ευαίσθητων χαρακτηριστικών.
- Η ℓ-Διαφορετικότητα δεν απαιτεί από τον εκδότη του RT να έχει την γνώση που θα έχει ο επιτιθέμενος. Όσο μεγαλύτερη είναι η αξία του ℓ, τόσο περισσότερες πληροφορίες είναι απαραίτητες για να αποκλειστούν πιθανές τιμές της ευαίσθητης ιδιότητας.
- Η Γνώση σε επίπεδο οντότητας (Ο γιος του Μπομπ λέει στην Άλικη ότι ο Μπαμπ δεν έχει διαβήτη) καλύπτεται επαρκώς, δίχως περεταίρω ενέργειες.
- Διαφορετικοί επιτιθέμενοι μπορεί να έχουν διαφορετικό γνωστικό υπόβαθρο που θα τους οδηγήσει σε διαφορετικά συμπεράσματα. Η ℓ-ποικιλότητα ταυτόχρονα προστατεύει ενάντια σε όλες τις οντότητες, χωρίς την ανάγκη για

τον έλεγχο που συμπερασμάτων μπορεί να γίνει με τον οποίο τα επίπεδα υποβάθρου της γνώσης.

Ο Machanavajjhala ορίζει το πως εφαρμόζεται η έννοια της l-ποικιλομορφίας σε ένα δείγμα:

1. **Διακριτή l-ποικιλομορφία**: αποτελεί την πιο απλή μορφή όπου εξασφαλίζει ότι υπάρχουν τουλάχιστον l διακριτές τιμές της αισθητής ιδιότητας σε κάθε ισοδύναμη κλάση. Αυτή η μορφή δεν προφυλάσσει το δείγμα μας από επιθέσεις πιθανοθεωρητικές, καθώς μια ισοδύναμη κλάση μπορεί να έχει τιμές που εμφανίζονται αρκετά συχνά από κάποιες άλλες. Από αυτό μπορεί να συμπεράνει κάποιος ότι είναι η πιο πιθανή περίπτωση να έχει κάποιος αυτή την ευαίσθητη ιδιότητα.
2. **Εντροπία l-διαφορετικότητα:** Ένα δείγμα θεωρείται ότι έχει εντροπία l-διαφορετικότητα εάν για κάθε ισοδύναμη κλάση E, Εντροπία (E) ≥ log(l). Αυτό σημαίνει ότι και η εντροπία ολοκλήρου του δείγματος θα πρέπει να είναι τουλάχιστον Log(l). Κάτι τέτοιο μπορεί όμως να είναι αρκετά περιοριστικό, καθώς η εντροπία του δείγματος ενδέχεται να είναι μικρή εάν κάποιες τιμές εμφανίζονται αρκετά συχνά.
3. **Αναδρομική (c,l) διαφορετικότητα:** Εξασφαλίζει ότι οι πιο συχνές τιμές δεν εμφανίζονται τόσο συχνά και οι λιγότερο συχνές όχι τόσο σπάνια.

Ευπάθειες

Η l-διαφορετικότητα αποτελεί εάν σημαντικό βήμα ως προς την ανωνυμοποίηση ιδιοτήτων που δεν προστατεύει η k-ανωνυμία. Παρόλα αυτά έχει αρκετές ελλείψεις και ευπάθειες που πρέπει να μελετήσουμε.

☛Η l-διαφορετικότητα μπορεί να είναι δύσκολη και όχι αναγκαία προϋπόθεση για να επιτευχθεί. Ας υποθέσουμε ότι το δείγμα μας περιέχει μόνο ευαίσθητα δεδομένα για έναν ιό με δυνατές τιμές θετικό ή αρνητικό. Επιπλέον ας υποθέσουμε ότι το δείγμα μας έχει 10.000 εγγραφές με το 99% αυτών αρνητικές και μόνο το 1% θετικές. Σε

αυτή την περίπτωση η 2-διαφορετικότητα δεν είναι αναγκαία για μια ισοδύναμη κλάση που περιέχει έγγραφες που είναι αρνητικές. Για να έχουμε διακριτή 2-διαφορετικοτητα τότε θα πρέπει να υπάρχουν το πολύ 10.000*1%=100 ισοδύναμες κλάσεις και κατά συνέπεια η απώλεια πληροφορίας θα ήταν μεγάλη. Ακόμη βλέπουμε ότι η εντροπία της ευαίσθητης ιδιότητας είναι πολύ μικρή, άρα εάν χρησιμοποιηθεί εντροπία θα πρέπει η τιμή της να είναι επίσης μικρή.

☛ Η l-διαφορετικότητα είναι ανεπαρκής ώστε να εμποδίσει την αποκάλυψη ιδιοτήτων. Ας δούμε δυο παραδείγματα αυτής της επίθεσης.

Ασύμμετρη επίθεση

Όταν η συνολική διανομή είναι ασύμμετρη, η l-διαφορετικότητα δεν προστατεύει την αποκάλυψη τιμών μιας ιδιότητας. Ας θεωρήσουμε ότι στο προηγούμενο παράδειγμα με τον ιό, υπήρχε στο δείγμα ακριβώς ο ίδιος αριθμός θετικών και αρνητικών τιμών και άρα ικανοποιείται το κριτήριο της 2-διαφορετικοτητας. Όμως αυτό δεν προστατεύει την ιδιωτικότητα καθώς καθένας στο δείγμα μας έχει 50% πιθανότητα να είναι είτε θετικός είτε αρνητικός.

Ας θεωρήσουμε τώρα ότι μια ισοδύναμη κλάση έχει 49 θετικές και 1 αρνητική εγγραφή. Αυτή η κλάση θα αποτελεί μια διακριτή 2-διαφορετικοτητα και θα έχει μεγαλύτερη εντροπία από ότι το συνολικό δείγμα μας και επιπλέον καθένας στην καλή αυτή θα έχει 98%πιθανοτητα να είναι θετικός.

Ταυτόσημη επίθεση

Όταν οι ευαίσθητες ιδιότητες σε μια ισοδύναμη κλάση είναι διακριτές αλλά σχεδόν ταυτόσημες, κάποιος μπορεί να συνάγει σημαντικές πληροφορίες. Ας υποθέσουμε ότι έχουμε το παρακάτω δείγμα σχήμα 3.14, ενώ το σχήμα 3.15 αποτελεί την ανωνυμοποίηση εκδοχή του δείγματος ικανοποιώντας διακριτή και εντροπία 3-διαφορετικοτητα..

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Σχήμα 3.14

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Σχήμα 3.15

Έχουμε δυο ευαίσθητες ιδιότητες, τον μισθό και την ασθένεια. Εάν υποθέσουμε ότι γνωρίζουμε ότι μια οντότητα είναι στις τρεις πρώτες έγγραφες τότε συμπεραίνουμε ότι ο μισθός του είναι [3K-5K] και είναι σχετικά μικρός. Επίσης γνωρίζουμε για το άτομο αυτό ότι πάσχει από στομαχικές διαταραχές καθώς και οι τρεις αυτές τιμές για την ασθένεια είναι σχετικές. Αυτό συμβαίνει γιατί η 3-διαφορετικοτητα, εξετάζει την διαφορετικότητα σε κάθε ισοδύναμη κλάση, ενώ δεν λαμβάνει καθόλου υπόψη τη σημασιολογική εγγύτητα των τιμών αυτών.

t-Εγγύτητα (t-Closeness)

Σαν αποτέλεσμα των πιο πάνω, κατανομές οι οποίες έχουν το ίδιο επίπεδο ποικιλομορφίας προσφέρουν διαφορετικά επίπεδα ιδιωτικότητας, ανάλογα (α) με τις σημασιολογικές σχέσεις ανάμεσα στις ευαίσθητες τιμές τους, (β) τα διαφορετικά επίπεδα ευαισθησίας των πεδίων και (γ) την ολική κατανομή δεδομένων στη βάση.

Εισαγωγή

Διαισθητικά η μέτρηση της ιδιωτικότητας, γίνεται από κάποιον όταν παρατηρήσει το τελικό δείγμα, από την πληροφόρηση που μπορεί να συνάγει για μια οντότητα. Η t-Closeness είναι μια περαιτέρω βελτίωση της l-διαφορετικότητας, που χρησιμοποιείται για προστατεύει την ιδιωτικότητα σε δείγματα, μειώνοντας την αναλυτικότητα ης πληροφορίας. Η μείωση αυτή είναι ένας συμβιβασμός που οδηγεί σε κάποια απώλεια των δεδομένων, προκειμένου να έχουμε προστασία της ιδιωτικότητας του δείγματος. Το μοντέλο t-Closeness επεκτείνει την l-διαφορετικότητα μετατρέποντας τις τιμές ενός χαρακτηριστικού σε διακριτές, λαμβάνοντας υπόψη την διανομή των τιμών δεδομένων για το εν λόγω χαρακτηριστικό.

Μια ισοδύναμη κλάση λεμέ ότι έχει t-Closeness εάν η διάφορα διανομής της ευαίσθητης ιδιότητας μέσα στην κλάση και η διανομή της ιδιότητας αυτής σε όλο το δείγμα δεν ξεπερνά το όριο t. Ένα δείγμα λεμέ ότι ικανοποίει t-Closeness εάν όλες οι ισοδύναμες κλάσεις του εφαρμόζουν t-Closeness.

Όσο πιο μικρή είναι η τιμή του t, τόσο πιο κοντά βρίσκονται οι δύο κατανομές.

Ορισμοί

Απόσταση μετακίνησης μάζας (EMD)

Το EMD [22] βασίζεται στην ελάχιστη δυνατή εργασία που απαιτείται για να μετασχηματιστεί μια διανομή σε μια άλλη, μετακινώντας το σύνολο των διανομών μεταξύ τους. Μια διανομή θεωρείται σαν μια μάζα απλωμένη σε έναν χώρο και η άλλη σαν μια συλλογή από τρύπες στο ίδιο χώρο. Το EMD μετρά την λιγότερη εργασία που απαιτείται για να γεμίσουν αυτές οι τρύπες από την μάζα αυτή.

Ισότιμη απόσταση

Η απόσταση μεταξύ δυο οποιονδήποτε τιμών μιας ιδιότητας κατηγορίας ορίζεται ότι είναι ίση με την μονάδα. Καθώς λοιπόν η απόσταση δυο τιμών είναι ίση με 1, η μια θα πρέπει να μεταφερθεί σε κάποια άλλη κλάση.

Ιεραρχική απόσταση

Η απόσταση μεταξύ δυο οποιονδήποτε τιμών μιας ιδιότητας κατηγορίας ορίζεται ως το ελάχιστο επίπεδο που οι δυο αυτές τιμές είναι γενικευμένες, στην ιδία τιμή συμφώνα με την ιεραρχία.

Ας δούμε τώρα πως εξετάζει η t-Closeness με EMD το δείγμα του σχήματος 3.14 και πως διαχειρίζεται τα προβλήματα που είχαμε με την l-διαφορετικότητα.

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Σχήμα 3.16

Το παραπάνω Σχήμα 3.16 μας δείχνει το αποτέλεσμα του δείγματος κατόπιν εφαρμογής 0.167-εγγύτητας στον μισθό και 0.278-εγγύτητα στην ασθένεια. Το δείγμα παρέχει προστασία ενάντια στην ταυτόσημη επίθεση, σε αντίθεση με την l-διαφορετικότητα. Επίσης η t-εγγύτητα προστατεύει ευαίσθητες ιδιότητες και όχι ολόκληρη την οντότητα, συνεπώς θα είναι καλό να χρησιμοποιούμε k-ανωνυμοποίηση μαζί με την t-εγγύτητα.

4. Παραδείγματα

Θα δούμε κάποια παραδείγματα σχετικά με την ανωνυμοποίηση ενός δείγματος και προβλήματα που παρουσιάζονται ανάλογα με το δείγμα και με την χρήση για την οποία προορίζεται. Οι τεχνικές που θα εφαρμόσουμε βασίζονται στην παραπάνω θεωρία με διαφορές παραλλαγές που θα εξετάσουμε αναλυτικά. Οι δύο βασικοί άξονες που θα μελετήσουμε αφορούν cross-sectional και longtitudinal ιατρικά δείγματα που αποτελούν και τους βασικούς άξονες στατιστικής ανάλυσης, με τα δεύτερα να απαιτούν περισσότερο χρόνο για την συλλογή τους. Τα μεν πρώτα αφορούν οντότητες που επισκέφτηκαν μια φορά ένα νοσοκομείο και κάθε εγγραφή του δείγματος εξετάζεται ότι αφορά μια μοναδική οντότητα. Στα longtitudinal δείγματα η κάθε οντότητα έχει πάνω από μια εγγραφές στο δείγμα οι οποίες με κάποιο τρόπο συνδέονται μεταξύ τους.

1. Παράδειγμα 1 (Cross-Sectional Data)

Ας εξετάσουμε την περίπτωση που κάποιος ερευνητής χρειάζεται να αναλύσει ένα δείγμα με νεογέννητα (στο οποίο περιέχονται και οι μητέρες τους). Ο κεντρικός οργανισμός που διατηρεί αυτές τις έγγραφες δεν συγκρατεί άμεσους προσδιοριστές, αλλά μόνο έμμεσους. Ο ερευνητής συγκεντρώνει όλες τις πληροφορίες που χρειάζεται, χωρίς να θέσει κάποιο διάστημα από το 2005-2011 που το κέντρο έχει 919.710 χιλ. έγγραφες. Τα δεδομένα που μας αφορούν είναι άκρως ευαίσθητα αφού η ερεύνα βασίζεται στις γενετικές ανωμαλίες και παίζει σημαντικό ρολό στον προσδιορισμό του ρίσκου μας. Τα στοιχειά που αιτήθηκαν φαίνονται στο παρακάτω σχήμα:

Πεδίο	**Περιγραφή**
BSex	Φύλο του νεογέννητου
MDOB	Ημ. Γέννησης της Μητέρας
BDOB	Ημ. Γέννησης του νεογέννητου
MPC	Τ.Κ. διαμονής της Μητέρας

Σχήμα 4.1

Όπως είναι λογικό δεν μπορούμε να γενικεύσουμε το φύλο, αλλά μπορούμε να γενικεύσουμε όλες τις άλλες μεταβλητές όπως βλέπουμε παρακάτω.

Πεδίο	Ιεραρχία Γενίκευσης
MDOB	dd/mm/yyyy →Εβδ./Έτος→mm/Έτος→Τρίμηνο/Έτος→Έτος→5-χρόνια διάστημα→10-χρόνια διάστημα
BDOB	dd/mm/yyyy → Εβδ./Έτος →mm/Έτος→Τρίμηνο/Έτος→Έτος→5-χρόνια διάστημα→10-χρόνια διάστημα
MPC	Αποκοπή του(ων) τελευταίων x χαρακτήρα(ων), όπου x είναι 1→2→3→4→5

Σχήμα 4.2

Ρίσκο

Όπως μπορούμε αν δούμε και από το Σχήμα 2.2 ένα μέσο όριο ρίσκου της τάξης του 0.1-0.05 είναι αποδεκτό στην περίπτωση μας. Εάν θεωρήσουμε ότι τα δεδομένα μας είναι ευαίσθητα, το όριο θα πρέπει να τείνει στο 0.05.

Απειλές

Ας εξετάσουμε τις 3 πιθανές επιθέσεις από γνωστό επιτιθέμενο.

Α) Στην περίπτωση της σκόπιμης προσπάθειας επανα-προσδιορισμού πρέπει να υπολογίσουμε την πιθανότητα της προσπάθειας Pr(προσπάθεια). Στην περίπτωση του ερευνητή, θα πρέπει να έχει υπογράφει κάποια συνθήκη ασφάλειας για τα προσωπικά δεδομένα μεταξύ αυτού και του κέντρου και επίσης στον τομέα αυτό δεν υπάρχει κάποιος προφανής λόγο για να επαναπροσδιορίσει κάποιος τα δεδομένα αυτά. Συνεπώς μιας και τα δεδομένα μας δεν θεωρούνται ευαίσθητα μπορούμε να δώσουμε μια χαμηλή τιμή στην Pr(προσπαθεια)=0,4.

Β) Μια ακουσία προσπάθεια προϋποθέτει κάποιος να έχει έναν γνωστό μέσα στο δείγμα. Η πιθανότητα λοιπόν Pr(*αναγνώρισης*) είναι να γνωρίζει κάποιος μια γυναικά που να έχει παΐδι στο διάστημα 2005-2011. Για το έτος 2008 η *είχαμε 119.785 γεννήσεις σε σύνολο 4.478.500 γυναικών, άρα p=119.785/4.478.500=0,027. Συνεπώς Pr(αναγνώρισης)* = $1-(1-\mathrm{p})^{150/2}$ = 0,87

C) Η πιθανότητα *Pr(απώλειας)=0,27 η οποία είναι* βασισμένη σε στατιστικά δεδομένα.

Συνεπώς η συνολική μας πιθανότητα είναι Pr(επανα-προσδ., T) = *Pr*(T) × *Pr*(επανα-προσδ. | T) , όπου *Pr*(T) είναι μια από τις παραπάνω πιθανότητες. Επιλεγούμε την μεγαλύτερη πιθανότητα, που είναι αυτή της ακούσιας προσπάθειας, που είναι και η πιο πιθανή και έχουμε: *Pr*(επανα-προσδ., *αναγνώρισης*) = *0,87* × *Pr*(επανα-προσδ. | T)≤0,1 βάση του ρίσκου που θέσαμε παραπάνω. Άρα *Pr*(επανα-προσδ. | *αναγνώρισης*)≤0,115

Αποτελέσματα

Έχοντας υπολογίσει το όριο του ρίσκου για το δείγμα μας και με την γενίκευση που έχουμε καθορίσει., απο-προσδιορίζουμε το δείγμα μας ελαχιστοποιώντας την απόκρυψη κελιων[20]. Γενικεύουμε λοιπόν MDOM στο έτος, το DBOB στο Εβδ./έτος και τα MPC's στον ένα χαρακτήρα και έχουμε τα παρακάτω αποτελέσματα.

	Cell missingness	**Record missingness**	**Entropy**
Πρίν	0.02%	0.08%	
Μετά	0.75%	0.79%	58.26%

Σχήμα 4.3

Υποθέτουμε ότι χρειαζόμαστε μεγαλύτερη ακρίβεια στον T.K. και όχι τόσο μεγάλη ακρίβεια στο MDOM. Δίνουμε λοιπόν ακρίβεια 3-ψηφίων στον T.K. και γενικεύουμε το MDOM χωρίς να μεγαλώσουμε το ρίσκο. Στην δεύτερη αυτή προσπάθεια από-προσδιορισμού γενικεύουμε την MDOM στα 10 χρόνια ακρίβεια, το DBOB στο Τρίμηνο/έτος και τα MPC's στους τρεις χαρακτήρες και έχουμε τα παρακάτω αποτελέσματα.

	Cell missingness	Record missingness	Entropy
Πρίν	0.02%	0.08%	
Μετά	0.02%	0.08%	64.24%

Σχήμα 4.4

Η εντροπία αυξήθηκε λόγω της μεγάλης μεταβολής στην MDOB. Παρατηρούμε ότι δεν υπήρξε αλλαγή στην απώλεια πληροφορίας, οπότε ας δοκιμάσουμε να αυξήσουμε την ακρίβεια στην BDOB. Στην τρίτη αυτή προσπάθεια από-προσδιορισμού γενικεύουμε την MDOM στα 10 χρόνια ακρίβεια, το DBOB στο μήνα/έτος και τα MPC's στους τρεις χαρακτήρες και έχουμε τα παρακάτω αποτελέσματα

	Cell missingness	Record missingness	Entropy
Πρίν	0.02%	0.08%	
Μετά	26.7%	26.7%	59.59%

Σχήμα 4.5

Όπως βλέπουμε η απώλεια πληροφορίας είναι αρκετά μεγάλη στο τελευταίο μας παράδειγμα.

2. Παράδειγμα 2 (Longitudinal Data)

Ας υποθέσουμε ότι έχουμε δεδομένα από ασθενείς, για ένα δοσμένο χρονικό διάστημα που έχουν επισκεφθεί μια κλινική για πάνω από μια φορές καθώς ενδεχομένως πάσχουν από χρόνιες ασθένειες. Μπορούμε να δούμε ότι εάν εφαρμόσουμε τις τεχνικές που χρησιμοποιήσαμε στο Παράδειγμα 1 το αποτέλεσμα θα είναι είτε να χάσουμε μεγάλο μέρος της πληροφορίας είτε να μην προστατεύσουμε σωστά την ιδιωτικότητα του δείγματος.

Έχουμε 6 ασθενείς στον Πίνακα (a) του παρακάτω σχήματος και οι επισκέψεις μαζί με τους Τ.Κ. καταγράφονται στον πινάκα (b) σε περίπτωση που οι ασθενείς αλλάξουν διεύθυνση.

(a)

PID	DoB	Gender
10	2000/08/07	M
11	1975/01/01	F
12	1975/06/24	F
13	1975/08/17	F
14	1975/18/09	F
15	2000/02/12	M

(b)

PID	Visit Date	Postal Code
10	2009/01/01	K7G2C3
10	2009/01/14	K7G2C3
10	2009/04/18	K7G2C4
11	2009/01/01	K1V7E6
11	2009/01/20	K1V7E8
11	2009/02/22	K1V7E8
12	2008/12/15	K1Y4L5
12	2009/01/20	K1V7E8
13	2008/12/22	K1Z5H9
14	2009/01/13	K1Y4L5
15	2009/04/20	K7G2G5

(c)

PID	DoB	Gender	Visit Date	Postal Code
10	2000	M	2009/01	K7G
10	2000	M	2009/01	K7G
10	2000	M	2009/04	K7G
11	1975	F	2009/01	K1V
11	1975	F	2009/01	---
11	---	---	---	---
12	1975	F	2008/12	K1Y
12	1975	F	2009/01	K1V
13	1975	F	2008/12	K1Y
14	1975	F	2009/01	---
15	2000	M	2009/04	K7G

(d)

		Visit Date									Postal Code							
DoB	Sex	2009/01/01	2009/14/01	2009/04/18	2009/01/20	2009/02/22	2008/12/15	2008/12/22	2009/01/12	2009/04/20	K7G2C3	K7G2C4	K1V7E6	K1V7E8	K1Y4L5	K1V7E8	K1Z5H9	K7G2G5
2000/08/07	M	X	X	X							X	X						
1975/01/01	F	X			X	X							X	X				
1975/06/24	F				X		X								X	X		
1975/08/17	F							X									X	
1975/09/18	F								X						X			
2000/02/12	M									X								X

Σχήμα 4.6

Σε αυτού του τύπου τα δεδομένα οι τεχνικές της κ-ανωνυμοποίησης με χρήση generalization και suppression δεν λειτουργούν όπως θα περιμέναμε.

- Απόκρυψη σε έγγραφες

  Αν ενώσουμε τους πινάκες (a) και (b) του παραπάνω σχήματος καταλήγουμε στον πινάκα (c) που μας δείχνει πως θα είναι το δείγμα μας με χρήση 2-ανωνυμοποίησης. Για τον ασθενή 11 ο Τ.Κ. έχει φύγει στην 2η επίσκεψη και

όλοι οι άμεσοι προσδιοριστές έχουν εξαφανιστεί στην 3η επίσκεψη. Η κ-ανωνυμοποίηση αγνοεί το PID καθώς αντιμετωπίζει κάθε γραμμή σαν ανεξάρτητη εγγραφή. Έτσι κάποιος μπορεί να παρατηρήσει ότι οι δυο πρώτες έγγραφες έγιναν τον ίδιο μηνά και πιθανότητα ο Τ.Κ. δεν θα έχει αλλάξει. Άρα η 2-ανωνυμοποιηση δεν λειτούργει όπως θα περιμέναμε.

- Ομογενείς ισοδύναμες κλάσεις

  Οι δυο πρώτες έγγραφες όπως είδαμε αντιστοιχούν στον ίδιο ασθενή και ονομάζονται ισοδύναμες κλάσεις. Η πιθανότητα επανα-προσδιορισμού είναι 1 παρόλο που έχουμε εφαρμόσει 2-ανωνυμοποίηση.

- Γνώση ιστορικότητας

  Ας υποθέσουμε ότι κάποιος γνωρίζει ένα άτομο μέσα στο δείγμα και ξέρει ότι έχει κάνει 2 επισκέψεις καθώς και τις ημερομηνίες αυτών. Στο δείγμα μας έχουμε δυο οντότητες με 2 επισκέψεις, σε διαφορετικές ημερομηνίες, οπότε αφού κάποιος γνωρίζει το ποτέ, μπορεί να προσδιορίσει την οντότητα μοναδικά.

Ένας άλλος τρόπος για να αναπαραστήσουμε τα δείγμα μας είναι να δημιουργήσουμε μια μεταβλητή για κάθε άμεσο προσδιοριστή όπως φαίνεται στο Σχήμα 4.6 (d). Οι τεχνικές κ-ανωνυμοποίησης για αυτού του τύπου δεδομένα, μας εξασφαλίζουν ότι υπάρχουν τουλάχιστον κ ισοδύναμες κλάσεις. Όπως είναι προφανές κάτι τέτοιο θα οδηγήσει σε μεγάλο αριθμό μεταβλητών και το τελικό δείγμα μας θα έχει πολλές διαστάσεις. Αυτό με την σειρά του θα μας οδηγήσει σε μεγάλη απώλεια πληροφορίας [21].

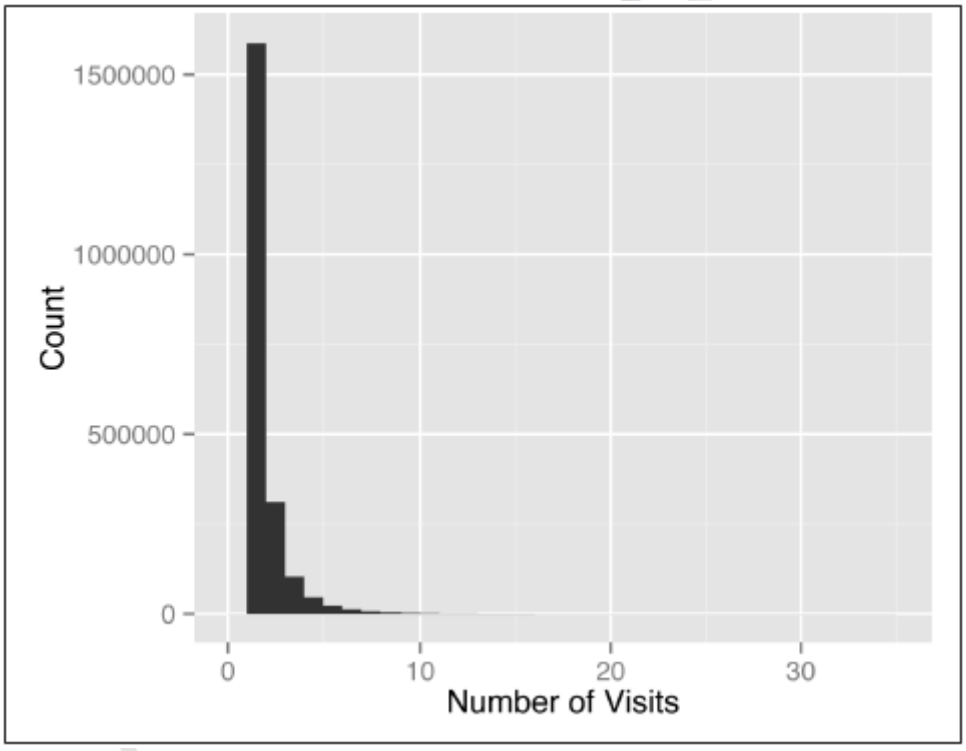
Ας θεωρήσουμε ότι έχουμε ένα δημοσιοποιημένο δείγμα με τους παρακάτω αμέσους-προσδιοριστές:

Επίπεδο	**Πεδίο**	**Περιγραφή**
1	Gender	Το φύλο του ασθενή
1	BirthYear	Έτος Γέννησης του ασθενή
2	AdmissionYear	Πρώτο έτος εισαγωγής του ασθενή

2	DaySinceLastService	Αριθμός ημερών από την τελευταία επίσκεψη του ασθενή
2	LenthOfStay	Αριθμός ημερών νοσηλείας του ασθενή στο νοσοκομείο

Σχήμα 4.7

Οι περισσότεροι ασθενείς έχουν μια ή δυο επισκέψεις που σημαίνει ότι υπάρχουν μια ή δυο έγγραφες στο δείγμα μας για κάθε ασθενή. Όσο πιο άρρωστος είναι κάποιος τόσο περισσότερες φορές θα νοσηλευτεί και κατά συνέπεια τόσο περισσότερες εγγραφές θα έχει στο δείγμα μας.



Σχήμα 4.8

Δεν μπορούμε να γενικεύσουμε το Φύλο, αλλά μπορούμε να γενικεύσουμε όλες τις άλλες μεταβλητές, όπως βλέπουμε στο σχήμα 4.9. Χρησιμοποιούνται ανώτατα και κατώτατα όρια, που σημαίνει ότι για την Ημ. Γέννησης, όταν είναι μικρότερη του 1910 αλλάζει και γίνεται 1910.

Πεδίο	Ιεραρχία Γενίκευσης
BirthYear	Κατώτατο όριο 1910: Έτος→5-χρόνια διάστημα→10-χρόνια διάστημα
AdmissionYear	Κατώτατο όριο 2006: 1-χρόνο διάστημα → 2-χρόνια διάστημα
DaysSinceLastService	Μέχρι 6 ημέρες, μετά Εβδομάδες, ανώτατο όριο 182+ → Κατώτατο όριο 7-, μετά 28-ημέρες διάστημα, ανώτατο όριο 182+
LengthOfStay	Μέχρι 6 ημέρες, μετά Εβδομάδες, ανώτατο όριο 182+ → Κατώτατο όριο 7-, μετά 28-ημέρες διάστημα, ανώτατο όριο 182+ (Ομοίως με το DaysSinceLastService)

Σχήμα 4.9

Ρίσκο

Οποιοσδήποτε μπορεί να έχει πρόσβαση σε δημοσιά δεδομένα, οπότε μπορούμε να υποθέσουμε ότι κάποιος θα προσπαθήσει να επανα-προσδιορίσει τα δεδομένα. Συνεπώς μπορούμε να ορίσουμε ένα όριο της τάξης του 0,09 για τον από-προσδιορισμό του δείγματος όπως έχουμε προαναφέρει.

Απειλές

Έχουμε μόνο ένα είδος επίθεσης και από την στιγμή που τα δεδομένα μας είναι δημοσιοποιημένα, θα χρησιμοποιήσουμε το μέγιστο δυνατό ρίσκο Pr(επαν.-προσδιορισμού)≤0.09 (το οποίο είναι ισοδύναμο με την 11-ανωνυμοποίηση).

Αποτελέσματα

Θεωρήσαμε ότι κάποιος έχει γνώση του δείγματος και για τον απο-προσδιορισμό του ελαχιστοποιήσαμε την απόκρυψη πληροφοριών. Με την ιεραρχία γενίκευσης που ορίσαμε το αποτέλεσμα του δείγματος μας έχει κατώτατο όριο «Έτους Γέννησης» το 1910, με διάστημα 5-ετίας, το «Έτος εισαγωγής» δεν έχει αλλάξει καθόλου και οι «Ημέρες από την τελευταία επίσκεψη» έχει κατώτατο όριο 7-, με διάστημα 28 ήμερων και ανώτατο όριο 182+ και το «διάστημα διαμονής» όπως και οι «Ημέρες από την τελευταία επίσκεψη» επειδή είναι άμεσα συσχετισμένοι σαν Άμεσοι-Προσδιοριστές. Το ποσοστό απώλειας και η εντροπία φαίνεται στο Σχήμα 4.10

	Cell missingness	**Record missingness**	**Entropy**
Πριν	0.05%	3.14%	
Μετα	15.3%	28.5%	44.7%

Σχήμα 4.10

3. Παράδειγμα 3

Πήραμε ένα δείγμα από το CDC Wonder και από επίσημα στοιχειά του 2011 για τα αιτία θανάτου στην περιοχή της South Carolina.Το δείγμα που λάβαμε ήταν γενικευμένο, από το οποίο κατασκευάσαμε με τυχαιοποίηση ένα αναλυτικό cross-sectional δείγμα 3.872 ασθενών όπου και θα μελετήσουμε ως προς την ανωνυμοποίηση του χρησιμοποιώντας τις διαφορετικές τεχνικές που αναφέραμε. Θεωρούμε ότι έχουμε ένα δείγμα με στοιχεία ασθενών: ΑΜΚΑ, ΤΚ, Ηλικία, Ασθένεια. Θα εξετάσουμε το παραπάνω δείγμα ως προς την προστασία και ασφάλεια της πληροφορίας χρησιμοποιώντας μεμονωμένα τις τεχνικές ανωνυμοποίησης που αναφέραμε παραπάνω.

- k-Anonymity
- l-Divercity
- t-Closeness

Αρχικά θα αναλύσουμε τους τύπους των δεδομένων που διαθέτουμε. Ως Άμεσους προσδιοριστές ορίζουμε τα πεδία: ΑΜΚΑ ενώ ως έμμεσους: Ηλικία, Φύλο, ΤΚ και ως ευαίσθητα πεδία: Ασθένεια.

A	B	C	D	E
ΑΜΚΑ	**ΗΛΙΚΙΑ**	**ΦΥΛΟ**	**Τ.Κ.**	**ΑΣΘΕΝΕΙΑ**
1067131	55	Γυναίκα	45003	C34.9
1067132	70	Άνδρας	45003	C34.9
1067133	82	Άνδρας	45003	C34.9
1067134	79	Γυναίκα	45003	J18.9
1067135	79	Γυναίκα	45003	C34.9
1067136	77	Άνδρας	45003	C34.9
1067137	76	Άνδρας	45003	C34.9
1067138	82	Γυναίκα	45003	C34.9
1067139	77	Γυναίκα	45003	C34.9
1067140	80	Γυναίκα	45003	C34.9
1067141	83	Άνδρας	45003	C34.9
1067142	77	Γυναίκα	45003	C34.9
1067143	79	Γυναίκα	45003	C34.9
1067144	76	Γυναίκα	45003	C34.9
1067145	76	Γυναίκα	45003	C34.9
1067146	81	Γυναίκα	45003	C34.9
1067147	80	Άνδρας	45003	C34.9
1067148	77	Γυναίκα	45003	C34.9
1067149	83	Γυναίκα	45003	C34.9
1067150	77	Γυναίκα	45003	C34.9
1067151	77	Γυναίκα	45003	C34.9
1067152	82	Άνδρας	45003	G30.9
1067153	80	Άνδρας	45003	G30.9
1067154	82	Γυναίκα	45003	G30.9
1067155	79	Άνδρας	45003	G30.9
1067156	79	Γυναίκα	45003	G30.9
1067157	82	Άνδρας	45003	G30.9
1067158	81	Γυναίκα	45003	G30.9
1067159	77	Άνδρας	45003	G30.9

Σχήμα 4.11

Το αρχικό δείγμα ήταν της παρακάτω μορφής το οποίο και μετασχηματίστηκε κατάλληλα:

ΗΛΙΚΙΑ	ΠΕΡΙΟΧΗ	ΑΣΘΕΝΕΙΑ	ΠΛΗΘΟΣ ΑΣΘΕΝΩΝ
15-24 years	Charleston County, SC (45019)	X95 (Assault by other and unspecified firearm discharge)	10
25-34 years	Greenville County, SC (45045)	X44 (Accidental poisoning by and exposure to other and unspecified drugs, medicaments and biological substances)	10
45-54 years	Anderson County, SC (45007)	C34.9 (Bronchus or lung, unspecified - Malignant neoplasms)	11
45-54 years	Anderson County, SC (45007)	I21.9 (Acute myocardial infarction, unspecified)	11
45-54 years	Anderson County, SC (45007)	X44 (Accidental poisoning by and exposure to other and unspecified drugs, medicaments and biological substances)	15
45-54 years	Charleston County, SC (45019)	C34.9 (Bronchus or lung, unspecified - Malignant neoplasms)	17
45-54 years	Cherokee County, SC (45021)	I51.8 (Other ill-defined heart diseases)	13
45-54 years	Darlington County, SC (45031)	I21.9 (Acute myocardial infarction, unspecified)	14
45-54 years	Florence County, SC (45041)	C34.9 (Bronchus or lung, unspecified - Malignant neoplasms)	15
45-54 years	Florence County, SC (45041)	I46.9 (Cardiac arrest, unspecified)	12
45-54 years	Greenville County, SC (45045)	C34.9 (Bronchus or lung, unspecified - Malignant neoplasms)	21
45-54 years	Greenville County, SC (45045)	C50.9 (Breast, unspecified - Malignant neoplasms)	12
45-54 years	Greenville County, SC (45045)	I25.0 (Atherosclerotic cardiovascular disease, so described)	13
45-54 years	Greenville County, SC (45045)	I25.1 (Atherosclerotic heart disease)	19
45-54 years	Horry County, SC (45051)	C34.9 (Bronchus or lung, unspecified - Malignant neoplasms)	22

Σχήμα 4.12

Ρίσκο

Ένα μέσο όριο ρίσκου της τάξης του 0,1-0,05 είναι αποδεκτό στην περίπτωσή μας. Εάν θεωρήσουμε ότι τα δεδομένα μας είναι ευαίσθητα, το όριο θα πρέπει να τείνει στο 0,05.

Απειλές

Ας εξετάσουμε τις 3 πιθανές επιθέσεις από γνωστό επιτιθέμενο.

Α) Στην περίπτωση της σκόπιμης προσπάθειας επανα-προσδιορισμού πρέπει να υπολογίσουμε την πιθανότητα της προσπάθειας Pr(προσπάθεια). Στην περίπτωση του ερευνητή, θα πρέπει να έχει υπογράφει κάποια συνθήκη ασφάλειας για τα προσωπικά δεδομένα μεταξύ αυτού και του κέντρου και επίσης στον τομέα αυτό δεν υπάρχει κάποιος προφανής λόγο για να επαναπροσδιορίσει κάποιος τα δεδομένα αυτά. Συνεπώς μιας και τα δεδομένα μας δεν θεωρούνται ευαίσθητα μπορούμε να δώσουμε μια χαμηλή τιμή στην Pr(προσπαθεια)=0,4.

Β) Μια ακουσία προσπάθεια προϋποθέτει κάποιος να έχει έναν γνωστό μέσα στο δείγμα. Η πιθανότητα λοιπόν Pr(*αναγνώρισης*) είναι να γνωρίζει κάποιος κάποιον που να έχει πεθάνει το 2011. Για το έτος 2011 η *είχαμε από τα δημογραφικά στοιχεία της πολιτείας 4.626.875 πληθυσμό, άρα p=3.872/4.626.875=0,0008368. Συνεπώς Pr(αναγνώρισης)* = $1-(1-\mathrm{p})^{150}$ = 0,12

C) Η πιθανότητα *Pr(απώλειας)=0,27 η οποία είναι* βασισμένη σε στατιστικά δεδομένα.

Συνεπώς η συνολική μας πιθανότητα είναι Pr(επανα-προσδ., T) = *Pr*(T) × *Pr*(επανα-προσδ. | T) , όπου *Pr*(T) είναι μια από τις παραπάνω πιθανότητες. Επιλεγούμε την μεγαλύτερη πιθανότητα, που είναι αυτή της ακούσιας προσπάθειας, που είναι και η πιο πιθανή και έχουμε: *Pr*(επανα-προσδ., *αναγνώρισης*) = *0,4* × *Pr*(επανα-προσδ. | T)≤0,1 βάση του ρίσκου που θέσαμε παραπάνω. Άρα *Pr*(επανα-προσδ. | *αναγνώρισης*)≤0,25

Λογισμικό ARX

Πρόκειται για ένα εργαλείο ανωνυμοποίησης, που χρησιμοποιεί όλες τις προαναφερόμενες τεχνικές ανωνυμοποίησης όπως k-ανωνυμοποίηση, l-διαφορετικότητα και t-Closeness και έχει ως σκοπό να μετατρέψει ένα δείγμα σε ανώνυμο, ορίζοντας τα κατάλληλα πεδία δεδομένων και προσδιορίζοντας το κατάλληλο ρίσκο και ποσοστό απώλειας.

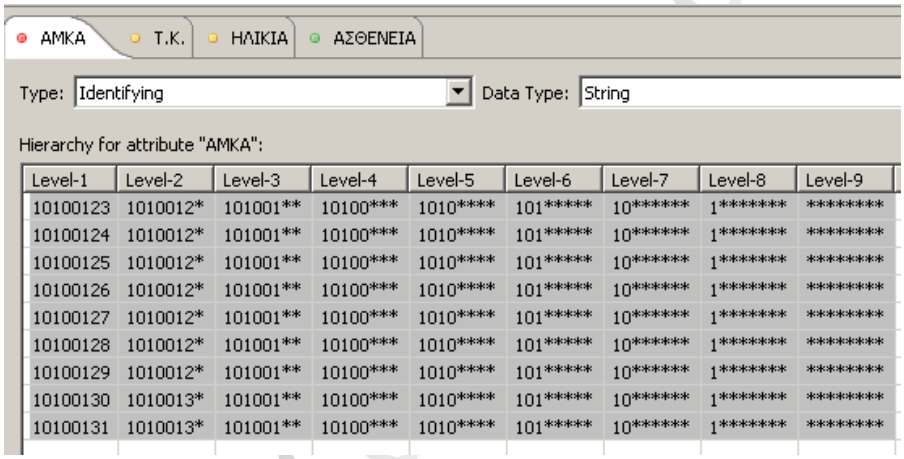
Κατά την ανωνυμοποίηση των δεδομένων, το πλαίσιο διακρίνει μεταξύ τεσσάρων διαφορετικών τύπων χαρακτηριστικά:

- Identifier χαρακτηριστικά (όπως όνομα) αφαιρούνται από το σύνολο δεδομένων.
- Quasi-αναγνωριστικά (όπως ηλικία ή Τ.Κ.) γενικευμένη εφαρμόζοντας τις παρεχόμενες ιεραρχίες γενίκευσης.
- Τα ευαίσθητα χαρακτηριστικά διατηρούνται όπως είναι και μπορεί να χρησιμοποιηθεί για να αντλήσει t-Close ή l-Diverce μεταμορφώσεις.
- Insensitive χαρακτηριστικά διατηρούνται όπως είναι.

Επί του παρόντος, οι μετρήσεις που υποστηρίζονται για την απώλεια πληροφοριών περιλαμβάνουν μονοτονική και μη-μονότονη με διαφερετικές παραλλαγές Ύψος, Ακριβεία, Ποιότητα και Εντροπία.

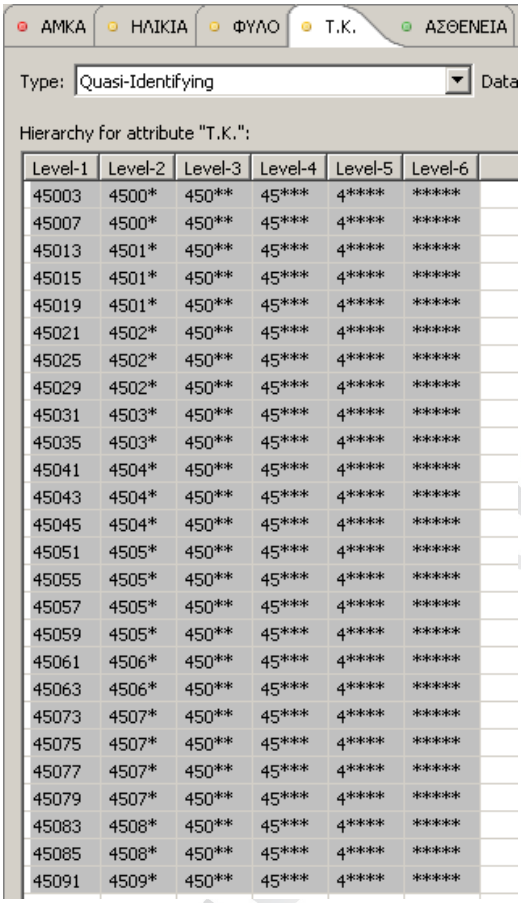
Τύποι χαρακτηριστικών

Ορίζουμε τους κατάλληλους τύπους χαρακτηριστικών του δείγματος μας και κατασκευάζουμε ιεραρχίες γενίκευσης για τόση Άμεσους και Εμέσους προσδιοριστές μας. Όπως παρατηρούμε στα σχήματα 4.12 και 4.13 για τον Άμεσο προσδιοριστή ΑΜΚΑ και τον Άμεσο προσδιοριστή Τ.Κ. έχουν δημιουργηθεί οι ιεραρχίες γενίκευσης βάση των όποιων θα καθορίσουμε το όριο του ρίσκου και θα εφαρμόσουμε το κατάλληλο επίπεδο γενίκευσης (Level-#).



Level-1	Level-2	Level-3	Level-4	Level-5	Level-6	Level-7	Level-8	Level-9
10100123	1010012*	101001**	10100***	1010****	101*****	10******	1*******	********
10100124	1010012*	101001**	10100***	1010****	101*****	10******	1*******	********
10100125	1010012*	101001**	10100***	1010****	101*****	10******	1*******	********
10100126	1010012*	101001**	10100***	1010****	101*****	10******	1*******	********
10100127	1010012*	101001**	10100***	1010****	101*****	10******	1*******	********
10100128	1010012*	101001**	10100***	1010****	101*****	10******	1*******	********
10100129	1010012*	101001**	10100***	1010****	101*****	10******	1*******	********
10100130	1010013*	101001**	10100***	1010****	101*****	10******	1*******	********
10100131	1010013*	101001**	10100***	1010****	101*****	10******	1*******	********

Σχήμα 4.13

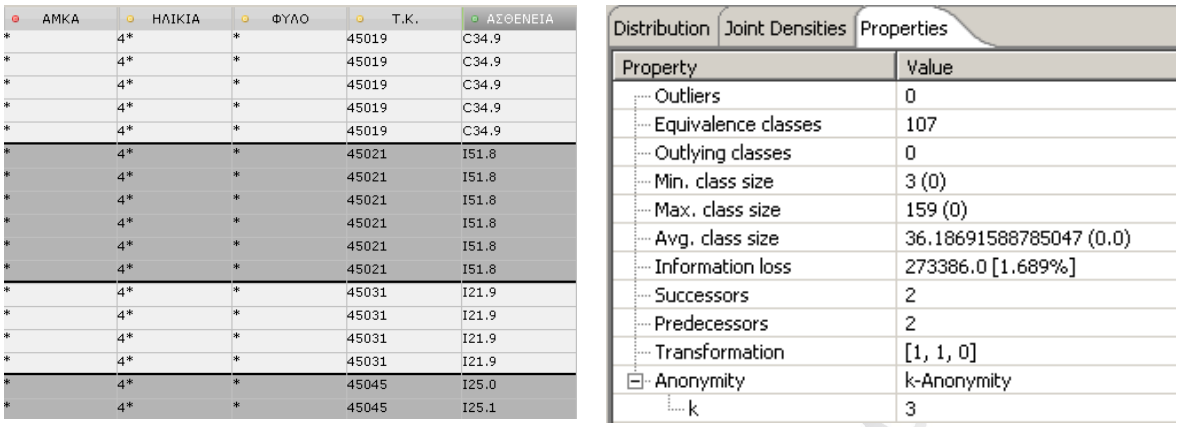


Level-1	Level-2	Level-3	Level-4	Level-5	Level-6
45003	4500*	450**	45***	4****	*****
45007	4500*	450**	45***	4****	*****
45013	4501*	450**	45***	4****	*****
45015	4501*	450**	45***	4****	*****
45019	4501*	450**	45***	4****	*****
45021	4502*	450**	45***	4****	*****
45025	4502*	450**	45***	4****	*****
45029	4502*	450**	45***	4****	*****
45031	4503*	450**	45***	4****	*****
45035	4503*	450**	45***	4****	*****
45041	4504*	450**	45***	4****	*****
45043	4504*	450**	45***	4****	*****
45045	4504*	450**	45***	4****	*****
45051	4505*	450**	45***	4****	*****
45055	4505*	450**	45***	4****	*****
45057	4505*	450**	45***	4****	*****
45059	4505*	450**	45***	4****	*****
45061	4506*	450**	45***	4****	*****
45063	4506*	450**	45***	4****	*****
45073	4507*	450**	45***	4****	*****
45075	4507*	450**	45***	4****	*****
45077	4507*	450**	45***	4****	*****
45079	4507*	450**	45***	4****	*****
45083	4508*	450**	45***	4****	*****
45085	4508*	450**	45***	4****	*****
45091	4509*	450**	45***	4****	*****

Σχήμα 4.14

3-Ανωνυμοποίηση

Αν εφαρμόσουμε στο δείγμα μας 3-ανωνυμοποιηση (Σχήμα 4.15) έχουμε απώλεια πληροφορίας της τάξης του 1.689% για τους γενικευμένους προσδιοριστές μας και έχουν δημιουργηθεί ισοδύναμες κλάσεις εφαρμόζοντας την [1,1,0] ιεραρχία γενίκευσης.



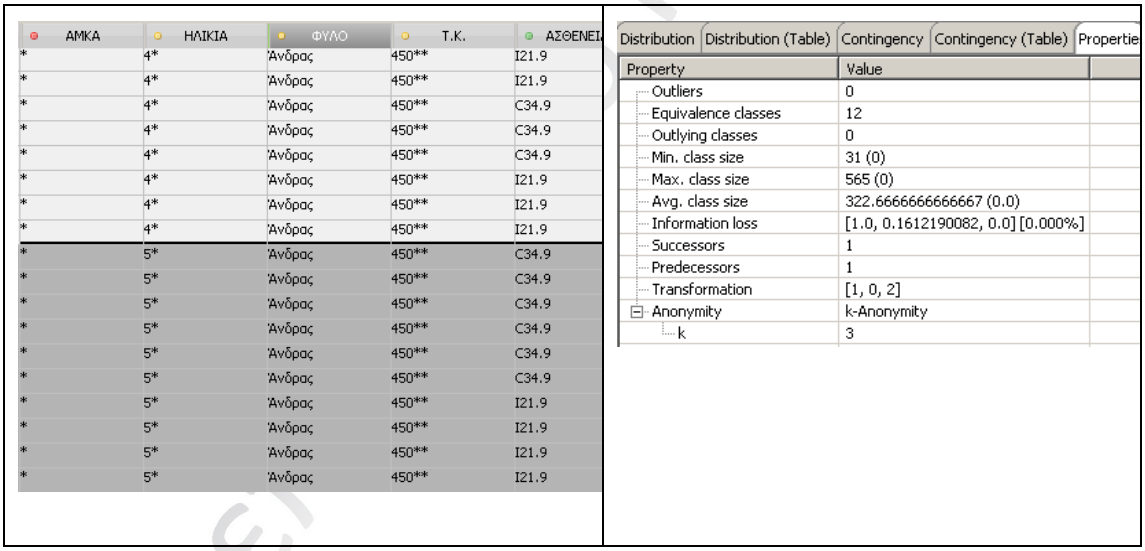
ΑΜΚΑ	ΗΛΙΚΙΑ	ΦΥΛΟ	Τ.Κ.	ΑΣΘΕΝΕΙΑ
*	4*	*	45019	C34.9
*	4*	*	45019	C34.9
*	4*	*	45019	C34.9
*	4*	*	45019	C34.9
*	4*	*	45019	C34.9
*	4*	*	45021	I51.8
*	4*	*	45021	I51.8
*	4*	*	45021	I51.8
*	4*	*	45021	I51.8
*	4*	*	45021	I51.8
*	4*	*	45021	I51.8
*	4*	*	45031	I21.9
*	4*	*	45031	I21.9
*	4*	*	45031	I21.9
*	4*	*	45031	I21.9
*	4*	*	45045	I25.0
*	4*	*	45045	I25.1

Distribution | Joint Densities | Properties

Property	Value
Outliers	0
Equivalence classes	107
Outlying classes	0
Min. class size	3 (0)
Max. class size	159 (0)
Avg. class size	36.18691588785047 (0.0)
Information loss	273386.0 [1.689%]
Successors	2
Predecessors	2
Transformation	[1, 1, 0]
Anonymity	k-Anonymity
k	3

Σχήμα 4.15 (α)

Εάν υποθέσουμε ότι μας ενδιαφέρει το φύλο τότε θα πρέπει να αλλάξουμε ιεραρχία γεννίκευσης. Έτσι έχουμε τα παρακάτω αποτελέσματα:



ΑΜΚΑ	ΗΛΙΚΙΑ	ΦΥΛΟ	Τ.Κ.	ΑΣΘΕΝΕΙ
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	5*	Άνδρας	450**	C34.9
*	5*	Άνδρας	450**	C34.9
*	5*	Άνδρας	450**	C34.9
*	5*	Άνδρας	450**	C34.9
*	5*	Άνδρας	450**	C34.9
*	5*	Άνδρας	450**	C34.9
*	5*	Άνδρας	450**	I21.9
*	5*	Άνδρας	450**	I21.9
*	5*	Άνδρας	450**	I21.9
*	5*	Άνδρας	450**	I21.9

Distribution | Distribution (Table) | Contingency | Contingency (Table) | Propertie

Property	Value
Outliers	0
Equivalence classes	12
Outlying classes	0
Min. class size	31 (0)
Max. class size	565 (0)
Avg. class size	322.6666666666667 (0.0)
Information loss	[1.0, 0.1612190082, 0.0] [0.000%]
Successors	1
Predecessors	1
Transformation	[1, 0, 2]
Anonymity	k-Anonymity
k	3

Σχήμα 4.15 (β)

Lattice | List

Transformation	Anonymity	Min. Info. Loss	Max. Info. Loss
[1, 1, 0]	ANONYMOUS	273386.0 [1.689%]	273386.0 [1.689%]
[0, 1, 2]	ANONYMOUS	393850.0 [2.493%]	393850.0 [2.493%]
[0, 1, 3]	ANONYMOUS	393850.0 [2.493%]	393850.0 [2.493%]
[0, 1, 4]	ANONYMOUS	393850.0 [2.493%]	393850.0 [2.493%]
[2, 0, 1]	ANONYMOUS	970598.0 [6.346%]	970598.0 [6.346%]
[1, 0, 2]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 3]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 4]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 5]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[2, 0, 2]	ANONYMOUS	1565886.0 [10.322%]	7496194.0 [49.931%]
[2, 0, 3]	ANONYMOUS	1565886.0 [10.322%]	7496194.0 [49.931%]
[2, 0, 4]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[2, 1, 0]	ANONYMOUS	273386.0 [1.689%]	1.4992384E7 [100.000%]
[1, 1, 1]	ANONYMOUS	273386.0 [1.689%]	1.4992384E7 [100.000%]
[2, 1, 1]	ANONYMOUS	970598.0 [6.346%]	1.4992384E7 [100.000%]
[1, 1, 2]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 2]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[1, 1, 3]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 3]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[1, 1, 4]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[0, 1, 5]	ANONYMOUS	393850.0 [2.493%]	1.4992384E7 [100.000%]
[2, 1, 4]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 0, 5]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[1, 1, 5]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 5]	ANONYMOUS	1.4992384E7 [100.000%]	1.4992384E7 [100.000%]

Σχήμα 4.16

Σχήμα 4.17

Όπως φαίνεται στο Σχήμα 4.16 η ιεραρχία Γενίκευσης που χρησιμοποιήθηκε είναι μεταξύ των επιπέδων γενίκευσης [1,1,0] με απώλεια πληροφορίας 1.689%

Η ευπάθεια της k-ανωνυμοποίησης μπορεί να παρατηρηθεί στο παραπάνω σχήμα όπου και διακρίνεται η επίθεση, της γνώσης στοιχείων μιας οντότητας. Εάν γνωρίζει λοιπόν κάποιος την ηλικία ενός ατόμου (40-49) και τον Τ.Κ. 45031 και εάν γνωρίζει ότι δεν έχει πρόβλημα με την I21.9-1, τότε συμπεραίνει ότι πάσχει από I21.9-2.

Η 3-ανωνυμοποίηση στο δείγμα μας δεν είναι επαρκής βάση του ρίσκου που έχουμε ορίσει. Για το λόγο αυτό χρησιμοποιούμε μεγαλύτερο k, ώστε σε κάθε ισοδύναμη κλάση να έχουμε ελάχιστο μέγεθος αρκετό ώστε να πληροί το όριο του ρίσκου που θέσαμε.

4-Ανωνυμοποίηση

Αν εφαρμόσουμε στο δείγμα μας 4-ανωνυμοποιηση έχουμε απώλεια πληροφορίας της τάξης του 2.493% για τους γενικευμένους προσδιοριστές μας και έχουν δημιουργηθεί ισοδύναμες κλάσεις εφαρμόζοντας την [0,1,2] ιεραρχία γενίκευσης.

ΑΜΚΑ	ΗΛΙΚΙΑ	ΦΥΛΟ	Τ.Κ.	ΑΣΘΕΝΕΙΑ
*	45	*	450**	I51.8
*	45	*	450**	I51.8
*	45	*	450**	I25.0
*	45	*	450**	I25.1
*	45	*	450**	I25.0
*	45	*	450**	C34.9
*	45	*	450**	C34.9
*	46	*	450**	C34.9
*	46	*	450**	C34.9
*	46	*	450**	I51.8
*	46	*	450**	I21.9
*	46	*	450**	I25.1
*	46	*	450**	C34.9
*	46	*	450**	I21.9
*	46	*	450**	I21.9
*	46	*	450**	C34.9
*	46	*	450**	C34.9
*	46	*	450**	C34.9
*	46	*	450**	I21.9
*	46	*	450**	I25.0
*	46	*	450**	C34.9
*	46	*	450**	I21.9
*	46	*	450**	I21.9
*	47	*	450**	C34.9

Distribution | Joint Densities | Properties

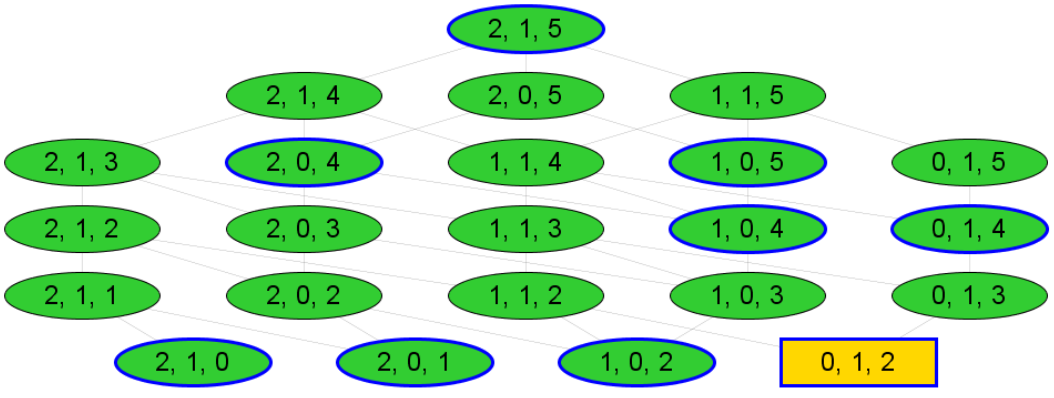
Property	Value
Outliers	0
Equivalence classes	51
Outlying classes	0
Min. class size	7 (0)
Max. class size	171 (0)
Avg. class size	75.92156862745098 (0.0)
Information loss	393850.0 [2.493%]
Successors	2
Predecessors	2
Transformation	[0, 1, 2]
Anonymity	k-Anonymity
k	4

Σχήμα 4.18

Lattice | List

Transformation	Anonymity	Min. Info. Loss	Max. Info. Loss
[0, 1, 2]	ANONYMOUS	393850.0 [2.493%]	393850.0 [2.493%]
[0, 1, 3]	ANONYMOUS	393850.0 [2.493%]	393850.0 [2.493%]
[0, 1, 4]	ANONYMOUS	393850.0 [2.493%]	393850.0 [2.493%]
[2, 0, 1]	ANONYMOUS	970598.0 [6.346%]	970598.0 [6.346%]
[2, 1, 0]	ANONYMOUS	1278208.0 [8.400%]	1278208.0 [8.400%]
[1, 0, 2]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 3]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 4]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 5]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[2, 0, 2]	ANONYMOUS	1565886.0 [10.322%]	7496194.0 [49.931%]
[2, 0, 3]	ANONYMOUS	1565886.0 [10.322%]	7496194.0 [49.931%]
[2, 0, 4]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[2, 1, 1]	ANONYMOUS	1278208.0 [8.400%]	1.4992384E7 [100.000%]
[1, 1, 2]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 2]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[1, 1, 3]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 3]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[1, 1, 4]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[0, 1, 5]	ANONYMOUS	393850.0 [2.493%]	1.4992384E7 [100.000%]
[2, 1, 4]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 0, 5]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[1, 1, 5]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 5]	ANONYMOUS	1.4992384E7 [100.000%]	1.4992384E7 [100.000%]

Σχήμα 4.19



Σχήμα 4.20

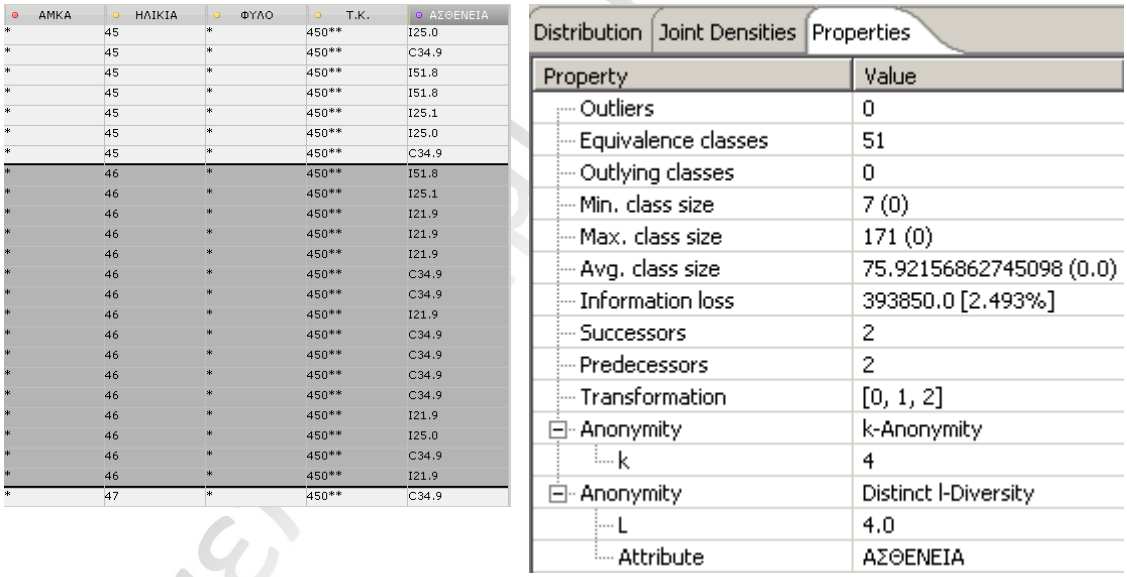
Συγκριτικά με την 3-ανωνυμοποίηση παρατήσουμε ότι έχουμε λιγότερο αριθμό ισοδύναμων κλάσεων με μεγαλύτερη απώλεια πληροφορίας αφού έχει εφαρμοστεί μεγαλύτερη ιεραρχία γενίκευσης.

4-διαφορετικότητα

Η εφαρμογή της 4-διαφορετικοτητας αναφέρεται στο ότι πρέπει να υπάρχουν τουλάχιστον 4 καλές αναπαραστάσεις σε κάθε μια από τις ισοδύναμες κλάσεις για κάθε ευαίσθητη ιδιότητα (ΑΣΘΕΝΕΙΑ).

Διακριτή 4-διαφορετικότητα

Όπως προαναφέραμε αυτός ο τύπος δεν προστατεύει το δείγμα μας από το να συνάγει κάποιος συμπερασματικά πληροφορίες για κάποιον μέσα σε ένα δείγμα.



ΑΜΚΑ	ΗΛΙΚΙΑ	ΦΥΛΟ	Τ.Κ.	ΑΣΘΕΝΕΙΑ
*	45	*	450**	I25.0
*	45	*	450**	C34.9
*	45	*	450**	I51.8
*	45	*	450**	I51.8
*	45	*	450**	I25.1
*	45	*	450**	I25.0
*	45	*	450**	C34.9
*	46	*	450**	I51.8
*	46	*	450**	I25.1
*	46	*	450**	I21.9
*	46	*	450**	I21.9
*	46	*	450**	I21.9
*	46	*	450**	C34.9
*	46	*	450**	C34.9
*	46	*	450**	I21.9
*	46	*	450**	C34.9
*	46	*	450**	C34.9
*	46	*	450**	C34.9
*	46	*	450**	C34.9
*	46	*	450**	I21.9
*	46	*	450**	I25.0
*	46	*	450**	C34.9
*	46	*	450**	I21.9
*	47	*	450**	C34.9

Distribution | Joint Densities | Properties

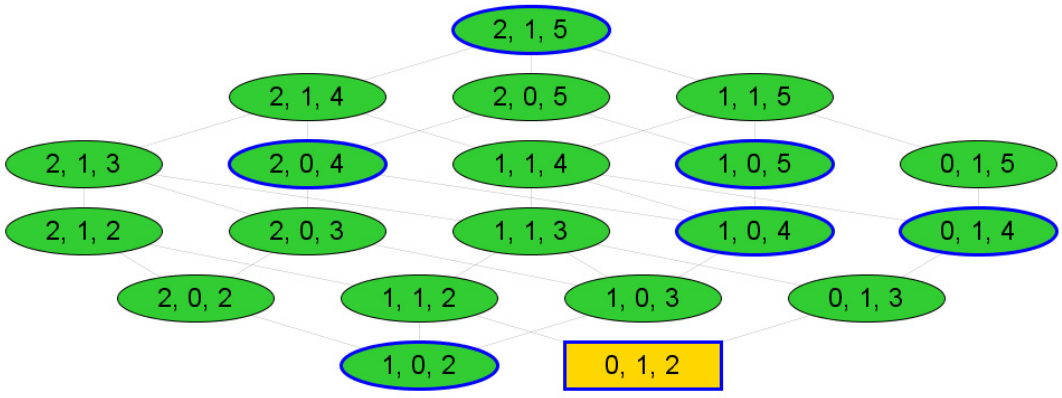
Property	Value
Outliers	0
Equivalence classes	51
Outlying classes	0
Min. class size	7 (0)
Max. class size	171 (0)
Avg. class size	75.92156862745098 (0.0)
Information loss	393850.0 [2.493%]
Successors	2
Predecessors	2
Transformation	[0, 1, 2]
Anonymity	k-Anonymity
k	4
Anonymity	Distinct l-Diversity
L	4.0
Attribute	ΑΣΘΕΝΕΙΑ

Σχήμα 4.21

Lattice | List

Transformation	Anonymity	Min. Info. Loss	Max. Info. Loss
[0, 1, 2]	ANONYMOUS	393850.0 [2.493%]	393850.0 [2.493%]
[0, 1, 3]	ANONYMOUS	393850.0 [2.493%]	393850.0 [2.493%]
[0, 1, 4]	ANONYMOUS	393850.0 [2.493%]	393850.0 [2.493%]
[1, 0, 2]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 3]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 4]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 5]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[2, 0, 2]	ANONYMOUS	1565886.0 [10.322%]	7496194.0 [49.931%]
[2, 0, 3]	ANONYMOUS	1565886.0 [10.322%]	7496194.0 [49.931%]
[2, 0, 4]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[1, 1, 2]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 2]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[1, 1, 3]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 3]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[1, 1, 4]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[0, 1, 5]	ANONYMOUS	393850.0 [2.493%]	1.4992384E7 [100.000%]
[2, 1, 4]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 0, 5]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[1, 1, 5]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 5]	ANONYMOUS	1.4992384E7 [100.000%]	1.4992384E7 [100.000%]

Σχήμα 4.22



Σχήμα 4.23

Όπως φαίνεται στο Σχήμα 4.18 η ιεραρχία Γενίκευσης που χρησιμοποιήθηκε είναι μεταξύ των επιπέδων [0,1,2] με απώλεια πληροφορίας 2.493%

Εντροπία 4-διαφορετικότητα

Η εντροπία 4-διαφορετικότητα ορίζει ότι στο δείγμα μας θα πρέπει σε κάθε ισοδύναμη κλάση η Εντροπία ≥log(4). Όπως αναφέραμε η εντροπία αποτελεί πιο ισχυρό ορισμό από την διακριτή και όπως φαίνεται στο σχήμα 4.19 η ιεραρχία γενίκευσης μεγάλωσε.

ΑΜΚΑ	ΗΛΙΚΙΑ	ΦΥΛΟ	Τ.Κ.	ΑΣΘΕΝΕΙΑ
*	4*	*	450**	C34.9
*	4*	*	450**	C34.9
*	4*	*	450**	C34.9
*	4*	*	450**	C34.9
*	4*	*	450**	C34.9
*	4*	*	450**	I51.8
*	4*	*	450**	I51.8
*	4*	*	450**	I51.8
*	4*	*	450**	I51.8
*	4*	*	450**	I51.8
*	4*	*	450**	I51.8
*	4*	*	450**	I21.9
*	4*	*	450**	I21.9
*	4*	*	450**	I21.9
*	4*	*	450**	I21.9
*	4*	*	450**	I25.0
*	4*	*	450**	I25.1
*	4*	*	450**	I25.1
*	4*	*	450**	C34.9
*	4*	*	450**	I25.1
*	4*	*	450**	I25.1
*	4*	*	450**	I25.1
*	4*	*	450**	I25.0
*	4*	*	450**	I25.0

Distribution | Joint Densities | Properties

Property	Value
Outliers	0
Equivalence classes	6
Outlying classes	0
Min. class size	76 (0)
Max. class size	1085 (0)
Avg. class size	645.3333333333334 (0.0
Information loss	3126786.0 [20.747%]
Successors	2
Predecessors	3
Transformation	[1, 1, 2]
Anonymity	k-Anonymity
k	4
Anonymity	Entropy l-Diversity
l	4.0
Attribute	ΑΣΘΕΝΕΙΑ

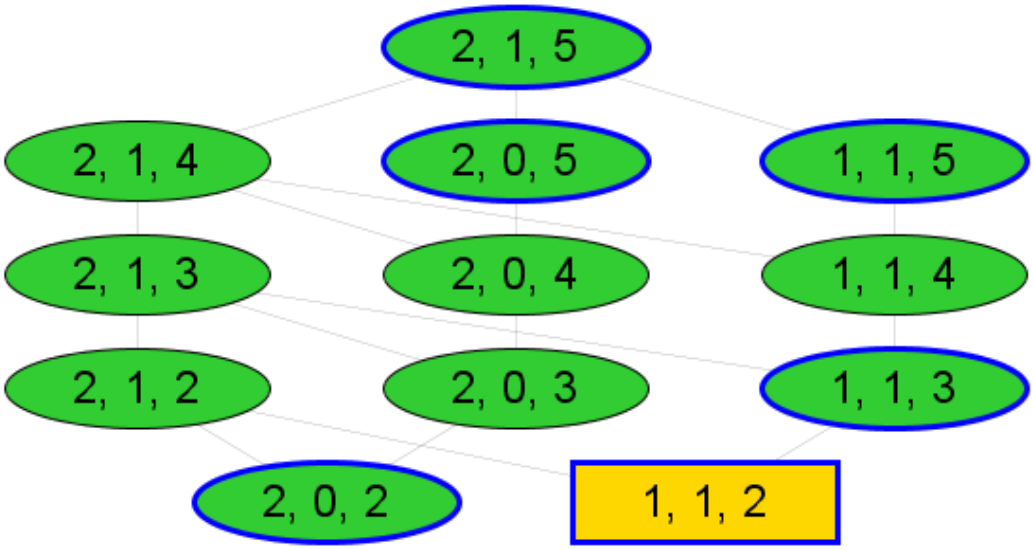
Σχήμα 4.24

Η βέλτιστη γενίκευση είναι [1,1,2] με απώλειας πληροφορίας 20.747% όπως βλέπουμε παρακάτω.

Lattice | List

Transformation	Anonymity	Min. Info. Loss	Max. Info. Loss
[1, 1, 2]	ANONYMOUS	3126786.0 [20.747%]	3126786.0 [20.747%]
[1, 1, 3]	ANONYMOUS	3126786.0 [20.747%]	3126786.0 [20.747%]
[1, 1, 4]	ANONYMOUS	3126786.0 [20.747%]	3126786.0 [20.747%]
[1, 1, 5]	ANONYMOUS	3126786.0 [20.747%]	3126786.0 [20.747%]
[2, 0, 2]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[2, 0, 3]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[2, 0, 4]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[2, 0, 5]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[2, 1, 2]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 1, 3]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 1, 4]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 1, 5]	ANONYMOUS	1.4992384E7 [100.000%]	1.4992384E7 [100.000%]

Σχήμα 4.25



Σχήμα 4.26

Η εντροπία όπως παρατηρούμε θέτει υψηλές απαιτήσεις και υπάρχει περίπτωση η εντροπία του δείγματος μας να είναι μικρή εάν έχουμε μερικές ευαίσθητες τιμές αρκετά συχνά.

Αναδρομική (4,5)-διαφορετικοτητα

Ορίζει ότι οι πιο συχνές τιμές δεν εμφανίζονται τόσο συχνά και οι λιγότερες συχνές όχι τόσο σπάνια. Εφαρμόζοντας στο δείγμα μας αυτό τον τύπο δεν μπορούμε να παράγουμε ανώνυμο δείγμα όπως βλέπουμε. Εφαρμόζοντας γενίκευση επίπεδου [1,0,2] έχουμε τα παρακάτω αποτελέσματα

ΑΜΚΑ	ΗΛΙΚΙΑ	ΦΥΛΟ	Τ.Κ.	ΑΣΘΕΝΕΙΑ
*	4*	Άνδρας	450**	I25.1
*	4*	Άνδρας	450**	I25.1
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Γυναίκα	450**	C34.9
*	4*	Γυναίκα	450**	C34.9
*	4*	Γυναίκα	450**	C34.9
*	4*	Γυναίκα	450**	I51.8
*	4*	Γυναίκα	450**	I51.8

Distribution | Joint Densities | Properties

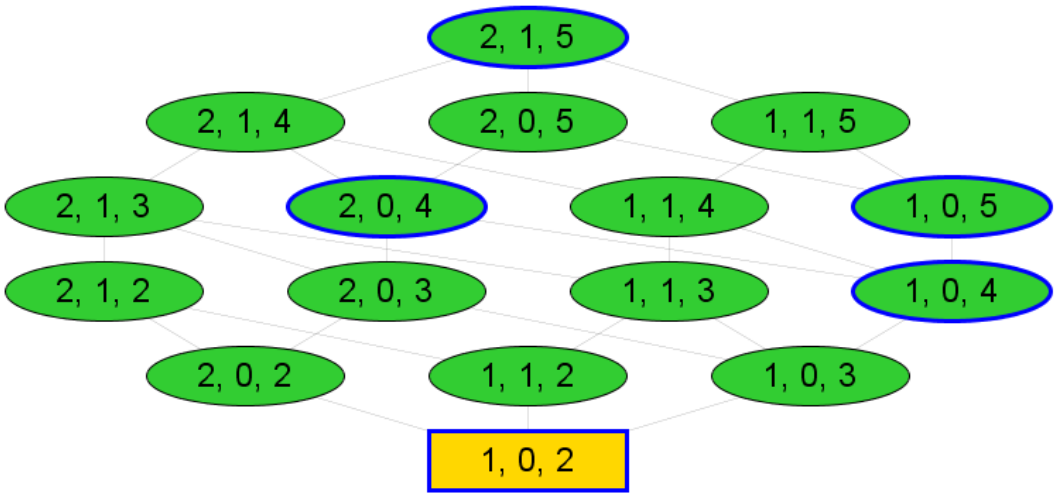
Property	Value
Outliers	0
Equivalence classes	12
Outlying classes	0
Min. class size	31 (0)
Max. class size	565 (0)
Avg. class size	322.6666666666667 (0.0)
Information loss	1565886.0 [10.322%]
Successors	3
Predecessors	2
Transformation	[1, 0, 2]
⊟ Anonymity	k-Anonymity
k	4
⊟ Anonymity	Recursive-(c,l)-Diversity
c	4.9
l	4.0
Attribute	ΑΣΘΕΝΕΙΑ

Σχήμα 4.27

Lattice | List

Transformation	Anonymity	Min. Info. Loss	Max. Info. Loss
[1, 0, 2]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 3]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 4]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 5]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[2, 0, 2]	ANONYMOUS	1565886.0 [10.322%]	7496194.0 [49.931%]
[2, 0, 3]	ANONYMOUS	1565886.0 [10.322%]	7496194.0 [49.931%]
[2, 0, 4]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[1, 1, 2]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 2]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[1, 1, 3]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 3]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[1, 1, 4]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 4]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 0, 5]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[1, 1, 5]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 5]	ANONYMOUS	1.4992384E7 [100.000%]	1.4992384E7 [100.000%]

Σχήμα 4.28



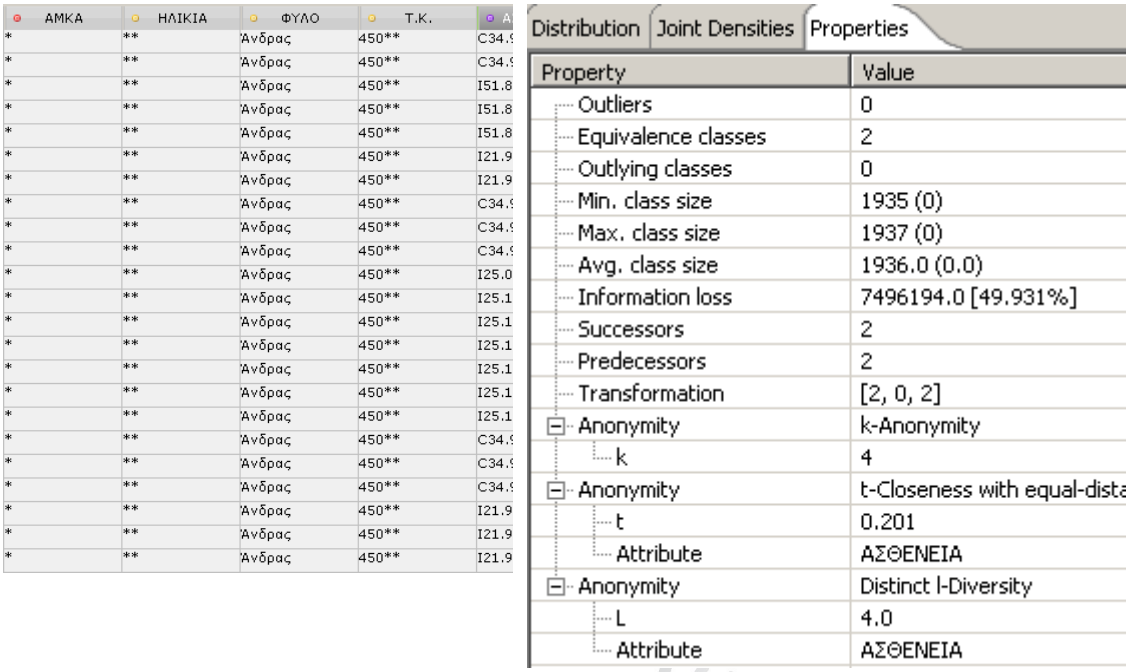
Σχήμα 4.29

4-anonymization με 0.2-closeness

Η closeness όπως έχουμε προαναφέρει προστατεύει πληροφορίες και όχι εγγραφές. Για αυτό το λόγο χρησιμοποιείται μαζί με την 4-anonymization.

Ισοδύναμη EMD

Η 4-anonymization ορίζει ότι θα έχουμε τουλάχιστον 4 ίδιες αναπαραστάσεις σε κάθε ισοδύναμη κλάση. Η 0.2 – closeness μας ορίζει για την ευαίσθητη ιδιότητα μας, βάση της ιεραρχίας γενίκευσης, ότι η σχετική απόσταση κάθε τιμής της σε κάθε ισοδύναμη κλάση αλλά και σε ολόκληρο το δείγμα δεν θα είναι μεγαλύτερη από 0.2. Βάση αυτών παρατηρούμε μεγάλη απώλεια πληροφορίας 49.93% καθώς η εφαρμογή του ορίου του 0.2 είναι αρκετά μικρή.



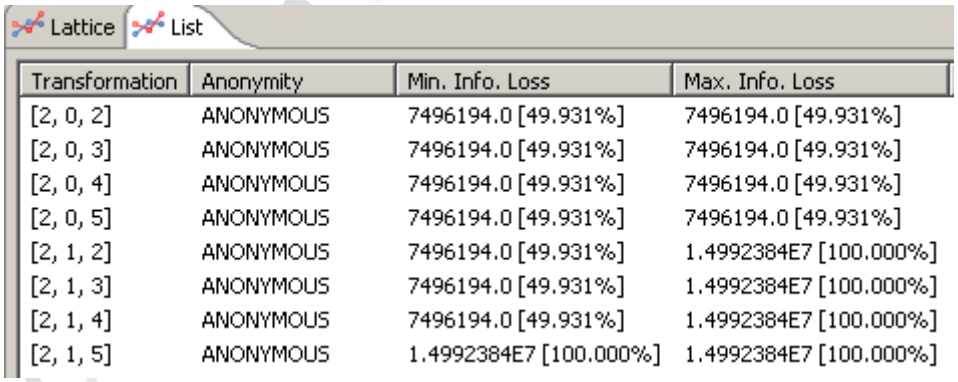
ΑΜΚΑ	ΗΛΙΚΙΑ	ΦΥΛΟ	Τ.Κ.	Α
*	**	Άνδρας	450**	C34.9
*	**	Άνδρας	450**	C34.9
*	**	Άνδρας	450**	I51.8
*	**	Άνδρας	450**	I51.8
*	**	Άνδρας	450**	I51.8
*	**	Άνδρας	450**	I21.9
*	**	Άνδρας	450**	I21.9
*	**	Άνδρας	450**	C34.9
*	**	Άνδρας	450**	C34.9
*	**	Άνδρας	450**	C34.9
*	**	Άνδρας	450**	I25.0
*	**	Άνδρας	450**	I25.1
*	**	Άνδρας	450**	I25.1
*	**	Άνδρας	450**	I25.1
*	**	Άνδρας	450**	I25.1
*	**	Άνδρας	450**	I25.1
*	**	Άνδρας	450**	I25.1
*	**	Άνδρας	450**	C34.9
*	**	Άνδρας	450**	C34.9
*	**	Άνδρας	450**	C34.9
*	**	Άνδρας	450**	I21.9
*	**	Άνδρας	450**	I21.9
*	**	Άνδρας	450**	I21.9

Distribution | Joint Densities | Properties

Property	Value
Outliers	0
Equivalence classes	2
Outlying classes	0
Min. class size	1935 (0)
Max. class size	1937 (0)
Avg. class size	1936.0 (0.0)
Information loss	7496194.0 [49.931%]
Successors	2
Predecessors	2
Transformation	[2, 0, 2]
Anonymity	k-Anonymity
k	4
Anonymity	t-Closeness with equal-dista
t	0.201
Attribute	ΑΣΘΕΝΕΙΑ
Anonymity	Distinct l-Diversity
L	4.0
Attribute	ΑΣΘΕΝΕΙΑ

Σχήμα 4.30

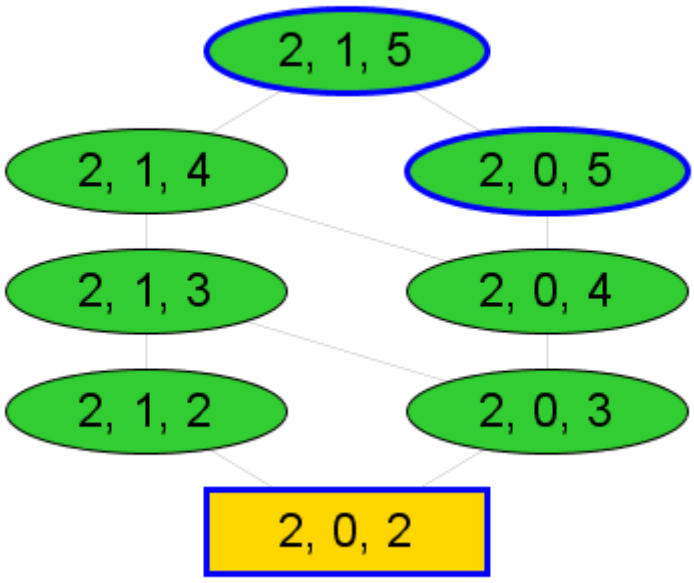
Παρατηρούμε ακόμη πως διαμορφώνεται η απώλεια, βάση της ιεραρχίας γενίκευσης.



Lattice | List

Transformation	Anonymity	Min. Info. Loss	Max. Info. Loss
[2, 0, 2]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[2, 0, 3]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[2, 0, 4]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[2, 0, 5]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[2, 1, 2]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 1, 3]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 1, 4]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 1, 5]	ANONYMOUS	1.4992384E7 [100.000%]	1.4992384E7 [100.000%]

Σχήμα 4.31



Σχήμα 4.32

4-ανωνυμοποιηση με 0.5-Closeness

Ισοδύναμη EMD

ΑΜΚΑ	ΗΛΙΚΙΑ	ΦΥΛΟ	Τ.Κ.	ΑΣΘΕ
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	I51.8
*	4*	Άνδρας	450**	I51.8
*	4*	Άνδρας	450**	I51.8
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	I25.0
*	4*	Άνδρας	450**	I25.1
*	4*	Άνδρας	450**	I25.1
*	4*	Άνδρας	450**	I25.1
*	4*	Άνδρας	450**	I25.1
*	4*	Άνδρας	450**	I25.1
*	4*	Άνδρας	450**	I25.1
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	C34.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9
*	4*	Άνδρας	450**	I21.9

Distribution | Joint Densities | Properties

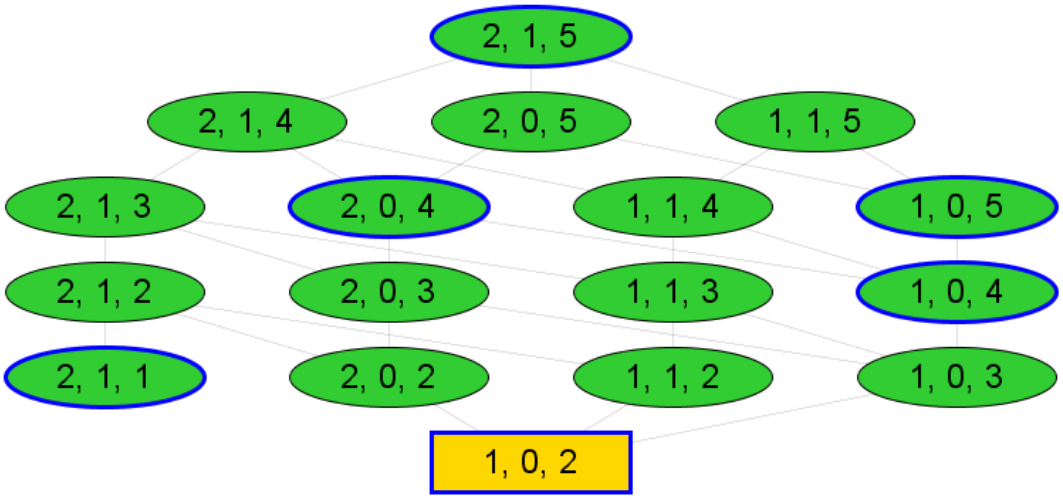
Property	Value
Outliers	0
Equivalence classes	12
Outlying classes	0
Min. class size	31 (0)
Max. class size	565 (0)
Avg. class size	322.6666666666667 (0.0)
Information loss	1565886.0 [10.322%]
Successors	3
Predecessors	2
Transformation	[1, 0, 2]
Anonymity	k-Anonymity
k	4
Anonymity	t-Closeness with equal-distanc
t	0.5
Attribute	ΑΣΘΕΝΕΙΑ
Anonymity	Distinct l-Diversity
L	4.0
Attribute	ΑΣΘΕΝΕΙΑ

Σχήμα 4.33

Lattice | List

Transformation	Anonymity	Min. Info. Loss	Max. Info. Loss
[1, 0, 2]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 3]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 4]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[1, 0, 5]	ANONYMOUS	1565886.0 [10.322%]	1565886.0 [10.322%]
[2, 1, 1]	ANONYMOUS	1937776.0 [12.805%]	1937776.0 [12.805%]
[2, 0, 2]	ANONYMOUS	1565886.0 [10.322%]	7496194.0 [49.931%]
[2, 0, 3]	ANONYMOUS	1565886.0 [10.322%]	7496194.0 [49.931%]
[2, 0, 4]	ANONYMOUS	7496194.0 [49.931%]	7496194.0 [49.931%]
[1, 1, 2]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 2]	ANONYMOUS	1937776.0 [12.805%]	1.4992384E7 [100.000%]
[1, 1, 3]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 3]	ANONYMOUS	1937776.0 [12.805%]	1.4992384E7 [100.000%]
[1, 1, 4]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 4]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[2, 0, 5]	ANONYMOUS	7496194.0 [49.931%]	1.4992384E7 [100.000%]
[1, 1, 5]	ANONYMOUS	1565886.0 [10.322%]	1.4992384E7 [100.000%]
[2, 1, 5]	ANONYMOUS	1.4992384E7 [100.000%]	1.4992384E7 [100.000%]

Σχήμα 4.34



Σχήμα 4.35

Συμπεράσματα

Συγκεντρωτικά έχουμε τα παρακάτω αποτελέσματα για το δείγμα μας. Όπως αναφέραμε το αποδεκτό όριο του ρίσκου μας καλύπτεται σε όλες τις περιπτώσεις εκτός από την πρώτη (3 anonymity).

	3anonymity	**4anonymity**	**Διακριτή 4anonymity**	**Εντροπία 4anonymity**	**Αναδρομική (4,5)anonymity**	**4anonymity with 0.2closeness**	**4anonymity with 0.5closeness**
Outliers	0	0	0	0	0	0	0
Equivalence Class	107	51	51	6	12	2	12
Outlying Classes	0	0	0	0	0	0	0
Min. Class Size	3(0)	7(0)	7(0)	76(0)	31(0)	1935(0)	31(0)
Max. Class Size	159	171	171	1085	565	1937	565

Avg. Class Size	36.19	75.92	75.92	645.33	322.66	1936	322.66
Information Loss	1.69%	2. 49%	2.49%	20.75%	10.32%	49.93%	10.32%
Successors	2	2	2	2	3	2	3
Predecessors	2	2	2	3	2	2	2
Transformation	[1,1,0]	[0,1,2]	[0,1,2]	[1,1,2]	[1,0,2]	[2,0,2]	[1,0,2]

Από τα αποτελέσματα που έχουμε βλέπουμε ότι η κάθε μέθοδος χρησιμοποιεί τις δικές της ιεραρχίες γενίκευσης, χωρίς αυτό βέβαια να αποκλείει ότι θα έχουμε το ίδιο βαθμό ανωνυμοποίησης του δείγματος χρησιμοποιώντας κάποια άλλη.

Παρατηρούμε ακόμη στους μετασχηματισμούς, όσο μεγαλύτερη είναι η γενίκευση τόσο μεγαλύτερη είναι και η απώλεια. Οι μέθοδοι λοιπόν 4 και 6 με ιεραρχίες γενίκευσης [1,1,2] και με την [2,0,2] έχουν το μεγαλύτερο ποσοστό απωλειών, το οποίο είναι λογικό αφού η πληροφορία των quasi-Identifier γενικεύεται σε μεγαλύτερο βαθμό.

Η μέθοδος που θα πρέπει να ακολουθήσουμε εξαρτάται κάθε φορά από αρκετούς παράγοντες. Ο πιο σημαντικός είναι ο καθορισμός του ρίσκου και κατόπιν να προσδιορίσουμε την πληροφορία που πρέπει να εξαχθεί π.χ. για κάποιον μπορεί να είναι σημαντική η ηλικία, αλλά όχι η μεγάλη ακρίβεια στην ιεραρχία του Τ.Κ. ορίζοντας έτσι σωστά την ιεραρχία γενίκευσης που θα πρέπει να ακολουθήσουμε. Αφού γίνουν αυτά και ανάλογα με τη φύση δείγματος μας, εφαρμόζουμε τα κατάλληλα φίλτρα των μεθόδων που παρουσιάστηκαν και κρίνουμε εάν είναι κατάλληλο για δημοσιοποίηση ή όχι.

Βιβλιογραφία

1. K. El Emam, E. Jonker, E. Moher, and L. Arbuckle, "A Review of Evidence on Consent Bias in Research," American Journal of Bioethics 13:4 (2013): 42–44.
2. K. El Emam, A Guide to the De-identification of Personal Health Information, (Boca Raton, FL: CRC Press/Auerbach, 2013).
3. Landweher vs. AOL Inc. Case No. 1:11-cv-01014-CMH-TRJ in the District Court in the Eastern District of Virginia.
4. L. Sanches, "2012 HIPAA Privacy and Security Audits," Office for Civil Rights, Department of Health and Human Services.
5. F. Dankar and K. El Emam, "Practicing Differential Privacy in Health Care: A Review," Transactions on Data Privacy 6:1 (2013): 35–67.
6. K. El Emam, A Guide to the De-identification of Personal Health Information (Boca Raton, FL: CRC Press/Auerbach, 2013).
7. en.wikipedia.org/wiki/Dunbar's_number
8. Symantec, Symantec Global Internet Threat Report—Trends for July-December 07 (Symantec Enterprise Security, 2008).
9. B. Krebs, "Hackers Break into Virginia Health Professions Database, Demand Ransom," Washington Post, 4 May 2009.
10. P. Samarati and L. Sweeney, "Protecting Privacy When Disclosing Information: k-anonymity and Its Enforcement Through Generalisation and Suppression," Technical Report SRI-CSL-98-04 (Menlo Park, CA: SRI International, 1998).
11. Tiancheng Li and Ninghui Li, On the Tradeoff Between Privacy and Utility in Data Publishing
12. L. Sweeney, k-anonymity a model for protecting privacy. Int. J. Uncertain. Fuzz., 10(5):557–570, 2002.
13. National Association of Health Data Organizations, A Guide to State-Level Ambulatory Care Data Collection Activities (Falls Church: National Association of Health Data Organizations, Oct. 1996).
14. Group Insurance Commission testimony before the Massachusetts Health Care Committee. SeeSession of the Joint Committee on Health Care, Massachusetts State Legislature , (March 19, 1997).
15. Cambridge Voters List Database. City of Cambridge, Massachusetts .Cambridge: February 1997.
16. Hyoungmin Park and Kyuseok Shim *Approximate Algorithms for* k- anonymity
17. V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, k - anonymity
18. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu, *Anonymizing Tables*
19. Pierangela Samarati and Latanya Sweeney *Generalizing Data to Provide Anonymity when Disclosing Information*
20. K. El Emam, F. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J. Bottomley, "A Globally Optimal k-anonimity Method for the De-identification of Health Data," Journal of the American Medical Informatics Association 16:5 (2009): 670–682.

21. Aggarwal, "On k-anonymity and the Curse of Dimensionality," Proceedings of the 31st International Conference on Very Large Data Bases (VLDB Endowment, 2005).
22. Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for ımage retrieval. Int. J. Comput. Vision, 40(2):99–121, 2000.
23. Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam (March 2007). "l-Diversity: Privacy Beyond k-Anonymity"
24. Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian (2007). "t Closeness: Privacy beyond k-anonymity and l-diversity"