



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

**ΕΡΓΑΛΕΙΟ ΑΝΩΝΥΜΟΠΟΙΗΣΗΣ ΓΙΑ
ΔΗΜΟΣΙΕΥΣΕΙΣ ΔΕΔΟΜΕΝΩΝ**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

των

**Καρρά Σωτήρη
Πανοπούλου Φανή-Βασιλική**

Επιβλέπων : Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2014

Η σελίδα αυτή είναι σκόπιμα λευκή.



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΛΕΙΟ ΑΝΩΝΥΜΟΠΟΙΗΣΗΣ ΓΙΑ ΔΗΜΟΣΙΕΥΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

των

Καρρά Σωτήρη
Πανοπούλου Φανή-Βασιλική

Επιβλέπων : Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 29^η Ιουλίου 2014.

.....
Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

.....
Κωνσταντίνος Κοντογιάννης
Καθηγητής Ε.Μ.Π.

.....
Θεόδωρος Δαλαμάγκας
Ερευνητής Β' ΙΠΣΥ/Ε.Κ. "Αθηνά"

Αθήνα, Ιούλιος 2014

.....

Καρράς Σωτήριος

Ηλεκτρολόγος Μηχανικός
και Μηχανικός Η/Υ ΕΜΠ

.....

Πανοπούλου Φανή -Βασιλική

Ηλεκτρολόγος Μηχανικός
και Μηχανικός Η/Υ ΕΜΠ

© 2014 – All rights reserved

Περίληψη

Ο σκοπός της διπλωματικής εργασίας είναι ο σχεδιασμός και ανάπτυξη ενός ολοκληρωμένου εργαλείου ανωνυμοποίησης δεδομένων. Μέσω της εφαρμογής, προσφέρεται η δυνατότητα χρήσης αρκετών από τους πλέον διαδεδομένους αλγορίθμους ανωνυμοποίησης δεδομένων, με στόχο την προστασία των προσωπικών πληροφοριών των ατόμων που βρίσκονται σε προς δημοσίευση σύνολα δεδομένων. Ταυτόχρονα, παρέχονται στον χρήστη στατιστικές πληροφορίες και διαγράμματα για το ποσοστό γενίκευσης τιμών, το χρόνο εκτέλεσης του επιλεγμένου αλγορίθμου και την απώλεια πληροφορίας που προκύπτει λόγω της ανωνυμοποίησης, δίνοντας έτσι την δυνατότητα σύγκρισης και επιλογής του καταλληλότερου αλγορίθμου για την κάθε περίπτωση. Όλες οι παραπάνω δυνατότητες γίνονται εφικτές μέσω ενός πλούσιου και εύχρηστου γραφικού περιβάλλοντος.

Πιο συγκεκριμένα, μετά από μελέτη των μεθόδων ανωνυμοποίησης επιλέχθηκαν να υλοποιηθούν αυτοί της k -ανωνυμίας και της l -ποικιλομορφίας, ως οι χαρακτηριστικότεροι και πιο γενικής χρήσης μέθοδοι προστασίας.

Η αρχιτεκτονική του εργαλείου έχει σχεδιαστεί με τέτοιο τρόπο ώστε να είναι εύκολα επεκτάσιμο, καθιστώντας ιδιαίτερα εύκολη την υλοποίηση και προσθήκη περαιτέρω αλγορίθμων ανωνυμοποίησης.

Στο τελικό στάδιο της εργασίας, η ορθή λειτουργία της εφαρμογής δοκιμάστηκε με ρεαλιστικά σύνολα δεδομένων με ιδιαίτερα θετικά αποτελέσματα.

Λέξεις Κλειδιά:

προστασία ιδιωτικότητας, ανωνυμοποίηση δεδομένων, βάσεις δεδομένων, επιθέσεις ιδιωτικότητας, γενίκευση, k -ανωνυμία, l -ποικιλομορφία

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

The purpose of this thesis was the design and development of a complete data anonymization tool. The application offers the usage of a variety of anonymization algorithms, aiming to protect the personal data of individuals present in collections of data publications. At the same time, the user is presented with statistical information and diagrams concerning the generalization percentage, execution time of the selected algorithm and information loss, thus allowing the comparison and, ultimately, the selection of the most suited algorithm for each case. All of the aforementioned features can be accessed through a rich and user-friendly graphical interface.

Specifically, as the most typical and generic methods of protection k-anonymity and l-diversity were considered the best choice for implementation, after an extensive study of existing anonymization methods and algorithms. The tool's architecture is particularly designed, in order to be easily extensible, making it easy to implement and include additional anonymization algorithms.

In the final stages of the application development, its proper functionality was tested using real-life datasets, resulting in genuinely good output datasets.

Keywords:

privacy protection, data anonymization, databases, privacy attacks, generalization, k-anonymity, l-diversity

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

Θα θέλαμε να ευχαριστήσουμε τον καθηγητή μας κ. Ιωάννη Βασιλείου για την ευκαιρία που μας έδωσε να ασχοληθούμε με ένα τόσο ενδιαφέρον αντικείμενο.

Επίσης θα θέλαμε να δώσουμε τις εγκάρδιες ευχαριστίες μας στην επιβλέπουσα κυρία Όλγα Γκουντούνα για την εξαιρετική συνεργασία που είχαμε μαζί της, για την συνεχή της υποστήριξη και καθοδήγηση που μας προσέφερε καθ' όλη την διάρκεια της εκπόνησης της εργασίας αυτής.

Τέλος θα θέλαμε να ευχαριστήσουμε τους γονείς μας για τη στήριξη τους, καθώς και τους φίλους μας που με τις συζητήσεις που κάναμε μαζί τους στα διάφορα στάδια αυτής της εργασίας, άφησαν και αυτοί το στίγμα τους πάνω της.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

1	Εισαγωγή.....	1
1.1	Υλοποίηση εργαλείου ανωνυμοποίησης δεδομένων	2
1.1.1	Συνεισφορά	2
1.2	Οργάνωση κειμένου.....	4
2	Θεωρητικό υπόβαθρο	5
3	Σχετικές εργασίες.....	9
3.1.1	Επιθέσεις σύνδεσης (<i>linking attacks</i>) και η τεχνική της <i>k</i> -ανωνυμίας.....	9
3.1.2	Επιθέσεις ομογενών δεδομένων και οι τεχνικές της <i>l</i> -ποικιλομορφίας και ανατομίας.....	16
3.1.3	Επιθέσεις παρουσίας. Η τεχνική της <i>δ</i> -παρουσίας	24
3.1.4	Επιθέσεις ανομοιομορφων δεδομένων/ομοιότητας. Η τεχνική της <i>t</i> -εγγύτητας. ..	26
3.1.5	Επιθέσεις γνώσης σε πίνακες με συχνές ενημερώσεις. Η τεχνική της <i>m</i> -αμεταβλητότητας	32
3.1.6	Επίθεση γνώσης ενάντια πληροφορία οργανωμένη σε σύνολα (<i>background attack against set-valued data</i>)	39
3.2	Εργαλείο ανωνυμοποίησης του πανεπιστημίου του Texas.....	41
4	Ανάλυση Συστήματος.....	42
4.1	Αρχιτεκτονική.....	42
4.2	Περιγραφή Λειτουργιών	44
4.2.1	Εισαγωγή δεδομένων/παραμέτρων και διεπαφή με τον χρήστη	44
4.2.2	Ανωνυμοποίηση.....	49
4.2.3	Στατιστικά και Διαγράμματα	52
4.2.4	Διαχείριση ιεραρχιών γενίκευσης	60
5	Υλοποίηση	62
5.1	Πλατφόρμες και προγραμματιστικά εργαλεία	62
5.2	Λεπτομέρειες υλοποίησης.....	65
5.2.1	Κλάση FileData.....	65

5.2.2	Κλάση statistics και κλάσεις που την κληρονομούν.....	67
5.2.2.1	Κλάση AnatomyStatistics	68
5.2.2.2	Κλάση kStatistics	68
5.2.3	Κλάση AnatomyData	68
5.2.3	Κλάση kData.....	72
5.2.4	Κλάσεις για στατιστικά δειγμάτων	76
5.2.5	Κλάσεις παραθύρων.....	78
6	Έλεγχος.....	79
6.1	Μεθοδολογία ελέγχου.....	79
6.1.1	<i>Ανατομία</i>	80
6.1.2	<i>k-ανωνυμία</i>	82
7	Επίλογος	85
7.1	Σύνοψη και συμπεράσματα.....	85
7.2	Μελλοντικές επεκτάσεις	86
8	Βιβλιογραφία.....	87

1

Εισαγωγή

Ιδιωτικότητα και ανωνυμία δημοσιευμένων δεδομένων

Τα τελευταία χρόνια, έχει σημειωθεί εκθετική αύξηση του όγκου των πληροφοριών που είναι διαθέσιμες στο ευρύ κοινό, καθώς οι ταχύτητες και η συνδεσιμότητα στο Διαδίκτυο αλλά και ο αποθηκευτικός χώρος γίνονται ολοένα και πιο διαθέσιμα. Το γεγονός αυτό, σε συνδυασμό με την προσωπική φύση του μεγαλύτερου όγκου αυτής της πληροφορίας, έχουν οδηγήσει στην ανάπτυξη τόσο νομοθεσίας, όσο και τεχνικών “ανωνυμοποίησης” των δεδομένων προς δημοσίευση με σκοπό την προστασία της ταυτότητας του ατόμου, αλλά και πιθανών “ευαίσθητων” γνωρισμάτων (sensitive attributes ή sensitive data), όπως για παράδειγμα η ασθένεια σε μία δημοσιευμένη βάση ενός νοσοκομείου. Πειράματα όμως που πραγματοποιήθηκαν στις ΗΠΑ [1] απόδειξαν πως η τεχνική της απαλοιφής τέτοιων γνωρισμάτων δεν ήταν αρκετή για την προστασία των προσωπικών δεδομένων. Το χαρακτηριστικότερο τέτοιο παράδειγμα είναι η σύνδεση ενός κυβερνήτη της πολιτείας Μασαχουσέτης με τα ιατρικά του στοιχεία, χρησιμοποιώντας την λίστα υποψηφίων σε συνδυασμό με τις φαινομενικά ανωνυμοποιημένες πληροφορίες του οργανισμού GIC ο οποίος είναι υπεύθυνος για την ιατρική ασφάλιση των εργαζομένων της πολιτείας, δύο δημοσιευμένα και φαινομενικά ανεξάρτητα σύνολα δεδομένων. Ως αποτέλεσμα στις αρχές της δεκαετίας του 2000, άρχισαν να αναπτύσσονται πολυπλοκότερες μέθοδοι προστασίας της ιδιωτικότητας με πρώτη αυτή της k-ανωνυμίας [1]. Τα επόμενα χρόνια χαρακτηρίστηκαν τόσο από επεκτάσεις της k-ανωνυμίας όσο και από την δημιουργία νέων τεχνικών για την διασφάλιση της ιδιωτικότητας των προσώπων σε σύνολα δεδομένων, έχοντας ως συνέπεια την δημιουργία ενός νέου κλάδου της

επιστήμης των υπολογιστών και ειδικότερα των βάσεων δεδομένων ο οποίος είναι πλέον γνωστός ως κλάδος “ιδιωτικότητας δεδομένων – ανωνυμοποίησης δεδομένων” (data privacy – data anonymization).

1.1 Υλοποίηση εργαλείου ανωνυμοποίησης δεδομένων

Στόχος της παρούσας εργασίας είναι η δημιουργία μίας εφαρμογής, η οποία θα επιτρέπει την χρήση πολλαπλών αλγορίθμων ανωνυμοποίησης δεδομένων, προσφέροντας ταυτόχρονα ένα πλούσιο και εύχρηστο γραφικό περιβάλλον. Με την επίτευξη της ανωνυμοποίησης, το εργαλείο προβάλλει στον χρήστη στατιστικά που αφορούν τόσο τους αλγορίθμους όσο και το ασφαλές σύνολο πληροφοριών που έχει προκύψει από την εφαρμογή τους (διαγράμματα απώλειας πληροφορίας, χρόνος εκτέλεσης αλγορίθμου κτλ.). Επιπλέον, δίνεται η δυνατότητα εφαρμογής των διαφόρων αλγορίθμων σε δείγμα του συνολικού όγκου δεδομένων για την ευκολότερη επιλογή του καταλληλότερου για κάθε περίπτωση αλγορίθμου, απαλλάσσοντας έτσι τον χρήστη από το να εφαρμόζει, στον εν δυνάμει μεγάλο όγκο πληροφορίας, χρονοβόρες δοκιμές πάνω στα δεδομένα του. Τέλος, μετά την διασφάλιση της ιδιωτικότητας, ο χρήστης μπορεί να εξάγει τα τροποποιημένα δεδομένα σε μορφή για εύκολη δημοσίευση είτε άμεσα είτε σε μία βάση δεδομένων.

Με αυτόν τον τρόπο, προσφέρεται μία εφαρμογή, η οποία όχι μόνο προσφέρει μία βιβλιοθήκη αλγορίθμων ανωνυμοποίησης αλλά και το περιβάλλον χρήσης τους.

1.1.1 Συνεισφορά

Πιο αναλυτικά, η συνεισφορά της παρούσας εργασίας μπορεί να συνοψιστεί σε 7 σημεία ως εξής:

1. Εκτενής μελέτη της υπάρχουσας βιβλιογραφίας, με στόχο την πλήρη κατανόηση των αναγκών για ανωνυμοποίηση και των πρακτικών προβλημάτων που θα πρέπει να αντιμετωπίζει μία εφαρμογή προστασίας ευαίσθητων δεδομένων.

2. Συγγραφή βιβλιογραφικής εργασίας για την συνοπτική παρουσίαση και ευκολότερη κατανόηση των βασικών ιδεών της ανωνυμοποίησης δεδομένων. Κατά την διάρκεια της μελέτης και συγγραφής της βιβλιογραφικής εργασίας αποκρυσταλλώθηκαν οι απαιτήσεις του συστήματος και προσδιορίστηκαν τα τεχνικά χαρακτηριστικά της εφαρμογής. Μία συνοπτική παρουσίαση του θεωρητικού υπόβαθρου δίνεται στην ενότητα 3.
3. Υλοποιήθηκαν δύο από τους κυριότερους αλγορίθμους ανωνυμοποίησης δεδομένων, αυτοί της Mondrian πολυδιάστατης k-ανωνυμίας [16] και του anatomy [3]. Μαζί με αυτούς υλοποιήθηκαν και επεκτάσιμες κλάσεις που μπορούν να χρησιμοποιηθούν για την μελλοντική υλοποίηση επιπλέον αλγορίθμων, αφού προσφέρουν τις βασικές τεχνικές (k-ανωνυμία και l-ποικιλομορφία) πάνω στις οποίες στηρίζονται και οι μεταγενέστερες τεχνικές (t-εγγύτητα, m-αμεταβλητότητα).
4. Οι αλγόριθμοι ενσωματώθηκαν σε πρότυπο σύστημα με εύχρηστο γραφικό περιβάλλον για την εύκολη διαχείριση των ευαίσθητων δεδομένων.
5. Υλοποιήθηκαν κλάσεις για την παραγωγή στατιστικών και διαγραμμάτων, με στόχο τόσο την αξιολόγηση του κάθε αλγορίθμου ξεχωριστά, για την εκτίμηση και υπολογισμό των παραμέτρων ανωνυμοποίησης, όσο και την συγκριτική αποτίμηση των αποτελεσμάτων που προσφέρουν διαφορετικοί αλγόριθμοι πάνω στο ίδιο σύνολο δεδομένων.
6. Ταυτόχρονα, έχει προστεθεί η δυνατότητα εξαγωγής τόσο των ανωνυμοποιημένων στοιχείων όσο και των στατιστικών που τα αφορούν σε .csv και .txt αρχεία αντίστοιχα.
7. Τέλος, έχει προστεθεί και η επιλογή για την εισαγωγή και εμφάνιση σε δεντρική μορφή ορισμένων από τον χρήστη ιεραρχιών γενίκευσης από .txt αρχεία, με σκοπό την εν τέλει πλήρη παραμετροποίηση των τεχνικών ανωνυμοποίησης.

1.2 Οργάνωση κειμένου

Σε αυτό το σημείο είναι απαραίτητο να δοθεί μία συνοπτική παρουσίαση των κεφαλαίων που ακολουθούν. Το Κεφάλαιο 2 δίνει το βασικό θεωρητικό υπόβαθρο για την κατανόηση του γνωστικού πεδίου της ανωνυμοποίησης δεδομένων. Το Κεφάλαιο 3 παρουσιάζει αναλυτικά τις εργασίες που έχουν ήδη δημοσιευτεί και αφορούν τεχνικές ιδιαιτερότητας καθώς και συνοπτική παρουσίαση προσπαθειών υλοποίησής τους. Το κεφάλαιο 4 ξεκινά μία συζήτηση σχετικά με θέματα σχεδιασμού της εφαρμογής, χωρίζοντας το εργαλείο σε υποσυστήματα και μελετώντας τις δυνατότητες που προσφέρει το καθένα. Γίνεται εκτενής ανάλυση της λειτουργικότητας του κάθε υποσυστήματος και παρέχονται στιγμιότυπα από τις οθόνες της εφαρμογής για την καλύτερη κατανόηση των λειτουργιών της. Στο Κεφάλαιο 5 παρουσιάζονται θέματα υλοποίησης της εφαρμογής και οργάνωσης του κώδικα. Γίνεται εισαγωγή στο περιβάλλον ανάπτυξης και στα εργαλεία που χρησιμοποιήθηκαν καθώς και εκτενής παρουσίαση των κλάσεων και των μεθόδων που βρίσκονται στον πυρήνα της λειτουργικότητας της εφαρμογής. Στο Κεφάλαιο 6 παρουσιάζονται τα αποτελέσματα, της ολοκληρωμένης πλέον εφαρμογής, της ανωνυμοποίησης πάνω σε σύνολα δεδομένων. Τελικά, γίνεται αποτίμηση της ορθής λειτουργίας και της επίτευξης των στόχων που έχουν τεθεί για το εργαλείο ανωνυμοποίησης. Στο Κεφάλαιο 7 δίνεται ο επίλογος, με μία περίληψη και ανασκόπηση των θεμάτων που μελετήθηκαν στην εργασία. Τέλος, στο Κεφάλαιο 8 παρουσιάζεται αναλυτικά η βιβλιογραφία που μελετήθηκε για την συγγραφή και ανάπτυξη του έργου.

2

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.

- **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 $f_g: T \rightarrow U'$ 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:
$$\exists p_i \in U : f_g(f_c(p_i)[Q_t]) = p_i$$

- **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.

3

Σχετικές εργασίες

Σε αυτό το σημείο, θα ήταν καλό να αναλύσουμε εκτενέστερα τις εργασίες τόσο του τομέα της ιδιωτικότητας των δεδομένων όσο και των αλγορίθμων που προέκυψαν από την προσπάθεια επίλυσης των προβλημάτων που έθεσε αυτός. Για αυτόν τον σκοπό, θα ξεκινήσουμε με την ανάλυση και παρουσίαση των επιθέσεων που κάθε μία από αυτές προσπάθησε να αντιμετωπίσει. Θα επεκταθούμε τόσο στους αλγορίθμους που έχουν υλοποιηθεί στα πλαίσια της εργασίας, όσο και σε επεκτάσεις αυτών. Στο τέλος του κεφαλαίου, υπάρχει αναφορά και σε μία παρόμοια προσπάθεια υλοποίησης αλγορίθμων ανωνυμοποίησης για κοινή χρήση.

3.1 Επιθέσεις κατά της ιδιωτικότητας

3.1.1 Επιθέσεις σύνδεσης (*linking attacks*) και η τεχνική της *k*-ανωνυμίας.

Όπως αναφέρθηκε και στο κεφάλαιο 1, η πρώτη τεχνική προστασίας που προτάθηκε για την ανωνυμοποίηση δεδομένων ήταν αυτή της *k*-ανωνυμίας, η οποία σχεδιάστηκε για την προστασία από μία κατηγορία επιθέσεων που χαρακτηρίζεται ως “επιθέσεις σύνδεσης”. Ας την δούμε λίγο πιο αναλυτικά δίνοντας ένα παράδειγμα. Στον Πίνακα 1, δείχνουμε ένα μέρος των ιατρικών αρχείων ενός νοσοκομείου που δημοσιεύτηκαν για ερευνητικούς σκοπούς. Παρατηρούμε, πως στον πίνακα έχει εφαρμοστεί η τεχνική της απαλοιφής και έχουν διαγραφεί γνωρίσματα, όπως το όνομα και το τηλέφωνο, για να μην είναι δυνατή η αντιστοίχιση της ταυτότητας κάποιου ατόμου με τα δημοσιευμένα γνωρίσματα.

Φαινομενικά, λοιπόν, έχουμε ανώνυμα δεδομένα μιας και γνωρίζοντας μόνο τον πίνακα των ιατρικών αρχείων και χωρίς να έχουμε πρότερη γνώση για κανένα από τα άτομα των οποίων τα γνωρίσματα δημοσιεύτηκαν (δεν γνωρίζουμε για παράδειγμα πως ο Bob είχε πνευμονία), δεν μπορούμε να εξάγουμε καμία περαιτέρω πληροφορία για την ταυτότητα των νοσηλευθέντων.

Ας θεωρήσουμε την συνέχεια, πως μία εταιρία πώλησης αυτοκινήτων δημοσιεύει την ίδια περίοδο ένα πίνακα με τις πωλήσεις υβριδικών αυτοκινήτων την περίοδο 2010-2011. Ένα μέρος του πίνακα αυτού δίνεται στον *Πίνακα 2*. Με μία πρώτη ματιά στον *Πίνακα 2* κάνουμε δύο πολύ βασικές παρατηρήσεις:

1. Συμπεριλαμβάνεται το γνώρισμα *Όνομα*, οπότε έχουμε επώνυμη παρουσία των ατόμων στον πίνακα.
2. Οι δύο δημοσιευμένοι πίνακες έχουν τρία κοινά γνωρίσματα τα {Φύλο, Ταχυδρομικό Κώδικα, Ημερομηνία Γέννησης}.

Με μία πιο προσεκτική μελέτη των δύο δημοσιευμένων πινάκων, μπορούμε εύκολα να δούμε πως χρησιμοποιώντας τις τιμές των παραπάνω κοινών πεδίων, μπορούμε να αντιστοιχήσουμε τον Bob (και μάλιστα μονοσήμαντα) με την τελευταία εγγραφή του πίνακα των ιατρικών στοιχείων, παρόλο που οι τιμές των τριών κοινών γνωρισμάτων εμφανίζονται παραπάνω από μία φορές.

Διάγνωση	Ταχυδρομικός Κώδικας	Φύλο	Ημερομηνία Γέννησης	Ημερομηνία Εξιτηρίου
Γρίπη	11234	Άντρας	12/08/1970	03/04/2010
Πνευμονία	11238	Γυναίκα	03/06/1990	16/07/2010
Γρίπη	19732	Γυναίκα	30/04/1978	30/11/2010
Κάταγμα ποδιού	19737	Άντρας	17/09/1974	17/12/2010
Υπέρταση	19734	Γυναίκα	16/03/1980	09/01/2011
Γαστρεντερίτιδα	11232	Γυναίκα	25/12/1989	05/05/2011
Πνευμονία	11222	Άντρας	30/04/1980	13/05/2011

Πίνακας 1

Όνομα	Ημερομηνία Αγοράς	Φύλο	Ημερομηνία Γέννησης	Ταχυδρομικός Κώδικας
Bob	12/05/2010	Άντρας	30/04/1980	11222
Alice	15/09/2010	Άντρας	14/10/1960	16598
Katherine	23/12/2011	Γυναίκα	19/04/1970	12399
Andy	17/01/2011	Γυναίκα	04/04/1968	12994

Πίνακας 2

Διάγνωση	Ταχυδρομικός Κώδικας	Φύλο	Ημερομηνία Γέννησης	Ημερομηνία Εξιτηρίου
Γρίπη	112**	Άντρας	1970-1980	03/04/2010
Πνευμονία	112**	Άντρας	1970-1980	13/05/2011
Γαστρεντερίτιδα	1123*	Γυναίκα	1980-1990	05/05/2011
Πνευμονία	1123*	Γυναίκα	1980-1990	16/07/2010
Κάταγμα ποδιού	1973*	Γυναίκα	1970-1980	17/12/2010
Υπέρταση	1973*	Γυναίκα	1970-1980	09/01/2011
Γρίπη	1973*	Γυναίκα	1970-1980	30/11/2010

Πίνακας 3

Αποδεικνύεται λοιπόν πως η απαλοιφή ως μέθοδος ανωνυμοποίησης είναι ανεπαρκής. Για να αντιμετωπιστούν τέτοιες επιθέσεις σύνδεσης, προτάθηκε η έννοια της k-ανωνυμίας.

Ορισμός (k-ανωνυμία)

Έστω πίνακας $RT(A_1, \dots, A_n)$ και QI_{RT} τα ψευδο-αναγνωριστικά που σχετίζονται με αυτόν. Θα λέμε ότι ο RT θα ικανοποιεί την k-ανωνυμία αν κάθε ακολουθία τιμών στον $RT[QI_{RT}]$ εμφανίζεται τουλάχιστον k φορές ή ισοδύναμα κάθε σύνολο τιμών.

Ας επεκτείνουμε το παραπάνω παράδειγμα για να κατανοήσουμε τον ορισμό που μόλις δώσαμε. Στον Πίνακα 3 έχουμε μία 2-ανώνυμη μορφή του Πίνακα 1 των ιατρικών αρχείων αν θεωρήσουμε ως ψευδο-αναγνωριστικά το σύνολο {Ταχυδρομικός Κώδικας, Φύλο, Ημερομηνία Γέννησης}. Όπως φαίνεται, τα σύνολα τιμών {112**, Άντρας, 1970-1980} και {1123*, Γυναίκα, 1980-1990} εμφανίζονται δύο φορές το καθένα, ενώ οι τιμές {1973*, Γυναίκα, 1970-1980} εμφανίζονται τρεις φορές στον 2-ανώνυμο πίνακα.

Ας δούμε τώρα αν γνωρίζοντας των Πίνακα 2 των πωλήσεων και τον 2-ανώνυμο Πίνακα 3 των ιατρικών στοιχείων μπορούμε να αποκαλύψουμε την ταυτότητα του Bob. Ο επιτιθέμενος, γνωρίζει πως τα ψευδο-αναγνωριστικά του Bob έχουν τιμές {11222, Άντρας, 30/04/1980} από τον Πίνακα 2. Πηγαίνοντας όμως στον Πίνακα 3, θα συνειδητοποιήσει πως υπάρχουν δύο εγγραφές που ταιριάζουν με αυτές τις τιμές οι {Γρίπη, 112**, Άντρας, 1970-1980, 03/04/2010} και {Πνευμονία, 112**, Άντρας, 1970-1980, 13/05/2011} οπότε και δεν μπορεί να προσδιορίσει ποια από αυτές αντιστοιχεί στον Bob.

Όμως για να επιτευχθεί η 2-ανωνυμία, χρειάστηκε να κάνουμε γενίκευση των τιμών των ψευδο-αναγνωριστικών (μετατρέποντας για παράδειγμα τον ταχυδρομικό κώδικα από 11222 σε 112**) γεγονός το οποίο μας οδηγεί σε απώλεια πληροφορίας. Αυτός ο συμβιβασμός μεταξύ της ισχύος της ανωνυμίας και της πληροφορίας που θα χαθεί, είναι ένα θέμα που καλούνται να αντιμετωπίσουν όλες οι τεχνικές ανωνυμοποίησης,

προσπαθώντας να επιτύχουν όσο το δυνατόν ισχυρότερη προστασία, χάνοντας όσο το δυνατόν λιγότερη πληροφορία.

Στο σημείο αυτό θα πρέπει να σημειωθεί ότι η γενίκευση για την επιτυχία της k-ανωνυμίας μπορεί να γίνει κυρίως με χρήση της ιεραρχίας γενίκευσης όπως προαναφέρθηκε. Θεωρούμε μια συνάρτηση γενίκευσης φ η οποία δέχεται ως όρισμα μία τιμή και με βάση την ιεραρχία γενίκευσης δίνει μια γενίκευση της τιμής αυτής. Οι αλγόριθμοι που υλοποιούν την γενίκευση μπορούν να χωρισθούν σε μονοδιάστατους (Single Dimensional Algorithm – SDA) και βασισμένους στο κελί (Cell Based Algorithm – CBA) [5]. Σημειώνεται ως $A[c][r]$ η τιμή που βρίσκεται στον πίνακα A στην σειρά r και στην στήλη c.

Ένας μονοδιάστατος αλγόριθμος, δεδομένου ενός πίνακα PT δημιουργεί έναν k-ανώνυμο πίνακα RT εφαρμόζοντας την συνάρτηση φ σε κάθε γνώρισμα του PT έτσι ώστε $RT[c][r] = \varphi(PT[c][r])$. Αυτό, σημαίνει πρακτικά ότι αν για να γίνει k-ανώνυμος ένας πίνακας πρέπει σε μία ομάδα εγγραφών η τιμή a να γενικευθεί στην τιμή b τότε αυτό θα γίνει για όλες τις εγγραφές του πίνακα, ανεξάρτητα από το αν αυτό είναι απαραίτητο ή όχι.

Ένας αλγόριθμος βασισμένος στο κελί δεδομένου ενός πίνακα PT δημιουργεί έναν k-ανώνυμο πίνακα RT έτσι ώστε για κάθε c και r του PT η τιμή $RT[c][r]$ να είναι μία γενίκευση του $PT[c][r]$ ως προς το γνώρισμα που προσδιορίζεται από τον c. Στην περίπτωση αυτή μπορεί να γενικευθεί σε διαφορετικό βαθμό η ίδια τιμή ενός γνωρίσματος ανάλογα με την γραμμή στην οποία βρίσκεται.

Για παράδειγμα δίδεται ο παρακάτω πίνακας με τις γενικεύσεις που γίνονται από τους δύο αυτούς αλγορίθμους:

Ηλικία	Φύλο	Χώρα	Ασθένεια
25	Άνδρας	Βραζιλία	Γρίπη
27	Άνδρας	Η.Π.Α	Πνευμονία
42	Γυναίκα	Καναδάς	Γαστρεντερίτιδα
47	Άνδρας	Η.Π.Α	Υπέρταση

Αρχικό σύνολο δεδομένων

Ηλικία	Φύλο	Χώρα	Ασθένεια
2*	*	Αμερική	Γρίπη
2*	*	Αμερική	Πνευμονία
4*	*	Αμερική	Γαστρεντερίτιδα
4*	*	Αμερική	Υπέρταση

Πίνακας που προέκυψε από SDA

Ηλικία	Φύλο	Χώρα	Ασθένεια
25	Άνδρας	Αμερική	Γρίπη
27	Άνδρας	Αμερική	Πνευμονία
42	*	Β.Αμερική	Γαστρεντερίτιδα
47	*	Β.Αμερική	Υπέρταση

Πίνακας που προέκυψε από CBA

Στο παράδειγμα αυτό υπάρχουν δύο κλάσεις ισοδυναμίας που περιέχουν τις δύο πρώτες και δύο τελευταίες εγγραφές αντίστοιχα. Στην περίπτωση του SDA επειδή για την δεύτερη ομάδα ισοδυναμίας η τιμή “Άνδρας” έπρεπε να γενικευθεί, η γενίκευση αυτή εφαρμόστηκε και στην πρώτη ομάδα όπου η k ανωνυμία θα διατηρούνταν ακόμα κι αν κρατούσαμε την πρωτότυπη τιμή. Αντίστοιχα, επειδή η Βραζιλία και οι Η.Π.Α ανήκουν σε διαφορετικό κομμάτι της Αμερικής δεν μπορούσαν παρά να γενικευθούν στην τιμή “Αμερική”. Με τον SDA αυτό επηρέασε και την δεύτερη κλάση ισοδυναμίας ενώ στην περίπτωση του CBA οι τιμές της δεύτερης κλάσης γενικεύθηκαν σε μικρότερο βαθμό απεικονιζόμενες στην τιμή “Β. Αμερική”. Όπως μπορεί κανείς να παρατηρήσει η γενίκευση που εφαρμόζεται στον CBA είναι η απαραίτητη και εξαρτάται από τις τιμές που απαντώνται σε κάθε γνώρισμα στην ομάδα ισοδυναμίας που θέλουμε να δημιουργηθεί.

Για τον ίδιο σκοπό χρησιμοποιούνται και οι CDGH αλγόριθμοι που πρακτικά είναι CBA αλγόριθμοι οι οποίοι αρχικά χρησιμοποιούν μια κατάτμηση των δεδομένων και ανωνυμοποιούν ξεχωριστά κάθε τμήμα που προκύπτει. Ακόμη μια κατηγορία είναι οι αλγόριθμοι που δεν βασίζονται στην καθορισμένη από τον χρήστη ιεραρχία DGH αλλά σε ιεραρχίες γενίκευσης φυσικού πεδίου (Natural Domain Generalization Hierarchy - NDGH). Η ιεραρχία NDGH ορίζεται για ένα σύνολο τιμών U ως ένας κατευθυνόμενος γράφος $G(V,E)$ όπου $(V=U^*)$ και για κάθε $u,v \in V$ με $(u,v) \in E$ αν $u \subseteq v$. Οι αλγόριθμοι στους οποίους αν δοθεί ένα σύνολο δεδομένων και ένας ακέραιος k σαν είσοδος, δημιουργούν ένα k -ανώνυμο σύνολο δεδομένων το οποίο βρίσκεται μέσα στην NDGH του αρχικού συνόλου ονομάζονται αλγόριθμοι NDGH. Οι αλγόριθμοι CNDGH είναι αλγόριθμοι NDGH που χρησιμοποιούν κατατμήσεις των δεδομένων σε ομάδες [5].

Κλείνοντας την ανάλυση της k-ανωνυμίας, αξίζει να αναφερθούν δύο σημαντικές υπο-περιπτώσεις της επίθεσης σύνδεσης, η *επίθεση ταξινόμησης* και η *επίθεση συμπληρωματικής επίθεσης*.

Η επίθεση ταξινόμησης, βασίζεται στην σειρά με την οποία εμφανίζονται οι εγγραφές στον δημοσιευμένο πίνακα. Έστω ότι έχουμε δύο πίνακες GT_1 και GT_2 οι οποίοι βασίζονται στον μη δημοσιευμένο πίνακα PT και είναι και δύο k-ανώνυμοι. Αν οι δύο πίνακες περιέχουν εν μέρη διαφορετικές πληροφορίες και γνωρίζουμε πως η θέση με την οποία εμφανίζονται οι εγγραφές αντιστοιχούν στην θέση των εγγραφών στον PT , τότε έχουμε διαρροή πληροφορίας. Ο τρόπος αντιμετώπισης είναι προφανώς, εκτός από την εφαρμογή της k-ανωνυμίας, η τυχαία τοποθέτηση των εγγραφών του προς δημοσίευση πίνακα προτού αυτός δημοσιευτεί. Παρατηρούμε, πως η επίθεση ταξινόμησης είναι επίθεση αποκάλυψης γνώρισματος και μπορεί να αντιμετωπιστεί από την k-ανωνυμία παρόλο που η τελευταία σχεδιάστηκε για την αντιμετώπιση επιθέσεων αποκάλυψης ταυτότητας.

Τέλος, ας υποθέσουμε ότι δημοσιεύουμε πληροφορία από ένα πίνακα PT , σε k-ανώνυμη μορφή. Αν στην συνέχεια δημοσιεύσουμε ένα άλλο μέρος του πίνακα PT περιλαμβάνοντας πιθανώς και διαφορετικά γνώρισματα (ή ακόμη και τα ίδια γνώρισματα με διαφορετική μέθοδο γενίκευσης), τότε έχουμε κίνδυνο διαρροής πληροφορίας ακόμα κι αν ο δεύτερος πίνακας είναι και αυτός σε k-ανώνυμη μορφή. Η λύση, στο πρόβλημα της συμπληρωματικής δημοσίευσης είναι να θεωρούμε κάθε επιπλέον γνώρισμα ή πληροφορία που δημοσιεύεται σαν ψευδο-αναγνωριστικό και να επαναπροσδιορίζουμε την k-ανωνυμία χρησιμοποιώντας το νέο σύνολο QI .

3.1.2 Επιθέσεις ομογενών δεδομένων και οι τεχνικές της l-ποικιλομορφίας και ανατομίας.

Οι επιθέσεις ομογενών δεδομένων, εντοπίζονται στην περίπτωση που μια κλάση ισοδυναμίας έχει τα ίδια ή παρόμοια ευαίσθητα δεδομένα. Για παράδειγμα, ας θεωρήσουμε την περίπτωση ενός επιτιθέμενου που θέλει να εντοπίσει την ασθένεια κάποιου, μέσα από έναν πίνακα k -ανωνυμοποιημένων δεδομένων, ο οποίος αντιστοιχεί στους ασθενείς ενός νοσοκομείου. Έστω ότι ο επιτιθέμενος γνωρίζει αρκετά στοιχεία ώστε να εντοπίσει k εγγραφές που μπορεί να αντιστοιχούν στο άτομο που αναζητά.

Αν και οι k εγγραφές έχουν την ίδια τιμή για τα ευαίσθητα δεδομένα, τότε ο επιτιθέμενος μπορεί να συμπεράνει με βεβαιότητα ότι το άτομο που τον ενδιαφέρει έχει την τιμή αυτή. Είναι φανερό λοιπόν, ότι τέτοιου τύπου επιθέσεις οδηγούν σε αποκάλυψη γνωρισμάτων. Η επίθεση ομογενών δεδομένων όπως περιγράφηκε παραπάνω αντιστοιχεί σε θετική αποκάλυψη δεδομένων.

Στην ίδια κατηγορία επιθέσεων ανήκει και η περίπτωση που ενώ δεν έχουν όλες οι εγγραφές της ομάδας ισοδυναμίας την ίδια τιμή για το ευαίσθητο γνώρισμα, κάποια τιμή εμφανίζεται σε μεγάλο ποσοστό αυτών ή το σύνολο των τιμών που παρατηρούνται είναι πολύ μικρό. Στην περίπτωση αυτή αν και δεν μπορεί να συμπεραθεί με βεβαιότητα η αναζητούμενη από τον επιτιθέμενο τιμή, δύναται να αλλάξει η αρχική πεποίθηση του επιτιθέμενου για την περίπτωση η τιμή του ζητούμενου γνωρίσματος να είναι s .

Ας θεωρήσουμε για παράδειγμα τον παρακάτω 4-ανώνυμο πίνακα και έναν επιτιθέμενο που αναζητά για τον Bob την ασθένεια που είχε γνωρίζοντας την ηλικία, τον ταχυδρομικό κώδικα και το γεγονός ότι υπάρχει εγγραφή για αυτόν στον δημοσιευμένο πίνακα.

Ταχυδρομικός Κώδικας	Ηλικία	Ασθένεια
145**	30-39	Βρογχίτιδα
145**	30-39	Βρογχίτιδα
145**	30-39	Βρογχίτιδα
145**	30-39	Βρογχίτιδα
112**	40-45	Γαστρίτιδα
112**	40-45	Γαστρίτιδα
112**	40-45	Γρίπη
112**	40-45	Γρίπη
114**	47-49	Καρκίνος
114**	47-49	Γαστρίτιδα
114**	47-49	Πνευμονία
114**	47-49	Βρογχίτιδα

Αν ο Bob ανήκει στην πρώτη κλάση ισοδυναμίας πρόκειται για την πρώτη περίπτωση που αναφέρθηκε και επειδή και τα τέσσερα άτομα που ανήκουν στην κλάση αυτή πάσχουν από βρογχίτιδα ο επιτιθέμενος μπορεί με απόλυτη βεβαιότητα να συμπεράνει ότι αυτή είναι η ασθένεια του Bob. Η δεύτερη περίπτωση είναι ο Bob να ανήκει στην δεύτερη κλάση ισοδυναμίας. Αυτή την φορά ο επιτιθέμενος δεν μπορεί να συμπεράνει με βεβαιότητα την ασθένεια του Bob μπορεί όμως να συμπεράνει ότι κατά 50% έχει γαστρίτιδα και κατά 50% έχει γρίπη. Έτσι, μετά την παρατήρηση αυτού του πίνακα έχει καταλήξει σε δύο πιθανές τιμές. Αντιθέτως, αν ο Bob ανήκε στην τελευταία κλάση ισοδυναμίας, λόγω των τεσσάρων διαφορετικών τιμών δεν κερδίζει τόσο σημαντικές πληροφορίες αφού για καμία από τις ασθένειες που παρουσιάζονται στην κλάση αυτή δεν μπορεί να θεωρήσει ότι αντιστοιχεί στην ασθένεια του Bob με ποσοστό πάνω από 25%.

Για την αντιμετώπιση επιθέσεων τέτοιου τύπου, η πρώτη μέθοδος που προτάθηκε ήταν αυτή της 1-ποικιλομορφίας [2]. Η κεντρική ιδέα της μεθόδου αυτής είναι η απαίτηση κάθε κλάση ισοδυναμίας να περιέχει τουλάχιστον 1 διαφορετικές τιμές ευαίσθητων δεδομένων.

Αρχή της l-ποικιλομορφίας (l-diversity principle):

Μια κλάση ισοδυναμίας είναι l-ποικιλόμορφη αν περιέχει τουλάχιστον l “αντιπροσωπευτικές” τιμές για το ευαίσθητο γνώρισμα S. Ένας πίνακας είναι l-ποικιλόμορφος αν κάθε κλάση ισοδυναμίας είναι l-ποικιλόμορφη.

Ορισμός (l-ποικιλομορφία)

Ορίζεται έλλειψη ποικιλομορφίας η περίπτωση που σχεδόν όλες οι πλειάδες έχουν την ίδια τιμή s για το ευαίσθητο γνώρισμα S και έτσι ο επιτιθέμενος είναι σχεδόν σίγουρος ότι αυτή είναι η σωστή τιμή.

Ο όρος αντιπροσωπευτικές μπορεί να μεταφρασθεί με διαφορετικούς τρόπους, οι οποίοι αντιστοιχούν στις τρεις βασικές εκδοχές της l-ποικιλομορφίας. Οι εκδοχές αυτές είναι οι εξής:

l-ποικιλομορφία διαφορετικότητας (distinct l-diversity)

Στην περίπτωση αυτή η λέξη “αντιπροσωπευτικές” σημαίνει ότι υπάρχουν τουλάχιστον l τιμές του ευαίσθητου γνωρίσματος στην κλάση ισοδυναμίας. Η θεώρηση αυτή είναι η πιο απλή και αν και αντιμετωπίζει στην γενική περίπτωση το αρχικό πρόβλημα αφήνει περιθώρια σε αρκετές περιπτώσεις στον επιτιθέμενο να αυξήσει την γνώση του. Πιο συγκεκριμένα δεν εμποδίζει από επιθέσεις που εκμεταλλεύονται πιθανοτικά χαρακτηριστικά για την εξαγωγή συμπερασμάτων και την αύξηση της γνώσης του επιτιθέμενου. Αυτό συμβαίνει για παράδειγμα στις περιπτώσεις που η συχνότητα εμφάνισης μιας τιμής μέσα στην κλάση ισοδυναμίας είναι κατά πολύ μεγαλύτερη από τις υπόλοιπες.

l-ποικιλομορφία εντροπίας (entropy l-diversity)

Ορίζουμε την εντροπία μιας κλάσης ισοδυναμίας E την ποσότητα

$$\text{Εντροπία}(E) = - \sum_{s \in S} p(E, s) \log p(E, s)$$

όπου S είναι το πεδίο του ευαίσθητου γνωρίσματος και p είναι το ποσοστό των εγγραφών στην κλάση E οι οποίες έχουν την τιμή s στο γνώρισμα αυτό.

Ένας πίνακας είναι 1-ποικιλόμορφος εντροπίας αν για κάθε κλάση ισοδυναμίας E ισχύει η σχέση:

$$\text{Εντροπία}(E) \geq \log l$$

Η εκδοχή αυτή προστατεύει από τις επιθέσεις που δεν κάλυπτε η προηγούμενη, όμως η αυστηρότητά της μπορεί να οδηγήσει σε υπερβολική γενίκευση των δεδομένων, ειδικά στις περιπτώσεις όπου μια τιμή εμφανίζεται πολύ συχνά. Για να μπορεί να εφαρμοσθεί ο ορισμός αυτός και να ισχύσει η απαίτηση που τέθηκε για κάποια τιμή του ευαίσθητου γνωρίσματος στις κλάσεις ισοδυναμίας, πρέπει η εντροπία του γνωρίσματος αυτού σε ολόκληρο τον πίνακα να είναι τουλάχιστον $\log l$.

Αναδρομική (c,l)-ποικιλομορφία (recursive (c,l)-diversity)

Στην περίπτωση αυτή ελέγχεται αν η πιο συχνά εμφανιζόμενη τιμή του ευαίσθητου γνωρίσματος εμφανίζεται μέσα σε μια κλάση ισοδυναμίας υπερβολικά συχνά καθώς και αν οι λιγότερο συχνές τιμές δεν εμφανίζονται υπερβολικά σπάνια.

Αν μια κλάση ισοδυναμίας έχει m τιμές και r_i ο αριθμός των φορών που εμφανίζεται η i-οστή συχνότερη τιμή της κλάσης, τότε για να είναι μια κλάση αναδρομικά (c,l)-ποικιλόμορφη πρέπει να ισχύει η σχέση:

$$r_1 < c(r_1 + r_{i+1} + \dots + r_m)$$

Ένας πίνακας έχει αναδρομική (c,l) - ποικιλομορφία αν όλες οι κλάσεις του είναι αναδρομικά (c,l) - ποικιλόμορφες.

Με βάση τον ορισμό αυτό ορίστηκαν και η αναδρομική (c,l)-ποικιλομορφία θετικής αποκάλυψης (positive disclosure recursive (c,l)-diversity), αναδρομική (c,l)-ποικιλομορφία αρνητικής/θετικής αποκάλυψης (negative/positive disclosure recursive (c,l)-diversity).

Τέλος μια ξεχωριστή περίπτωση είναι η περίπτωση να υπάρχουν πολλαπλά ευαίσθητα γνωρίσματα. Στην περίπτωση αυτή ορίζεται η πολυγνωρισματική 1-ποικιλομορφία (multi-attribute 1-diversity) ως εξής:

Έστω T ένας πίνακας με μη-ευαίσθητα γνωρίσματα Q_1, \dots, Q_m και ευαίσθητα γνωρίσματα S_1, \dots, S_n . Ο πίνακας T θα λέγεται l -ποικιλύμορφος αν για κάθε ένα από τα ευαίσθητα γνωρίσματα S_i , ο πίνακας T είναι l -ποικιλύμορφος θεωρώντας το S_i ως μοναδικό ευαίσθητο γνώρισμα .

Σε ότι αφορά την αλγοριθμική υλοποίηση της προσέγγισης αυτής, τροποποιείται ο αλγόριθμος της k -ανωνυμίας ώστε αντί να ελέγχεται αν η γενίκευση T^* ενός πίνακα T πληροί την k -ανωνυμία, ελέγχεται αν πληροί την l -ποικιλομορφία. Επειδή αυτό στηρίζεται στην μέτρηση των ευαίσθητων γνωρισμάτων στις κλάσεις ισοδυναμίας, ο αλγόριθμος αυτός είναι αποδοτικός.

Συνολικά η προσέγγιση της l -ποικιλομορφίας είναι πρακτική, απλή και είναι ένα σημαντικό βήμα στην ανωνυμοποίηση των δεδομένων μετά την k -ανωνυμία, αντιμετωπίζοντας κάποια από τα προβλήματα που αυτή αφήνει ανοιχτά, όπως μερικές επιθέσεις ομογενών δεδομένων. Με τις διαφορετικές εκδοχές που προτάθηκαν για την l -ποικιλομορφία σταμάτησε να είναι απαραίτητη η πλήρης κατανομή των γνωρισμάτων.

Για να μην χάνεται πληροφορία λόγω της γενίκευσης των τιμών των ψευδο-αναγνωριστικών προτάθηκε η τεχνική της ανατομίας (anatomy) [3]. Ας δούμε αρχικά ένα παράδειγμα που η γενίκευση αυτή δημιουργεί πρόβλημα στην στατιστική επεξεργασία. Έστω ο παρακάτω πίνακας:

Ηλικία	Φύλο	Ταχυδρομικός Κώδικας	Διάγνωση
33	Γυναίκα	11234	Γρίπη
36	Γυναίκα	11238	Πνευμονία
42	Γυναίκα	11232	Βρογχίτιδα
57	Γυναίκα	11237	Πνευμονία
60	Άντρας	19732	Γρίπη
67	Άντρας	19737	Δυσπεψία
70	Άντρας	19734	Υπέρταση
72	Άντρας	19739	Γρίπη

και η 3-ποικιλόμορφη δημοσίευση αυτού:

Ηλικία	Φύλο	Ταχυδρομικός Κώδικας	Διάγνωση
[31-50]	Γυναίκα	1123*	Γρίπη
[31-50]	Γυναίκα	1123*	Πνευμονία
[31-50]	Γυναίκα	1123*	Βρογχίτιδα
[31-50]	Γυναίκα	1123*	Πνευμονία
[61-70]	Άντρας	1973*	Γρίπη
[61-70]	Άντρας	1973*	Δυσπεψία
[61-70]	Άντρας	1973*	Υπέρταση
[61-70]	Άντρας	1973*	Γρίπη

Ας θεωρήσουμε τώρα έναν ερευνητή που κάνει μια έρευνα για την πνευμονία σε άτομα ηλικίας 31-40. Από τον παραπάνω πίνακα αυτό που μπορεί να δει είναι ότι δύο άτομα με ηλικία μεταξύ 31-50 έχουν πνευμονία και έτσι πρέπει να υπολογίσει την πιθανότητα τα άτομα αυτά να ανήκουν στο επιθυμητό εύρος ηλικιών. Το εύρος της κλάσης ισοδυναμίας είναι 31-50 και δεδομένου ότι ψάχνει ηλικίες 31-40, αν δεν έχει κανένα άλλο στοιχείο θα θεωρήσει την πιθανότητα αυτή $1/3$ ενώ στην πραγματικότητα το $1/2$ των τιμών μέσα στην κλάση αυτή ανήκουν στο ζητούμενο

εύρος τιμών. Με βάση λοιπόν τέτοιου είδους προβλήματα αναπτύχθηκε η τεχνική της ανατομίας, που αποφεύγει την γενίκευση γνωρισμάτων.

Η τεχνική της ανατομίας δημοσιεύει δύο ξεχωριστούς πίνακες, έναν πίνακα ψευδο-αναγνωριστικών (QIT - quasi identifier table) και έναν πίνακα ευαίσθητου γνωρίσματος (ST - sensitive table). Αρχικά χωρίζονται τα δεδομένα σε κλάσεις ισοδυναμίας και αποδίδεται ένας αριθμός (id) σε κάθε μια από αυτές. Έπειτα, κατασκευάζεται ο QIT που περιέχει όλες τις ακριβείς τιμές των ψευδο-αναγνωριστικών και έναν αριθμό που προσδιορίζει σε ποια κλάση ισοδυναμίας ανήκει η εγγραφή. Ο ST περιέχει για κάθε κλάση ισοδυναμίας τις τιμές του ευαίσθητου γνωρίσματος που συναντώνται μέσα στην κλάση, καθώς και το πόσες εγγραφές της κλάσης αντιστοιχούν σε κάθε μία.

Έτσι η δημοσίευση των δεδομένων του παραδείγματος με χρήση της τεχνικής της ανατομίας θα γινόταν με τους δύο παρακάτω πίνακες:

Ηλικία	Φύλο	Ταχυδρομικός Κώδικας	Κλάση ισοδυναμίας
33	Γυναίκα	11234	1
36	Γυναίκα	11238	1
42	Γυναίκα	11232	1
57	Γυναίκα	11237	1
60	Άντρας	19732	2
67	Άντρας	19737	2
70	Άντρας	19734	2
72	Άντρας	19739	2

QIT

Κλάση ισοδυναμίας	Διάγνωση	Μετρητής
1	Γρίπη	1
1	Πνευμονία	2
1	Βρογχίτιδα	1
2	Γρίπη	2
2	Δυσπεψία	1
2	Υπέρταση	1

ST

Στην περίπτωση αυτή ο ερευνητής του παραπάνω παραδείγματος ξέρει ακριβώς ποια είναι η πιθανότητα μια εγγραφή της πρώτης κλάσης ισοδυναμίας να έχει την ζητούμενη ηλικία και μπορεί να εξάγει πιο ακριβή και έγκυρα συμπεράσματα. Με βάση τα παραπάνω παρουσιάζουμε τον τυπικό ορισμό της τεχνικής αυτής.

Ορισμός (Ανατομία)

Για έναν πίνακα T και μια εγγραφή αυτού t συμβολίζουμε $A_1^{qi}, A_2^{qi}, \dots, A_m^{qi}$ τα γνωρίσματα που αποτελούν ψευδο-αναγνωριστικά, τις τιμές αυτών $t[1], t[2], \dots, t[d]$ και A^s το ευαίσθητο γνώρισμα με τιμή $t[d+1]$. Θεωρούμε μια διαμέριση του πίνακα με QI_1, QI_2, \dots, QI_m κλάσεις ισοδυναμίας που κάθε μια από αυτές πληροί τις απαιτήσεις που έχουν τεθεί από την l -ποικιλομορφία. Η τεχνική της ανατομίας παράγει έναν πίνακα ψευδο-αναγνωριστικών QIT και έναν πίνακα ευαίσθητου γνωρίσματος ST με σχήματα: $(A_1^{qi}, A_2^{qi}, \dots, A_m^{qi}, \text{κλάση ισοδυναμίας})$ και $(\text{κλάση ισοδυναμίας}, A^s, \text{Μετρητής})$ αντίστοιχα.

Για κάθε κλάση ισοδυναμίας QI_j ($1 \leq j \leq m$), και για κάθε πλειάδα t που ανήκει σε αυτήν ο QIT έχει μια εγγραφή της μορφής $(t[1], t[2], \dots, t[d], j)$.

Για κάθε κλάση ισοδυναμίας QI_j και κάθε διαφορετική τιμή v του ευαίσθητου γνωρίσματος, ο ST περιέχει μία εγγραφή της μορφής $(j, v, c_j(v))$, όπου $c_j(v)$ είναι ο αριθμός των πλειάδων $t \in QI_j$ για τις οποίες $t[d+1] = v$. Οι πίνακες QIT και ST δεν περιλαμβάνουν κανένα άλλο δεδομένο εκτός από αυτά που ορίστηκαν παραπάνω.

Η τεχνική της ανατομίας προστατεύει την ιδιωτικότητα γιατί με τον τρόπο αυτό, αν και δημοσιεύονται οι ακριβείς τιμές των προσδιοριστών, δεν υπάρχει στα δημοσιευμένα δεδομένα η ακριβής αντιστοίχισή τους με το ευαίσθητο γνώρισμα, ικανοποιώντας ταυτόχρονα την l -ποικιλομορφία με ότι προστασία αυτή συνεπάγεται. Επιπλέον οι παραπάνω πίνακες είναι εύκολο να κατασκευαστούν, δίνοντας στην μέθοδο μια απλή υλοποίηση με μικρή πολυπλοκότητα συγκριτικά με τις υπόλοιπες μεθόδους. Παρ'όλα αυτά με την μέθοδο αυτή φαίνεται να είναι πιο εύκολο για τον επιτιθέμενο να μάθει κάποια στοιχεία και οι δημοσιεύσεις που παράγονται να του προσφέρουν πιο ακριβή γνώση από ότι έχει ήδη για τα γνωρίσματα του ψευδο-αναγνωριστικού με κίνδυνο να μπορεί να τα χρησιμοποιήσει σε μια επίθεση σύνδεσης ή να βεβαιώσει την παρουσία κάποιου ατόμου στον πίνακα που δημοσιεύθηκε. Αυτό το μειονέκτημα παρουσιάζεται γιατί ουσιαστικά η ανάπτυξη της παραπάνω μεθόδου έχει γίνει με βάση δύο υποθέσεις. Κατά πρώτον θεωρείται ότι ο επιτιθέμενος έχει γνώση όλων των τιμών του ψευδο-αναγνωριστικού και αναζητά την τιμή του

ευαίσθητου δεδομένου. Κατά δεύτερον, ο επιτιθέμενος γνωρίζει με βεβαιότητα ότι το άτομο που αναζητά σχετίζεται με αυτά τα δεδομένα, δηλαδή ότι υπάρχει εγγραφή που του αντιστοιχεί. Παρά λοιπόν την αποδοτικότητα και την μη απώλεια δεδομένων, η τεχνική αυτή αφήνει μεγάλο περιθώριο στον επιτιθέμενο να κερδίσει γνώση.

3.1.3 Επιθέσεις παρουσίας. Η τεχνική της δ-παρουσίας.

Μια διαφορετική πρόταση για την αντιμετώπιση των επιθέσεων ομογενών δεδομένων είναι η προσέγγιση της **δ-παρουσίας (δ-presence)** [8]. Η προσέγγιση αυτή επικεντρώνεται στην γνώση της παρουσίας ενός ατόμου σε έναν δημοσιευμένο πίνακα. Εξετάζει δηλαδή κυρίως περιπτώσεις, που με βάση ένα πίνακα P δημοσιεύεται ένας πίνακας T που περιέχει το υποσύνολο των εγγραφών που πληρούν κάποια ιδιότητα. Είναι λοιπόν φανερό ότι αν και στην περίπτωση αυτή δεν δημοσιεύεται ευαίσθητο γνώρισμα, η τιμή του είναι γνωστή και κοινή για όλα τα άτομα που αντιστοιχούν σε εγγραφή του πίνακα.

Για παράδειγμα ας θεωρηθεί το σύνολο δεδομένων P και το ερευνητικό υποσύνολο T. Έστω ότι το P περιέχει όλα τα άτομα που έλαβαν μέρος σε μια έρευνα για τον διαβήτη. Το ερευνητικό υποσύνολο T δεν περιέχει το γνώρισμα αυτό, περιέχει όμως όλα τα άτομα που έχουν διαβήτη. Δημοσιοποιείται εξ' αυτών ο πίνακας P και η γενίκευση του πίνακα T. Σκοπός του επιτιθέμενου σε αυτή την περίπτωση είναι να καθορίσει την παρουσία ή μη κάποιου ατόμου στον πίνακα T.

Η βασική ιδέα της εργασίας που αφορά την δ-παρουσία είναι ότι η ανωνυμοποίηση ενός συνόλου δεδομένων τέτοιου τύπου σημαίνει ότι ο επιτιθέμενος δεν θα πρέπει να είναι σε θέση να προσδιορίσει κανένα άτομο ως παρόν στον πίνακα T ή ως απόν.

Ορισμός (δ-παρουσία)

Έστω ένας εξωτερικός δημόσιος πίνακας P , ένας ιδιωτικός πίνακας T και η γενικευμένη δημοσίευση αυτού T^* . Ικανοποιείται η δ-παρουσία για την γενίκευση του T , αν η πιθανότητα να ανήκει οποιαδήποτε πλειάδα του P στον T γνωρίζοντας τον πίνακα T^* είναι μεταξύ δύο τιμών δ_{\min} και δ_{\max} .

Αυτό σημαίνει ότι το διάστημα $\delta = (\delta_{\min}, \delta_{\max})$ είναι το διάστημα των αποδεκτών τιμών για την πιθανότητα μια πλειάδα να ανήκει στον T δεδομένου του T^* . Για να επιτευχθεί μια καλή ανωνυμοποίηση με την μέθοδο αυτή πρέπει να επιλεγεί το μέγιστο δυνατό δ_{\min} και το ελάχιστο δυνατό δ_{\max} . Δύο αλγόριθμοι έχουν προταθεί για την επιτυχία της δ-παρουσίας. Ο πρώτος είναι ο αλγόριθμος μονοδιάστατης παρουσίας (Single-Dimensional Presence Algorithm – SPALM) ο οποίος είναι κατάλληλος για σύνολα δεδομένων στα οποία μπορεί να εφαρμοστεί πλήρης γενίκευση πεδίου (full domain generalization). Ο αλγόριθμος αυτός παράγει τέτοιες γενικεύσεις που να πληρούν την δ-παρουσία, μεγιστοποιώντας μια μετρική ακρίβειας. Το πρόβλημα που εμφανίζεται με την χρήση αυτού του αλγορίθμου είναι ότι παρά την βέλτιστη γενίκευση που εφαρμόζεται, με τις μονοδιάστατες γενικεύσεις αν σε μια πλειάδα γενικευθεί μία τιμή, τότε όλες οι όμοιες τιμές του πίνακα γενικεύονται. Το πρόβλημα αυτό έρχεται να αντιμετωπίσει ο δεύτερος αλγόριθμος που ονομάζεται αλγόριθμος πολυδιάστατης παρουσίας (Multi-Dimensional Presence Algorithm MPALM) . Αυτός επιτρέπει σε μια ομάδα πλειάδων να γενικεύονται, αφήνοντας κάποια άλλη ομάδα για την οποία δεν υπάρχει ανάγκη γενίκευσης στην αρχική της μορφή. Μπορεί αυτός ο αλγόριθμος να παράγει λιγότερο βέλτιστες πολυδιάστατες γενικεύσεις, αλλά οδηγούν τελικά σε μικρότερη αλλοίωση της πληροφορίας.

Για να αντιστοιχηθεί η προσέγγιση αυτή με αυτή της 1-ποικιλομορφίας θα πρέπει να προστεθεί ένα ευαίσθητο γνώρισμα στον πίνακα P , του οποίου η τιμή φανερώνει την παρουσία ή απουσία του ατόμου από τον πίνακα T . Στο παράδειγμα που περιγράφηκε παραπάνω αυτό θα ήταν τιμή 0 ή 1 που προσδιορίζει για κάθε άτομο αν έχει την ασθένεια ή όχι.

Ο καθορισμός της δ-παρουσίας θεωρείται πως υπερέρχει της 1-ποικιλομορφίας ως προς την μεγαλύτερη ακρίβεια που επιτυγχάνεται με την χρήση των δύο παραμέτρων δ_{\min} και δ_{\max} . Η πιο ακριβής εκδοχή της 1-ποικιλομορφίας είναι αυτή της αναδρομικής (c,1)-ποικιλομορφίας, η οποία παρόλα αυτά δεν είναι κατάλληλη να δώσει την δ-παρουσία, λόγω της μιας παραμέτρου c που χρησιμοποιεί, σε αντίθεση με τις δύο παραμέτρους της δ-παρουσίας και της αδυναμίας να αντιστοιχισθεί έστω σε μία από τις δύο αυτές παραμέτρους. Παρόλα αυτά σε περιπτώσεις που δεν είναι τόσο ειδικές όσο αυτή που περιγράφεται στην εργασία της δ-παρουσίας, η προσέγγιση της 1-ποικιλομορφίας, με τις διάφορες εκδοχές που αυτό ορίζεται, είναι αυτή που υποδεικνύει τον τρόπο αντιμετώπισης επιθέσεων ομογενών δεδομένων.

Μια ριζικά διαφορετική περίπτωση δεδομένων στα οποία μπορεί να εμφανιστεί η επίθεση ομογενών δεδομένων είναι αυτή των δεδομένων που οι τιμές τους είναι σύνολα (set-valued data). Θα επανέλθουμε στις επιθέσεις εναντίον τέτοιων δεδομένων αργότερα στην μελέτη μας, όταν θα χρειαστεί να άρουμε την υπόθεση κατά την οποία τα δεδομένα ακολουθούν το σχεσιακό μοντέλο οργάνωσης (υπόθεση την οποία έχουμε κάνει “σιωπηλά” από την αρχή της μελέτης).

3.1.4 Επιθέσεις ανομοιομορφων δεδομένων/ομοιότητας. Η τεχνική της t-εγγύτητας.

Είναι πιθανό ένας επιτιθέμενος να κερδίσει πληροφορία για ένα ευαίσθητο γνώρισμα, γνωρίζοντας την συνολική κατανομή των τιμών του γνωρίσματος αυτού στον πληθυσμό που μελετάται. Αυτό συμβαίνει όταν η συνολική κατανομή είναι ανομοιομορφη. Στις περιπτώσεις αυτές λέμε ότι έχουμε επίθεση ανομοιομορφων δεδομένων (skewness attack). Έχει παρατηρηθεί ότι η 1-ποικιλόμορφη ανωνυμοποίηση των δεδομένων δεν αρκεί για να αντιμετωπίσει το πρόβλημα και με αφορμή τις παρατηρήσεις αυτές προτάθηκε η ανωνυμοποίηση της t-εγγύτητας [7].

Ορισμός (t-εγγύτητα)

Σύμφωνα με την θεώρηση της t-εγγύτητας μια κλάση ισοδυναμίας είναι t-εγγύς αν η απόσταση μεταξύ της κατανομής P ενός ευαίσθητου γνωρίσματος στην κλάση αυτή και της κατανομής Q του γνωρίσματος αυτού στον συνολικό πίνακα δεν είναι μεγαλύτερη από ένα κατώφλι t. Ένας πίνακας δεδομένων είναι t-εγγύς αν όλες οι κλάσεις ισοδυναμίας του είναι t-εγγύς.

Για την μέτρηση της απόστασης μεταξύ δύο κατανομών χρησιμοποιείται η μετρική EMD (απόσταση μετακίνησης γης - Earth Mover Distance).

Ας θεωρήσουμε σαν παράδειγμα ένα σύνολο δεδομένων που περιέχει τα αποτελέσματα μιας εξέτασης για έναν ιό. Το ευαίσθητο γνώρισμα σε αυτή την περίπτωση αντιστοιχεί στον αν βρέθηκε θετικός ή αρνητικός στον ιό ο εξεταζόμενος. Επίσης υποθέτουμε ότι το 99% των ατόμων που περιέχονται στα δεδομένα αυτά ήταν αρνητικοί στον ιό, ενώ μόνο το 1% ήταν θετικοί. Είναι φανερό πως δεν δημιουργεί πρόβλημα σε κάποιον να θεωρηθεί από τον επιτιθέμενο αρνητικός στον ιό, αφού αυτό ισχύει για την συντριπτική πλειοψηφία του πληθυσμού. Αν έχουμε μια κλάση ισοδυναμίας η οποία αποτελείται από ισάριθμες εγγραφές θετικών και αρνητικών ατόμων, κάθε άτομο που ανήκει στην κλάση αυτή θα θεωρηθεί από τον επιτιθέμενο θετικό στον ιό με πιθανότητα 50%. Αυτή η πιθανότητα είναι πολύ μεγαλύτερη από την αρχική πιθανότητα 1% και επομένως ο επιτιθέμενος έχει κερδίσει κάποια πληροφορία που δεν ήταν επιθυμητό να έχει. Στην περίπτωση αυτή η κατανομή P διέφερε σημαντικά από την κατανομή Q και η κλάση ισοδυναμίας δεν ήταν t-εγγύς.

Όπως έχει ήδη αναφερθεί το κέρδος πληροφορίας ενός επιτιθέμενου μπορεί να μετρηθεί ως η διαφορά της τελικής και αρχικής πεποίθησής του. Το κέρδος αυτό λαμβάνεται σε δύο στάδια. Το πρώτο αφορά το κέρδος πληροφορίας από τον συνολικό πληθυσμό και οδηγεί σε μια πεποίθηση B_1 . Το δεύτερο αφορά την πληροφορία για συγκεκριμένα άτομα και οδηγεί σε μια πεποίθηση B_2 . Αν B_0 είναι η αρχική πεποίθηση η I-ποικιλομορφία εξετάζει την διαφορά μεταξύ B_2 και B_0 , ενώ στην περίπτωση της t-εγγύτητας εξετάζεται η διαφορά μεταξύ B_2 και B_1 . Η προσέγγιση αυτή στηρίζεται στην θεώρηση ότι η κατανομή του ευαίσθητου γνωρίσματος στον συνολικό πίνακα είναι δημόσιο δεδομένο, αφού ακόμα και στην περίπτωση της πλήρους απαλοιφής η κατανομή αυτή θα δημοσιευόταν. Οριοθετώντας

λοιπόν την απόσταση των κατανομών P και Q , οριοθετείται και η διαφορά των B_2 και B_1 .

Είναι απαραίτητη η χρήση μιας μετρικής για τις τιμές των γνωρισμάτων έτσι ώστε η απόσταση εδάφους (ground distance) να ορίζεται μεταξύ οποιουδήποτε ζεύγους τιμών. Είναι επιθυμητό επίσης η απόσταση μεταξύ δύο πιθανοτικών κατανομών πάνω στις τιμές να εξαρτάται από την απόσταση αυτή μεταξύ των τιμών. Οι παραπάνω απαιτήσεις οδήγησαν στην χρήση της EMD. Αυτή βασίζεται στο ελάχιστο ποσό έργου που χρειάζεται για να μετασχηματιστεί μια κατανομή σε μια άλλη μετακινώντας μάζα κατανομής μεταξύ αυτών. Για αριθμητικά γνωρίσματα χρησιμοποιείται η διατεταγμένη απόσταση (ordered distance). Η διατεταγμένη απόσταση μεταξύ δύο τιμών υπολογίζεται με βάση το πλήθος των τιμών ανάμεσά τους. Για κατηγορικά (categorical) δεδομένα υπάρχουν δύο μετρικές: της ίσης απόστασης (equal distance) και της ιεραρχικής απόστασης (hierarchical distance). Στη πρώτη μετρική η απόσταση μεταξύ οποιονδήποτε κατηγορικών ορισμάτων ορίζεται ίση με ένα, ενώ στην δεύτερη περίπτωση βασίζεται στο ελάχιστο επίπεδο στο οποίο μπορούν οι τιμές αυτές να γενικευθούν από κοινού σύμφωνα με την ιεραρχία.

Δύο βασικές ιδιότητες της t -εγγύτητας είναι αυτή της γενίκευσης κι αυτή του υποσυνόλου.

Ιδιότητα γενίκευσης:

Έστω A και B οι γενικεύσεις ενός πίνακα T και η A είναι πιο γενική από την B . Αν ο T ικανοποιεί την t -εγγύτητα χρησιμοποιώντας την γενίκευση B , τότε την ικανοποιεί και χρησιμοποιώντας την γενίκευση A .

Ιδιότητα υποσυνόλου:

Αν C ένα σύνολο γνωρισμάτων ενός πίνακα T και ο T ικανοποιεί την t -εγγύτητα για τα γνωρίσματα του C , τότε ο T ικανοποιεί την t -εγγύτητα για οποιοδήποτε υποσύνολο γνωρισμάτων του C .

Η μέθοδος ανωνυμοποίησης της t -εγγύτητας αντιμετωπίζει αποδοτικά την επίθεση ανομοιομορφων δεδομένων. Ο διαχωρισμός της διαδικασίας απόκτησης κέρδους από τον επιτιθέμενο σε δύο φάσεις είχε σαν αποτέλεσμα την εστίαση στο πραγματικά προβληματικό μέρος της διαδικασίας αυτής. Η χρήση της EMD παρείχε την κατάλληλη μετρική που ήταν απαραίτητη για την υλοποίηση. Παρ' όλα αυτά η απαίτηση της εγγύτητας των κατανομών που περιγράφηκε μπορεί να περιορίσει την ποσότητα χρήσιμων πληροφοριών αφού μειώνει την συσχέτιση μεταξύ των ψευδο-αναγνωριστικών και των ευαίσθητων γνωρισμάτων. Τέλος, η χρήση της μετρικής EMD παρουσιάζει το μειονέκτημα ότι δεν είναι ξεκάθαρη η σχέση μεταξύ της τιμής t της t -εγγύτητας και του κέρδους πληροφορίας που μπορεί να έχει ο επιτιθέμενος.

Σε παρόμοιο κλίμα, η τεχνική της t -εγγύτητας είναι ιδιαίτερα αποτελεσματική σε επιθέσεις ομοιότητας, οι οποίες αφορούν την περίπτωση όπου τιμές του ευαίσθητου γνωρίσματος είναι διαφορετικές αλλά νοηματικά όμοιες και επομένως ο επιτιθέμενος μπορεί να μάθει σημαντικές πληροφορίες.

Για την κατηγορία επιθέσεων αυτή θα θεωρήσουμε δύο περιπτώσεις. Στην μια περίπτωση το ευαίσθητο γνώρισμα παίρνει αριθμητικές τιμές, ενώ στην άλλη κατηγορικές τιμές. Το πρόβλημα της επίθεσης ομοιότητας μπορεί να εμφανισθεί και στις δύο περιπτώσεις.

Θεωρούμε έναν πίνακα ο οποίος αφορά κάποιους εργαζόμενους και λαμβάνουμε ως ευαίσθητο γνώρισμα τον μισθό αυτών. Βλέπουμε την παρακάτω δημοσίευση αυτού που είναι 1-ποικιλόμορφη με $l=3$.

Ταχυδρομικός Κώδικας	Ηλικία	Μισθός
145**	25-30	700
145**	25-30	750
145**	25-30	730
145**	25-30	750
112**	31-50	1000
112**	31-50	1700
112**	31-50	700
112**	31-50	1000

Παρατηρούμε ότι ο πίνακας αυτός έχει δύο κλάσεις ισοδυναμίας. Στην πρώτη από αυτές η τιμή του μισθού των εργαζομένων που ανήκουν στην πρώτη από αυτές παίρνει τρεις τιμές (700, 750, 730) οι οποίες είναι σχετικά κοντινές, δηλαδή περιγράφουν το ίδιο μισθολογικό επίπεδο. Στην δεύτερη κλάση αυτό δεν συμβαίνει και δεν μπορούμε να πούμε ότι οι τιμές αυτές ανήκουν όλες σε κάποιο επίπεδο. Έτσι αν ο επιτιθέμενος γνωρίζει ότι κάποιος αντιστοιχεί σε εγγραφή που ανήκει στην πρώτη κλάση ισοδυναμίας μπορεί να συμπεράνει ότι το επίπεδο του μισθού του είναι ανάμεσα στα 700-750 ευρώ και επομένως έχει κερδίσει σημαντική πληροφορία για την μισθολογική του κατάσταση. Ο επιτιθέμενος λοιπόν, αύξησε την γνώση που είχε για την τιμή του ευαίσθητου γνωρίσματος παρά την 1-ποικιλομορφία του πίνακα, λόγω του ότι οι τιμές ανήκουν σε ένα μικρό διάστημα συγκριτικά με τις τιμές που θα μπορούσε να λάβει το γνώρισμα αυτό και αυτές που υπάρχουν συνολικά στον πίνακα. Αν η εγγραφή του ατόμου που αναζητούσε ο επιτιθέμενος ανήκε στην δεύτερη κλάση ισοδυναμίας δεν θα μπορούσε να συμπεράνει κάτι παρόμοιο λόγω της μεγάλης απόστασης των τιμών του ευαίσθητου γνωρίσματος.

Όπως αναφέρθηκε παραπάνω το πρόβλημα αυτό δεν εντοπίζεται μόνο στην περίπτωση των ευαίσθητων γνωρισμάτων με αριθμητικές τιμές. Για παράδειγμα ας θεωρήσουμε τον παρακάτω δημοσιευμένο πίνακα:

Ταχυδρομικός κώδικας	Φύλλο	Ηλικία	Ασθένεια
141**	Άντρας	50-60	Καρκίνος του ήπατος
141**	Άντρας	50-60	Καρκίνος του λάρυγγα
141**	Άντρας	50-60	Καρκίνος των νεφρών
121**	Γυναίκα	20-30	Γαστρίτιδα
121**	Γυναίκα	20-30	Έλκος στομάχου
121**	Γυναίκα	20-30	Καρκίνος στομάχου
137**	Άντρας	35-40	Πνευμονία
137**	Άντρας	35-40	Γρίπη
137**	Άντρας	35-40	Γαστρίτιδα

Θεωρώντας ως ευαίσθητο γνώρισμα την ασθένεια, ο πίνακας αυτός έχει τρεις κλάσεις ισοδυναμίας. Είναι φανερό ότι αν ο επιτιθέμενος γνωρίζει αρκετά στοιχεία ώστε να κατατάξει το άτομο που τον ενδιαφέρει στην πρώτη ομάδα, μπορεί με βεβαιότητα να συμπεράνει ότι πάσχει από κάποια μορφή καρκίνου. Αυτή είναι μια σημαντική πληροφορία που κέρδισε χωρίς να μπορεί να αντιστοιχήσει την ακριβή τιμή του ευαίσθητου γνωρίσματος στο άτομο που αναζητά. Αν το άτομο που αναζητά ο επιτιθέμενος ανήκε στην δεύτερη κλάση ισοδυναμίας και πάλι θα μπορούσε να βγάλει κάποιο συμπέρασμα. Οι ασθένειες που παρατηρούνται στην κλάση αυτή σχετίζονται όλες με το στομάχι. Επομένως ο επιτιθέμενος γνωρίζει ότι το άτομο που τον ενδιαφέρει έχει κάποιο πρόβλημα με το στομάχι του και αυτό είναι ένα αρκετά μεγάλο κέρδος πληροφορίας. Στις επιθέσεις λοιπόν αυτές ο επιτιθέμενος αν και δεν βρίσκει την ακριβή τιμή του ευαίσθητου γνωρίσματος βρίσκει μια περιοχή τιμών στην οποία ανήκει αυτή, δηλαδή μαθαίνει περίπου ποια είναι η τιμή ή κάποια κατηγορία στην οποία ανήκει αυτή.

3.1.5 Επιθέσεις γνώσης σε πίνακες με συχνές ενημερώσεις. Η τεχνική της *m*-αμεταβλητότητας.

Ας θεωρήσουμε τώρα πως έχουμε έναν δημοσιευμένο πίνακα, ο οποίος ανά τακτά διαστήματα ενημερώνεται τόσο με την προσθήκη νέων εγγραφών όσο και με την διαγραφή κάποιων παλαιότερων. Σε αυτό το σημείο μπορούμε να παρατηρήσουμε πως οι τεχνικές που έχουμε μελετήσει ως τώρα (*k*-ανωνυμία, *l*-ποικιλομορφία) αφορούν στατικά δημοσιευμένα δεδομένα. Αυτό όμως, μπορεί να δημιουργήσει διαρροή πληροφορίας αν επιθυμούμε ο δημοσιευμένος πίνακας να ανανεώνεται ανάλογα με το πως μεταβάλλονται οι ιδιωτικές πληροφορίες (για παράδειγμα ένα νοσοκομείο μπορεί να αλλάζει τα δεδομένα του δημοσιευμένου πίνακα καθώς αλλάζουν και οι ασθενείς που νοσηλεύονται σε αυτό). Ας δούμε ένα παράδειγμα για να κατανοήσουμε καλύτερα το φαινόμενο.

Στον Πίνακα 4α είναι τα ιατρικά δεδομένα προς δημοσίευση ενός νοσοκομείου. Πριν την δημοσίευσή τους, εφαρμόζουμε σε αυτόν τις τεχνικές της *k*-ανωνυμίας και της *l*-ποικιλομορφίας. Αποτέλεσμα είναι ο 2-ανώνυμος και 2-ποικιλόμορφος Πίνακας 4β. Την χρονική στιγμή t τα ιατρικά δεδομένα έχουν αλλάξει και παριστάνονται στον Πίνακα 5α. Θα πρέπει ανάλογα λοιπόν να αλλάξουν και τα δημοσιευμένα δεδομένα ώστε να είναι συνεπή με τα ιδιωτικά. Για να το κάνουμε αυτό, παρατηρούμε πως από τα αρχικά δεδομένα έχουν διαγραφεί οι εγγραφές των Alice, Andy, Helen, Paul και Ken ενώ έχουν προστεθεί αυτές των Emily, Mary, Ray, Tom και Vince (φαίνονται με bold στον Πίνακα 5α). Δημοσιοποιείται λοιπόν, ο Πίνακας 5β ο οποίος είναι και αυτός 2-ανώνυμος και 2-ποικιλόμορφος. Υποθέτουμε, τώρα, πως ένας επιτιθέμενος γνωρίζει τα γνωρίσματα Ηλικία και Ταχυδρομικός Κώδικας του Bob και πως υπάρχει εγγραφή στον δημοσιευμένο πίνακα που να αντιστοιχεί σε αυτόν. Προσπαθώντας να αποσπάσει από τον Πίνακα 4β την ασθένεια του Bob δεν θα τα καταφέρει, μιας και θα βρει εκεί δύο εγγραφές που ταιριάζουν με την γνώση που διαθέτει. Το μόνο που θα μπορέσει να συμπεράνει είναι πως ο Bob πάσχει είτε από Δυσπεψία είτε από Βρογχίτιδα.

Την χρονική στιγμή t όμως, έχουμε δημοσίευση του Πίνακα 5β. Ο επιτιθέμενος, γνωρίζοντας πως η εγγραφή του Bob συνεχίσει να βρίσκεται στα δημοσιευμένα δεδομένα και ελπίζοντας πως αυτήν την φορά οι πληροφορίες δεν έχουν ανωνυμοποιηθεί προσπαθεί ξανά να ανακαλύψει την ασθένεια του Bob. Και πάλι (μιας και ο Πίνακας 5β είναι 2-ανώνυμος και 2-ποικιλόμορφος) αποτυγχάνει, καταλήγοντας πως ο Bob πάσχει είτε από Δυσπεψία είτε από Γαστρίτιδα. Ή μήπως δεν έχει αποτύχει; Επανεξετάζοντας την πληροφορία που αποκόμισε από τις δύο δημοσιεύσεις συνειδητοποιεί ξαφνικά πως η ασθένεια του Bob βρίσκεται στην τομή των {Δυσπεψία, Βρογχίτιδα} και {Δυσπεψία, Γαστρίτιδα}, δηλαδή μπορεί να εξάγει με απόλυτη βεβαιότητα πως ο Bob έχει Δυσπεψία, παρόλο που όλοι οι πίνακες που δημοσιεύτηκαν ικανοποιούν τις αρχές της k -ανωνυμίας και l -διαφορετικότητας.

Όνομα	Ηλικία	Ταχυδρομικός Κώδικας	Ασθένεια
Bob	21	12000	Δυσπεψία
Alice	22	14000	Βρογχίτιδα
Andy	24	18000	Γρίπη
David	23	25000	Γαστρίτιδα
Gary	41	20000	Γρίπη
Helen	36	27000	Γαστρίτιδα
Jane	37	33000	Δυσπεψία
Ken	40	35000	Γρίπη
Linda	43	26000	Γαστρίτιδα
Paul	52	33000	Δυσπεψία
Steve	56	34000	Γαστρίτιδα

Πίνακας 4α

Κλάση Ισοδυναμίας	Ηλικία	Ταχυδρομικός Κώδικας	Ασθένεια
1	[21,22]	[12000,14000]	Δυσπεψία
	[21,22]	[12000,14000]	Βρογχίτιδα
2	[23,24]	[18000,25000]	Γρίπη
	[23,24]	[18000,25000]	Γαστρίτιδα
3	[36,41]	[20000,27000]	Γρίπη
	[36,41]	[20000,27000]	Γαστρίτιδα
4	[37,43]	[26000,35000]	Δυσπεψία
	[37,43]	[26000,35000]	Γρίπη
5	[52,56]	[33000,34000]	Δυσπεψία
	[52,56]	[33000,34000]	Γαστρίτιδα

Πίνακας 4β

Όνομα	Ηλικία	Ταχυδρομικός Κώδικας	Ασθένεια
Bob	21	12000	Δυσπεψία
David	23	25000	Γαστρίτιδα
<i>Emily</i>	25	21000	Γρίπη
Jane	37	33000	Δυσπεψία
Linda	43	26000	Γαστρίτιδα
Gary	41	20000	Γρίπη
<i>Mary</i>	46	30000	Γαστρίτιδα
<i>Ray</i>	54	31000	Δυσπεψία
Steve	56	34000	Γαστρίτιδα
<i>Tom</i>	60	44000	Γαστρίτιδα
<i>Vince</i>	65	36000	Γρίπη

Πίνακας 5α

Κλάση Ισοδυναμίας	Ηλικία	Ταχυδρομικός Κώδικας	Ασθένεια
1	[21,23]	[12000,25000]	Δυσπεψία
	[21,23]	[12000,25000]	Γαστρίτιδα
2	[25,43]	[21000,33000]	Γρίπη
	[25,43]	[21000,33000]	Δυσπεψία
	[25,43]	[21000,33000]	Γαστρίτιδα
3	[41,46]	[20000,30000]	Γρίπη
	[41,46]	[20000,30000]	Γαστρίτιδα
4	[54,56]	[31000,34000]	Δυσπεψία
	[54,56]	[31000,34000]	Γαστρίτιδα
5	[60,65]	[36000,44000]	Γαστρίτιδα
	[60,65]	[36000,44000]	Γρίπη

Πίνακας 5β

Καταλήγουμε, λοιπόν, πως οι τεχνικές που μέχρι τώρα έχουν προταθεί για την ανωνυμοποίηση, δεν έχουν προβλέψει μη-στατικά δημοσιευμένα δεδομένα. Για αυτό τον σκοπό διατυπώθηκε η αρχή της m-αμεταβλητότητας (**m-invariance**) [6]. Πριν δώσουμε τον σχετικό ορισμό ας παρουσιάσουμε πρώτα μερικές ακόμα σημαντικές έννοιες:

Χρονόσημα(timestamp):

Χρονόσημα μίας πλειάδας j ενός πίνακα T ονομάζουμε τον ακέραιο t_i ο οποίος χαρακτηρίζει την i -οστή δημοσίευση της πλειάδας j .

Ιστορική Ένωση(Historical Union):

Έστω χρονική στιγμή $t_i > 1$. Η ιστορική ένωση $U(t_i)$ ενός πίνακα T , είναι το σύνολο όλων των πλειάδων του T με χρονόσημα $1, 2, \dots, t_i$ αντίστοιχα. Φορμαλιστικά έχουμε:

$$U(t_i) = \bigcup_{t_i=1}^n t_i$$

Υπογραφή (Signature):

Έστω ένας k -ανώνυμος πίνακας τον οποίο συμβολίζουμε $T^*(t)$, με QI^* μία κλάση ισοδυναμίας του κάποια χρονική στιγμή $t \in [1, n]$. Η υπογραφή της QI^* είναι το σύνολο των διακριτών ευαίσθητων δεδομένων στην QI^* .

Διάρκεια ζωής (Lifespan):

Έστω μία εγγραφή j ενός πίνακα T (θεωρούμε πως στον T γίνονται συχνές ενημερώσεις, καθεμία από τις οποίες χαρακτηρίζουμε με την χρονική στιγμή που συνέβησαν) την χρονική στιγμή t . Θα ονομάζουμε διάρκεια ζωής της j την πλειάδα $[x, y]$ όπου x η χρονική στιγμή στην οποία εμφανίζεται για πρώτη φορά η εγγραφή j στον πίνακα T και y η χρονική στιγμή στην οποία εμφανίζεται για τελευταία φορά σε αυτόν.

Τώρα που δώσαμε τους παραπάνω ορισμούς, ας επιστρέψουμε στο παράδειγμά μας και ας προσπαθήσουμε να δούμε τι οδήγησε σε αυτήν την διαρροή πληροφορίας. Εύκολα, μπορεί κάποιος να καταλήξει πως ο λόγος για τον οποίο ο επιτιθέμενος κατάφερε να αποσπάσει την ευαίσθητη πληροφορία, ήταν η απουσία της τιμής Βρογχίτιδας από το ευαίσθητο γνώρισμα Ασθένεια στην δεύτερη δημοσίευση του πίνακα. Αυτό το φαινόμενο ονομάζεται κρίσιμη απουσία (critical absence) και είναι αυτό το οποίο προσπαθεί να αντιμετωπίσει η m-αμεταβλητότητα της οποίας ήρθε η στιγμή να δώσουμε τον ορισμό.

m-αμεταβλητότητα (m-invariance):

Θα λέμε ότι ένας δημοσιευμένος k-ανώνυμος πίνακας $T^*(t)$ (για χρονικές στιγμές $t \in [1, n]$) είναι **m-μοναδικός (m-unique)** αν κάθε κλάση ισοδυναμίας σε αυτόν περιέχει το λιγότερο m εγγραφές και όλες οι εγγραφές στην κλάση έχουν διαφορετικές τιμές στα ευαίσθητα γνωρίσματα. Μία ακολουθία από δημοσιευμένους πίνακες $T^*(1), \dots, T^*(n), n \geq 1$ ονομάζεται m-αμετάβλητη αν ισχύουν οι παρακάτω προϋποθέσεις :

1. Ο πίνακας $T^*(t)$ είναι m-μοναδικός $\forall t \in [1, n]$
2. Για κάθε εγγραφή $j \in U(n)$ με διάρκεια ζωής $[x, y]$, τα $j.QI^*(x), j.QI^*(x+1), \dots, j.QI^*(y)$ έχουν την ίδια υπογραφή, όπου το $j.QI^*(t), t \in [x, y]$ είναι η κλάση ισοδυναμίας στην οποία ανήκει η εγγραφή j την χρονική στιγμή t .

Ας μελετήσουμε λίγο τους παραπάνω ορισμούς (m-μοναδικότητα, m-αμεταβλητότητα). Μπορούμε, κατ' αρχήν, να παρατηρήσουμε πως ο ορισμός της m-μοναδικότητας είναι μία πιο ισχυρή μορφή της l-ποικιλομορφίας μιας και αυτός απαιτεί όλες οι εγγραφές μίας κλάσης ισοδυναμίας να έχουν διαφορετικές τιμές στα ευαίσθητα γνωρίσματα. Οπότε, ένας πίνακας T ο οποίος ικανοποιεί την m-μοναδικότητα είναι και m-ποικιλόμορφος ενώ το αντίστροφο δεν ισχύει.

Συνεχίζουμε με τον ορισμό της m -αμεταβλητότητας ο οποίος είναι αυτός που μας προσφέρει τελικά άμυνα ενάντια στις επιθέσεις γνώσης σε πίνακες με συχνές ενημερώσεις. Η συνθήκη 1 του παραπάνω ορισμού είναι αρκετά ξεκάθαρη. Πρέπει ο δημοσιευμένος πίνακας T καθώς και κάθε πίνακας που προκύπτει από ενημέρωση αυτού (προσθήκες ή/και διαγραφές εγγραφών) να είναι m -μοναδικός. Η συνθήκη 2 χρειάζεται λίγη περισσότερη ανάλυση. Η κύρια πληροφορία που μας δίνεται από την συνθήκη 2 είναι ότι αν μία εγγραφή j δημοσιευτεί $(y - x + 1)$ φορές, κάθε κλάση ισοδυναμίας που την περιέχει θα πρέπει να έχει την ίδια υπογραφή. Αυτό προφανώς, είναι αδύνατον αν σκεφτούμε το φαινόμενο της κρίσιμης απουσίας που είδαμε παραπάνω, κατά το οποίο μία τιμή στην κλάση ισοδυναμίας χάνεται τελείως λόγω της διαγραφής μιας εγγραφής. Πρέπει, λοιπόν, να βρούμε ένα τρόπο “αναπαράστασης” των εγγραφών των οποίων η απουσία των τιμών από το σύνολο των ευαίσθητων γνωρισμάτων προκαλούν κρίσιμη απουσία. Αυτό γίνεται, με την προσθήκη στον δημοσιευμένο πίνακα *πλαστών εγγραφών* (*counterfeit tuples*) οι οποίες θα αναπληρώνουν τις τιμές των ευαίσθητων γνωρισμάτων που διαφορετικά ίσως προκαλούσαν διαρροή πληροφορίας. Ας επιστρέψουμε όμως στο αρχικό μας παράδειγμα για να γίνει πιο ξεκάθαρο το πως λειτουργούν οι πλαστές εγγραφές.

Όνομα	Κλάση Ισοδυναμίας	Ηλικία	Ταχυδρομικός Κώδικας	Ασθένεια
Bob	1	[21,22]	[12000,14000]	Δυσπεψία
c_1		[21,22]	[12000,14000]	Βρογχίτιδα
David	2	[23,25]	[21000,25000]	Γαστρίτιδα
Emily		[23,25]	[21000,25000]	Γρίπη
Jane	3	[37,43]	[26000,33000]	Δυσπεψία
c_2		[37,43]	[26000,33000]	Γρίπη
Linda		[37,43]	[26000,33000]	Γαστρίτιδα
Gary	4	[41,46]	[20000,30000]	Γρίπη
Mary		[41,46]	[20000,30000]	Γαστρίτιδα
Ray	5	[54,56]	[31000,34000]	Δυσπεψία
Steve		[54,56]	[31000,34000]	Γαστρίτιδα
Tom	6	[60,65]	[36000,44000]	Γαστρίτιδα
Vince		[60,65]	[36000,44000]	Γρίπη

Πίνακας 6α

Κλάση Ισοδυναμίας	#Πλαστών Εγγραφών
1	1
3	1

Πίνακας 6β

Ας υποθέσουμε πως θέλουμε να δημοσιεύουμε τα περιεχόμενα του *Πίνακα 4α* καθώς αυτά ανανεώνονται, ικανοποιώντας ταυτόχρονα την αρχή της m -αμεταβλητότητας. Το πρώτο βήμα που κάνουμε είναι να βεβαιωθούμε πως ο πίνακας ο οποίος δημοσιεύσαμε αρχικά είναι m -μοναδικός. Όντως, όπως μπορούμε εύκολα να δούμε από τον *Πίνακα 4β*, ο πρώτος πίνακας που δημοσιεύουμε είναι 2-μοναδικός μιας και όλες οι κλάσεις ισοδυναμίας του περιέχουν τουλάχιστον 2 εγγραφές και κάθε μία από αυτές έχουν διαφορετική τιμή στο ευαίσθητο γνώρισμα, το οποίο στην περίπτωσή μας είναι η Ασθένεια. Στην συνέχεια, σύμφωνα με τον ορισμό που δώσαμε παραπάνω, θα πρέπει να δημοσιεύσουμε έναν πίνακα που θα είναι και αυτός m -μοναδικός και κάθε εγγραφή που παραμένει σε αυτόν από την προηγούμενη δημοσίευση θα πρέπει να ανήκει σε μία κλάση ισοδυναμίας με τις ίδιες τιμές ευαίσθητων γνωρισμάτων. Δημοσιεύουμε, λοιπόν, τον *Πίνακα 6α*. Παρατηρούμε, πως για να καταφέρουμε να ικανοποιήσουμε την 2^η συνθήκη του ορισμού της m -αμεταβλητότητας έχουμε προσθέσει δύο νέες πλαστές εγγραφές, τις c_1 και c_2 . Επιτυγχάνουμε, λοιπόν, ο *Πίνακας 6α* να είναι 2-αμετάβλητος. Ας δούμε, κατ' αντιστοιχία με το προηγούμενο σενάριο, τι πληροφορία θα αποσπάσει αυτήν τη φορά ο επιτιθέμενος. Από την πρώτη δημοσίευση (*Πίνακας 4α*) αποσπά την ίδια πληροφορία με προηγουμένως (υποθέτοντας ότι έχει την ίδια πρότερη γνώση), ότι δηλαδή ο Bob έχει είτε Δυσπεψία είτε Βρογχίτιδα. Περιμένει, λοιπόν, την 2^η δημοσίευση για περισσότερες πληροφορίες. Μόλις γίνεται διαθέσιμος, ο *Πίνακας 6α* συνειδητοποιεί πως πάλι οι πιθανές ασθένειες που μπορεί να έχει ο Bob είναι Δυσπεψία ή Βρογχίτιδα. Αν η δημοσίευση των πινάκων συνεχίσει να ικανοποιεί την m -αμεταβλητότητα αυτό θα συμβαίνει σε κάθε πίνακα ο οποίος θα περιέχει την εγγραφή του Bob, καθιστώντας αδύνατη την πιθανή ταυτοποίηση του χωρίς επιπλέον απόκτηση γνώσης από τον επιτιθέμενο!

Κάποιος, όμως, μπορεί να παρατηρήσει πως προσθέτοντας τις πλαστές εγγραφές στον δημοσιευμένο πίνακα έχουμε αποκλίνει αρκετά από τον αρχικό στόχο της δημοσίευσης των πληροφοριών, ο οποίος ήταν να παρέχουμε δημοσίως στοιχεία για έρευνα και στατιστικές μελέτες. Με την προσθήκη των πλαστών εγγραφών έχουμε αλλοιώσει σημαντικά την πραγματική πληροφορία και κατά προέκταση κάθε έρευνα που θα βασιστεί σε αυτά τα στοιχεία δεν θα μπορεί να θεωρηθεί αξιόπιστη. Για να περιορίσουμε την αλλοίωση την πληροφορίας, μαζί με κάθε πίνακα που περιέχει

πλαστές εγγραφές, δημοσιεύουμε άλλον έναν πίνακα ο οποίος αντιστοιχεί κάθε κλάση ισοδυναμίας με τον αριθμό των πλαστών εγγραφών που περιέχονται σε αυτήν. Στο παράδειγμά μας, αυτός ο πίνακας είναι ο *Πίνακας 6β*.

Κλείνοντας αυτήν την κατηγορία επιθέσεων, πρέπει να σημειώσουμε πως παρόλη την ισχύ που μας προσφέρει η m -αμεταβλητότητα, τόσο η απώλεια πληροφορίας που έχουμε από τις συνεχείς γενικεύσεις των τιμών των γνωρισμάτων των ψευδο-αναγνωριστικών όσο και πιθανώς η σταδιακή συσσώρευση των πλαστών εγγραφών καθώς αυξάνεται ο αριθμός των δημοσιεύσεων μπορεί να οδηγήσει σε πληροφορία η οποία είναι αδύνατον να χρησιμοποιηθεί για οποιουδήποτε είδους μελέτης ή έρευνας. Μία πιθανή λύση, θα ήταν μετά από κάποιο χρονικό διάστημα ο εκάστοτε οργανισμός/πρόσωπο που θέλει να δημοσιεύσει ανώνυμα δεδομένα που υπόκεινται σε ενημερώσεις, να αποσύρει τα παλαιότερα δημοσιευμένα δεδομένα και να δημοσιοποιεί διαφορετικό (εννοώντας ξένο προς τα προηγούμενα) σύνολο.

3.1.6 Επίθεση γνώσης ενάντια πληροφορία οργανωμένη σε σύνολα (background attack against set-valued data)

Μέχρι τώρα, κατά την μελέτη της ιδιωτικότητας έχουμε υποθέσει πως τα δεδομένα τα οποία δημοσιεύονται ακολουθούν το σχεσιακό μοντέλο οργάνωσης. Φυσικά, μία τέτοια υπόθεση, αποκλείει ένα μεγάλο μέρος πληροφορίας το οποίο δεν είναι οργανωμένο με αυτόν τον τρόπο, από την προστασία τεχνικών όπως η k -ανωνυμία. Για αυτόν τον σκοπό και πιο ειδικά για πληροφορία που έχει οργανωθεί σε σύνολα με βάση κάποιο κριτήριο έχουν προταθεί τεχνικές προέκτασης των γνωστών τεχνικών για την προστασία της ανωνυμίας. Πριν αναφερθούμε σε αυτές όμως, ας παρουσιάσουμε ένα παράδειγμα παραβίασης της ιδιωτικότητας, σε τέτοιου είδους δεδομένα.

Ας υποθέσουμε, ότι μία μεγάλη αλυσίδα καταστημάτων αποφασίζει να δώσει στην δημοσιότητα πληροφορίες σχετικές με το τι προϊόντα αγόρασαν οι καταναλωτές την τελευταία εβδομάδα. Ένα παράδειγμα τέτοιας δημοσίευσης παρουσιάζεται στον *Πίνακα 7*. Ας υποθέσουμε τώρα πως ο γείτονας του Bob, πήγε την ίδια μέρα με

αυτόν, στο ίδιο πολυκατάστημα για να κάνει τα ψώνια του. Κατά τη διάρκεια των αγορών του, συναντάει τον Bob και βλέπει ότι στο καλάθι του περιέχονται μεταλλικό νερό και οδοντόκρεμα. Όταν δημοσιεύονται οι πληροφορίες, ο γείτονας του Bob μπορεί χωρίς μεγάλη δυσκολία να αναγνωρίσει τον Bob στην δημοσιευμένη λίστα καθώς και να αποκαλύψει τα υπόλοιπα προϊόντα που αυτός αγόρασε.

Bob	{ Μεταλλικό νερό, γάλα, οδοντόκρεμα }
Alice	{ Αεριούχο νερό, γάλα }
Gary	{ Αεριούχο νερό, οδοντόκρεμα, γάλα }
Andy	{ Μεταλλικό νερό, Αεριούχο νερό, οδοντόκρεμα }

Πίνακας 7

Για αυτό τον σκοπό αναπτύχθηκε η τεχνική της k^m -ανωνυμίας (k^m -anonymity) [4] η οποία προεκτείνει την αρχή της k -ανωνυμίας.

Ορισμός (k^m -ανωνυμία):

Έστω ότι ένας επιτιθέμενος έχει γνώση m το πολύ αντικειμένων, σε δεδομένα οργανωμένα σε σύνολα. Αν για κάθε σύνολο m ή λιγότερων αντικειμένων, υπάρχουν τουλάχιστον k δοσοληψίες που να περιέχουν αυτό το σύνολο, θα λέμε πως τα δεδομένα ικανοποιούν την αρχή της k^m -ανωνυμίας.

Για να καταφέρουμε να επιτύχουμε k^m -ανωνυμία θα πρέπει να χρησιμοποιήσουμε ιεραρχίες γενικεύσεις στις οποίες έχουμε αναφερθεί παραπάνω. Ας επιστρέψουμε όμως στο παράδειγμά μας, υποθέτοντας πως έχουμε μία ιεραρχία γενίκευσης {μεταλλικό νερό, αεριούχο νερό} \rightarrow {εμφιαλωμένο νερό}. Εφαρμόζουμε αυτήν την γενίκευση στον Πίνακα 7 και παραθέτουμε το αποτέλεσμα στον Πίνακα 8. Ας μελετήσουμε τώρα ξανά την περίπτωση που ο επιτιθέμενος γνωρίζει πως ο Bob έχει αγοράσει μεταλλικό νερό και οδοντόκρεμα. Κοιτάζοντας τον Πίνακα 8, εντοπίζει τρεις δοσοληψίες που περιέχουν αυτά τα προϊόντα (του Bob και του Gary). Τα δεδομένα μας σύμφωνα με τον ορισμό της k^m -ανωνυμίας είναι 3^2 – ανώνυμα και ο

επιτιθέμενος δεν μπορεί να προσδιορίσει ποια από τις τρεις δοσοληψίες ανήκει στον Bob.

Bob	{ Εμφιαλωμένο νερό, γάλα, οδοντόκρεμα }
Alice	{ Εμφιαλωμένο νερό, γάλα }
Gary	{ Εμφιαλωμένο νερό, γάλα, οδοντόκρεμα }
Andy	{ Εμφιαλωμένο νερό, οδοντόκρεμα }

Πίνακας 8

3.2 Εργαλείο ανωνυμοποίησης του πανεπιστημίου του Texas

Στα πλαίσια του ερευνητικού έργου του πανεπιστημίου του Texas στο Dallas (UTD), έχουν υλοποιηθεί από το τμήμα Ασφάλειας Δεδομένων και Ιδιωτικότητας, μία σειρά αλγορίθμων ανωνυμοποίησης για την κοινή χρήση από ερευνητές. Οι αλγόριθμοι αυτοί περιλαμβάνουν τους: Datafly, Mondrian Multidimensional k-Anonymity, Incognito, Incognito with l-diversity, Incognito with t-closeness και Anatomy. Σε αυτό το σημείο, αξίζει να σημειωθεί πως οι αλγόριθμοι μπορούν να χρησιμοποιηθούν είτε απ'ευθείας πάνω σε ένα σύνολο δεδομένων είτε σαν συναρτήσεις βιβλιοθήκης στα πλαίσια μίας εκτενέστερης εφαρμογής. Το εγχείρημα είναι ανοιχτού λογισμικού και έχει υλοποιηθεί στην γλώσσα προγραμματισμού Java [12].

4

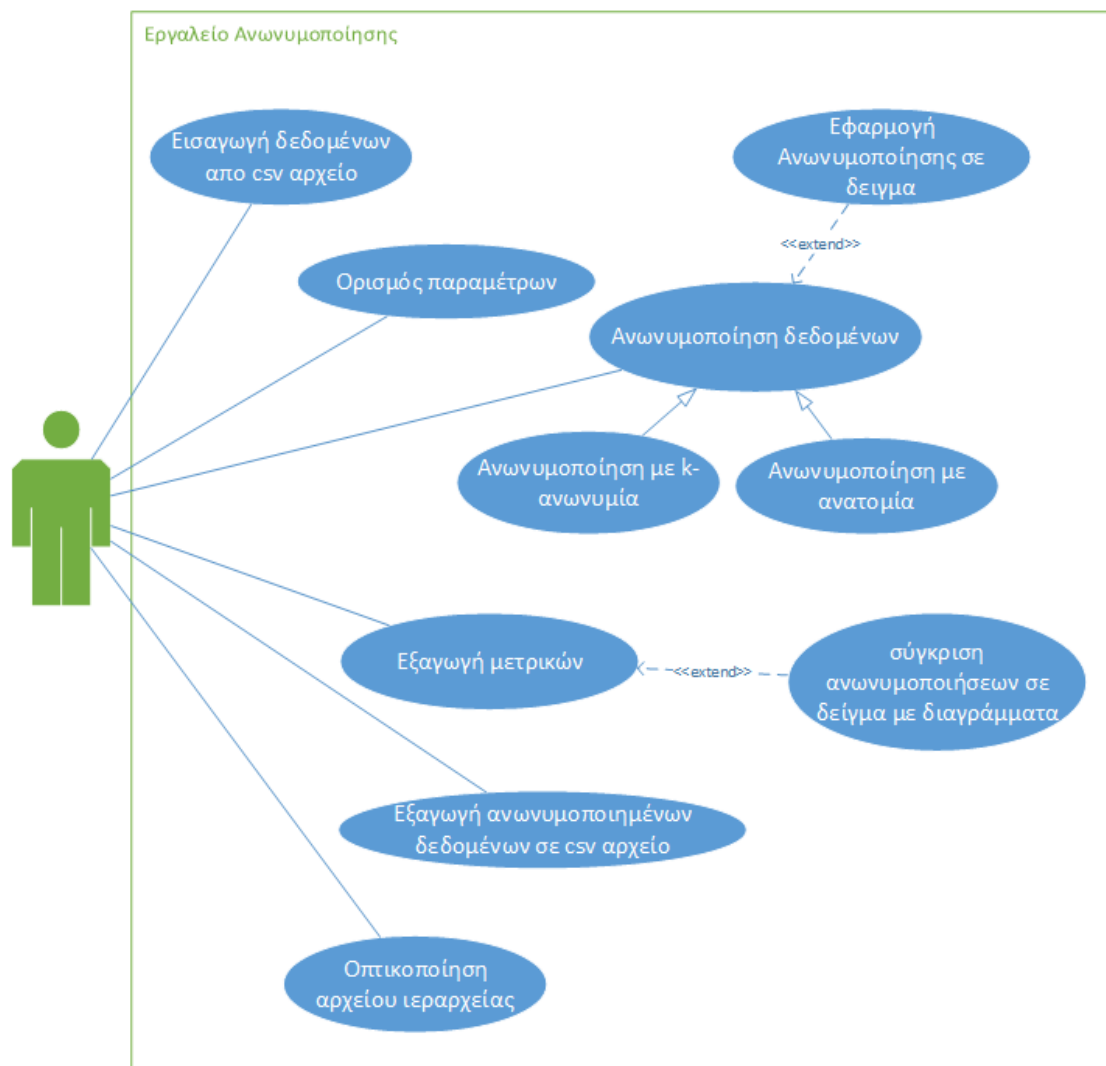
Ανάλυση Συστήματος

4.1 Αρχιτεκτονική

Η εφαρμογή που υλοποιήθηκε αποτελείται από τέσσερα υποσυστήματα:

- Εισαγωγή δεδομένων/παραμέτρων και διεπαφή με τον χρήστη
- Ανωνυμοποίηση
- Στατιστικά και Διαγράμματα
- Διαχείριση ιεραρχιών γενίκευσης.

Στο *διάγραμμα 1* που ακολουθεί, περιλαμβάνονται οι βασικές λειτουργίες που μπορεί να εκτελέσει χρήστης, καθώς και μία συνοπτική παρουσίαση των υποσυστημάτων όπως αυτά αναφέρθηκαν παραπάνω.

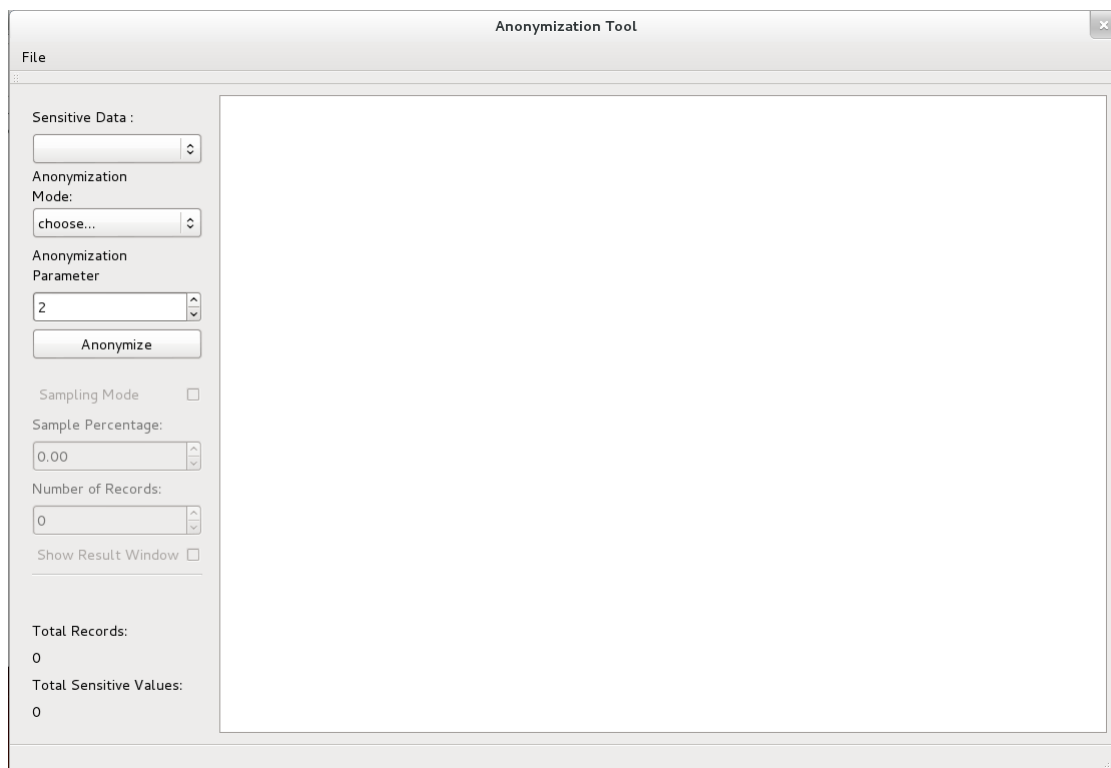


Διάγραμμα 1: Διάγραμμα περιπτώσεων χρήσης

4.2 Περιγραφή Λειτουργιών

4.2.1 Εισαγωγή δεδομένων/παραμέτρων και διεπαφή με τον χρήστη

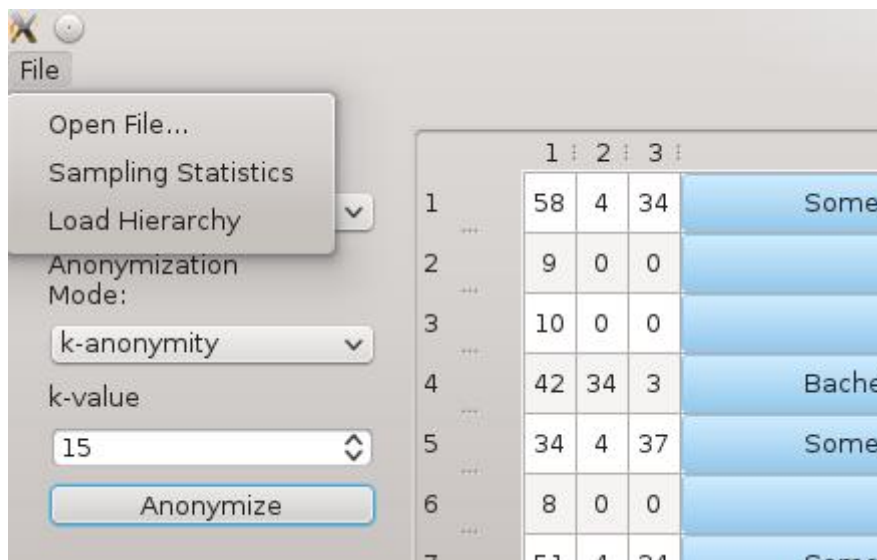
Η διεπαφή με τον χρήστη έχει ως κεντρικό σημείο αναφοράς την αρχική οθόνη, η οποία αποτελεί την αφετηρία για τις λειτουργίες που προσφέρει το αναπτυχθέν εργαλείο (Εικόνα 4.1).



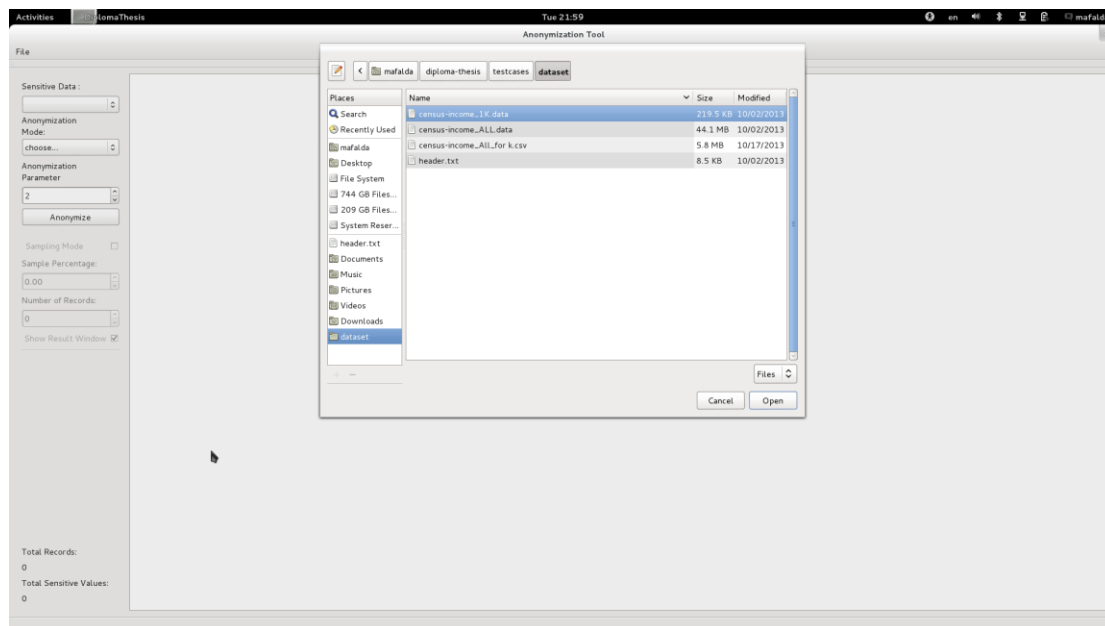
Εικόνα 4.1 - Αρχική Οθόνη

Το πρώτο μέρος της εφαρμογής που πρέπει να μελετηθεί, είναι η διαδικασία εισαγωγής των δεδομένων και παραμετροποίησης του αλγορίθμου ανωνυμοποίησης ώστε να ικανοποιεί τις ανάγκες του χρήστη. Για αυτόν τον σκοπό, έχει σχεδιαστεί ένα πλήρες, εύχρηστο γραφικό περιβάλλον το οποίο παρέχει τις εξής λειτουργίες:

- Εύκολη πλοήγηση για την εύρεση των αρχείων δεδομένων προς ανωνυμοποίηση. Το επιλεγμένο αρχείο πρέπει να είναι ένα αρχείο κειμένου με τιμές οριοθετημένες με κόμματα και μπορεί προαιρετικά να περιλαμβάνει στην πρώτη γραμμή τον τίτλο κάθε πεδίου. Για τον σκοπό αυτό ο χρήστης έχει την δυνατότητα από το στοιχείο «File» που βρίσκεται στην γραμμή καταλόγου της αρχικής οθόνης να επιλέξει την ενέργεια «Open File..» (Εικόνα 4.2). Με την επιλογή αυτή εμφανίζεται ο προεπιλεγμένος για το σύστημα στο οποίο τρέχει η εφαρμογή file manager και αναμένεται από τον χρήστη να γίνει μέσω αυτού η επιλογή του επιθυμητού αρχείου (Εικόνα 4.3). Αν δεν επιλεγθεί αρχείο ή αν η μορφή του δεν ήταν η αναμενόμενη (για παράδειγμα αν μια γραμμή έχει παραπάνω αριθμό πεδίων από τις υπόλοιπες) εμφανίζεται το κατάλληλο ενημερωτικό μήνυμα προς τον χρήστη

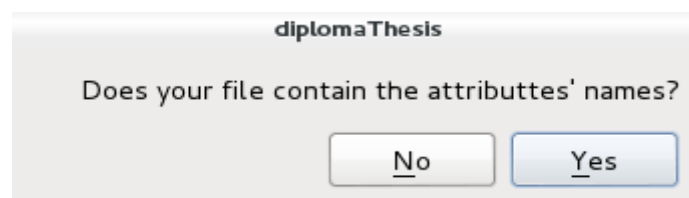


Εικόνα 4.2 – Load Hierarchy



*Εικόνα 4.3 –Επιλογή αρχείου δεδομένων προς ανωνυμοποίηση
(με Nautilus έκδοση 3.4.2)*

- Μετά την επιλογή του αρχείου, ο χρήστης καλείται να ορίσει αν το αρχείο προς ανωνυμοποίηση περιλαμβάνει τους τίτλους των δεδομένων πεδίων και έπειτα τα δεδομένα παρουσιάζονται στον χρήστη σε μορφή πίνακα στον κεντρικό χώρο της αρχικής οθόνης ενώ ταυτόχρονα, ενεργοποιούνται οι δυνατότητες επιλογής ευαίσθητων δεδομένων και αλγορίθμων ανωνυμοποίησης (Εικόνες 4.4 - 4.5). Αν οι τίτλοι των δεδομένων δεν παρέχονται από το επιλεγμένο αρχείο τότε τα ονόματα των στηλών αντικαθίστανται από τον αριθμό της θέσης της κάθε στήλης (Εικόνα 4.6). Στο κάτω αριστερά μέρος τις οθόνης εμφανίζεται ο αριθμός των εγγραφών που φορτώθηκαν.



Εικόνα 4.4 – Καθορισμός πρώτης γραμμής αρχείου ως γραμμή τίτλων

	Id	SSN	Age	Problem
1	1	12345	36	Flu
2	2	54321	12	Short Breath
3	3	21345	23	Allergy
4	4	45321	78	Cancer
5	5	32145	45	Stomachache
6	6	43451	55	Headache
7	7	14532	34	Sore Throat
8	8	98675	28	Broken Ribs
9	9	46783	15	Eye Problem
10	10	42235	80	Flu

Εικόνα 4.5 – Τίτλοι δεδομένοι στην πρώτη γραμμή αρχείου

	1	2	3	4	5	6	7	8
1	58	Self-employed-not_incorporated	4	34	Some_college_but_no_degree	0	Not_in_universe	Divorced
2	9	Not_in_universe	0	0	Children	0	Not_in_universe	Never_married
3	10	Not_in_universe	0	0	Children	0	Not_in_universe	Never_married
4	42	Private	34	3	Bachelors_degree(BA_AB_BS)	0	Not_in_universe	Married-civilian_spouse_present
5	34	Private	4	37	Some_college_but_no_degree	0	Not_in_universe	Married-civilian_spouse_present
6	8	Not_in_universe	0	0	Children	0	Not_in_universe	Never_married
7	51	Private	4	34	Some_college_but_no_degree	0	Not_in_universe	Married-civilian_spouse_present
8	46	Private	37	31	High_school_graduate	0	Not_in_universe	Divorced
9	13	Not_in_universe	0	0	Children	0	Not_in_universe	Never_married
10	39	Not_in_universe	0	0	10th_grade	0	Not_in_universe	Married-civilian_spouse_present
11	12	Not_in_universe	0	0	Children	0	Not_in_universe	Never_married
12	27	Self-employed-not_incorporated	4	34	Some_college_but_no_degree	0	Not_in_universe	Married-civilian_spouse_present

Εικόνα 4.6 – Αριθμημένες στήλες αντί για τίτλους πεδίων

- Στα αριστερά του πίνακα των δεδομένων, έχει υλοποιηθεί ένα toolbox το οποίο περιέχει όλες τις δυνατότητες επιλογής και παραμετροποίησης των τεχνικών ανωνυμοποίησης. Όπως φαίνεται στην Εικόνα 4.7 ο χρήστης μπορεί να επιλέξει ανάμεσα στους αλγορίθμους που θέλει και να θέσει τις τιμές των k ή l παραμέτρων.

The screenshot shows the 'Anonymization Tool' window. On the left is a sidebar with controls: 'Sensitive Data' (set to 25), 'Anonymization Mode' (set to Anatomy), 'l-value' (set to 20), 'Anonymize' button, 'Sampling Mode' (checkbox), 'Sample Percentage' (set to 0.00), 'Number of Records' (set to 0), and 'Show Result Window' (checkbox). The main area is a table with 29 rows and 25 columns. The columns are numbered 13 to 25. The table contains various demographic and status data for each record.

	13	14	15	16	17	18	19	20	21	22	23	24	25	
1	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Head_of_household	South	Arkansas	Householder	Householder	1053.55	MSA
2	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Nonfiler	Not_in_universe	Not_in_universe	Child.<18_never_marr_not_in_subfamily	Child_under_18_never_married	1758.14	Non
3	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Nonfiler	Not_in_universe	Not_in_universe	Child.<18_never_marr_not_in_subfamily	Child_under_18_never_married	1069.10	Non
4	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	5178	0	0	Joint_both_under_65	Not_in_universe	Not_in_universe	Householder	Householder	1535.86	Non
5	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_under_65	Not_in_universe	Not_in_universe	Householder	Householder	1146.77	Non
6	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Nonfiler	Not_in_universe	Not_in_universe	Child.<18_never_marr_not_in_subfamily	Child_under_18_never_married	2066.24	Non
7	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_under_65	Not_in_universe	Not_in_universe	Householder	Householder	2441.22	Non
8	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	1590	0	Single	Not_in_universe	Not_in_universe	Householder	Householder	978.16	Non
9	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Nonfiler	Not_in_universe	Not_in_universe	Child.<18_never_marr_not_in_subfamily	Child_under_18_never_married	1520.08	Non
10	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_under_65	Not_in_universe	Not_in_universe	Spouse_of_householder	Spouse_of_householder	1274.04	Non
11	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Nonfiler	South	Utah	Child.<18_never_marr_not_in_subfamily	Child_under_18_never_married	155.02	MSA
12	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_under_65	Not_in_universe	Not_in_universe	Householder	Householder	1004.69	Non
13	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	1977	100	Joint_both_under_65	Not_in_universe	Not_in_universe	Householder	Householder	899.48	Non
14	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_under_65	Not_in_universe	Not_in_universe	Householder	Householder	1483.69	Non
15	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Nonfiler	Not_in_universe	Not_in_universe	Child.<18_never_marr_not_in_subfamily	Child_under_18_never_married	1660.53	Non
16	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	1669	700	Single	Not_in_universe	Not_in_universe	Secondary_individual	Nonrelative_of_householder	1311.15	Non
17	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_under_65	Not_in_universe	Not_in_universe	Householder	Householder	1629.02	Non
18	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Single	Not_in_universe	Not_in_universe	Child.18+_never_marr_Not_in_a_subfamily	Child.18_or_older	1398.03	Non
19	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Single	Not_in_universe	Not_in_universe	Householder	Householder	2492.74	Non
20	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	1000	Joint_both_under_65	Not_in_universe	Not_in_universe	Householder	Householder	780.1	Non
21	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Nonfiler	Not_in_universe	Not_in_universe	Child.<18_never_marr_not_in_subfamily	Child_under_18_never_married	498.25	Non
22	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Nonfiler	Midwest	Minnesota	Householder	Householder	1773.08	MSA
23	Male	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Nonfiler	Not_in_universe	Not_in_universe	Child.<18_never_marr_not_in_subfamily	Child_under_18_never_married	485.84	Non
24	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_65+	Not_in_universe	Not_in_universe	Spouse_of_householder	Spouse_of_householder	4183.26	Non
25	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_under_65	Not_in_universe	Not_in_universe	Spouse_of_householder	Spouse_of_householder	1040.15	Non
26	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_under_65	Not_in_universe	Not_in_universe	Spouse_of_householder	Spouse_of_householder	1407.39	Non
27	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Nonfiler	West	Alaska	Child.<18_never_marr_not_in_subfamily	Child_under_18_never_married	235.92	MSA
28	Female	Not_in_universe	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_under_65	Not_in_universe	Not_in_universe	Spouse_of_householder	Spouse_of_householder	1869.47	Non
29	Male	No	Not_in_universe	Children_or_Armed_Forces	0	0	0	Joint_both_under_65	Midwest	Kansas	Householder	Householder	1812	NonMSA

Εικόνα 4.7- Απεικόνιση συνόλου δεδομένων και επιλογές παραμέτρων

- Ταυτόχρονα, στο ίδιο toolbox δίνεται η επιλογή για sampling mode, με στόχο την εξαγωγή στατιστικών πάνω σε ένα μέρος των δεδομένων, επιτρέποντας την γρήγορη εκτέλεση και σύγκριση της τεχνικής και των παραμέτρων. Με τον τρόπο αυτό διευκολύνεται και επιταχύνεται η διαδικασία επιλογής του βέλτιστου συνδυασμού μεθόδου-τιμής παραμέτρου που ταιριάζει καλύτερα στο κάθε σύνολο δεδομένων. Περισσότερες λεπτομέρειες για αυτή την επιλογή, θα δοθούν στην υποενότητα 4.2.3 Στατιστικά και διαγράμματα.

- Αφού έχουν φορτωθεί τα προς προστασία δεδομένα, ο χρήστης μπορεί να επιλέξει την στήλη του ευαίσθητου γνωρίσματος είτε από το combobox το οποίο βρίσκεται στο toolbox είτε επιλέγοντας την στήλη του πίνακα χρησιμοποιώντας το αριστερό κλικ του ποντικιού. (Εικόνα 4.7). Σε αυτό το σημείο, στο κάτω αριστερό μέρος του παραθύρου εμφανίζονται οι διακριτές τιμές του ευαίσθητου γνωρίσματος.

4.2.2 Ανωνυμοποίηση

Το δεύτερο υποσύστημα που θα μελετηθεί, είναι αυτό της ανωνυμοποίησης και αποτελεί και τον πυρήνα της παρούσας διπλωματικής εργασίας. Σε αυτό, στα δεδομένα που έχουν εισαχθεί από τον χρήστη, εφαρμόζονται οι αλγόριθμοι ιδιωτικότητας και παράγονται τα ανωνυμοποιημένα δεδομένα. Στα παράθυρα παρουσίασης των ανωνυμοποιημένων δεδομένων, δίνεται η επιλογή για εξαγωγή των ανωνυμοποιημένων πλέον δεδομένων σε αρχεία .csv, για την εύκολη εισαγωγή σε οποιοδήποτε σύστημα διαχείρισης βάσεων δεδομένων και την περαιτέρω επεξεργασία τους. Ακόμη δίνεται η δυνατότητα εξαγωγής στατιστικών για την διαδικασία που εκτελέστηκε. Πιο αναλυτικά, το κομμάτι της Ανωνυμοποίησης χωρίζεται σε δυο μέρη με βάση τους αλγόριθμους που έχουν υλοποιηθεί: αυτό της k -ανωνυμίας και αυτό της ανατομίας. Στην συνέχεια, θα αναλυθούν με λεπτομέρεια αυτά τα δύο κομμάτια.

4.2.2.1 k -ανωνυμία

Αφού έχει επιλεγθεί το ευαίσθητο γνώρισμα, ο χρήστης καλείται να εισάγει την παράμετρο k . Η τιμή του k , καθορίζει το μέγεθος των κλάσεων ισοδυναμίας και εν τέλει την αναλογία μεταξύ της απώλειας πληροφορίας και ισχύος της ανωνυμίας των ανωνυμοποιημένων δεδομένων. Αφού ο χρήστης δώσει την εντολή για ανωνυμοποίηση γίνεται έλεγχος της τιμής k . Αν αυτή είναι μεγαλύτερη από τον αριθμό των εγγραφών του συνόλου δεδομένων προς ανωνυμοποίηση, τότε η τιμή αυτή του k είναι μη αποδεκτή καθότι δεν μπορεί να δημιουργηθεί ομάδα εγγραφών με μέγεθος τουλάχιστον ίσο με το ζητούμενο. Στην περίπτωση αυτή, ενημερώνεται ο χρήστης με το απαραίτητο μήνυμα, διαφορετικά ξεκινάει η εκτέλεση του modrian πολυδιάστατου αλγορίθμου.

QI Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	[2-12]	0	Children	0	0	0	0	[199.63 - 210.67]	0	0	0	0	0	0	94
2	[2-12]	0	Children	0	0	0	0	[199.63 - 210.67]	0	0	0	0	0	0	94
3	[2-12]	0	Children	0	0	0	0	[199.63 - 210.67]	0	0	0	0	0	0	94
4	[2-12]	0	Children	0	0	0	0	[199.63 - 210.67]	0	0	0	0	0	0	94
5	[2-12]	0	Children	0	0	0	0	[199.63 - 210.67]	0	0	0	0	0	0	94
6	[2-12]	0	Children	0	0	0	0	[199.63 - 210.67]	0	0	0	0	0	0	94
7	[2-12]	0	Children	0	0	0	0	[199.63 - 210.67]	0	0	0	0	0	0	94
8	[2-12]	0	Children	0	0	0	0	[199.63 - 210.67]	0	0	0	0	0	0	94
9	[0-13]	0	Children	0	0	0	0	[211.33 - 229.01]	0	0	0	0	0	0	94
10	[0-13]	0	Children	0	0	0	0	[211.33 - 229.01]	0	0	0	0	0	0	94
11	[0-13]	0	Children	0	0	0	0	[211.33 - 229.01]	0	0	0	0	0	0	94
12	[0-13]	0	Children	0	0	0	0	[211.33 - 229.01]	0	0	0	0	0	0	94
13	[0-13]	0	Children	0	0	0	0	[211.33 - 229.01]	0	0	0	0	0	0	94
14	[0-13]	0	Children	0	0	0	0	[211.33 - 229.01]	0	0	0	0	0	0	94
15	[0-13]	0	Children	0	0	0	0	[211.33 - 229.01]	0	0	0	0	0	0	94
16	[0-13]	0	Children	0	0	0	0	[211.33 - 229.01]	0	0	0	0	0	0	94
17	[2-14]	0	Children	0	0	0	0	[236.52 - 242.91]	0	0	0	0	0	0	94
18	[2-14]	0	Children	0	0	0	0	[236.52 - 242.91]	0	0	0	0	0	0	94
19	[2-14]	0	Children	0	0	0	0	[236.52 - 242.91]	0	0	0	0	0	0	94
20	[2-14]	0	Children	0	0	0	0	[236.52 - 242.91]	0	0	0	0	0	0	94
21	[2-14]	0	Children	0	0	0	0	[236.52 - 242.91]	0	0	0	0	0	0	94
22	[2-14]	0	Children	0	0	0	0	[236.52 - 242.91]	0	0	0	0	0	0	94
23	[2-14]	0	Children	0	0	0	0	[236.52 - 242.91]	0	0	0	0	0	0	94
24	[2-14]	0	Children	0	0	0	0	[236.52 - 242.91]	0	0	0	0	0	0	94
25	[1-11]	0	Children	0	0	0	0	[245.17 - 254.64]	0	0	0	0	0	0	94
26	[1-11]	0	Children	0	0	0	0	[245.17 - 254.64]	0	0	0	0	0	0	94
27	[1-11]	0	Children	0	0	0	0	[245.17 - 254.64]	0	0	0	0	0	0	94
28	[1-11]	0	Children	0	0	0	0	[245.17 - 254.64]	0	0	0	0	0	0	94
29	[1-11]	0	Children	0	0	0	0	[245.17 - 254.64]	0	0	0	0	0	0	94
30	[1-11]	0	Children	0	0	0	0	[245.17 - 254.64]	0	0	0	0	0	0	94
31	[1-11]	0	Children	0	0	0	0	[245.17 - 254.64]	0	0	0	0	0	0	94

Εικόνα 4.8- Αποτέλεσμα k-ανωνυμίας

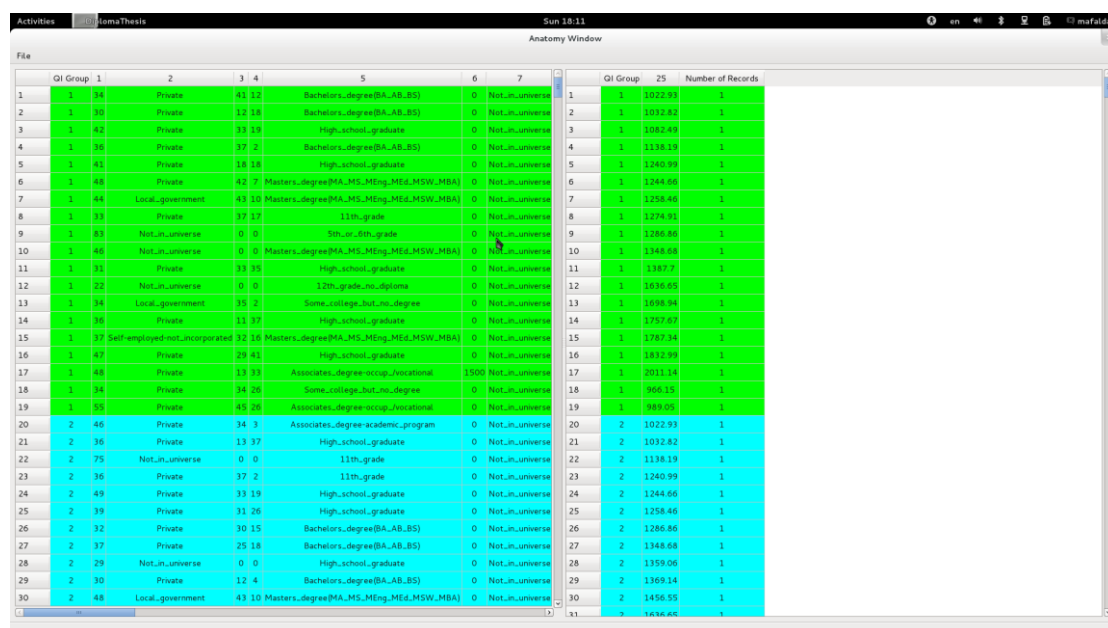
Όταν ολοκληρωθεί ο αλγόριθμος, το πρόγραμμα εμφανίζει το παράθυρο k-anonymity Window, το οποίο παρουσιάζει τα δεδομένα μετά την ανωνυμοποίηση (Εικόνα 4.8).

Οι πληροφορίες εμφανίζονται ως εξής:

- Τα δεδομένα μπαίνουν σε πίνακα ο οποίος περιέχει τον ίδιο αριθμό γραμμών με τον πίνακα εισόδου και ένα επιπλέον πεδίο.
- Το επιπλέον πεδίο έχει ονομασία QI Group και αποτελεί αναγνωριστικό της κλάσης ισοδυναμίας στην οποία ανήκουν τα δεδομένα.
- Για την καλύτερη εποπτεία των κλάσεων, διαδοχικές κλάσεις έχουν χρωματιστεί με διαφορετικά χρώματα.
- Οι τιμές των ψευδο-αναγνωριστικών, γενικεύονται σε εύρη τιμών και παρουσιάζονται στην μορφή [ελάχιστη τιμή – μέγιστη τιμή] για κάθε πεδίο της κλάσης ισοδυναμίας.

4.2.2.2 Ανατομία

Σε αντιστοιχία με τον αλγόριθμο της k-ανωνυμίας έτσι και ο αλγόριθμος της ανατομίας, απαιτεί αρχικά την επιλογή του ευαίσθητου γνωρίσματος και στην συνέχεια τον ορισμό της παραμέτρου ποικιλομορφίας l . Η τιμή του l , καθορίζει τον αριθμό των διαφορετικών τιμών του ευαίσθητου γνωρίσματος, όπως αυτό εμφανίζεται στις εγγραφές μίας κλάσης ισοδυναμίας. Αφού ο χρήστης δώσει την εντολή για ανωνυμοποίηση, ελέγχεται αν είναι αποδεκτή η δοθείσα τιμή της παραμέτρου l και ξεκινάει η εκτέλεση του αλγορίθμου. Σημειώνεται ότι η τιμή l θεωρείται αποδεκτή αν είναι τουλάχιστον ίση με το πλήθος των διαφορετικών ευαίσθητων τιμών και στην περίπτωση που δεν ισχύει αυτή η προϋπόθεση ο χρήστης ενημερώνεται για το κατάλληλο μήνυμα και δεν ξεκινά η εκτέλεση του αλγορίθμου της ανατομίας.



The screenshot shows a window titled 'Anatomy Window' with a table of data. The table has columns for 'QI Group' and 'Number of Records'. The data is organized into groups, with the first group containing 30 rows and the second group containing 11 rows. The rows are color-coded in a repeating pattern of green, yellow, and blue. The table content is as follows:

QI Group	1	2	3	4	5	6	7	QI Group	25	Number of Records
1	1	34	Private	61.12	Bachelors.degree(BA_AB_BS)	0	Not_in_universe	1	1022.93	1
2	1	30	Private	12.18	Bachelors.degree(BA_AB_BS)	0	Not_in_universe	2	1032.82	1
3	1	62	Private	33.19	High_school_graduate	0	Not_in_universe	3	1082.49	1
4	1	36	Private	37.2	Bachelors.degree(BA_AB_BS)	0	Not_in_universe	4	1138.19	1
5	1	41	Private	18.18	High_school_graduate	0	Not_in_universe	5	1240.99	1
6	1	48	Private	42.7	Masters.degree(MA_MS_MEng_MED_MSW_MBA)	0	Not_in_universe	6	1244.66	1
7	1	44	Local-government	43.10	Masters.degree(MA_MS_MEng_MED_MSW_MBA)	0	Not_in_universe	7	1258.46	1
8	1	31	Private	37.13	11th_grade	0	Not_in_universe	8	1278.31	1
9	1	83	Not_in_universe	0.0	9th_to_10th_grade	0	Not_in_universe	9	1328.86	1
10	1	46	Not_in_universe	0.0	Masters.degree(MA_MS_MEng_MED_MSW_MBA)	0	Not_in_universe	10	1348.68	1
11	1	31	Private	33.35	High_school_graduate	0	Not_in_universe	11	1387.7	1
12	1	22	Not_in_universe	0.0	12th_grade_no_diploma	0	Not_in_universe	12	1638.65	1
13	1	34	Local-government	35.2	Some_college_but_no_degree	0	Not_in_universe	13	1658.94	1
14	1	36	Private	11.37	High_school_graduate	0	Not_in_universe	14	1757.67	1
15	1	37	Self-employed-not-incorporated	32.16	Masters.degree(MA_MS_MEng_MED_MSW_MBA)	0	Not_in_universe	15	1787.34	1
16	1	47	Private	29.41	High_school_graduate	0	Not_in_universe	16	1812.99	1
17	1	48	Private	13.33	Associates_degree-occupational	1500	Not_in_universe	17	2011.14	1
18	1	34	Private	34.26	Some_college_but_no_degree	0	Not_in_universe	18	2050.15	1
19	1	55	Private	45.26	Associates_degree-occupational	0	Not_in_universe	19	2339.05	1
20	2	46	Private	34.3	Associates_degree-academic-program	0	Not_in_universe	20	2022.93	1
21	2	36	Private	13.37	High_school_graduate	0	Not_in_universe	21	2032.82	1
22	2	75	Not_in_universe	0.0	11th_grade	0	Not_in_universe	22	2138.19	1
23	2	36	Private	37.2	11th_grade	0	Not_in_universe	23	2240.99	1
24	2	49	Private	33.19	High_school_graduate	0	Not_in_universe	24	2244.66	1
25	2	39	Private	31.26	High_school_graduate	0	Not_in_universe	25	2258.46	1
26	2	32	Private	30.15	Bachelors.degree(BA_AB_BS)	0	Not_in_universe	26	2286.86	1
27	2	37	Private	25.18	Bachelors.degree(BA_AB_BS)	0	Not_in_universe	27	2348.66	1
28	2	29	Not_in_universe	0.0	High_school_graduate	0	Not_in_universe	28	2359.06	1
29	2	30	Private	12.4	Bachelors.degree(BA_AB_BS)	0	Not_in_universe	29	2369.14	1
30	2	48	Local-government	43.10	Masters.degree(MA_MS_MEng_MED_MSW_MBA)	0	Not_in_universe	30	2456.55	1
31	2	46	Private	34.3	Associates_degree-academic-program	0	Not_in_universe	31	2456.65	1

Εικόνα 4.9– Αποτέλεσμα anatomy

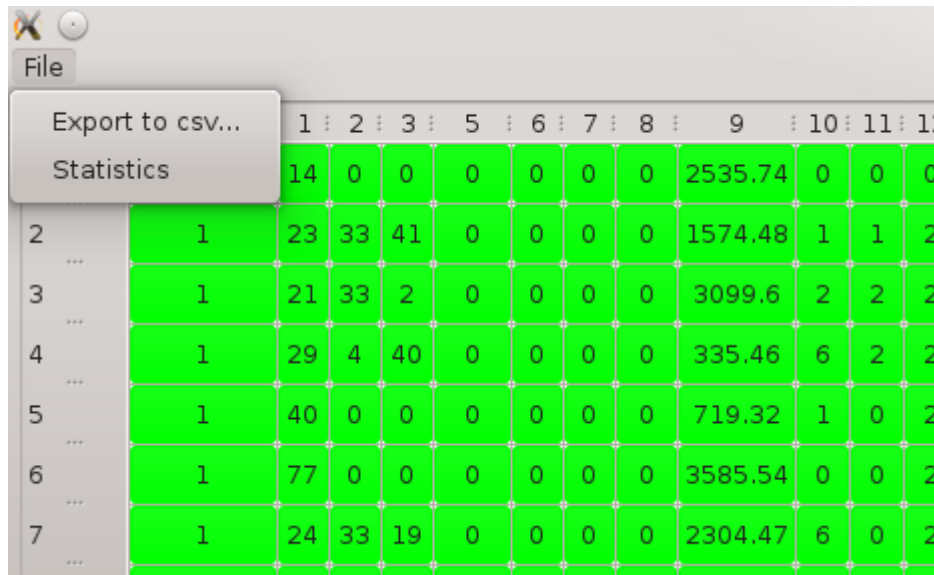
Όταν ολοκληρωθεί η εκτέλεση, το πρόγραμμα εμφανίζει το παράθυρο anatomy Window, το οποίο παρουσιάζει τα δεδομένα μετά την ανωνυμοποίηση (Εικόνα 4.9). Οι πληροφορίες εμφανίζονται ως εξής:

- Το παράθυρο χωρίζεται σε δύο μέρη. Στο αριστερό μισό, εμφανίζεται ο πίνακας των κλάσεων ισοδυναμίας, ο οποίος περιέχει όλες τις εγγραφές των αρχικών δεδομένων εκτός από την στήλη με το ευαίσθητο γνώρισμα η οποία έχει αντικατασταθεί από το αναγνωριστικό της κλάσης ισοδυναμίας (*QI Group*). Το δεξί μισό, περιλαμβάνει ένα πίνακα τριών στηλών, την στήλη *QI Group*, την στήλη του ευαίσθητου γνωρίσματος και την στήλη *Number Of Records*. Η πρώτη στήλη, δίνει στον χρήστη την πληροφορία του σε ποια κλάση ισοδυναμίας ανήκει η εγγραφή με το ευαίσθητο γνώρισμα που βρίσκεται στην δεύτερη στήλη. Η τρίτη στήλη περιέχει τον αριθμό των εγγραφών με αυτό το ευαίσθητο γνώρισμα που βρίσκονται στην κλάση ισοδυναμίας.
- Για την καλύτερη εποπτεία των κλάσεων, διαδοχικές κλάσεις έχουν χρωματιστεί με διαφορετικά χρώματα και στους δύο πίνακες του αποτελέσματος.

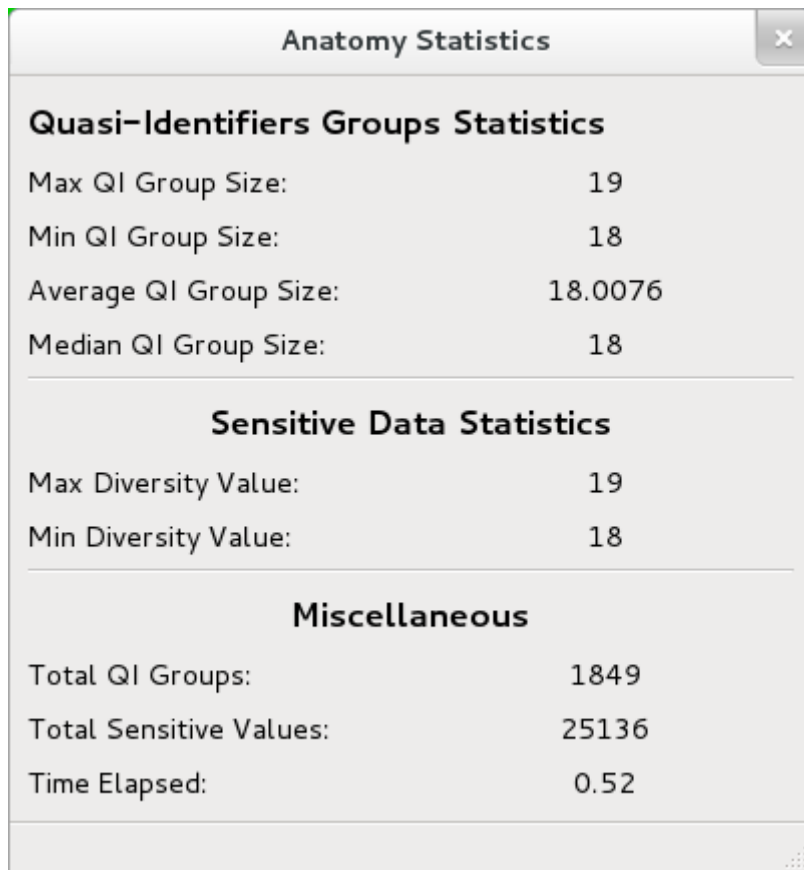
4.2.3 Στατιστικά και Διαγράμματα

Το δεύτερο κυριότερο υποσύστημα του εργαλείου ανωνυμοποίησης που υλοποιήσαμε, αφορά τα στατιστικά και διαγράμματα τα οποία μοντελοποιούν τα αποτελέσματα των αλγορίθμων ιδιωτικότητας.

Μετά την εμφάνιση των αποτελεσμάτων ανωνυμοποίησης, στο παράθυρο που εμφανίζεται, δίνεται στον χρήστη η επιλογή πατώντας *File -> Statistics* (Εικόνα 4.10), να του εμφανιστούν αποτελέσματα που αφορούν, τον χρόνο που έτρεξε ο αλγόριθμος, στατιστικές τιμές πάνω στο μέγεθος των κλάσεων ισοδυναμίας, το ποσοστό απώλειας πληροφορίας σε σχέση με το αρχικό σύνολο δεδομένων καθώς και το πλήθος των διακριτών τιμών του ευαίσθητου γνωρίσματος (Εικόνα 4.11, Εικόνα 4.12).



Εικόνα 4.10 - Εξαγωγή Στατιστικών



Εικόνα 4.11- Στατιστικά αλγορίθμου ανατομίας

Quasi-Identifiers Groups Statistics

Max QI Group Size:	6
Min QI Group Size:	5
Average QI Group Size:	5
Median QI Group Size:	6

Information Loss

Information Loss:	1.77%
-------------------	-------

Miscellaneous

Total QI Groups:	16384
Total Sensitive Values:	17
Time Elapsed:	8.83 sec

Εικόνα 4.12– Στατιστικά k-ανωνυμίας

Εκτός όμως από τα στατιστικά του κάθε αλγορίθμου, η εφαρμογή δίνει την δυνατότητα συλλογής συγκριτικών στατιστικών πάνω σε ένα δείγμα των συνολικών δεδομένων, το μέγεθος του οποίου επιλέγεται από τον χρήστη (Εικόνα 4.13).

The screenshot shows the 'Anonymization Tool' interface. On the left, there is a sidebar with the following controls:

- Sensitive Data:** A dropdown menu set to '4'.
- Anonymization Mode:** A dropdown menu set to 'k-anonymity'.
- k-value:** A dropdown menu set to '18'.
- Anonymize:** A button.
- Sampling Mode:** A radio button that is selected.
- Sample Percentage:** A dropdown menu set to '35.00'.
- Number of Records:** A dropdown menu set to '33296'.
- Show Result Window:** A checked checkbox.

The main window displays a table with 15 columns (labeled 1 to 15) and 30 rows of data. Each row represents a record with its group size and various quasi-identifier values. The table content is as follows:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	58	4	34	Some_college_but_no_degree	0	0	0	0	1053.55	1	0	2	52	94	~50000
2	9	0	0	Children	0	0	0	0	1758.14	0	0	0	0	0	94 ~50000
3	10	0	0	Children	0	0	0	0	1069.16	0	0	0	0	0	94 ~50000
4	42	34	3	Bachelors_degree(BA_AB_BS)	0	5178	0	0	1535.86	6	0	2	52	94	~50000
5	34	4	37	Some_college_but_no_degree	0	0	0	0	1146.79	6	0	2	52	94	~50000
6	8	0	0	Children	0	0	0	0	2466.24	0	0	0	0	0	94 ~50000
7	51	4	34	Some_college_but_no_degree	0	0	0	0	2441.22	3	0	2	52	94	~50000
8	46	37	31	High_school_graduate	0	0	1590	0	978.16	6	0	2	52	94	~50000
9	13	0	0	Children	0	0	0	0	1520.08	0	0	0	0	0	94 ~50000
10	39	0	0	10th_grade	0	0	0	0	1274.04	0	0	2	0	0	94 ~50000
11	12	0	0	Children	0	0	0	0	455.02	0	0	0	0	0	94 ~50000
12	27	4	34	Some_college_but_no_degree	0	0	0	0	1004.69	6	1	2	52	94	~50000
13	46	45	12	Masters_degree(MA_MS_MEng_MEd_MSW_MBA)	0	0	1977	100	999.46	1	0	2	52	94	~50000
14	55	0	0	Some_college_but_no_degree	0	0	0	0	1483.69	0	0	2	0	0	94 ~50000
15	2	0	0	Children	0	0	0	0	1660.53	0	0	0	0	0	94 ~50000
16	37	4	2	Bachelors_degree(BA_AB_BS)	0	0	1669	700	1331.35	6	0	2	52	94	~50000
17	46	4	34	Some_college_but_no_degree	0	0	0	0	1629.02	1	0	2	52	94	~50000
18	25	45	23	High_school_graduate	0	0	0	0	1998.03	1	0	2	52	94	~50000
19	46	32	42	High_school_graduate	0	0	0	0	2492.74	3	0	2	52	94	~50000
20	39	34	2	Bachelors_degree(BA_AB_BS)	0	0	0	1000	980.1	6	0	2	52	94	~50000
21	11	0	0	Children	0	0	0	0	498.25	0	0	0	0	0	94 ~50000
22	30	0	0	High_school_graduate	0	0	0	0	1773.08	0	0	2	0	0	94 ~50000
23	7	0	0	Children	0	0	0	0	885.84	0	0	0	0	0	94 ~50000
24	66	33	19	7th_and_8th_grade	0	0	0	0	4183.26	1	0	2	38	94	~50000
25	26	0	0	High_school_graduate	0	0	0	0	1040.15	0	0	2	0	0	94 ~50000
26	52	0	0	High_school_graduate	0	0	0	0	1407.39	0	0	2	0	0	94 ~50000
27	5	0	0	Children	0	0	0	0	235.92	0	0	0	0	0	94 ~50000
28	35	0	0	High_school_graduate	0	0	0	0	1669.47	0	0	2	0	0	94 ~50000
29	36	33	34	High_school_graduate	0	0	0	0	1812	3	0	2	48	94	~50000
30	42	44	10	Bachelors_degree(BA_AB_BS)	0	0	0	0	2635.89	6	0	2	52	94	~50000

Εικόνα 4.13 - Ενεργοποίηση Sampling mode και επιλογή ποσοστού

Δίνοντας την εντολή για δειγματοληψία η εφαρμογή εκτελεί και τους δύο υλοποιημένους αλγορίθμους πάνω σε ένα τυχαία επιλεγμένο δείγμα των δεδομένων, του οποίου το μέγεθος συμφωνεί με αυτό που έχει εισάγει ο χρήστης. Όσο η δειγματοληψία είναι ενεργοποιημένη, τα αποτελέσματα των αλγορίθμων για τις διάφορες παραμέτρους k και l αποθηκεύονται. Μετά το πέρας της ανωνυμοποίησης, το πρόγραμμα εμφανίζει το παράθυρο *sampling statistics*, το οποίο περιέχει 6 tabs, με συγκεντρωμένα στατιστικά για όλα τα μεγέθη που αφορούν τους δύο αλγορίθμους. Ας δούμε λίγο αναλυτικότερα το κάθε ένα από αυτά.

1. Το πρώτο tab, ονομάζεται *Anatomy Grid* και περιέχει τα αποτελέσματα του αλγορίθμου της ανατομίας πάνω στο sample των δεδομένων (Εικόνα 4.14). Σε αυτά περιέχονται τα μέγιστο, ελάχιστο, μέσο, διάμεσος μεγέθη των κλάσεων ισοδυναμίας, το εύρος τιμών των διακριτών τιμών του ευαίσθητου γνωρίσματος μέσα σε αυτές αλλά και συνολικά στο δείγμα και ο χρόνος εκτέλεσης του αλγορίθμου, σε ένα φάσμα τιμών της παραμέτρου l .

	5	6	7	9	10	12	15
Max QI Group Size	61	72	79	101	118	130	165
Min QI Group Size	21	30	37	52	64	78	105
Average QI Group Size	38.1397	47.702	57.2096	76.367	85.8144	105.035	133.719
Median QI Group Size	38	48	57	76	86	104	134
Max Diversity Value	6	7	8	10	11	13	17
Min Diversity Value	5	6	7	9	10	12	15
Total QI Groups	873	698	582	436	388	317	249
Total Sensitive Values	597	597	597	597	597	597	597
Time Elapsed	0.18	0.18	0.17	0.17	0.18	0.16	0.16

Εικόνα 4.14 - Anatomy Grid

2. Η επόμενη καρτέλα ονομάζεται k-anonymity Grid και σε αντιστοιχία με την πρώτη μας, παρουσιάζει τα ίδια στατιστικά μεγέθη με την προσθήκη της μετρικής GCP με την οποία αποτιμάμε την απώλεια πληροφορίας (για τον ορισμό της, αναφερόμαστε στο *Κεφάλαιο 3: Θεωρητικό Υπόβαθρο*)

Sampling Statistics

File

Number of Records in Sample: 33296 (35.00%)

Anatomy Grid k-anonymity Grid Anatomy Diversity Chart k-anonymity GCP chart QI Groups Size Chart

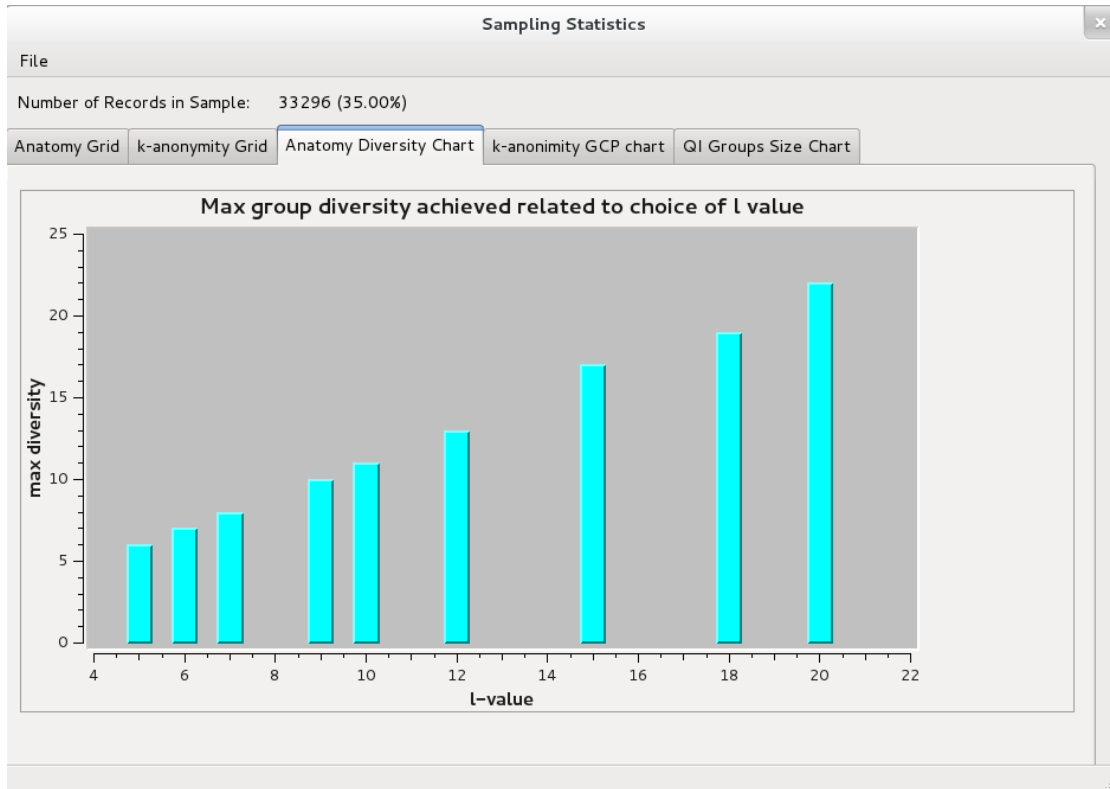
k-value

	5	7	9	11	11	15	18
Max QI Group Size	9	9	17	17	17	17	33
Min QI Group Size	8	8	16	16	16	16	32
Average QI Group Size	8	8	16	16	16	16	32
Median QI Group Size	8	8	16	16	16	16	33
GCP	0.0247186	0.0247186	0.0272819	0.0272819	0.0272819	0.0272819	0.0293809
Total QI Groups	4096	4096	2048	2048	2048	2048	1024
Total Sensitive Values	597	597	597	597	597	597	597
Time Elapsed	2.79	2.78	2.56	2.55	2.55	2.55	2.35

Εικόνα 4.15 - k-anonymity Grid

3. Στην επόμενη καρτέλα (Anatomy Diversity Chart), παρουσιάζεται ένα ιστόγραμμα το οποίο συγκρίνει την μέγιστη τιμή ποικιλομορφίας σε μία κλάση ισοδυναμίας με της τιμές παραμέτρου l (Εικόνα 4.16).
4. Στην συνέχεια, παρουσιάζεται πάλι ένα ιστόγραμμα, το οποίο αυτή την φορά συγκρίνει το μέγεθος της μετρικής GCP με την τιμή της παραμέτρου k (Εικόνα 4.17).
5. Στην καρτέλα QI Group Size Chart μπορούμε να δούμε το μέγεθος των κλάσεων ισοδυναμίας σε σχέση με το μέγεθος του συνόλου δεδομένων τόσο ξεχωριστά για κάθε αλγόριθμο, όσο και μαζί στο ίδιο ιστόγραμμα για την μεταξύ τους σύγκριση (Εικόνα 4.18). Η εμφάνιση ή απόκρυψη της σειράς που αντιστοιχεί σε έναν αλγόριθμο γίνεται πατώντας πάνω στον αντίστοιχο τίτλο της δεξιά του διαγράμματος. Στο σημείο εκείνο γίνεται φανερό και το χρώμα με το οποίο απεικονίζεται η κάθε σειρά τιμών στο γράφημα.
6. Στην τελευταία καρτέλα Time Elapsed, φαίνεται ο χρόνος που χρειάστηκε για να ολοκληρωθεί η εκτέλεση των αλγόριθμων σε σχέση με τις παραμέτρους k και l που δόθηκαν από τον χρήστη (Εικόνα 4.19). Όπως και στο QI Group Size Chart έτσι και εδώ μπορεί να γίνει σύγκριση μεταξύ των δύο αλγορίθμων.

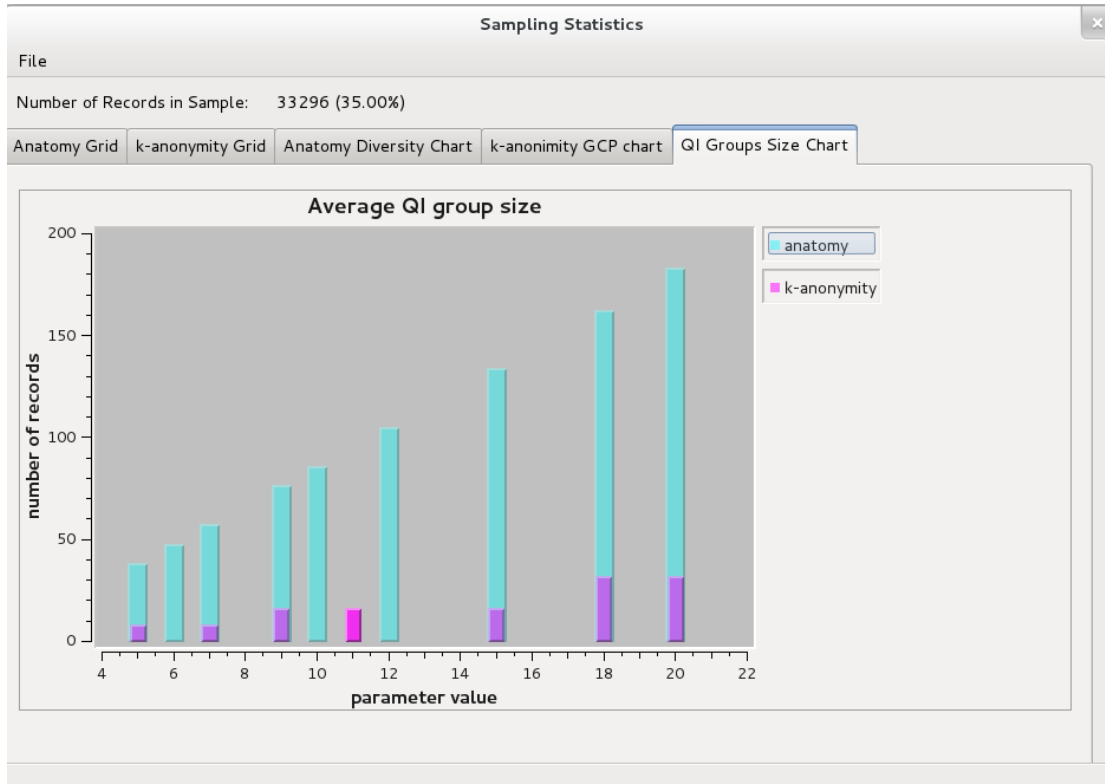
Τέλος, το sampling mode πέρα από την εξαγωγή των στατιστικών και τον σχεδιασμό των διαγραμμάτων, δίνει και την επιλογή στον χρήστη για την εξαγωγή των αριθμητικών τιμών των στατιστικών σε .csv αρχεία για την περαιτέρω μελέτη και ανάλυση των αναγκών ανωνυμοποίησης του κάθε συνόλου δεδομένων.



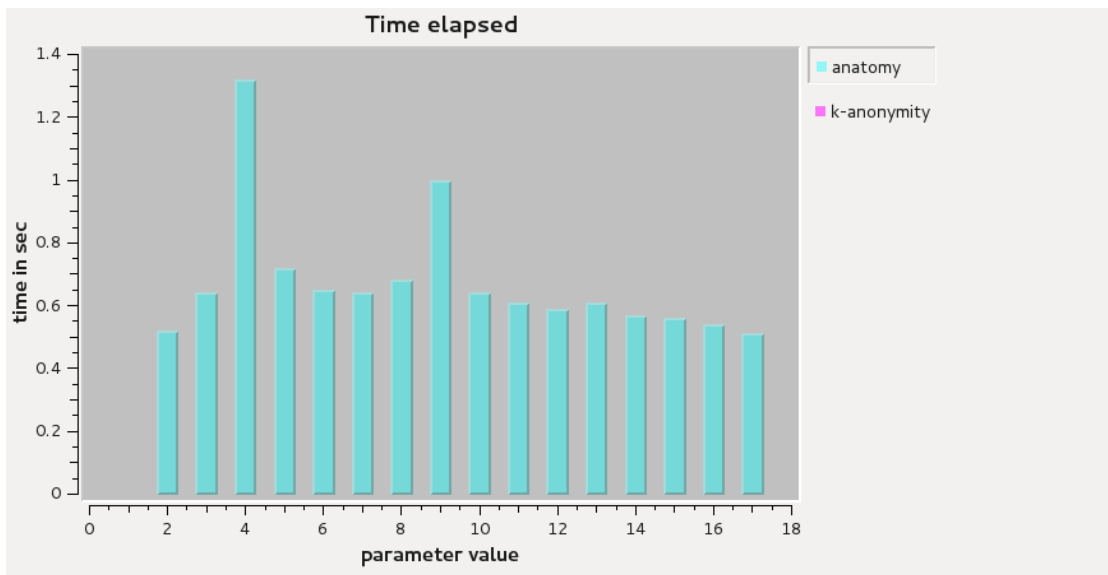
Εικόνα 4.16 – Μέγιστη ποικιλομορφία ανά τιμή l



Εικόνα 4.17- Ιστόγραμμα σύγκρισης GCP με k -value



Εικόνα 4.18 - Σύγκριση μεγέθους κλάσεων ισοδυναμίας μεταξύ των αλγορίθμων



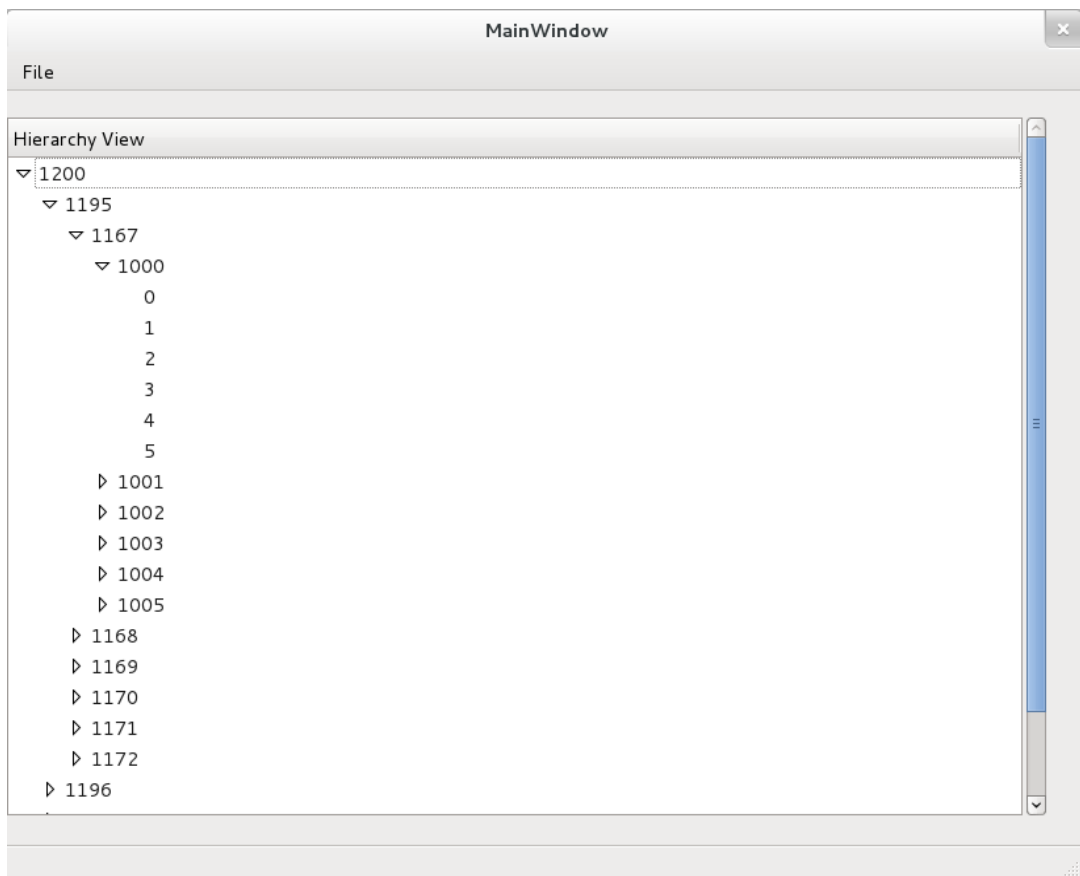
Εικόνα 4.19- Διάγραμμα χρόνου εκτέλεσης

4.2.4 Διαχείριση ιεραρχιών γενίκευσης

Το τελευταίο κομμάτι της εφαρμογής, εστιάζει στην εισαγωγή και εποπτεία ιεραρχιών γενίκευσης κατασκευασμένων από τον χρήστη με την χρήση της επιλογής *Load Hierarchy* (Εικόνα 4.2). Η ιεραρχία γενίκευσης δίνεται σε .txt αρχεία και παρουσιάζεται σε δεντρική μορφή.

Ο τρόπος απεικόνισης έχει επιλεγθεί με τέτοιο τρόπο ώστε να είναι ευκολονόητα τα στάδια της γενίκευσης (Εικόνα 4.20). Ξεκινώντας από την ρίζα του δέντρου γενίκευσης μπορούμε να πατήσουμε πάνω στον απεικονισμένο κόμβο και να εμφανιστούν κάτω από αυτό και στο αμέσως εσωτερικότερο επίπεδο στοίχισης οι τιμές που αντιστοιχούν στα παιδιά κατά το δέντρο της ιεραρχίας. Με τον ίδιο τρόπο μπορούμε να εμφανίσουμε τα παιδιά οποιουδήποτε κόμβου φτάνοντας στο τελευταίο επίπεδο το οποίο αντιστοιχεί στα φύλλα του δέντρου και δεν μπορεί να αναπτυχθεί περαιτέρω. Τα εικονίδια σε σχήμα βέλους απεικονίζουν το αν ο κόμβος στον οποίο βρίσκονται αυτά είναι ανοιγμένος και έχουν αναπτυχθεί από κάτω του τα παιδιά αυτού ή είναι κλειστός και πρέπει να πατηθεί για να εξερευνηθούν τα πιο κάτω επίπεδα.

Η εν λόγω δυνατότητα, έχει υλοποιηθεί στην εφαρμογή για να διευκολύνει την μελλοντική ανάπτυξη και ενσωμάτωση περισσότερων αλγορίθμων ανωνυμοποίησης και για την προσφορά ενός πλήρους περιβάλλοντος για κάθε ανάγκη προστασίας δεδομένων.



Εικόνα 4.20 - Ιεραρχία γενίκευσης όπως εισάγεται από τον χρήστη

5

Υλοποίηση

Σε αυτό το κεφάλαιο παρουσιάζονται θέματα που αφορούν το τεχνικό κομμάτι της ανάπτυξης της εφαρμογής. Πιο συγκεκριμένα, θα γίνει αρχικά αναφορά στα εργαλεία με τα οποία αναπτύχθηκε η εφαρμογή καθώς και κάποιων τεχνικών που τα διέπουν και επηρέασαν κάποια τμήματα της υλοποίησης. Στην συνέχεια θα γίνει ανάλυση των κλάσεων που απαρτίζουν την εφαρμογή και των αλγορίθμων που υλοποιήθηκαν, καθώς και σχολιασμός των επιλογών που έγιναν κατά την υλοποίηση ή τον σχεδιασμό αυτών.

5.1 Πλατφόρμες και προγραμματιστικά εργαλεία

Για την ανάπτυξη του το εν λόγω εργαλείου χρησιμοποιήθηκε το Qt framework (έκδοση Qt 4.8.2) και η γλώσσα C++. Το Qt πρόκειται για ένα framework ανοιχτού κώδικα για την ανάπτυξη cross-platform εφαρμογών γραφικού περιβάλλοντος [13]. Το ολοκληρωμένο περιβάλλον ανάπτυξης (IDE) το οποίο επιλέχθηκε ήταν το Qt Creator καθώς υποστηρίζει πληρέστερα το Qt Framework και τον σχεδιασμό παραθυρικών εφαρμογών, έχοντας ενσωματωμένο τον Qt Designer και δίνοντας άμεση πρόσβαση στο Qt SDK. Επιπρόσθετα, για την οπτικοποίηση ορισμένων μετρικών, ενσωματώθηκαν στην εφαρμογή αυτή οι κατάλληλες γραφικές παραστάσεις. Για τον σκοπό αυτό έγινε χρήση της βιβλιοθήκης Qwt (έκδοση 6.0.0-1.2) η οποία περιλαμβάνει Qt Widgets για τεχνικές εφαρμογές [14].

5.1.1 Qt Framework

Στο σημείο αυτό θα γίνει μια συνοπτική αναφορά σε κάποια από τα βασικότερα χαρακτηριστικά του Qt Framework που χρησιμοποιήθηκαν στην ανάπτυξη των κλάσεων που αφορούν την διεπαφή με τον χρήστη.

Qt Widgets

Τα κυριότερα στοιχεία για την δημιουργία διεπαφών με τον χρήστη στην Qt λέγονται widgets. Παρέχουν τις βασικότερες δυνατότητες που αφορούν την απεικόνιση δεδομένων, την παρουσίαση κάποια συγκεκριμένης κατάστασης, την λήψη εισόδου από τον χρήστη της εφαρμογής. Τα widgets μπορούν να περιέχουν άλλα widgets φροντίζοντας για την ομαδοποίηση τους και την διάταξη τους με συγκεκριμένους τρόπους. Κάθε widget έχει ως γονιό το widget που το περιέχει, με εξαίρεση το παράθυρο (window) που αποτελούν τα μόνα widget που δεν έχουν γονείς.

Signals and Slots

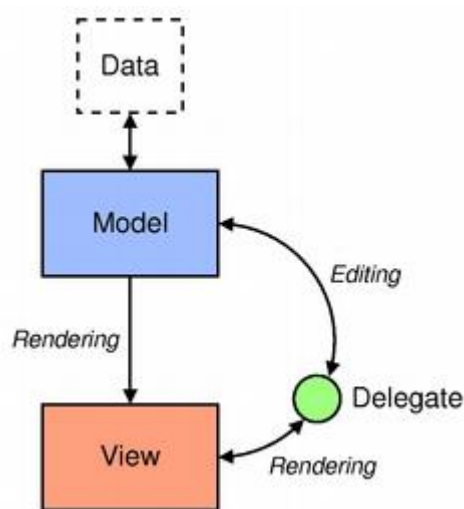
Όπως σε κάθε εφαρμογή γραφικού περιβάλλοντος είναι απαραίτητη η επικοινωνία μεταξύ αντικειμένων. Για παράδειγμα, όταν ο χρήστης πατάει ένα κουμπί θα θέλαμε να καλείται μια συγκεκριμένη μέθοδος. Η επικοινωνία μεταξύ των αντικειμένων στο Qt Framework γίνεται με signals και slots. Όταν η κατάσταση ενός αντικειμένου αλλάξει, τότε αυτό μπορεί να εκπέμψει ένα signal. Τα slots είναι συναρτήσεις που συνδέονται με κάποιο συγκεκριμένο signal και καλούνται κάθε φορά που το αντίστοιχο signal εκπέμπεται.

Προγραμματισμός Model/View

Η λογική του προτύπου Model/View έχει ως βάση το γνωστότερο πρότυπο Model-View-Controller (MVC), το οποίο αποτελείται από τρία ξεχωριστά μέρη:

- το μοντέλο (Model) που είναι το αντικείμενο της εφαρμογής.
- την όψη (View) που απεικονίζει το μοντέλο με τον επιθυμητό τρόπο στον χρήστη.
- τον ελεγκτή (Controller) που περιλαμβάνει τους τρόπους με τους οποίους αλληλοεπιδρά ο χρήστης με το μοντέλο αλλάζοντας την κατάστασή του ή τον τρόπο με τον οποίο βλέπει την όψη.

Στην Model/View αρχιτεκτονική έχει ενοποιηθεί η έννοια του ελεγκτή σε αυτή της όψη. Το αποτέλεσμα είναι το μοντέλο να αναλαμβάνει την επικοινωνία με τα δεδομένα, παίζοντας τον ρόλο της διεπαφής προς τις υπόλοιπες συνιστώσες. Η όψη μπορεί να ανακτήσει δεδομένα μέσω του model το οποίο της παρέχει αναφορές προς αυτά για αυτό το σκοπό, οι οποίες ονομάζονται model indexes. Την αλληλεπίδραση με τον χρήστη αναλαμβάνουν διάφοροι delegates. Η επικοινωνία μεταξύ των συνιστωσών της αρχιτεκτονικής αυτής γίνεται μέσω σημάτων (signals) και σχισμών (slots) [15].



Qt Model/View

5.2 Λεπτομέρειες υλοποίησης

Η εφαρμογή απαρτίζεται κυρίως από κλάσεις που εξυπηρετούν την διαδραστικότητα με τον χρήστη μέσω του γραφικού περιβάλλοντος, ενώ συνδυάζονται με δύο κλάσεις που αφορούν την υλοποίηση των δύο αλγορίθμων ανωνυμοποίησης (k-Mondrian και Anatomy) και αποτελούν τον πυρήνα της εφαρμογής. Ανεξάρτητο κομμάτι αποτελεί μια κλάση οπτικής απεικόνισης αρχείων ιεραρχίας.

5.2.1 Κλάση *FileData*

Η κλάση *FileData* αντιπροσωπεύει τα δεδομένα εισόδου των αλγορίθμων και περιλαμβάνει μεθόδους για την αρχικοποίηση τους είτε με βάση ένα αρχείο κειμένου με τιμές οριοθετημένες με κόμματα (.csv), είτε δειγματοληπτώντας από ένα άλλο αντικείμενο της κλάσης αυτής.

Τα δεδομένα αυτά κρατώνται στο πεδίο *initData* που ένα διάνυσμα που κάθε στοιχείο του είναι ένας σταθερού μεγέθους πίνακας από συμβολοσειρές και αντιστοιχεί σε μια γραμμή των αρχικών δεδομένων. Επίσης, περιέχει και τα εξής πεδία πληροφοριών σχετικών με τα δεδομένα:

- **numberOfRecords**: ο αριθμός των γραμμών.
- **numberOfAttributes**: ο αριθμός των γνωρισμάτων που μεταφράζονται σε στήλες.
- **created**: ορίζει αν έχουν αρχικοποιηθεί επιτυχώς τα δεδομένα της κλάσης.
- **hasHeaders**: ορίζεται αν τα δεδομένα εισόδου περιείχαν την ονομασία του κάθε γνωρίσματος.
- **attributeNames**: κρατείται η ονομασία των γνωρισμάτων. Αν δεν δίνεται από τα δεδομένα εισόδου τότε την αντικαθιστά ο αριθμός της στήλης στην οποία αντιστοιχεί το κάθε γνώρισμα.
- **sensitiveDataIndex**: ο αριθμός της στήλης στην όποια βρίσκεται το γνώρισμα που θεωρείται ευαίσθητο για την ανωνυμοποίηση.
- **isSample**: έχει τιμή *false* αν το αντικείμενο αντιστοιχεί στα δεδομένα εισόδου και *true* αν έχει δημιουργηθεί με δειγματοληψία.

Οι μέθοδοι αυτής της κλάσης είναι:

- **Initialize** (fileUri, hasHeaders): φορτώνει τα δεδομένα με βάση ένα csv αρχείο, του οποίου το URI δίδεται από το όρισμα fileUri. Το δεύτερο όρισμα καθορίζει αν στην πρώτη γραμμή του αρχείου περιέχονται τα ονόματα των πεδίων.
- **createSample** (from, records): με βάση ένα άλλο αντικείμενο της κλάσης FileData, δημιουργεί ένα τυχαίο δείγμα με μέγεθος που ορίζεται από το δεύτερο όρισμά της. Για τον σκοπό αυτό, δημιουργείται αρχικά ένας πίνακας με τις θέσεις όλων των γραμμών της initData από το αρχικό αντικείμενο που παίρνει η μέθοδος ως παράμετρο, με την μορφή ακεραίων. Στην συνέχεια γίνεται τυχαία αναδιάταξή τους, με την βοήθεια της συνάρτησης random_shuffle που δίνεται από την βιβλιοθήκη algorithm της C++. Η συνάρτηση αυτή, ανταλλάσσει την τιμή κάθε στοιχείου με ένα άλλο τυχαία επιλεγμένο και εκτελείται σε γραμμικό ως προς το πλήθος των στοιχείων χρόνο. Τέλος, στο διάνυσμα initData αντιγράφονται από τον αρχικό vector δείκτες στις γραμμές που υποδεικνύουν τα πρώτα records στον αριθμό στοιχεία του αναδιατεταγμένου πίνακα και τίθεται σε true το πεδίο isSample.
- **clearData()** : καθαρίζονται τα δεδομένα που περιείχε.
- **printData()**: τυπώνονται στο standard output τα δεδομένα που περιέχονται στο initData για διευκόλυνση της αποσφαλμάτωσης.

Επιπλέον, η κλάση αυτή αποτελεί και το μοντέλο για την απεικόνιση αυτών των δεδομένων σε ένα TableView που παίζει τον ρόλο της όψης. Για την σύνδεση της όψης με τα δεδομένα αυτά μέσω των QModelIndexes, η κλάση FileData κληρονομεί την abstract κλάση QAbstractModelTable. Αυτό μεταφράζεται στην υλοποίηση των παρακάτω μεθόδων:

- **rowCount**: επιστρέφει τον αριθμό των γραμμών του initData.
- **columnCount**: επιστρέφει τον αριθμό των στηλών του initData.
- **data**: επιστρέφει τα δεδομένα που αντιστοιχούν στον ζητούμενο δείκτη μοντέλου, με τρόπο που ορίζεται ανάλογα με το ItemDataRole που δίνεται ως δεύτερο όρισμα. Κυρίως, χρησιμοποιούμε τον ρόλο DisplayRole (με αντίστοιχη τιμή 0) ο οποίος επιστρέφει τα δεδομένα σε μορφή κειμένου, ενώ ρόλοι όπως οι TextColorRole, TextAlignmentRole, BackgroundRole κλπ

χρησιμοποιούνται για την διαχείριση των δεδομένων και του τρόπου που θα πρέπει να εμφανίζονται στις όψεις.

- **headerData**: η μέθοδος αυτή αφορά τους τίτλους των γραμμών ή των στηλών, ανάλογα με την τιμή της δεύτερης παραμέτρου. Στο πρώτο όρισμα δίνεται η εκάστοτε γραμμή ή στήλη ενώ το τελευταίο αντιστοιχεί και πάλι στον ItemDataRole ρόλο.

Η επιλογή να υλοποιηθεί ένας Model-View σχεδιασμός σε αυτή την περίπτωση, με απεικόνιση των δεδομένων τελικά σε ένα QTableView αντί για ένα απλούστερο QTableWidgetItem έγινε με βάση τον μικρότερο χρόνο φόρτωσης που παρατηρήθηκε συγκρίνοντας τις δύο μεθόδους. Το ίδιο ισχύει για όλα τα σημεία που έχει γίνει αυτή η επιλογή και οι συναρτήσεις που χρειάστηκε να υλοποιηθούν σε κάθε περίπτωση έχουν πάντοτε την ίδια μορφή με αυτή που αναφέρθηκε, καθώς πάντοτε η απεικόνιση αφορούσε δεδομένα σε μορφή πίνακα.

5.2.2 Κλάση statistics και κλάσεις που την κληρονομούν

Σκοπός της κλάσης statistics και των δύο κλάσεων AnatomyStatistics και kStatistics που την κληρονομούν είναι η συλλογή πληροφοριών και ο υπολογισμός μετρικών που αφορούν τα δεδομένα που ανωνυμοποιούνται, την εκτέλεση των αλγορίθμων και τα αποτελέσματα που προέκυψαν. Η κλάση statistics αποτελείται από τα πεδία τα οποία αφορούν στατιστικά κοινά και στις δύο περιπτώσεις ανωνυμοποίησης και είναι τα παρακάτω:

- **time**: ο χρόνος CPU εκτέλεσης του αλγορίθμου ανωνυμοποίησης σε δευτερόλεπτα.
- **sensitiveValuesNumber**: το πλήθος των διακριτών τιμών του ευαίσθητου γνωρίσματος.
- **maxQIGroupSize, minQIGroupSize**: το μέγιστο και ελάχιστο μέγεθος των κλάσεων ισοδυναμίας που σχηματίστηκαν.
- **medianQIGroupSize, averageQIGroupSize**: η διάμεσος κι ο μέσος όρος των μεγεθών των κλάσεων ισοδυναμίας.

- **calculated** : βοηθητικό πεδίο που καθορίζει αν έχει γίνει υπολογισμός των στατιστικών.

5.2.2.1 Κλάση *AnatomyStatistics*

Περιέχει τα επιπλέον πεδία **maxDiversityValue** και **minDiversityValue** με τύπο ακέραιο που αντιστοιχούν στο μέγιστο και ελάχιστο πλήθος διαφορετικών τιμών του ευαίσθητου γνωρίσματος που βρίσκεται σε μία κλάση ισοδυναμίας.

5.2.2.2 Κλάση *kStatistics*

Αυτή η κλάση εκτός των πεδίων που κληρονομεί από την κλάση *statistics* περιλαμβάνει το πεδίο **GCP** που αντιστοιχεί στο Συνολική Ποινή Βεβαιότητας και πρόκειται για μετρική απώλειας πληροφορίας. Ακόμη, περιλαμβάνει μια μέθοδο *calculate* η οποία υπολογίζει τις προαναφερθείσες μετρικές.

5.2.3 Κλάση *AnatomyData*

Η κλάση που χρησιμοποιείται για την ανωνυμοποίηση των δεδομένων με τον αλγόριθμο της ανατομίας είναι κλάση *AnatomyData*. Για την κατασκευή ενός αντικειμένου αυτής της κλάσης, είναι απαραίτητα δύο ορίσματα: τα αρχικά δεδομένα τύπου *FileData* που κρατώνται στο πεδίο *initData* της κλάσης *AnatomyData* και η τιμή της παραμέτρου *l* του αλγορίθμου της ανατομίας που αντιστοιχεί στο πεδίο *lvalue*. Η βασική λειτουργία της κλάσης αυτής υλοποιείται στην μέθοδος *anatomy()* της οποία θα γίνει και εκτενέστερη ανάλυση. Εκτός αυτής, η κλάση περιλαμβάνει πεδία που αφορούν τα αποτελέσματα αυτού του αλγορίθμου, καθώς και το πεδίο *stats* για την συλλογή στατιστικών σχετικών με τα δεδομένα και την εκτέλεση του αλγορίθμου και έχει τύπο *AnatomyStatistics*. Πριν την ανάλυση της υλοποίησης της μεθόδου της ανατομίας, είναι απαραίτητο να αναφερθούμε στα πεδία **QITableData** και **QISensitiveMap** που αντιστοιχούν στα δεδομένα που περιλαμβάνονται στον πίνακα ψευδο-αναγνωριστικών QIT και στον πίνακα ευαίσθητου γνωρίσματος ST, τα οποία είναι τα αποτελέσματα του αλγορίθμου.

Το πρώτο πεδίο έχει τύπο `AnatomyQITableData`, μια κλάση που δημιουργήθηκε για την απεικόνιση του πίνακα ψευδο-αναγνωριστικών και βασικό της πεδίο είναι το πεδίο `Data` που είναι ίδιας μορφής με το πεδίο `initData` της κλάσης `FileData`, με την διαφορά ότι αντί για το ευαίσθητο δεδομένο υπάρχει μια στήλη `QI Group` η οποία προσδιορίζει την κλάση ισοδυναμίας στην οποία κατατάχθηκε κάθε εγγραφή. Οι εγγραφές με τον ίδιο αριθμό σε αυτή τη στήλη ανήκουν στην ίδια κλάση ισοδυναμίας, πρόκειται δηλαδή για ένα αναγνωριστικό της κάθε ομάδας και στο εξής θα αναφέρεται ως `QIGroupId`. Επιπλέον αυτή η κλάση υλοποιεί, παρόμοια με την κλάση `FileData`, μεθόδους της `QAbstractTableModel`, την οποία κληρονομεί, με την διαφορά ότι στην μέθοδο `data` αυτής της κλάσης γίνεται ξεχωριστή αντιμετώπιση του ρόλου `BackgroundColorRole` για να επιτευχθεί ο χρωματισμός των κλάσεων ισοδυναμίας, με σκοπό την καλύτερη διαφοροποίηση τους στην παρουσίασή τους προς τον χρήστη.

Το πεδίο `QISensitiveMap` έχει τύπο `map` και πρόκειται ουσιαστικά για μια δομή ευρετηρίου. Κλειδί αυτής είναι ένα ζευγάρι που αποτελείται από έναν ακέραιο που αντιστοιχεί στο `QIGroupId` και ένα `QString` το οποίο αντιστοιχεί σε τιμή ευαίσθητου γνωρίσματος. Οι τιμές του ευρετηρίου αντιστοιχούν στο πλήθος των εγγραφών της κλάσης ισοδυναμίας, οι οποίες έχουν την δεδομένη τιμή ευαίσθητου γνωρίσματος. Για παράδειγμα, ας θεωρήσουμε την περίπτωση στην οποία αρχικά έχουμε έναν πίνακα ασθενών και το ευαίσθητο γνώρισμα είναι η ασθένεια. Τότε, αν το κλειδί είναι ένα ζεύγος (5, "πνευμονία") και η αντίστοιχη τιμή είναι 7 σημαίνει ότι η πέμπτη κλάση ισοδυναμίας που δημιουργήθηκε κατά την ανωνυμοποίηση περιλαμβάνει επτά εγγραφές που αντιστοιχούσαν στην ασθένεια πνευμονία.

Όπως προαναφέρθηκε, η βασική λειτουργία της κλάσης αντιπροσωπεύεται από την μέθοδο `anatomy`. Ο αλγόριθμος αυτός, όπως περιγράφεται και στην δημοσίευσή του, ξεκινά με την αρχικοποίηση δύο άδειων πινάκων `QIT` και `ST` που στην παρούσα εφαρμογή αντιπροσωπεύονται από τα πεδία που περιγράφηκαν προηγουμένως. Συνεχίζει με την δημιουργία κάδων, που κάθε κάδος αντιπροσωπεύει μια τιμή ευαίσθητου γνωρίσματος και περιέχει τις εγγραφές που έχουν την τιμή αυτή. Έπειτα, ακολουθούν οι διαδικασίες της δημιουργίας ομάδων και της ανάθεσης σε ομάδες των εγγραφών που απέμειναν.

Για την υλοποίηση του μέρους που αφορά τους κάδους, κατασκευάζεται ένα ευρετήριο (map), στο οποίο για κάθε τιμή ευαίσθητου γνωρίσματος αντιστοιχίζονται οι αριθμοί των γραμμών στις οποίες εμφανίζεται αυτό. Το ευρετήριο αυτό θα αναφέρεται στο εξής ως sensitiveMap. Για παράδειγμα αν οι γραμμές 1,3,7 και 8 αντιστοιχούν σε ασθενείς με πνευμονία, τότε η τιμή του sensitiveMap για το κλειδί “πνευμονία” θα είναι το διάνυσμα [1,3,7,8]. Αφού, λοιπόν, δημιουργηθεί αυτή η δομή με βάση τα αρχικά δεδομένα, ελέγχεται αν το πλήθος των διαφορετικών ευαίσθητων τιμών, το οποίο πλέον αντιστοιχεί και στο πλήθος των κλειδιών του sensitiveMap, είναι μεγαλύτερο από την τιμή της παραμέτρου I. Αν δεν είναι, όπως είναι λογικό δεν μπορεί να επιτευχθεί ανωνυμοποίηση για αυτή την τιμή της παραμέτρου με τα εν λόγω δεδομένα και έτσι η εκτέλεση σταματά εκεί επιστρέφοντας την τιμή -1 ώστε να γίνει στην συνέχεια η κατάλληλη διαχείριση αυτής της περίπτωσης.

Έπειτα, για την διευκόλυνση των φάσεων της δημιουργίας ομάδων και της κατάταξης σε ομάδες των υπόλοιπων εγγραφών, με βάση το sensitiveMap, δημιουργείται ένα διάνυσμα από buckets. Τα buckets είναι αντικείμενα τα οποία περιέχουν δύο πεδία: το key που είναι μια τιμή ευαίσθητου γνωρίσματος και το number_of_records που δείχνει το πλήθος των εγγραφών που έχουν αυτή την τιμή. Επιπλέον, στην κλάση Bucket έχει ορισθεί η υπερφόρτωση του τελεστή σύγκρισης < καθώς και της ανάθεσης. Σκοπός του διανύσματος των Buckets είναι να παρέχει ανά πάσα στιγμή τα ευαίσθητα γνωρίσματα ταξινομημένα με φθίνουσα σειρά ανάλογα με το πλήθος των εγγραφών που τους αντιστοιχούν, έτσι ώστε να είναι εύκολο να γίνει ο εντοπισμός των n πρώτων τέτοιων τιμών. Εκτός από την αρχικοποίηση και την αρχική ταξινόμηση των κάδων, στο σημείο αυτό γίνεται και τυχαία αναδιάταξη των αριθμών που περιέχονται σε κάθε εγγραφή του ευρετηρίου sensitiveMap, ώστε να λαμβάνεται άμεσα στην συνέχεια μια τυχαία εγγραφή που να αντιστοιχεί σε μια δεδομένη τιμή ορίσματος.

Αφού πλέον έχουν αρχικοποιηθεί οι απαραίτητες δομές ξεκινάει η φάση της δημιουργίας των ομάδων. Σε αυτή, όσο υπάρχουν τουλάχιστον I μη κενοί κάδοι διαλέγεται μία τυχαία εγγραφή από τους I μεγαλύτερους σε μέγεθος και σχηματίζεται μια ομάδα. Ταυτόχρονα λαμβάνεται μέριμνα για την ενημέρωση του QISensitiveMap,

καθώς και των μεγεθών των κάδων. Τέλος πρέπει το διάνυσμα των buckets να παραμείνει ταξινομημένο μετά από κάθε επανάληψη.

Στο σημείο αυτό έγινε μια προσπάθεια εύρεσης ενός αποδοτικού τρόπου με τον οποίο θα παραμένει ταξινομημένο το διάνυσμα. Μιας και στην πραγματικότητα, μετά από κάθε επανάληψη μειώνονται οι πρώτες l τιμές κατά ένα ο πίνακας θα είναι πάντοτε σχεδόν ταξινομημένος. Έτσι η πρώτη λογική επιλογή για την περίπτωση αυτή είναι ταξινόμηση με συγχώνευση (merge sort) με χρονική πολυπλοκότητα $O(n \log n)$ όπου n ο αριθμός των κάδων. Όμως, παρατηρώντας λίγο πιο προσεκτικά την διάταξη που έχουν τα στοιχεία του διανύσματος μετά από κάθε επανάληψη, μπορεί κανείς να διαπιστώσει ότι το μόνο που χρειάζεται για να αποκατασταθεί η διάταξη των στοιχείων είναι να γίνει κατά πρώτον ένας έλεγχος αν το στοιχείο στην θέση l έχει μεγαλύτερη τιμή από το στοιχείο στην θέση $l-1$ η οποία αντιστοιχεί και στον τελευταίο κάδο από τον οποίο αφαιρέθηκε στοιχείο. Αν δεν ισχύει το παραπάνω, τότε το διάνυσμα έχει παραμείνει ταξινομημένο και δεν χρειάζεται να αλλάξει θέση κανένα από τα στοιχεία του. Στην συνέχεια, γίνεται αναζήτηση του πρώτου στοιχείου που έχει τιμή ίση με αυτό στην θέση $l-1$, καθώς και του τελευταίου που έχει τιμή ίδια με την τιμή του στοιχείου στην θέση l . Δεδομένων αυτών των θέσεων, εύκολα πλέον μπορεί να γίνει ανταλλαγή των στοιχείων με τιμή $l-1$ με αυτά που έχουν τιμή l . Αυτό ισχύει γιατί λόγω της ελάττωσης κατά ένα των πρώτων l τιμών σε κάθε επανάληψη, όταν θα χρειαστεί η αναδιάταξη τα στοιχεία των θέσεων l και $l-1$ θα διαφέρουν ακριβώς κατά ένα. Έτσι επιτεύχθηκε το επιθυμητό αποτέλεσμα με γραμμική πολυπλοκότητα καθώς και οι δύο αναζητήσεις των θέσεων και η ανταλλαγή των στοιχείων είναι γραμμικές ως προς το χρόνο εκτέλεσης διαδικασίες.

Τέλος, στην φάση της ανάθεσης σε ομάδες των εγγραφών που έχουν μείνει σε κάδους από την προηγούμενη διαδικασία διαλέγεται για κάθε εγγραφή μια τυχαία ήδη δημιουργημένη ομάδα μέσω μιας γεννήτριας τυχαίων αριθμών με εύρος αυτό των GroupIds που έχουν παραχθεί. Έτσι προστίθεται στην διαλεγμένη ομάδα και ενημερώνεται το QISensitiveMap. Επιπλέον, κάποια από τα στατιστικά που αφορούν την ανωνυμοποίηση με ανατομία όπως το πλήθος των ομάδων που σχηματίστηκαν, ο αριθμός των διαφορετικών τιμών του ευαίσθητου γνωρίσματος καθώς και ο χρόνος εκτέλεσης συλλέγονται στην συνάρτηση anatomy.

5.2.3 Κλάση *kData*

Η κλάση που αφορά την ανωνυμοποίηση με τον αλγόριθμο της *k*-ανωνυμίας είναι η κλάση *kData*. Πιο συγκεκριμένα πρόκειται για τον πολυδιάστατο *mondrian* αλγόριθμο [16] για *k*-ανωνυμοποίηση και η παρούσα υλοποίηση αντιμετωπίζει αμιγώς αριθμητικά δεδομένα ψευδο-αναγνωριστικών, καθότι απαιτείται ορισμός της διάταξης στα πεδία και εύρεση διαμέσου για τον σχηματισμό των κλάσεων ισοδυναμίας.

Ο άπληστος πολυδιάστατος *mondrian* αλγόριθμος για *k*-ανωνυμοποίηση, περιλαμβάνει δύο βήματα. Το πρώτο περιλαμβάνει την διαδικασία δημιουργίας των κλάσεων ισοδυναμίας, μια διαδικασία που έχει ως βάση αυτή της δημιουργίας *kd*-δέντρων (*kd-trees* ή *k-dimensional trees*). Τα *kd*-δεντρα είναι δυαδικές δομές δεδομένων, τα οποία χρησιμοποιούνται για την οργάνωση πολυδιάστατων δεδομένων και ουσιαστικά παριστάνουν την διαδοχική υποδιαίρεση του χώρου σε υποχώρους.

Η κατασκευή των δέντρων αυτών γίνεται διαμερίζοντας αναδρομικά το σύνολο των δεδομένων ως προς κάποια διάσταση που επιλέγεται σε κάθε αναδρομική κλήση. Η τιμή του άξονα στην οποία θα γίνει η διαμέριση πρέπει υποχρεωτικά να έχει υπαρκτή αντιστοίχιση με τουλάχιστον ένα από τα δεδομένα του εξεταζόμενου συνόλου. Αυτή η μέθοδος εφαρμόζεται και πάνω στα προς ανωνυμοποίηση δεδομένα, επιλέγοντας σε κάθε βήμα την διάμεσο των τιμών του συνόλου που αντιστοιχούν στον επιλεγμένο άξονα. Η διαδικασία αυτή εφαρμόζεται σε κάθε σύνολο που παράγεται ξεχωριστά έως ότου να μην μπορεί να γίνει περεταίρω διαχωρισμός. Αυτό καθορίζεται από το γεγονός ότι η δημιουργία αυτών των συνόλων αποσκοπεί στον σχηματισμό των κλάσεων ισοδυναμίας, οι οποίες σύμφωνα με την *k*-ανωνυμία έχουν τον περιορισμό να μην έχουν μέγεθος μικρότερο από *k*. Το δεύτερο βήμα της *Mondrian k*-ανωνυμοποίησης, περιλαμβάνει τον ορισμό μιας αντιπροσωπευτικής του επιλεγμένου συνόλου εγγραφών τιμής για κάθε πεδίο που ανήκει στο ψευδο-αναγνωριστικό και πρόκειται για την γενίκευση των τιμών της κάθε κλάσης ισοδυναμίας. Για τον σκοπό αυτό, επιλέγεται είτε η στατιστική εύρους αντιπροσωπεύοντας ένα σύνολο τιμών με

το εύρος [ελάχιστο – μέγιστο] των τιμών αυτών, είτε η στατιστική μέσης τιμής αντιπροσωπεύοντας ένα σύνολο τιμών με την μέση τιμή του.

Ο κατασκευαστής των αντικειμένων αυτής της κλάσης δέχεται σαν ορίσματα τα δεδομένα προς ανωνυμοποίηση στην μορφή `FileData`, καθώς και την επιθυμητή τιμή της παραμέτρου `k` του αλγορίθμου. Η ανωνυμοποίηση επιτυγχάνεται με την μέθοδο `constructKdTree` και ξεκινά με την κλήση αυτής από τον κατασκευαστή της `kData`. Τα ανωνυμοποιημένα δεδομένα καταλήγουν στο πεδίο `kAnonData` το οποίο περιλαμβάνει έναν πίνακα για κάθε γραμμή των δεδομένων με τον ίδιο τρόπο που αντιστοιχούνται τα αρχικά δεδομένα στο πεδίο `initData` της κλάσης `FileData`. Τα δεδομένα αυτά διαφέρουν από τα αρχικά ως προς το ότι έχουν γενικευθεί κατά την διαδικασία της ανωνυμοποίησης.

Συμπληρωματικά με το πεδίο `kAnonData` στην περιγραφή του αποτελέσματος του αλγορίθμου λειτουργεί το πεδίο `QIGroups`. Αυτό το πεδίο περιλαμβάνει ουσιαστικά την αντιστοίχιση της κάθε γραμμής του `kAnonData` με μια κλάση ισοδυναμίας. Αυτή η πληροφορία παρέχεται μέσω ζευγών, των οποίων το πρώτο στοιχείο είναι ένας επαναλήπτης του διανύσματος `kAnonData` και αφορά την εκάστοτε γραμμή, ενώ το δεύτερο ένας ακέραιος ο οποίος αντιπροσωπεύει την κλάση ισοδυναμίας στην οποία κατέταξε ο αλγόριθμος την συγκεκριμένη γραμμή – εγγραφή. Ο αριθμός αυτός προσδίδεται στην κάθε κλάση με την σειρά δημιουργίας της κατά την εκτέλεση του αλγορίθμου. Ο τύπος του επαναλήπτη του διανύσματος `kAnonData` έχει ονομαστεί για την διευκόλυνση της περαιτέρω χρήσης του `QIGroupIt` και χρησιμοποιείται σε αρκετά μεγάλο βαθμό στην υλοποίηση του αλγορίθμου της `k`-ανωνυμίας.

Τελειώνοντας με ότι αφορά τα δεδομένα εξόδου, έχουν υλοποιηθεί οι μέθοδοι που υπαγορεύει το `QAbstractTableModel`, όπως έχουν περιγραφεί και σε προηγούμενες αντίστοιχες περιπτώσεις. Στην μέθοδο `data` η οποία περιλαμβάνεται σε αυτές υπεισέρχεται μια διαφοροποίηση λόγω των δύο δομών που φιλοξενούν τα δεδομένα εισόδου, τα οποία όμως τελικά είναι επιθυμητό να παρουσιαστούν τελικά μέσω ενός ενιαίου πίνακα. Πρέπει δηλαδή να παρουσιαστούν τα δεδομένα του `kAnonData` πεδίου, προσθέτοντας στην αρχή άλλη μια στήλη με τον αριθμό της κλάσης ισοδυναμίας. Για να επιτευχθεί αυτό, στο τμήμα της `data` το οποίο αφορά τον ρόλο

DisplayRole ελέγχεται μέσω του QModelIndex αν αναζητώνται δεδομένα της πρώτης στήλης οπότε και αυτά λαμβάνονται από το δεύτερο στοιχείο των ζευγών που περιλαμβάνονται στο QIGroups. Αν πρόκειται για οποιαδήποτε άλλη στήλη, τα δεδομένα λαμβάνονται από τον πίνακα-στοιχείο του kAnonData που υποδεικνύει ο επαναλήπτης του πρώτου στοιχείου των ζευγών του QIGroups. Η γραμμή λαμβάνεται πάντοτε από το QIGroups λόγω του ότι, όπως θα φανεί και παρακάτω στην ανάλυση του αλγορίθμου, το πεδίο αυτό γεμίζει καθώς η κάθε εγγραφή εντάσσεται σε κάποια κλάση και αυτό γίνεται με την ολοκλήρωση της δημιουργίας της κάθε κλάσης. Έτσι τα δεδομένα είναι τελικά ταξινομημένα ως προς την κλάση στο διάνυσμα αυτό και επιτρέπεται έτσι η άμεση παρουσίαση των δεδομένων με τέτοιο τρόπο ώστε να είναι συγκεντρωμένες οι εγγραφές που ανήκουν στην ίδια κλάση ισοδυναμίας.

Αφού λοιπόν περιγράφηκαν τα πεδία τα οποία περιλαμβάνουν την πληροφορία για το αποτέλεσμα της k-ανωνυμοποίησης, στο σημείο αυτό θα γίνει μια περιγραφή του τρόπου που γεμίζουν οι δομές αυτές μέσω της μεθόδου constructKdTree που είναι η κεντρική μέθοδος του αλγορίθμου, καθώς και των μεθόδων generalize και selectAxis που λειτουργούν βοηθητικά σε αυτή την διαδικασία. Η constructKdTree είναι μια αναδρομική διαδικασία η οποία λαμβάνει δύο ορίσματα: το πρώτο είναι ο QIGroupIt begin, ενώ το δεύτερο είναι ένας ακέραιος size και μαζί περιγράφουν το σύνολο των εγγραφών του οποίου θα διαχειριστεί η κλήση της συνάρτησης. Η πρώτη κλήση γίνεται στον κατασκευαστή και τα δεδομένα τα οποία καλείται να διαχειριστεί είναι τα δεδομένα όλου του πεδίου kAnonData, τα οποία τη στιγμή που γίνεται η πρώτη κλήση είναι ακριβές αντίγραφο των δεδομένων εισόδου.

Πρώτο βήμα είναι η επιλογή του άξονα, το οποίο αναλαμβάνει να εκτελέσει η μέθοδος selectAxis. Κάθε φορά επιλέγεται ως άξονας το πεδίο με το μέγιστο κανονικοποιημένο εύρος στο επιλεγμένο σύνολο. Έτσι η συνάρτηση selectAxis μέσω των ορισμάτων της, με τρόπο πανομοιότυπο με την μέθοδο constructKdTree, δέχεται το προς εξέταση σύνολο εγγραφών και υπολογίζοντας για κάθε στήλη που αντιστοιχεί στο ψευδο-αναγνωριστικό το μέγιστο κανονικοποιημένο εύρος, επιστρέφει τον αριθμό που αντιστοιχεί στην θέση της στήλης που βρέθηκε η μέγιστη τιμή.

Έπειτα, γίνεται ο έλεγχος ύπαρξης επιτρεπτής διαμέρισης, ο οποίος αποτελεί και την συνθήκη τερματισμού της αναδρομής. Ελέγχεται δηλαδή αν το μέγεθος των συνόλων που θα σχηματιστούν με την διχοτόμηση είναι μεγαλύτερο από k . Κάθε φορά, ως σημείο του διαχωρισμού επιλέγεται η διάμεσος, η οποία σαφώς πληροί την προϋπόθεση να υπάρχει εγγραφή που να αντιστοιχεί σε αυτή την τιμή, χωρίζοντας το αρχικό σύνολο εγγραφών στην μέση. Έτσι το μέγεθος των συνόλων που θα σχηματιστούν αν γίνει ο διαχωρισμός αυτός θα είναι το μισό από το μέγεθος του αρχικού συνόλου και αν δεν είναι μεγαλύτερο από k , αφού σταματά η αναδρομή έχει προκύψει ένα σύνολο που αποτελεί μια κλάση ισοδυναμίας. Τέλος, γίνεται γενίκευση αυτού με την στατιστική του εύρους τιμών μέσω της μεθόδου `generalize`.

Η μέθοδος `generalize` δέχεται ως ορίσματα τους επαναλήπτες τύπου `QIGroupIt` που αντιστοιχούν στην αρχή και στο τέλος του συνόλου των εγγραφών στο διάνυσμα `kAnonData`, καθώς και τον αριθμό που αντιστοιχεί στην κατασκευαζόμενη κλάση ισοδυναμίας. Στην `generalize` γίνεται ο υπολογισμός του εύρους των τιμών κάθε πεδίου εκτός του ευαίσθητου γνωρίσματος και δημιουργείται μια συμβολοσειρά στην μορφή `[min-max]` που αντικαθιστά τις αρχικές τιμές. Σε αυτή την συνάρτηση λαμβάνεται μέριμνα για την διαχείριση ενός άλλου πεδίου της κλάσης `kData`, του πεδίου `QIGroupRanges`. Το πεδίο αυτό είναι ένα διάνυσμα στοιχείων με τύπο `AttributeRangesStruct`, ο οποίος πρόκειται για μια δομή που περιλαμβάνει τον αριθμό που αντιπροσωπεύει μια κλάση ισοδυναμίας (`QIGroupId`) και ένα διάνυσμα με το εύρος των τιμών κάθε πεδίου της συγκεκριμένης κλάσης. Καθώς, λοιπόν στην μέθοδο `generalize` γίνεται υπολογισμός της μέγιστης και ελάχιστης τιμής κάθε πεδίου της κλάσης ισοδυναμίας, μπορεί να υπολογιστεί άμεσα η διαφορά τους και να αποθηκευθεί στο διάνυσμα που περιλαμβάνεται με το όνομα `Ranges` στην προαναφερθείσα δομή. Αφού και το `QIGroupId` έχει ληφθεί ως όρισμα της μεθόδου `generalize`, είναι και αυτό άμεσα διαθέσιμο και επομένως δημιουργείται ένα στοιχείο που προστίθεται στο διάνυσμα `QIGroupRanges`. Το πεδίο αυτό και η διαδικασία που μόλις περιεγράφηκε εξυπηρετεί τον μετέπειτα υπολογισμό των NCP-GCP μετρικών κατά την παραγωγή των στατιστικών που αφορούν την εκτέλεση του αλγορίθμου της k -ανωνυμοποίησης.

Αν η διαμέριση κριθεί επιτρεπτή, η εκτέλεση της μεθόδου `constructKdData` συνεχίζεται υπολογίζοντας την τιμή της διαμέσου και δημιουργώντας βάση αυτής δύο σύνολα. Το πρώτο σύνολο περιέχει τις εγγραφές που είναι μικρότερες από την διάμεσο των τιμών του αρχικού συνόλου για το επιλεγμένο πεδίο, ενώ το δεύτερο όσες είχαν μεγαλύτερη από την διάμεσο τιμή. Τέλος, η συνάρτηση ξανακαλείται μια φορά για το πρώτο σύνολο και μία για το δεύτερο.

Επιπρόσθετα όσων αναφέρθηκαν, η κλάση `kData` περιλαμβάνει ένα πεδίο τύπου `boolean`, το οποίο ονομάζεται `created` και του οποίου η τιμή επηρεάζεται από τον κατασκευαστή της κλάσης. Το πεδίο αυτό υποδυκνύει αν δημιουργήθηκε επιτυχώς και σωστά αρχικοποιημένο ένα αντικείμενο της κλάσης ή αν προέκυψε κάποιο πρόβλημα όπως η περίπτωση που η παράμετρος `k` υπερβαίνει τον αρχικό αριθμό εγγραφών και η επίτευξη `k`-ανωνυμίας είναι αδύνατη. Τέλος, το πεδίο `time` αντιστοιχεί στην διάρκεια εκτέλεσης του αλγορίθμου της `k`-ανωνυμίας σε δευτερόλεπτα και υπολογίζεται και αυτό στον κατασκευαστή του αντικειμένου, θέτοντας έναρξη πριν την κλήση της `constructKdTree` και λήξη μόλις τελειώσει συνολικά η αναδρομική εκτέλεση αυτής.

5.2.4 Κλάσεις για στατιστικά δειγμάτων

Πρόκειται για τις κλάσεις `AnatomySamplingStatistics` και `kSamplingStatistics`. Αυτές περιλαμβάνουν ένα πεδίο `stats` που αφορά τα συλλεγμένα στατιστικά και στην πρώτη κλάση είναι τύπου `AnatomyStatistics`, ενώ στην δεύτερη είναι τύπου `kStatistics`. Ακόμα περιλαμβάνουν ένα ακέραιο πεδίο που αντιστοιχεί στην τιμή της παραμέτρου `l` και `k` αντίστοιχα για τις οποίες υπολογίσθηκαν τα συγκεκριμένα στατιστικά. Οι κατασκευαστές των κλάσεων αυτών παίρνουν ως όρισμα τιμές που αντιστοιχούν στα δύο αυτά πεδία και περιλαμβάνουν μόνο την αρχικοποίηση αυτών. Τέλος ορίζεται η υπερφόρτωση του τελεστή σύγκρισης `<` για τις κλάσεις αυτές ώστε τα αντικείμενά τους να μπορούν να διαταχθούν άμεσα με βάση την τιμή `l` ή `k` αντίστοιχως, με σκοπό την καλύτερη οπτικοποίηση τους στους αντίστοιχους πίνακες αλλά και την ευκολότερη δημιουργία διαγραμμάτων. Σε ότι αφορά την χρήση αυτών, υπάρχουν

στην κλάση του κυρίου παραθύρου της εφαρμογής (**MainWindow**) δύο πεδία που αντιστοιχούν σε διανύσματα που το ένα περιλαμβάνει αντικείμενα της κλάσης **AnatomySamplingStatistics**, ενώ το άλλο αντικείμενα της κλάσης **kSamplingStatistics**. Τα διανύσματα αυτά αρχικοποιούνται κάθε φορά που δημιουργείται ένα δείγμα, ενώ προστίθενται αντικείμενα σε αυτά κάθε φορά που ο χρήστης τρέχει τον τύπο ανωνυμοποίησης που τους αντιστοιχεί στο δείγμα.

Η παρουσίαση αυτών γίνεται με δύο τρόπους. Ο ένας είναι σε πλέγματα τύπου **QTableWidget**. Ο δεύτερος είναι γραφήματα ράβδων που αφορούν το μέγεθος των κλάσεων ισοδυναμίας, την μετρική GCP για την k ανωνυμοποίηση, το πλήθος των διαφορετικών τιμών ανά κλάση ισοδυναμίας και τον χρόνο εκτέλεσης. Για να επιτευχθεί αυτό γίνεται χρήση των κλάσεων που προσφέρονται από το Qwt project και συγκεκριμένα της κλάσης **QwtPlotHistogram**, η οποία αφορά ιστογράμματα. Η δυνατότητα δημιουργίας ιστογραμμάτων είναι ότι πλησιέστερο στο ζητούμενο παρέχεται από την συγκεκριμένη έκδοση Qwt και για την χρήση της με τον επιθυμητό τρόπο δημιουργήθηκε η κλάση **BarChart** που την κληρονομεί.

Ο κατασκευαστής της δέχεται σαν ορίσματα τον τίτλο του διαγράμματος και το χρώμα των ράβδων. Αυτός καλεί τον κατασκευαστή της **QwtPlotHistogram** με τον δεδομένο από τα ορίσματα τίτλο, ορίζει μέσω της κληρονομούμενης **setStyle** πως η επιθυμητή απεικόνιση είναι σε στήλες και τέλος μέσω της μεθόδου **setColor** ορίζει παραμέτρους των ράβδων (π.χ. ότι θα έχουν σχήμα κουτιού) και τα χρώματα του γραφήματος. Περιέχει ακόμα μια public μέθοδο, την **setValues**. Αυτή παίρνει ως όρισμα ένα διάνυσμα το οποίο περιλαμβάνει τα ζεύγη τιμών που αντιστοιχούν στον οριζόντιο και κάθετο άξονα που πρέπει να απεικονιστούν στο γράφημα. Το δεύτερο όρισμα αυτής είναι το ζητούμενο πλάτος των ράβδων. Καθότι η κλάση που χρησιμοποιήθηκε είχε σκοπό την δημιουργία ενός ιστογράμματος και όχι διαγράμματος ράβδων δεν λαμβάνει μια μόνο τιμή για τον οριζόντιο άξονα αλλά ένα ζεύγος (x_1, x_2) που αντιστοιχεί σε κάθε περιοχή τιμών του ιστογράμματος. Έτσι για να δημιουργηθεί το διάγραμμα ράβδων από την τιμή x που λαμβάνεται από το πρώτο στοιχείο κάθε ζεύγους στο διάνυσμα των τιμών που θα απεικονιστούν, δημιουργείται ένα διάστημα ($x - \text{width}/2, x + \text{width}/2$) για κάθε ράβδο, όπου width το πλάτος που δίνεται σαν δεύτερο όρισμα. Το ύψος της κάθε ράβδου ορίζεται από το δεύτερο

στοιχείο του ζεύγους στο διάνυσμα των δεδομένων τιμών. Έτσι λοιπόν αυτή η κλάση παρέχει τις απαραίτητες μεθόδους για την δημιουργία των διαγραμμάτων που αφορούν τα στατιστικά των δειγμάτων.

5.2.5 Κλάσεις παραθύρων

Μια κατηγορία κλάσεων που περιλαμβάνει η εφαρμογή και αφορά όχι την εκτέλεση των αλγορίθμων αλλά την αλληλεπίδραση με τον χρήστη, είναι οι κλάσεις τύπου παραθύρου. Αυτές, έχουν το κοινό στοιχείο ότι κληρονομούν το `QMainWindow` και περιλαμβάνουν widgets όπως `menu`, κουμπιά, επιλογείς, πίνακες και όψεις πινάκων δίνοντας στο χρήστη την δυνατότητα να επιλέξει παραμέτρους και παρουσιάζοντας σε αυτόν δεδομένα και αποτελέσματα. Μέσα από τις κλάσεις αυτές γίνεται η δημιουργία αντικειμένων των κλάσεων που προαναφέρθηκαν. Η εκτέλεση της εφαρμογής ξεκινά με την δημιουργία και παρουσίαση του αντικειμένου `MainWindow` από το οποίο γίνεται η είσοδος των δεδομένων, η επιλογή των δειγμάτων, η επιλογή του αλγορίθμου ανωνυμοποίησης και η πλοήγηση στο παράθυρο των στατιστικών που αφορούν ένα δείγμα, την οπτικοποίηση της ιεραρχίας κλπ. Οι υπόλοιπες κλάσεις τέτοιου τύπου είναι οι εξής:

- `AnatomyWindow`: αφορά την εκτέλεση ανωνυμοποίησης με ανατομία, την παρουσίαση του αποτελέσματος και την δυνατότητα εξαγωγής αυτού σε `.csv` αρχεία. Ακόμη μέσω των επιλογών αυτού του παραθύρου είναι δυνατό να εμφανισθεί το παράθυρο που αφορά τα αντίστοιχα στατιστικά.
- `kWindow`: περιλαμβάνει αντίστοιχες λειτουργίες με την προηγούμενη κλάση με την διαφορά ότι αφορά την `k`-ανωνυμία.
- `HierarchyWindow`: αφορά την φόρτωση από αρχείο και οπτικοποίηση μιας ιεραρχίας σε δενδρική δομή. Στο παράθυρο αυτό, για την ανάγνωση του αρχείου και τη φόρτωση των δεδομένων της ιεραρχίας χρησιμοποιείται η κλάση `data` η οποία έχει ενσωματωθεί στο χώρο ονομάτων `DGH`. Την οπτικοποίηση της δεδομένης ιεραρχίας επιτρέπει το στοιχείο `QTreeWidget`.
- `SamplingStatsWindow`: περιλαμβάνει έξι `tabs` που απεικονίζουν στατιστικά συλλεγμένα για κάποιο δεδομένο δείγμα εγγραφών καθώς και την εξαγωγή αυτών σε αρχείο.

6

Έλεγχος

Σε αυτό το σημείο, έχουν αναλυθεί λεπτομερώς ζητήματα, που αφορούν τόσο την λειτουργικότητα της εφαρμογής όσο και της υλοποίησής της. Κρίνεται λοιπόν σκόπιμο να περιγραφεί ένα σενάριο λειτουργίας το οποίο θα εγγυάται την ορθή λειτουργία του προγράμματος δίνοντας βάση τόσο στα αποτελέσματα της ανωνυμοποίησης όσο και στον χρόνο ο οποίος χρειάστηκε για το πέρας της.

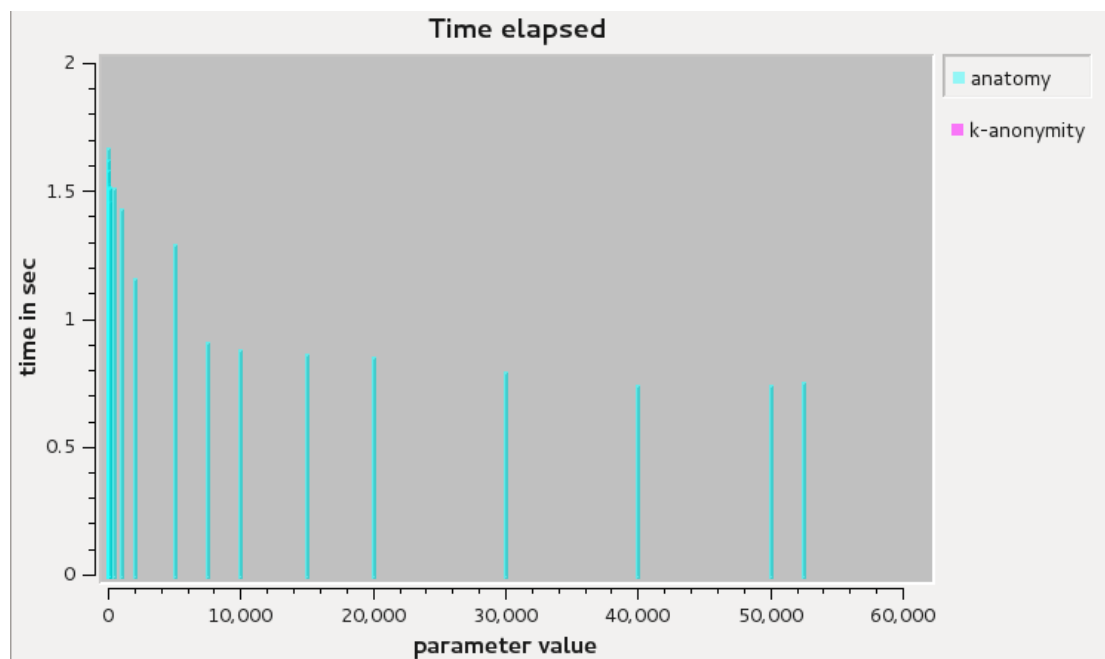
6.1 Μεθοδολογία ελέγχου

Για να πραγματοποιήσουμε τον έλεγχο της εφαρμογής, χρησιμοποιήθηκε ένα census dataset το οποίο παρέχει δύο .csv αρχεία μεγέθους 220KB και 45MB αντίστοιχα, για την πραγματοποίηση των πειραμάτων. Επειδή, όπως έχει ήδη αναφερθεί, ο αλγόριθμος της k-ανωνυμίας όπως έχει υλοποιηθεί στην παρούσα εργασία, εφαρμόζεται μόνο σε αριθμητικά δεδομένα, τα παραπάνω σύνολα δεδομένων έχουν τροποποιηθεί κατάλληλα για την δυνατότητα χρήσης τους από το εργαλείο ανωνυμοποίησης.

Στην συνέχεια, παρουσιάζονται τα διαγράμματα απώλειας πληροφορίας και χρόνου ολοκλήρωσης των αλγορίθμων που υλοποιήθηκαν.

6.1.1 Ανατομία

Στον αλγόριθμο της ανατομίας, δεν υπάρχει απώλεια πληροφορίας μιας και δεν εφαρμόζεται σε κανένα βήμα γενίκευση δεδομένων. Το ενδιαφέρον, λοιπόν, έγκειται στον χρόνο εκτέλεσης της ανωνυμοποίησης από το εργαλείο μας (Διάγραμμα 6.1).



Διάγραμμα 6.1 - Χρόνος ολοκλήρωσης ανατομίας στο census dataset συναρτήσει της παραμέτρου ποικιλομορφίας l

Στο παραπάνω διάγραμμα, έχουμε τους χρόνους εκτέλεσης του αλγορίθμου της ανατομίας, χρησιμοποιώντας τιμές της παραμέτρου ποικιλομορφίας l , οι οποίες κυμαίνονται από 5 μέχρι 52,500. Παρατηρώντας το διάγραμμα, μπορούμε να δούμε πως παρόλο τον μεγάλο όγκο των δεδομένων η διαδικασία ανωνυμοποίησής τους δεν διαρκεί ποτέ πάνω από 1.7 δευτερόλεπτα. Επίσης, είναι σαφές ότι, με εξαίρεση μικρές διακυμάνσεις, όσο αυξάνεται η τιμή του l , μειώνεται ταυτόχρονα και ο χρόνος εκτέλεσης του αλγορίθμου φτάνοντας μάλιστα σε τιμές αρκετά κοντά στο μισό δευτερόλεπτο.

Στην συνέχεια ακολουθούν οι ακριβείς τιμές από τις οποίες προέκυψε το παραπάνω διάγραμμα, όπως αυτές εξήχθησαν από την εφαρμογή:

I-value	QI Group Size				Max Diversity	Min Diversity	Total QI Groups	Total Sensitive Values	Time Elapsed
	Max	Min	Average	Median					
2	2	2	2	2	2	2	47565	52461	1.36
5	5	5	5	5	5	5	19026	52461	1.28
10	10	10	10	10	10	10	9513	52461	1.29
12	13	12	12.0008	12	13	12	7927	52461	1.66
19	20	19	19.0032	19	20	19	5006	52461	1.61
20	21	20	20.0021	20	21	20	4756	52461	1.52
30	30	30	30	30	30	30	3171	52461	1.24
40	41	40	40.0042	40	41	40	2378	52461	1.5
50	52	50	50.0158	50	52	50	1902	52461	1.57
100	102	100	100.032	100	102	100	951	52461	1.51
120	122	120	120.114	120	122	120	792	52461	1.45
150	152	150	150.047	150	152	150	634	52461	1.5
500	503	500	500.684	500	503	500	190	52461	1.5
1000	1005	1000	1001.37	1001	1005	1000	95	52461	1.42
2000	2039	2012	2024.04	2023	2039	2000	47	52461	1.15
5000	5010	5004	5006.84	5007	5005	5000	19	52461	1.28
7500	7955	7903	7927.5	7926	7903	7552	12	52461	0.9
10000	10612	10550	10570	10564	10437	10000	9	52461	0.87
15000	15881	15824	15855	15856	15228	15000	6	52461	0.85
20000	23844	23693	23782.5	23796	21695	21632	4	52461	0.84
30000	47574	47556	47565	47565	36358	36348	2	52461	0.78
40000	95130	95130	95130	95130	52461	52461	1	52461	0.73
50000	95130	95130	95130	95130	52461	52461	1	52461	0.73
52461	95130	95130	95130	95130	52461	52461	1	52461	0.74

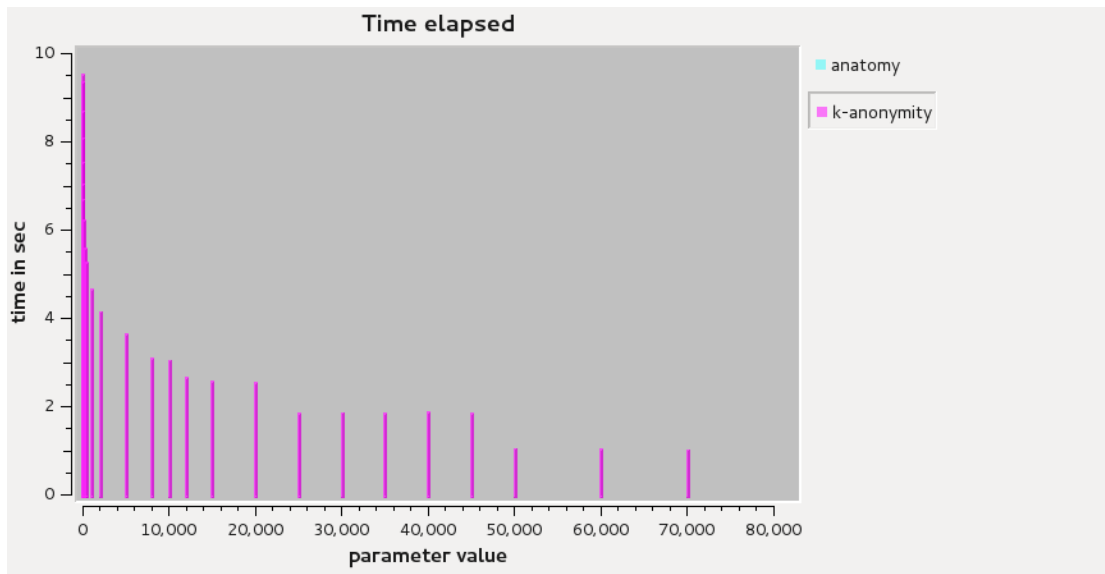
6.1.2 *k*-ανωνυμία

Στον αλγόριθμο της *k*-ανωνυμίας, σε αντίθεση με αυτόν της ανατομίας, έχουμε γενίκευση δεδομένων με αποτέλεσμα την απώλεια πληροφορίας από τα ανωνυμοποιημένα δεδομένα. Παράλληλα, λοιπόν με τον χρόνο εκτέλεσης της τεχνικής (Διάγραμμα 6.2) θα μελετήσουμε και το ποσοστό απώλειας πληροφορίας στα τελικά δεδομένα χρησιμοποιώντας ως μετρική το Συνολική Ποινή Βεβαιότητας (GCP) (Διάγραμμα 6.3).



Διάγραμμα 6.2 - Ποσοστό απώλειας δεδομένων συναρτήσει της παραμέτρου *k*

Τα παραπάνω αποτελέσματα έχουν προκύψει εφαρμόζοντας *k*-ανωνυμία πάνω στο census dataset, με τιμές της παραμέτρου *k* να κυμαίνονται μεταξύ 5 και 70,000. Μπορούμε εύκολα να δούμε πως για μικρές τιμές του *k* η απώλεια πληροφορίας είναι ελάχιστη (κοντά στο 1.2%), ενώ καθώς αυτή μεγαλώνει αυξάνεται και ο όγκος δεδομένων που χάνεται, φτάνοντας σε ένα μέγιστο για την τιμή *k*=50,000 λίγο κάτω από 7%. Κάτι τέτοιο είναι αναμενόμενο, καθώς με την αύξηση του μεγέθους των κλάσεων ισοδυναμίας έχουμε και περισσότερες τιμές να γενικεύονται στο ίδιο ανωνυμοποιημένο γνώρισμα.



Διάγραμμα 6.3 - Χρόνος ολοκλήρωσης ανατομίας στο census dataset συναρτήσει της παραμέτρου k

Τέλος, για τις ίδιες τιμές της παραμέτρου k με προηγουμένως προκύπτει το παραπάνω διάγραμμα χρόνου ολοκλήρωσης του αλγορίθμου της k -ανωνυμίας. Παρατηρούμε πως οι χρόνοι είναι σαφώς μεγαλύτεροι από αυτούς του αλγορίθμου της ανατομίας. Η αύξηση αυτή οφείλεται τόσο στην ανάγκη γενίκευσης τιμών όσο και στην διχοτόμηση των δεδομένων σε κλάσεις ισοδυναμίας. Πιο συγκεκριμένα βλέπουμε, πως ο αλγόριθμος δεν ξεπερνά ποτέ σε χρόνο εκτέλεσης τα 9 δευτερόλεπτα και φτάνει σε ελάχιστο λίγο πάνω από 1 δευτερόλεπτο.

Παρακάτω παραθέτονται οι ακριβείς τιμές των μετρικών που υπολογίσθηκαν για αυτό το σύνολο δεδομένων όπως εξήχθηκαν από την εφαρμογή:

k-value	QI Group Size				GCP	Total QI Groups	Total Sensitive Values	Time Elapsed
	Max	Min	Average	Median				
2	3	2	2	3	0.0147422	32768	17	9.48
3	5	3	3	3	0.0151062	29594	17	9.31
5	6	5	5	6	0.017657	16384	17	8.65
10	12	11	11	12	0.0199756	8192	17	8.04
20	24	23	23	23	0.0217716	4096	17	7.49
30	47	46	46	46	0.0236804	2048	17	7
70	93	92	92	93	0.0256756	1024	17	6.63
100	186	185	185	186	0.027763	512	17	6.18
300	372	371	371	372	0.0300027	256	17	5.54
500	744	743	743	743	0.0323859	128	17	5.21
1000	1487	1486	1486	1486	0.035177	64	17	4.62
2000	2973	2972	2972	2973	0.0380974	32	17	4.11
5000	5946	5945	5945	5946	0.0415694	16	17	3.6
8000	11892	11891	11891	11891	0.0469355	8	17	3.05
10000	11892	11891	11891	11891	0.0469355	8	17	3.01
12000	23783	23782	23782	23782	0.0541539	4	17	2.62
15000	23783	23782	23782	23782	0.0541539	4	17	2.52
20000	23783	23782	23782	23782	0.0541539	4	17	2.5
25000	47565	47565	47565	47565	0.0607382	2	17	1.81
30000	47565	47565	47565	47565	0.0607382	2	17	1.82
35000	47565	47565	47565	47565	0.0607382	2	17	1.81
40000	47565	47565	47565	47565	0.0607382	2	17	1.82
45000	47565	47565	47565	47565	0.0607382	2	17	1.81
50000	95130	95130	95130	95130	0.0666667	1	17	1.02
60000	95130	95130	95130	95130	0.0666667	1	17	0.98
70000	95130	95130	95130	95130	0.0666667	1	17	0.98

7

Επίλογος

7.1 Σύνοψη και συμπεράσματα

Συνοψίζοντας, έχει καταστεί πλέον σαφής η σημασία της χρήσης τεχνικών ανωνυμοποίησης για την προστασία της ταυτότητας των ατόμων με παρουσία σε δημοσιευμένα δεδομένα. Η παρούσα εργασία ικανοποιεί δύο σκοπούς: Την παρουσίαση του θεωρητικού υπόβαθρου των σημαντικότερων τεχνικών ανωνυμοποίησης την στιγμή συγγραφής αυτού του κειμένου καθώς και την δημιουργία ενός εύκολα επεκτάσιμου εργαλείου με υλοποιημένους δύο από τους βασικότερους αλγορίθμους ανωνυμοποίησης σε αυτή την βιβλιογραφία: την τεχνική της ανατομίας και αυτή της k-ανωνυμίας.

Επίσης, δίνεται η δυνατότητα στον χρήστη να τρέξει τις διάφορες μεθόδους, σε επιλεγμένο από αυτόν δείγμα των δεδομένων, ώστε να είναι σε θέση να διαλέξει τον αλγόριθμο που αναλογεί στην εκάστοτε περίπτωση.

Ταυτόχρονα, το εργαλείο έχει εμπλουτιστεί με κλάσεις και μεθόδους για τον υπολογισμό στατιστικών με στόχο την ανάλυση της απόδοσης του αλγορίθμου της ανωνυμοποίησης που χρησιμοποιήθηκε, καθώς και για την εύρεση των κατάλληλων παραμέτρων για τον υπολογισμό των προς δημοσίευση δεδομένων με την επιθυμητή αναλογία μεταξύ της πληροφορίας που χάνεται και του επιπέδου ασφαλείας που θέλουμε να εγγυηθούμε.

Τέλος, σε όλα τα παραπάνω, έχει προστεθεί η επιλογή για την δημιουργία διαγραμμάτων σε κάθε βήμα για την εποπτική εκτίμηση των αποτελεσμάτων.

Φυσικά, υπάρχει η ανάγκη για προσθήκη επιπλέον τεχνικών ανωνυμοποίησης. Κυριότερα παραδείγματα τέτοιων μεθόδων είναι η υλοποίηση της m-αμεταβλητότητας για την ανωνυμοποίηση δεδομένων τα οποία ενημερώνονται σε τακτά χρονικά διαστήματα, η k^m-ανωνυμία [4] για την προστασία δεδομένων τα οποία βρίσκονται σε οργανωμένα σε σύνολα και μέθοδοι διαφορικής ανωνυμίας για την προσθήκη στατιστικού θορύβου στις απαντήσεις των ερωτημάτων που γίνονται πάνω σε δημοσιευμένα δεδομένα. Ένα πρώτο βήμα για την περαιτέρω ανάπτυξη της εφαρμογής έχει ήδη γίνει, καθώς έχει υλοποιηθεί διαπροσωπία για την παρουσίαση ιεραρχιών γενίκευσης σε δενδρική μορφή, οι οποίες έχουν οριστεί από τον χρήστη.

7.2 Μελλοντικές επεκτάσεις

Πέρα από την προσθήκη των αλγορίθμων που αναφέρθηκαν στην παραπάνω ενότητα, σημαντική προσθήκη θα ήταν ο οποιοσδήποτε αλγόριθμος ανωνυμοποίησης της βιβλιογραφίας, με στόχο την δημιουργία ενός εργαλείου το οποίο θα καλύπτει όλες τις ανάγκες για προστασία δεδομένων. Επίσης, θα μπορούσαν να προστεθούν επιπλέον μετρικές, διαγράμματα και στατιστικά για την ακόμα καλύτερη κατανόηση και ευκολότερη παραμετροποίηση των αναγκών που ενδέχεται να παρουσιαστούν. Σημαντική προσθήκη θα είναι επίσης η δυνατότητα της απ' ευθείας σύνδεσης της εφαρμογής με την βάση δίνοντας και άλλες επιλογές για την εισαγωγή των δεδομένων, διευκολύνοντας και την διαδικασία επιλογής των εγγραφών οι οποίες θέλουμε να ανωνυμοποιηθούν.

8

Βιβλιογραφία

- [1] L. Sweeney.
“*k-Anonymity: A Model for Protecting Privacy.*”
International Journal of Uncertainty Fuzziness and Knowledge-Based Systems,
10(5):557{570, 2002.
- [2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian.
“*l-Diversity: Privacy Beyond k-Anonymity.*”
In *Proc. of ICDE*, 2006.
- [3] X. Xiao and Y. Tao.
“*Anatomy: Simple and effective privacy preservation.*”
In *VLDB*, pages 139–150, 2006.
- [4] M.Terrovitis, N.Mamoulis and P.Kalnis
“*Privacy preserving Anonymization of Set-valued Data.*”
In *Proc. Of VLDB*, 2008
- [5] M.E. Nergiz and C. Clifton.
“*Thoughts on k-Anonymization.*”
In *DKE*, 63(3):622{645, Elsevier Science,2007
- [6] X. Xiao and Y. Tao
“*m-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets.*”
In *Proc. ACM SIGMOD*, pp.689-700, 2007
- [7] N. Li, T. Li and S. Venkatasubramanian.
“*T-closeness: privacy beyond k-anonymity and l-diversity.*”
In *Proceedings of the IEEE ICDE 2007*, 2007
- [8] M.E. Nergiz, M. Atzori and C. Clifton,
“*Hiding the Presence of Individuals from Shared Databases.*”
Proc. ACM SIGMOD, pp.665-676, 2007.

- [9] L. Sweeney.
“*Uniqueness of Simple Demographics in the U.S. Population*” LIDAP-WP4.
Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh,
PA: 2000
- [10] R.J. Bayardo, Jr., and R. Agrawal.
“*Data Privacy through Optimal k-Anonymization.*”
In Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp.217-228, 2005.
- [11] K. LeFevre, D. DeWitt and R. Ramakrishnan.
“*Incognito: Efficient full-domain k-anonymity*”
In Proc. Of SIGMOD, 2005
- [12] UTD anonymization toolbox page:
<http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=home>
- [13] Qt project official site:
<http://qt-project.org/>
- [14] QWT sourceforge repository:
<http://qwt.sourceforge.net/index.html>
- [15] Qt Model/View:
<http://qt-project.org/doc/qt-4.8/model-view-programming.html>
- [16] LeFevre, K., DeWitt, D.J., Ramakrishnan, R.,
“*Mondrian Multidimensional K-Anonymity*”
Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference
on, vol., no., pp.25,25, 03-07 April 2006
- [17] Gabriel Ghinita, Panagiotis Karras, Panos Kalnis and Nikos Mamoulis
«Fast data anonymization with low information loss”
In *Proceedings of the 33rd international conference on Very large data bases*
(VLDB '07). VLDB Endowment 758-769