ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Προστασία ιδιωτικότητας από επιτιθέμενους με συναθροιστική γνώση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΑΙΚΑΤΕΡΙΝΗΣ ΛΕΠΕΝΙΩΤΗ

Επιβλέπων : Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2013

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Προστασία ιδιωτικότητας από επιτιθέμενους με συναθροιστική γνώση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΑΙΚΑΤΕΡΙΝΗΣ ΛΕΠΕΝΙΩΤΗ

Επιβλέπων : Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 6η Μαρτίου 2013.

.....	.....	.....
Ιωάννης Βασιλείου Καθηγητής Ε.Μ.Π.	Αντώνιος Συμβώνης Καθηγητής Ε.Μ.Π.	Κώστας Κοντογιάννης Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2013

.....

ΑΙΚΑΤΕΡΙΝΗ ΛΕΠΕΝΙΩΤΗ

Διπλωματούχος Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών Ε.Μ.Π.

Περίληψη

Σε πολλούς οργανισμούς, επιχειρήσεις ή δημόσιους φορείς, η συλλογή και διαχείριση προσωπικών δεδομένων αποτελεί ένα πολύτιμο εργαλείο. Με τη δημιουργία τέτοιων συλλογών συγκεντρώνεται σημαντική πληροφορία αναφορικά με τον πληθυσμό που συμμετέχει στα δεδομένα. Η δημοσίευση της πληροφορίας αυτής είναι ιδιαίτερα χρήσιμη για ερευνητικούς σκοπούς και στατιστικές μελέτες.

Η παρούσα εργασία ασχολείται με την διασφάλιση της ιδιωτικότητας σε συλλογές προσωπικών δεδομένων μέσω της *k*-ανωνυμίας. Εστιάζει σε σύνολα δεδομένων όπου τα διάφορα γνωρίσματα αντιπροσωπεύουν ένα κοινό πεδίο πληροφορίας, γεγονός που ερμηνεύεται με γνωρίσματα του ψευδο-αναγνωριστικού προερχόμενα από το ίδιο πεδίο τιμών.

Επιχειρούμε την προστασία της ιδιωτικότητας από απειλές με συναθροιστική γνώση πάνω στις τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού, η οποία εκφράζεται μέσω κάποιας συναθροιστικής συνάρτησης.

Για το λόγο αυτό αναπτύσσουμε και εξετάζουμε αναδρομικό αλγόριθμο που υλοποιεί την *k*-ανωνυμοποίηση του δοθέντος συνόλου δεδομένων, λαμβάνοντας υπόψη την συναθροιστική συνάρτηση για την εύρεση της κατάλληλης τοπικής γενίκευσης σε κάθε κλάση ισοδυναμίας.

Ο αλγόριθμος εγγυάται την ικανοποίηση της *k*-ανωνυμίας από τα δημοσιευμένα δεδομένα, ως προς την συναθροιστική συνάρτηση, την οποία αντίστοιχοι αλγόριθμοι αγνοούν. Διατηρεί περισσότερη πληροφορία στα δημοσιευμένα δεδομένα από άλλους αλγορίθμους *k*-ανωνυμοποίησης. Ακόμα, δίνει τη δυνατότητα επιλογής της θεωρούμενης γνώσης του επιτιθέμενου χαλαρώνοντας την εγγύηση της ανωνυμίας, έτσι ώστε να προσφέρει έναν αποδοτικό συνδυασμό ιδιωτικότητας και χρηστικότητας.

Λέξεις Κλειδιά:

k-ανωνυμία

αλγόριθμος ανωνυμοποίησης

συναθροιστική γνώση

προστασία ιδιωτικότητας

Abstract

In many organizations, enterprises or public services, collecting and managing personal data is a valuable tool. By creating such collections, important information is gathered, regarding the population that participates in the dataset. Releasing this information is particularly useful for research and statistical studies.

This paper deals with ensuring privacy protection in collections of personal data by applying *k*-anonymity. We focus on datasets where various attributes represent the same kind of information, which is interpreted by a quasi-identifier set of attributes from the same domain.

We attempt to protect privacy from attacks with aggregate knowledge over the values of attributes coming from the quasi-identifier set, expressed by some aggregate function.

For this reason we develop and test a recursive algorithm that implements *k*-anonymization of the given dataset by taking into account the aggregate function, in order to find the appropriate local generalization in every equivalence class.

This algorithm guarantees *k*-anonymity protection with respect to the aggregate function value, which is not taken into account by other algorithms.

The algorithm preserves more information in the published dataset in comparison with classic k-anonymization algorithms. Furthermore, it gives the option to vary the attacker's knowledge, relaxing the required security level, in order to return an efficient combination of privacy and utility.

Keywords:

k-anonymity

anonymization algorithm

aggregate knowledge

privacy protection

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ.Ιωάννη Βασιλείου για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα και να αποκομίσω ουσιαστικά προσόντα μέσα από αυτή την εργασία.
Επιπλέον θα ήθελα να ευχαριστήσω την κ.Όλγα Γκουντούνα για την όμορφη συνεργασία μας, την καθοδήγηση που μου προσέφερε και την συνεχή δίψα για βελτίωση που μου μετέδωσε, κατά την εκπόνηση της παρούσας διπλωματικής εργασίας.

Τέλος ευχαριστώ την οικογένεια και του φίλους μου για την στήριξή τους, μα πάνω από όλα την μητέρα μου στην οποία οφείλω κάθε βήμα της ζωής μου.

Πίνακας περιεχομένων

1 Εισαγωγή 1

1.1 Γενικά 1

1.2 Αντικείμενο διπλωματικής 3

1.2.1 Συνεισφορά 4

1.3 Οργάνωση κειμένου 5

2 Θεωρητικό υπόβαθρο 7

2.1 Οργάνωση δεδομένων 7

2.1.1 Ψευδο-αναγνωριστικό 8

2.1.2 Ευαίσθητα γνωρίσματα 9

2.1.3 Ιεραρχία γενίκευσης πεδίου τιμών γνωρίσματος 10

2.1.4 Απειλές κατά της ιδιωτικότητας 11

2.1.5 Απώλεια πληροφορίας 12

2.2 Αναγνώριση ταυτότητας 14

2.2.1 k-Ανωνυμία 14

2.2.2 Μοντέλα k-ανωνυμοποίησης 17

2.2.3 Αλγόριθμοι εύρεσης k-ανώνυμων πινάκων 19

2.2.3.1 Incognito 20

2.2.3.2 Mondrian 21

2.3 Αναγνώριση τιμής ευαίσθητων δεδομένων 23

2.3.1 Επιθέσεις κατά της k-ανωνυμίας 23

2.3.1.1 Επιθέσεις ομοιογένειας 24

2.3.1.2 Επιθέσεις με πρότερη γνώση 26

2.3.2 l-Διαφορετικότητα 27

2.3.3 t-Εγγύτητα 31

2.3.4 Ανατομία 35

2.4 Πολλαπλές δημοσιεύσεις 39

2.5 Ταυτοποίηση ύπαρξης 41

3 Ορισμός προβλήματος 45

3.1 Μοντέλο δεδομένων 47

3.2 Απειλές κατά της ιδιωτικότητας 48
3.3 Μετρική κόστους απώλειας πληροφορίας 49
3.4 Πιθανές λύσεις 50
3.4.1 Χρήση Incognito *51*
3.4.2 Χρήση Mondrian *52*
3.4.3 Χρήση αλγορίθμου με συναθροιστική συνάρτηση *53*
3.4.3.1 Χαλάρωση της εγγύησης της ανωνυμίας 54
4 Περιγραφή αλγορίθμου 57
4.1 Θεωρητικό Υπόβαθρο 57
4.2 Υλοποίηση 59
5 Αξιολόγηση 67
5.1 Παράμετροι αξιολόγησης 67
5.2 Οργάνωση πειραμάτων 68
5.2.1 Δεδομένα *69*
5.2.2 Διαδικασία πειραμάτων *69*
5.3 Αποτελέσματα 70
5.3.1 Χρόνος εκτέλεσης *70*
5.3.2 Κανονικοποιημένη Ποινή Βεβαιότητας *73*
5.4 Σύνοψη συμπερασμάτων αξιολόγησης 76
6 Τεχνικές λεπτομέρειες 77
6.1 Λεπτομέρειες υλοποίησης 77
6.1.1 Χαρακτηριστικά υλοποίησης *78*
6.1.1.1 Μορφή δεδομένων εισόδου - εξόδου 78
6.1.1.2 Κλήση της εφαρμογής 78
6.1.1.3 Δομές δεδομένων 78
6.1.1.4 Συναρτήσεις 79
6.1.2 Ανάλυση βασικών μεθόδων κώδικα *81*
6.1.2.1 Βασική Συνάρτηση 81
6.1.2.2 Συνάρτηση διαχωρισμού κλάσεων 82
6.1.2.3 Συνάρτηση αναδρομής 82
6.1.2.4 Συνάρτηση ανωνυμίας 83
6.1.2.5 Συνάρτηση υπολογισμού μετρικής 84
6.1.3 Ανάπτυξη χρήσιμων εργαλείων *85*

6.1.3.1 Αρχείο μετακίνησης δεδομένων ... 85
6.1.3.2 Αρχείο επιλογής γνωρισμάτων ... 86
6.1.3.3 Αρχείο δειγματοληψίας ... 86
6.1.3.4 Αρχείο υπολογισμού μετρικής Mondrian ... 86
6.1.4 Λεπτομέρειες λειτουργίας αλγορίθμου Mondrian ... 87
6.2 Πλατφόρμες και προγραμματιστικά εργαλεία ... 88
6.2.1 Εφαρμογή γραφικού περιβάλλοντος ... 88
7 Επίλογος ... 91
7.1 Σύνοψη και συμπεράσματα ... 91
7.2 Μελλοντικές επεκτάσεις ... 92
8 Βιβλιογραφία ... 95

1

Εισαγωγή

1.1 Γενικά

Στους περισσότερους πλέον τομείς της καθημερινότητας, από κοινωνικούς και ιδιωτικούς φορείς, όπως νοσοκομεία και τράπεζες μέχρι τις εμπορικές επιχειρήσεις, γίνεται συλλογή και αποθήκευση αναλυτικών προσωπικών δεδομένων των ατόμων που εξυπηρετούν. Με την εξέλιξη της τεχνολογίας και τη διαρκή ανταλλαγή πληροφοριών, η χωρίς έγκριση δημοσίευση και η επεξεργασία των δεδομένων αυτών είναι ανεξέλεγκτη ενώ ο καθένας δύναται να έχει πρόσβαση σε αυτά. Η μορφή της δημοσίευσης, ο τρόπος και η ορθότητα της επεξεργασίας αυτών των δεδομένων βρίσκονται κατά κύριο λόγο στη δικαιοδοσία όσων συλλέγουν τα δεδομένα, οι οποίοι καθορίζουν ποιες προϋποθέσεις ιδιωτικότητας πρέπει να πληρούνται και κατά πόσο αυτές συμβαδίζουν με το νομικό πλαίσιο που αφορά την προστασία των προσωπικών δεδομένων. Με την αυτόβουλη δημοσίευση προσωπικών δεδομένων ενέχει ο κίνδυνος εξόρυξης προσωπικής πληροφορίας ή και αναγνώρισης κάποιου ατόμου που εμφανίζεται στις συλλογές δεδομένων. Κάτι τέτοιο παρατηρείται και περιγράφεται από την [Swe02], όπου μέσω της σύνδεσης των εκλογικών καταλόγων και των ιατρικών προσωπικών δεδομένων του υπεύθυνου οργανισμού για την ασφάλιση των δημοσίων υπαλλήλων της Μασαχουσέτης, έγινε η ταυτοποίηση του κυβερνήτη της και των ιατρικών δεδομένων του. Μιας και η προστασία των προσωπικών δεδομένων και της ιδιωτικής ζωής αποτελεί ανθρώπινο δικαίωμα, για την αποφυγή της παραβίασης του απορρήτου τους απαιτείται η εύρεση της σωστής ισορροπίας μεταξύ της προστασίας της

ιδιωτικότητας των προσώπων και της ελεγχόμενης πρόσβασης των υπολοίπων σε αυτά. Προς την κατεύθυνση αυτή και για την αποτροπή της παράνομης επεξεργασίας ή της δημοσίευσης προσωπικών δεδομένων ξεκίνησε η προσπάθεια από τους αρμόδιους φορείς, όταν η Ευρωπαϊκή Ένωση το 1995 θέσπισε την ευρωπαϊκή οδηγία 95/46/ΕΚ, ένα κείμενο αναφοράς στα θέματα προστασίας των δεδομένων προσωπικού χαρακτήρα. Εκεί, καθορίζονται οι κατευθυντήριες αρχές που προσδιορίζουν τη νομιμότητα της επεξεργασίας των δεδομένων, ενώ κάθε κράτος μέλος προβλέπει μία ή περισσότερες ανεξάρτητες κρατικές αρχές οι οποίες επιφορτίζονται με την εποπτεία της εφαρμογής, στο εθνικό έδαφος, των ληφθέντων από τα κράτη μέλη μέτρων κατ' εφαρμογή της οδηγίας αυτής. Στην Ελλάδα, για τον παραπάνω σκοπό ιδρύθηκε η Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα, μια ανεξάρτητη διοικητική αρχή συνταγματικά κατοχυρωμένη, η οποία ξεκίνησε τη λειτουργία της στις 10 Νοεμβρίου 1997. Η Αρχή αυτή λειτουργεί ρυθμιστικά και ελεγκτικά, ενώ αποφασίζει για κάθε περίπτωση παραβίασης της ιδιωτικότητας την νομιμότητα ή μη των πράξεων κάθε πλευράς και συμβάλει στη διευθέτηση του προβλήματος.

Αν και η ανεξέλεγκτη δημοσίευση ή αλλοίωση προσωπικών δεδομένων παραβιάζει τα ανθρώπινα δικαιώματα, η ορθή συλλογή και η ελεγχόμενη δημοσίευσή τους είναι ιδιαίτερα χρήσιμη για την εξαγωγή πληροφορίας ή συμπερασμάτων αναφορικά με το σύνολο του πληθυσμού που καλύπτουν. Κάποια χαρακτηριστικά παραδείγματα είναι η χρησιμότητα αυτών των συλλογών σε ερευνητικά πλαίσια, δημογραφικές έρευνες αλλά και οικονομικές αναλύσεις.

Με την πάροδο του χρόνου στην επιστημονική κοινότητα δημιουργήθηκε ο τομέας της προστασίας της ιδιωτικότητας, όπου ερευνά τρόπους και ορίζει νέες έννοιες και τεχνικές επεξεργασίας των δεδομένων ώστε να εξασφαλίζεται η διαφύλαξη της προσωπικής πληροφορίας που περιέχουν, ενώ παράλληλα επιτρέπεται η ορθή χρήση τους. Ο τομέας της προστασίας της ιδιωτικότητας εξελίσσεται συνεχώς με σκοπό να διατηρεί την ισορροπία ανάμεσα στην εκμετάλλευση των προσωπικών δεδομένων και τον σεβασμό προς τα άτομα, και η παρούσα εργασία επιχειρεί να συμβάλει σε αυτό.

1.2 Αντικείμενο διπλωματικής

Η εργασία ασχολείται με μια διάσταση του προβλήματος της διασφάλισης της ιδιωτικότητας των ατόμων που εμφανίζονται σε συλλογές προσωπικών δεδομένων προς δημοσίευση, που μέχρι τώρα δεν έχει ερευνηθεί.

Το πρόβλημα που εξετάζεται αφορά βάσεις δεδομένων με γνωρίσματα από ένα κοινό πεδίο τιμών πάνω στα οποία δίνεται η δυνατότητα διεξαγωγής συναθροιστικής πληροφορίας. Κατά την δημοσίευση των δεδομένων, ο συνδυασμός των τιμών κάποιων γνωρισμάτων μπορεί να λειτουργήσει ως ψευδο-αναγνωριστικό και να οδηγήσει στην αναγνώριση της ταυτότητας κάποιας εγγραφής που συμμετέχει στα δεδομένα.

Θεωρούνται επιθέσεις αυτής της μορφής με γνωστικό υπόβαθρο αποκλειστικά βασισμένο σε συναθροιστική πληροφορία πάνω στις τιμές των γνωρισμάτων κάθε εγγραφής. Επιχειρείται η αποτροπή τους μέσω της τροποποίησης των δεδομένων πριν την δημοσίευσή τους. Χαρακτηριστικό παράδειγμα του εξεταζόμενου προβλήματος εμφανίζεται κατά τη δημοσίευση συλλογών φορολογικών δεδομένων με γνωρίσματα τα μερικά εισοδήματα φυσικών προσώπων. Σε περίπτωση της δημοσίευσης των αρχικών τιμών ενέχει ο κίνδυνος αναγνώρισης κάποιας εγγραφής ή ανακάλυψης κάποιου επιμέρους εισοδήματος από όποιον επιτιθέμενο γνωρίζει το συνολικό εισόδημα κάποιου φυσικού προσώπου. Βάσει αυτού μπορεί να το ταυτοποιήσει με όποια εγγραφή παρουσιάζει αντίστοιχο συνολικό εισόδημα.

Από την βιβλιογραφία που μελετήθηκε, η καλύτερη προτεινόμενη λύση για την διασφάλιση της ιδιωτικότητας των εγγραφών στο πρόβλημα αυτό εμφανίζεται η *k*-ανωνυμοποίηση των δεδομένων [Swe02], μέσω του αλγορίθμου πολυδιάστατης τοπικής ανακωδικοποίησης Mondrian [MGK+06]. Η *k*-ανωνυμοποίηση αναφέρεται στην τροποποίηση των αρχικών δεδομένων με τεχνικές γενίκευσης έτσι ώστε να ικανοποιείται η εγγύηση ιδιωτικότητας της *k*-ανωνυμίας, Με χρήση αυτού, στα ανωνυμοποιημένα δεδομένα κάθε εγγραφή δεν μπορεί να διακριθεί από τουλάχιστον *k* άλλες εγγραφές.

Η παρούσα εργασία προτείνει μια λύση που προφυλάσσει την ιδιωτικότητα των εγγραφών από επιθέσεις με στόχο την αναγνώριση της ταυτότητας κάποιας εγγραφής στην περίπτωση κατοχής συναθροιστικής γνώσης πάνω στις τιμές των γνωρισμάτων των δεδομένων από τον επιτιθέμενο. Επιχειρεί την ικανοποίηση της *k*-ανωνυμίας παράλληλα με την διατήρηση της μέγιστης δυνατής χρήσιμης πληροφορίας των δεδομένων κατά την τροποποίησή τους. Για το σκοπό αυτό, αναπτύσσεται αναδρομικός αλγόριθμος *k*-ανωνυμοποίησης που εφαρμόζει τοπική ανακωδικοποίηση. Η κύρια διαφορά με τους αλγόριθμους που έχουν οριστεί ως τώρα για την *k*-ανωνυμοποίηση δεδομένων εμφανίζεται στην εκμετάλλευση της συναθροιστικής συνάρτησης κατά τον υπολογισμό της βέλτιστης τοπικής ανακωδικοποίησης, την οποία θεωρείται πως χρησιμοποιεί ο επιτιθέμενος. Ο αλγόριθμος που προτείνεται, τροποποιεί τα

δεδομένα του προβλήματος εφαρμόζοντας την τεχνική της γενίκευσης και προσφέρει μια *k*-ανωνυμοποίηση αντίστοιχης ποιότητας με εκείνη του αλγορίθμου Mondrian, διατηρώντας όμως περισσότερη χρήσιμη πληροφορία στα ανωνυμοποιημένα δεδομένα.

Όπως αξιολογείται και από τα πειράματα, υπερέχει του αλγορίθμου Mondrian μιας και με τη χρήση του εξασφαλίζεται η διατήρηση της ανωνυμίας των εγγραφών που συμμετέχουν στα δημοσιευμένα δεδομένα. Παράλληλα, τα δεδομένα εμφανίζουν μεγαλύτερη χρηστικότητα για εκείνους στους οποίους απευθύνονται και στόχο έχουν την διεξαγωγή χρήσιμης προσωπικής πληροφορίας από το σύνολο των δεδομένων χωρίς την παραβίαση του προσωπικού τους χαρακτήρα. Ακόμα, ο αλγόριθμος δύναται να παρέχει πολύ μικρότερο χρόνο εκτέλεσης, εξαιτίας της διαχείρισης των δεδομένων ως ανεξάρτητες ομάδες εγγραφών, αν υλοποιηθεί έτσι ώστε να εξετάζει περισσότερες από μία ομάδες παράλληλα.

1.2.1 Συνεισφορά

Στην εργασία αυτή υλοποιήθηκε και ερευνήθηκε αναδρομικός αλγόριθμος γενίκευσης, τοπικής ανακωδικοποίησης, την οποία υπολογίζει λαμβάνοντας υπόψη την συναθροιστική συνάρτηση πάνω στις τιμές των γνωρισμάτων κάθε εγγραφής που θεωρητικά γνωρίζει ο επιτιθέμενος. Στόχος του αλγορίθμου είναι η *k*-ανωνυμοποίηση του συνόλου των δεδομένων, αντίστοιχη της προσφερόμενης από τους έως τώρα προτεινόμενους αλγορίθμους, Incognito και Mondrian. Επιπλέον παρέχει καλύτερη διαχείριση της χρήσιμης πληροφορίας που περιέχεται στα αρχικά δεδομένα.

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήθηκε η απαραίτητη βιβλιογραφία που αφορά την προστασία της ιδιωτικότητας των προς δημοσίευση δεδομένων με σκοπό την εύρεση της κατάλληλης αρχής ανωνυμίας για το πρόβλημα της εργασίας.
2. Αναπτύχθηκε και υλοποιήθηκε αλγόριθμος για την επίλυση του προβλήματος της βέλτιστης *k*-ανωνυμοποίησης των δεδομένων με στόχο την προστασία των προσωπικών δεδομένων από τις επιθέσεις με συναθροιστική γνώση, ως εφαρμογή τερματικού παραθύρου και ως εφαρμογή γραφικού περιβάλλοντος.
3. Εκτελέστηκαν πειράματα σύγκρισης του προτεινόμενου αλγορίθμου και του αλγορίθμου Mondrian για διαφορετικά σύνολα δεδομένων και τιμές των παραμέτρων.
4. Αξιολογήθηκαν τα αποτελέσματα των παραπάνω πειραμάτων και διαπιστώθηκε ως αποδοτικότερη λύση στο πρόβλημα της αποτροπής των επιθέσεων με γνώση συναθροιστικής πληροφορίας η ανωνυμοποίηση των δεδομένων με χρήση του προτεινόμενου αλγορίθμου.

1.3 Οργάνωση κειμένου

Η εργασία παρουσιάζεται σύμφωνα με τα παρακάτω κεφάλαια:

Στο **δεύτερο** κεφάλαιο αναλύεται η βιβλιογραφία που μελετήθηκε και αφορά έννοιες σχετικά με την προστασία της ιδιωτικότητας σε βάσεις δεδομένων προσωπικής πληροφορίας, τις εγγυήσεις ιδιωτικότητας που έχουν οριστεί και τους αντίστοιχους αλγορίθμους για την τροποποίηση δεδομένων με σκοπό την ανωνυμοποίηση τους.

Στο **τρίτο** κεφάλαιο ορίζεται το πρόβλημα της ανωνυμοποίησης δεδομένων με σκοπό την προστασία της ιδιωτικότητας από επιθέσεις με γνώση συναθροιστικής πληροφορίας το οποίο ερευνήθηκε, καθώς και οι πιθανές λύσεις που προσφέρουν οι αλγόριθμοι που ήδη έχουν οριστεί από την βιβλιογραφία.

Στο **τέταρτο** κεφάλαιο παρουσιάζεται ο αναδρομικός αλγόριθμος που λαμβάνει υπόψη την συναθροιστική συνάρτηση για την κατάλληλη *k*-ανωνυμοποίηση των δεδομένων, ο οποίος προτείνεται ως λύση του προβλήματος και αναλύεται η λειτουργία του.

Στο **πέμπτο** κεφάλαιο περιγράφεται η πειραματική διαδικασία που ακολουθήθηκε με σκοπό τη σύγκριση του αλγορίθμου με την έως τώρα προσφερόμενη βέλτιστη λύση για την ανωνυμοποίηση αντίστοιχων συνόλων δεδομένων, τον αλγόριθμο Mondrian καθώς και τα αποτελέσματα που προέκυψαν σχετικά με την απόδοση τους ως προς την διασφάλιση της ιδιωτικότητας και την διατήρηση της χρήσιμης πληροφορίας των αρχικών δεδομένων.

Στο **έκτο** κεφάλαιο καταγράφονται οι λεπτομέρειες υλοποίησης του αλγορίθμου ως εφαρμογή τερματικού παραθύρου, τα εργαλεία που αναπτύχθηκαν για την δειγματοληψία και την διεξαγωγή των πειραμάτων καθώς και οι λεπτομέρειες της εφαρμογής γραφικού περιβάλλοντος που δημιουργήθηκε.

Στο **έβδομο** κεφάλαιο συνοψίζονται τα αποτελέσματα της εργασίας αναφορικά με την βέλτιστη λύση του προβλήματος της διασφάλισης της ιδιωτικότητας από επιτιθέμενους με συναθροιστική γνώση και προτείνονται μελλοντικές επεκτάσεις του αλγορίθμου.

2

Θεωρητικό υπόβαθρο

Το ενδιαφέρον στην παρούσα εργασία, όπως και σε ένα μεγάλο τμήμα του τομέα της προστασίας της ιδιωτικότητας εστιάζεται σε συλλογές προσωπικών δεδομένων ξεχωριστών ατόμων. Με τον όρο προσωπικά δεδομένα, προσδιορίζουμε το σύνολο των πληροφοριών ενός ατόμου, όπως το όνομά του, η ηλικία του, το επάγγελμά του, τα οποία το καθορίζουν. Τα δεδομένα αυτά συγκεντρώνονται συχνά σε βάσεις δεδομένων, οι οποίες παρέχουν στους κατόχους των δεδομένων πολλές δυνατότητες μαζικής επεξεργασίας, μεταφοράς και διαχείρισής τους. Σε κάθε βάση δεδομένων το σύνολο των δεδομένων πιθανώς παρουσιάζει μια ιδιαίτερη μορφολογία, ανάλογα με τα γνωρίσματα που περιέχει, ενώ μπορεί να ικανοποιεί και κάποιες ξεχωριστές ιδιότητες. Για τον ορισμό αυτών χρησιμοποιούνται οι παρακάτω έννοιες.

2.1 Οργάνωση δεδομένων

Στη βιβλιογραφία που μελετήθηκε, τα προς δημοσίευση δεδομένα αφορούν προσωπικές πληροφορίες ανθρώπων και οργανώνονται στη μορφή πίνακα $RT(A_1, A_2, \ldots, A_n)$ σχεσιακής βάσης δεδομένων, όπου $A_1, A_2, \ldots, A_n$ είναι οι στήλες-γνωρίσματά του. Κάθε πλειάδα αφορά ένα άτομο και τις τιμές του στα αντίστοιχα πεδία πληροφορίας. Κάθε στήλη-γνώρισμα αντιπροσωπεύει μια κατηγορία πληροφορίας και έχει ένα σύνολο πιθανών τιμών, το πεδίο τιμών του γνωρίσματος. Για την επίτευξη της σωστής διαχείρισης της πληροφορίας που

αντιπροσωπεύει κάθε γνώρισμα, και εφόσον αυτά αναφέρονται σε προσωπική πληροφορία, σε πολλές περιπτώσεις τα γνωρίσματα είναι δυνατόν να διαχωριστούν βάσει του τι είναι απαραίτητο να δημοσιευθεί και τι πρέπει να αποκρυφθεί, ώστε να προστατεύονται τα προσωπικά δεδομένα των ατόμων του πίνακα. Κάποια πεδία πληροφορίας όπως ο Αριθμός Ταυτότητας ή ο Αριθμός Φορολογικού Μητρώου ενός ατόμου αποτελούν μοναδικά αναγνωριστικά μιας και προσδιορίζουν άμεσα κάποιο φυσικό πρόσωπο. Όταν κατά τη δημοσίευση της βάσης δεδομένων επιχειρείται η προστασία της ιδιωτικότητας των ατόμων που συμμετέχουν, τα γνωρίσματα αυτά δεν δημοσιεύονται μιας και οδηγούν κατευθείαν στην αναγνώριση του φυσικού προσώπου και κατά συνέπεια στην εξόρυξη της προσωπικής και πιθανώς απόρρητης πληροφορίας του. Τα γνωρίσματα που απαιτείται να δημοσιευθούν και περιέχουν προσωπική πληροφορία, αναλόγως με την πληροφορία που αντιπροσωπεύουν χωρίζονται σε δύο σύνολα, τα γνωρίσματα του *ψευδο-αναγνωριστικού* και τα *ευαίσθητα* γνωρίσματα.

2.1.1 Ψευδο-αναγνωριστικό

Ως *ψευδο-αναγνωριστικό (quasi-identifier)* του πίνακα δεδομένων $RT(A_1, A_2, \ldots, A_n)$ ορίζεται το ελάχιστο σύνολο γνωρισμάτων του, τα οποία σε συνδυασμό με εξωτερική πληροφορία μπορούν να οδηγήσουν στην αναγνώριση της ταυτότητας κάποιας εγγραφής [MGK+06, LDR05]. Το σύνολο των γνωρισμάτων του ψευδο-αναγνωριστικού αντιπροσωπεύει προσωπικές πληροφορίες οι οποίες μπορούν να προσδιορίσουν μοναδικά ένα άτομο όταν παρουσιάζουν δεδομένα που εμφανίζονται και σε άλλες ήδη δημοσιευμένες συλλογές. Με την εξακρίβωση της κοινής πληροφορίας στα δύο σύνολα κάποιος μπορεί να ανακαλύψει προσωπική πληροφορία και να την συσχετίσει με κάποιο φυσικό πρόσωπο.

Σύμφωνα με τους [Swe02] ένα χαρακτηριστικό σύνολο γνωρισμάτων που μπορεί να αποτελέσει ψευδο-αναγνωριστικό για άτομα που εμφανίζονται σε δημόσιους καταλόγους είναι το σύνολο {Όνομα, Ημερομηνία γεννήσεως, Διεύθυνση, Ταχυδρομικός κωδικός, Φύλο}.

Το σύνολο των γνωρισμάτων που πιθανόν να αποτελούν το ψευδο-αναγνωριστικό μπορεί να προσδιοριστεί από τον κάτοχο των δεδομένων, με αναζήτηση σε εξωτερικούς καταλόγους στους οποίους εμφανίζονται γνωρίσματα που περιέχονται στο προς δημοσίευση σύνολο δεδομένων. Το ποια από τα γνωρίσματα του πίνακα θα λειτουργήσουν στην πράξη ως ψευδο-αναγνωριστικό, απαντάται μόνο έχοντας γνώση όλων των εξωτερικών πληροφοριών που παρέχονται στον επιτιθέμενο, σε κάθε άτομο δηλαδή που θα προσπαθήσει να εξορύξει πληροφορία από τα δημοσιευμένα δεδομένα. Ένα παράδειγμα συνόλου γνωρισμάτων που μπορεί να λειτουργήσει ως ψευδο-αναγνωριστικό αποτελούν στον Πίνακα 2.1 τα γνωρίσματα

{Όνομα, Ταχυδρομικός κωδικός}, όπου τα δεδομένα του Πίνακα 2.1 υποθετικά αντιπροσωπεύουν μία δημοσίευση δεδομένων μισθοδοσίας. Αν αυτά συσχετιστούν με τους τοπικούς τηλεφωνικούς καταλόγους κάποιας περιοχής, μόνο κάποιος περιορισμένος αριθμός ατόμων θα εμφανίζει τις ίδιες τιμές στα κοινά γνωρίσματα των δύο συλλογών. Με τον τρόπο αυτό μπορεί να αναγνωριστεί η ταυτότητα κάποιας εγγραφής και συνεπώς να διεξαχθούν συμπεράσματα σχετικά με τις προσωπικές της πληροφορίες, όπως ο μισθός της, που αναπαριστώνται από τα γνωρίσματα του δημοσιευμένου συνόλου.

2.1.2 Ευαίσθητα γνωρίσματα

Όπως αναφέρεται από το [MGK+06], ως *ευαίσθητο γνώρισμα (sensitive attribute)* ορίζεται ένα γνώρισμα του οποίου η τιμή για κάθε άτομο που εμφανίζεται στη βάση δεδομένων επιβάλλεται να μην μπορεί να ανακαλυφθεί από όσους δεν έχουν άμεση πρόσβαση στα πρωτότυπα δεδομένα. Ο απόρρητος χαρακτήρας της ευαίσθητης τιμής κάθε εγγραφής διατηρείται στην περίπτωση της απόκρυψης της τιμής από κάθε δημοσίευση αλλά και στην περίπτωση δημοσίευσης της ευαίσθητης τιμής με τρόπο ώστε να αποκλείεται η συσχέτισή της με κάποια εγγραφή. Ένα παράδειγμα ευαίσθητου γνωρίσματος που συνήθως εμφανίζεται σε ιατρικές βάσεις δεδομένων αντιστοιχεί στην ασθένεια του ατόμου. Η ιατρική γνωμάτευση ή το συνολικό εισόδημα ενός ατόμου, όπως για παράδειγμα το γνώρισμα «Μισθός» στον Πίνακα 2.1, είναι μια προσωπική πληροφορία που θα πρέπει να μείνει απόρρητη και διασφαλισμένη απέναντι σε κάθε επιτιθέμενο που θα προσπαθήσει να την αποσπάσει, και συνήθως κρίνεται ως ευαίσθητο γνώρισμα. Η επιλογή των γνωρισμάτων που θα θεωρηθούν ευαίσθητα είναι επίσης στη δικαιοδοσία του κατόχου των δεδομένων και κρίνεται από τις συνθήκες ιδιωτικότητας που πρέπει να διασφαλίζουν τα δημοσιευμένα δεδομένα. Για το λόγο αυτό υπάρχουν περιπτώσεις που επιλέγονται κάποια γνωρίσματα ως ευαίσθητα χωρίς να αντιπροσωπεύουν κάποιο στοιχείο τόσο προσωπικό όπως μια ασθένεια, εφόσον ο ιδιοκτήτης της βάσης δεδομένων κρίνει πως αυτά δεν πρέπει να δημοσιευτούν γιατί μπορεί να οδηγήσουν στην αποκάλυψη πληροφορίας την οποία πρέπει να προστατέψει.

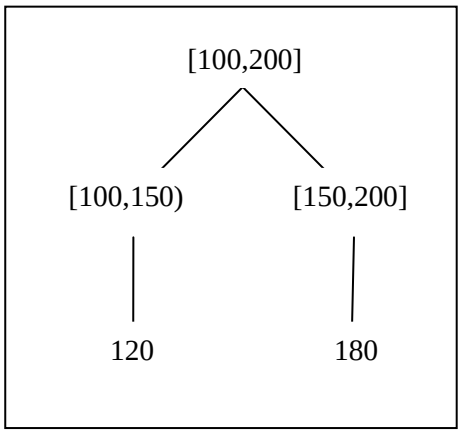
Α/Α	Όνομα	Ταχυδρομικός κωδικός	Ηλικία	Ύψος	Μισθός
1	Νίκος	18540	25	1,80	500
2	Τάκης	18530	27	1,73	900
3	Μιχάλης	14050	34	1,77	700
4	Μαρία	14244	31	1,67	900

Πίνακας 2.1

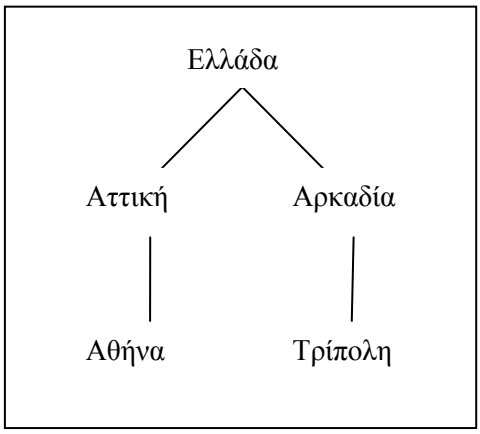
2.1.3 Ιεραρχία γενίκευσης πεδίου τιμών γνωρίσματος

Η τεχνική της γενίκευσης χρησιμοποιείται ευρέως στον τομέα της προστασίας της ιδιωτικότητας. Ως *γενίκευση (generalization)* ορίζεται η διαδικασία κατά την οποία η αρχική τιμή που εμφανίζεται στα δεδομένα αντικαθίστανται με μία πιο γενική τιμή που σημασιολογικά περιέχει την αρχική τιμή. Ως στόχο έχει την διατήρηση μέρους της πληροφορίας που περιέχει η αρχική τιμή χωρίς αυτή να αλλοιώνεται πλήρως, όπως γίνεται στην περίπτωση της απόκρυψης της αρχικής τιμής. Ακόμα, κάθε γενικευμένη τιμή δύναται να γενικευθεί περαιτέρω αναλόγως του πεδίου τιμών σε μια ακόμα πιο γενική σημασιολογικά τιμή. Σε μια σχεσιακή βάση δεδομένων κάθε γνώρισμα έχει ένα πεδίο τιμών. Από το πεδίο τιμών μπορεί να κατασκευαστεί ένα πιο γενικό πεδίο, όπου κάθε αρχική τιμή αντιστοιχίζεται σε μια γενικευμένη τιμή και με τον τρόπο αυτό σχηματίζεται ένα γενικευμένο πεδίο τιμών για το γνώρισμα. Τα διαφορετικά επίπεδα γενίκευσης του πεδίου τιμών ενός γνωρίσματος, στα οποία οδηγούνται οι αρχικές τιμές με κάθε γενίκευση συνήθως αποτυπώνονται με τη μορφή δένδρου, το οποίο ονομάζεται *ιεραρχία γενίκευσης του πεδίου τιμών, (Domain Generalization Hierarchy)*. Η τεχνική της γενίκευσης χρησιμοποιείται σε αριθμητικά αλλά και κατηγορικά πεδία τιμών. Στην περίπτωση ενός αριθμητικού πεδίου τιμών, μια αρχική τιμή συνήθως γενικεύεται σε ένα διάστημα τιμών, ενώ το διάστημα αυτό δύναται να γενικευθεί περαιτέρω σε ένα ακόμα μεγαλύτερο διάστημα τιμών. Ένα τέτοιο παράδειγμα εμφανίζεται στην Εικόνα 2.1. Εδώ, στην συγκεκριμένη ιεραρχία γενίκευσης, ο αριθμός 120 μπορεί να γενικευθεί στο διάστημα [100,150), ενώ το διάστημα αυτό μπορεί να γενικευθεί στο διάστημα [100,200].

Στην περίπτωση των κατηγορικών δεδομένων η γενίκευση εφαρμόζεται συνήθως βάσει της σημασιολογίας των αρχικών τιμών. Για παράδειγμα, αν το γνώρισμα αφορά την πληροφορία «Τόπος Κατοικίας», η αρχική τιμή της πόλης «Αθήνα» μπορεί να γενικευθεί στην τιμή του Νομού «Αττική» και εκείνη να γενικευθεί περαιτέρω στην τιμή της χώρας «Ελλάδα», όπως παρουσιάζεται στην Εικόνα 2.2.



Εικόνα 2.1: Ιεραρχία γενίκευσης αριθμητικού πεδίου τιμών γνωρίσματος



Εικόνα 2.2: Ιεραρχία γενίκευσης κατηγορικού πεδίου τιμών γνωρίσματος

2.1.4 Απειλές κατά της ιδιωτικότητας

Μέσω της δημοσίευσης προσωπικών δεδομένων, ακόμα και μετά την αφαίρεση των γνωρισμάτων που προσδιορίζουν μοναδικά κάθε φυσικό πρόσωπο, συνήθως εμφανίζονται γνωρίσματα που μπορούν να λειτουργήσουν ως ψευδο-αναγνωριστικό. Ακόμα και σε αυτή την περίπτωση υπάρχει η πιθανότητα εξαγωγής προσωπικής πληροφορίας για κάποιο άτομο που συμμετέχει στο σύνολο των δεδομένων. Η προσπάθεια κάποιου άλλου ατόμου που έχει πρόσβαση στα δημοσιευμένα δεδομένα να ανακαλύψει περαιτέρω προσωπικές πληροφορίες από εκείνες που γνωρίζει για κάποιο άτομο μέσω της δημοσίευσης, παραβιάζει το απόρρητο των προσωπικών δεδομένων και καλείται επίθεση στην ιδιωτικότητα του ατόμου. Αναλόγως των στοιχείων που έχει στη διάθεσή του, όπως για παράδειγμα δεδομένα από δημοσιεύσεις άλλων συλλογών δεδομένων ή προσωπική επαφή με το ζητούμενο άτομο, ο επιτιθέμενος, εκείνος που αναζητά την παραπάνω προσωπική πληροφορία, μπορεί να ταυτοποιήσει το άτομο με κάποια εγγραφή από τα δημοσιευμένα δεδομένα, να επιβεβαιώσει την παρουσία του στο συγκεκριμένο σύνολο, ή να ανακαλύψει την τιμή που αυτό παίρνει σε κάποιο γνώρισμα.

Στην περίπτωση ταυτοποίησης του ατόμου, ο επιτιθέμενος μπορεί να συγκρίνει πληροφορίες που ήδη έχει στην κατοχή του για κάποιο άτομο με εκείνες που εμφανίζουν οι εγγραφές στα δεδομένα. Από τις εγγραφές, πιθανώς κάποιες να έχουν αρκετά όμοιες τιμές στα αντίστοιχα γνωρίσματα έτσι ώστε με βεβαιότητα να συμπεράνει πως πρόκειται για το ίδιο άτομο. Στην περίπτωση που γνωρίζει ότι το άτομο βρίσκεται στο δημοσιευμένο σύνολο δεδομένων καθώς και κάποιες από τις τιμές που λαμβάνει σε κάποια από τα γνωρίσματα του συνόλου, ταυτοποιεί το άτομο με κάποια ή κάποιες εγγραφές και στη συνέχεια μπορεί να διεξάγει συμπεράσματα για τις τιμές του στα υπόλοιπα γνωρίσματα, με το ενδιαφέρον κυρίως στις τιμές των ευαίσθητων γνωρισμάτων, εκείνων που αντιπροσωπεύουν προσωπική και απόρρητη πληροφορία. Η κατάλληλη επεξεργασία των δεδομένων με στόχο την διαφύλαξη

από κάθε διαφορετική περίπτωση επίθεσης καθορίζεται από την προηγούμενη γνώση του επιτιθέμενου και από την γνώση που μπορεί να συμπεράνει συνδυάζοντάς την γνώση που έχει με τον δημοσιευμένο πίνακα δεδομένων. Στην παρούσα εργασία αναλύονται οι εγγυήσεις ιδιωτικότητας και οι τεχνικές που έχουν οριστεί βάσει των πληροφοριών που μπορεί να συμπεράνει ο επιτιθέμενος από τα δημοσιευμένα δεδομένα. Ο επιτιθέμενος κατά κύριο λόγο μπορεί να ανακαλύψει την ταυτότητα κάποιου συγκεκριμένου ατόμου που εμφανίζεται στα δεδομένα, ή να συμπεράνει την τιμή που αυτό λαμβάνει σε κάποιο συγκεκριμένο, ευαίσθητο γνώρισμα.

2.1.5 Απώλεια πληροφορίας

Τα σύνολα δεδομένων που περιέχουν προσωπική πληροφορία ζητείται να δημοσιευθούν κυρίως για την εκμετάλλευση της χρήσιμης πληροφορίας που περιέχουν σχετικά με το σύνολο του πληθυσμού που αντιπροσωπεύουν. Η πληροφορία που περιέχεται σε αυτά είναι πολύτιμη καθώς η μελέτη της μπορεί να αποδώσει σημαντικά αποτελέσματα και συμπεράσματα σε έρευνες ή στατιστικές αναλύσεις. Πριν τη δημοσίευση του συνόλου με στόχο την διαφύλαξη της ιδιωτικότητας των εγγραφών που συμμετέχουν απαιτείται η τροποποίηση των αρχικών δεδομένων. Σε κάθε περίπτωση τροποποίησης των δεδομένων οι αρχικές τιμές αντικαθίστανται με κάποιες άλλες έτσι ώστε μέρος της προσωπικής πληροφορίας να αποκρύπτεται στα δημοσιευμένα δεδομένα.

Κάτι τέτοιο έχει ως αποτέλεσμα την απώλεια μέρους της χρήσιμης πληροφορίας που υπήρχε στα αρχικά δεδομένα, η οποία αποτελεί τον βασικό λόγο για τον οποίο απαιτείται η δημοσίευσή τους. Συνεπώς σε κάθε περίπτωση τροποποίησης των δεδομένων με σκοπό την προστασία της ιδιωτικότητας σημαντικό ρόλο κατέχει και το ποσοστό χρήσιμης πληροφορίας που χάνεται.

Για να γίνει πιο κατανοητή η σημαντικότητα της απώλειας της πληροφορίας, θεωρούμε ως αρχικό σύνολο δεδομένων τα δεδομένα του Πίνακα 2.1. χωρίς το γνώρισμα «Όνομα». Μπορεί εύκολα να παρατηρηθεί στη δημοσίευση του Πίνακα 2.2 η διαφορά της χρήσιμης πληροφορίας που χάνεται με την τροποποίηση των δεδομένων συγκριτικά με τη δημοσίευση των δεδομένων με τη μορφή του Πίνακα 2.3.

Στην πρώτη περίπτωση όλες οι εγγραφές εμφανίζονται με τις ίδιες γενικευμένες τιμές στα γνωρίσματα «Ταχυδρομικός κωδικός» και «Ηλικία». Η δημοσίευση αυτή διατηρεί την ιδιωτικότητα των εγγραφών μιας και καμία εγγραφή δε μπορεί να αναγνωριστεί με βεβαιότητα, όμως χάνει την χρήσιμη πληροφορία που εμφανιζόταν από τις διαφορετικές αρχικές τιμές των εγγραφών.

Α/Α	Ταχυδρομικός κωδικός	Ηλικία	Ύψος	Μισθός
1	[14000,19000]	[20,40]	1,80	500
2	[14000,19000]	[20,40]	1,73	900
3	[14000,19000]	[20,40]	1,77	700
4	[14000,19000]	[20,40]	1,67	900

Πίνακας 2.2

Αντίστοιχα κατά τη δημοσίευση του Πίνακα 2.3 μέρος της αρχικής χρήσιμης πληροφορίας έχει αποκρυφθεί έτσι ώστε να διατηρείται η ιδιωτικότητα, όμως η διαφοροποίηση στις τιμές των εγγραφών στα δύο πρώτα γνωρίσματα έχει διατηρηθεί σε ικανοποιητικό βαθμό. Στην περίπτωση αυτή καμία εγγραφή δεν μπορεί να προσδιοριστεί μοναδικά κατά τη δημοσίευση, ενώ χρήσιμα συμπεράσματα μπορούν να διεξαχθούν αναφορικά με τις επιμέρους τιμές όπως για παράδειγμα η συσχέτιση μεταξύ της ηλικίας και του τόπου κατοικίας των ατόμων η οποία είναι εμφανής.

Α/Α	Ταχυδρομικός κωδικός	Ηλικία	Ύψος	Μισθός
1	[18500,18599]	[25,30]	1,80	500
2	[18500,18599]	[25,30]	1,73	900
3	[14000,14999]	[30,40]	1,77	700
4	[14000,14999]	[30,40]	1,67	900

Πίνακας 2.3

Κατανοώντας την σημαντικότητα της χρήσιμης πληροφορίας και της διατήρησής της κατά τη δημοσίευση τροποποιημένων δεδομένων, ο τομέας της προστασίας της ιδιωτικότητας εξετάζει τους αλγορίθμους που υλοποιούν της αρχές ανωνυμίας ως προς την αποδοτικότητά τους αναφορικά με αυτήν. Αναπτύσσει κατάλληλες μετρικές και εργαλεία έτσι ώστε να αξιολογεί τους αλγορίθμους και τις εγγυήσεις ιδιωτικότητας όχι μόνο βάσει της προστασίας που προσφέρουν αλλά και βάσει της χρήσιμης πληροφορίας που διατηρούν στα δημοσιευμένα δεδομένα. Σύμφωνα με αυτό το κριτήριο και με χρήση των κατάλληλων μετρικών διακρίνεται η βέλτιστη τροποποίηση των δεδομένων μεταξύ των προτεινόμενων τεχνικών.

2.2 Αναγνώριση ταυτότητας

Στην πρώτη περίπτωση επιθέσεων που εξετάζεται, ο επιτιθέμενος προσπαθεί άμεσα ή έμμεσα να ταυτοποιήσει κάποια εγγραφή του συνόλου δεδομένων με κάποιο φυσικό πρόσωπο. Χρησιμοποιώντας προηγούμενη γνώση ή συνδυάζοντας δημοσιευμένες συλλογές δεδομένων, μπορεί να συμπεράνει με ακρίβεια την ταυτότητα κάποιου ατόμου που εμφανίζεται στον δημοσιευμένο πίνακα δεδομένων με εξακρίβωση των τιμών του ψευδο-αναγνωριστικού του. Μπορεί δηλαδή να αναγνωρίσει ποια εγγραφή του πίνακα αντιστοιχεί σε ένα συγκεκριμένο άτομο, γνωρίζοντας από πριν κάποιες τιμές του ψευδο-αναγνωριστικού του. Ως παράδειγμα κατά την δημοσίευση του Πίνακα 2.1 ως έχει, ο επιτιθέμενος μπορεί να γνωρίζει ένα άτομο με τιμές στο σύνολο {Όνομα, Ταχυδρομικός κωδικός} τις {Τάκης, 18530} αντίστοιχα και να συμπεράνει πως το άτομο αυτό είναι η εγγραφή 2, αφού μόνο αυτή λαμβάνει το σύνολο των τιμών σε όλο τον πίνακα.

Μιας και τέτοιες επιθέσεις είναι πολύ πιθανές όταν το σύνολο των δεδομένων δημοσιεύεται με την αρχική του μορφή, ο τομέας της προστασίας της ιδιωτικότητας ασχολήθηκε αρχικά με αυτές, εξερευνώντας τρόπους ώστε να αποτρέπονται.

Για να επιτευχθεί κάτι τέτοιο, εφόσον δεν είναι δυνατόν να ελέγξει κάποιος πλήρως την πληροφορία που διαθέτει κάθε επιτιθέμενος ή όλους τους πιθανούς παραλήπτες των δημοσιευμένων δεδομένων, απαιτείται η κατάλληλη τροποποίηση των αρχικών δεδομένων έτσι ώστε τα δεδομένα που τελικά δημοσιεύονται να περιέχουν όση περισσότερη πληροφορία είναι δυνατό, αλλά να προστατεύουν τα συμμετέχοντα πρόσωπα από τις επιθέσεις που στόχο έχουν την αναγνώριση της ταυτότητάς τους.

Σε μεγάλο βαθμό τέτοιες επιθέσεις αποφεύγονται όταν στον δημοσιευμένο πίνακα ικανοποιείται η *k*-ανωνυμία, μια βασική αρχή γενίκευσης που ορίστηκε από [Swe02].

2.2.1 k-Ανωνυμία

Βάσει του ορισμού που έχει δοθεί από [Swe02, LDR05], ένας προς δημοσίευση πίνακας $RT(A_1, A_2, \ldots, A_n)$ με ψευδο-αναγνωριστικό $QI_{RT} = (A_1, A_2, .., A_j)$ το σύνολο των γνωρισμάτων $A_1, A_2, .., A_j$, ικανοποιεί την *k-ανωνυμία (k-anonymity)* αν κάθε ακολουθία τιμών στον πίνακα $RT[QI_{RT}]$ του ψευδο-αναγνωριστικού εμφανίζεται τουλάχιστον *k* φορές.

Κάτι τέτοιο πράγματι αποτρέπει σε ικανοποιητικό βαθμό επιθέσεις κατά τις οποίες επιχειρείται η αναγνώριση της ταυτότητας ενός ατόμου που συμμετέχει στα δεδομένα. Όταν τα δεδομένα που δημοσιεύονται ικανοποιούν την *k*-ανωνυμία, για κάθε συνδυασμό τιμών στα

γνωρίσματα του ψευδο-αναγνωριστικού θα υπάρχουν το λιγότερο *k* εγγραφές που θα τον περιέχουν. Κάθε ομάδα από πλειάδες που εμφανίζουν ταυτόσημες τιμές στα γνωρίσματα του ψευδο-αναγνωριστικού ονομάζεται *κλάση ισοδυναμίας (equivalence class).* Συνεπώς ένας πίνακας από δεδομένα όταν ικανοποιεί την *k*-ανωνυμία αποτελείται από κλάσεις ισοδυναμίας όπου σε καθεμία εμφανίζεται ένας συνδυασμός τιμών στα γνωρίσματα του ψευδο-αναγνωριστικού. Ο επιτιθέμενος πάνω σε δεδομένα που δημοσιεύονται με τέτοια μορφή δεν μπορεί με βεβαιότητα να αναγνωρίσει μέσω αυτών των τιμών των γνωρισμάτων μοναδικά μία εγγραφή, καθώς θα οδηγείται κάθε φορά σε τουλάχιστον *k* εγγραφές που παίρνουν τις ζητούμενες τιμές στα γνωρίσματα αυτά.

Επειδή είναι σπάνιο τα σύνολα δεδομένων που συλλέγονται να ικανοποιούν την *k*-ανωνυμία στην αρχική τους μορφή, ο τομέας της προστασίας της ιδιωτικότητας έχει αναπτύξει τεχνικές και αλγορίθμους ώστε να τροποποιούνται τα δεδομένα προς μια μορφή τέτοια ώστε να ικανοποιείται η *k*-ανωνυμία. Από τις διαδικασίες αυτές προκύπτει συνήθως μια νέα έκδοση του πίνακα δεδομένων.

Μία έκδοση του πίνακα, $RT^*(A_1, A_2, \ldots, A_n)$ λέγεται *k-ανωνυμοποίηση (k-anonymization)* του $RT(A_1, A_2, \ldots, A_n)$ αν τροποποιούνται ή αποκρύπτονται τα δεδομένα του πίνακα σύμφωνα με κάποιο μηχανισμό τέτοιο ώστε η έκδοση να ικανοποιεί την ιδιότητα της *k*-ανωνυμίας ως προς το σύνολο των γνωρισμάτων του ψευδο-αναγνωριστικού, όπως αναφέρεται στους [LDR05].

Στις μεθόδους της *k*-ανωνυμοποίησης χρησιμοποιούνται κατά κύριο λόγο δύο τεχνικές για την ανακωδικοποίηση, πάνω στις τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού:

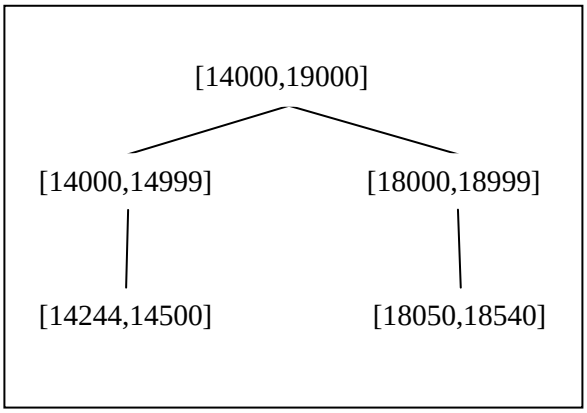
- *Γενίκευση (generalization):* Κατά την τεχνική της γενίκευσης όπως έχει ήδη αναφερθεί, η ιεραρχία γενίκευσης που χρησιμοποιείται για κάθε γνώρισμα αποτυπώνει τη σημασιολογική δομή του πεδίου τιμών του γνωρίσματος, και πώς κάθε έννοια ή τιμή που βρίσκεται σε αυτό μπορεί αντικατασταθεί από ένα πιο γενικό εύρος τιμών ή μια γενικευμένη τιμή που περιέχει την αρχική. Κατά την τεχνική της γενίκευσης οι αρχικές τιμές των γνωρισμάτων αντικαθίστανται με τις κατάλληλες γενικευμένες τιμές βάσει της συγκεκριμένης ιεραρχίας γενίκευσης κάθε γνωρίσματος του ψευδο-αναγνωριστικού. Με τον τρόπο αυτό δύο ή περισσότερες άνισες τιμές ενός γνωρίσματος στην κλάση ισοδυναμίας μπορούν να αντικατασταθούν με μία κοινή γενικευμένη τιμή έτσι ώστε οι εγγραφές της κλάσης ισοδυναμίας να παίρνουν την ίδια τιμή στο γνώρισμα και συνεπώς, αν αυτό ισχύει για κάθε γνώρισμα του ψευδο-αναγνωριστικού, να ικανοποιείται η *k*-ανωνυμία στην κλάση ισοδυναμίας.
- *Απόκρυψη (suppression)*: Σύμφωνα με την τεχνική της απόκρυψης επιλέγονται κάποιες τιμές των γνωρισμάτων των εγγραφών που αποκρύπτονται πλήρως στον δημοσιευμένο πίνακα, με σκοπό την ικανοποίηση της *k*-ανωνυμίας. Η απόκρυψη των

τιμών αυτών συνήθως επιλέγεται όταν στην κλάση ισοδυναμίας οι τιμές που παίρνει ένα γνώρισμα δεν οδηγούν σε μία περαιτέρω κοινή γενικευμένη τιμή, ώστε να μπορούν να αντικατασταθούν από αυτήν όπως συμβαίνει στο παράδειγμα του Πίνακα 2.5 για το γνώρισμα «Φύλο». Μιας και σημασιολογικά δεν υπάρχει τιμή στην οποία μπορούν να γενικευθούν οι τιμές «Άνδρας» και «Γυναίκα», επιλέγεται η απόκρυψη αυτών των τιμών, αφού αν παραμείνουν ως έχουν δεν θα ικανοποιούν την *k*-ανωνυμία.

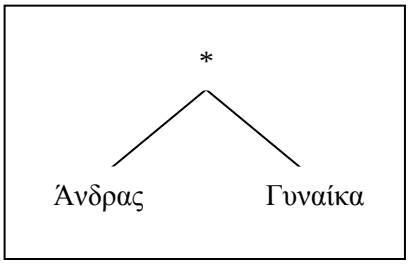
Στον Πίνακα 2.5 παρουσιάζεται η ιδιότητα της *k*-ανωνυμίας εφαρμοσμένη στα δεδομένα του Πίνακα 2.4. Ο Πίνακας 2.5 είναι μία 3-ανώνυμη έκδοση του αρχικού. Ως γνωρίσματα του ψευδο-αναγνωριστικού χρησιμοποιούνται τα γνωρίσματα {Ηλικία, Ταχυδρομικός κωδικός, Φύλο}, ως ευαίσθητο γνώρισμα εμφανίζεται ο «Μισθός» το οποίο παραμένει στην αρχική του μορφή κατά την εφαρμογή της *k*-ανωνυμίας, ενώ το γνώρισμα «Ύψος» δεν ανήκει σε καμία από τις δύο κατηγορίες, μιας και θεωρούμε ότι δεν αποτελεί κίνδυνο αναγνώρισης κάποιου ατόμου και μπορεί να δημοσιευθεί ως έχει. Χρησιμοποιούνται οι ιεραρχίες γενίκευσης όπως ενδεικτικά εμφανίζονται παρακάτω για τα αντίστοιχα γνωρίσματα. Συγκεκριμένα για την ανακωδικοποίηση του πεδίου «Ταχυδρομικός Κωδικός» χρησιμοποιείται η ιεραρχία γενίκευσης που παρουσιάζεται στην Εικόνα 2.3 ενώ για το πεδίο «Φύλο» χρησιμοποιείται η ιεραρχία γενίκευσης της Εικόνας 2.4, όπου παρατηρούμε ότι δεν υπάρχει περαιτέρω γενικευμένη τιμή για τις δύο τιμές και άρα η μόνη λύση είναι η απόκρυψη των αρχικών τιμών.

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο	Ύψος	Μισθός
1	25	18540	Γυναίκα	1,80	500
2	27	18530	Άνδρας	1,73	900
3	34	18050	Άνδρας	1,77	700
4	31	14244	Άνδρας	1,67	900
5	37	14430	Γυναίκα	1,55	800
6	30	14500	Γυναίκα	1,67	800

Πίνακας 2.4



Εικόνα 2.3: Ιεραρχία γενίκευσης γνωρίσματος «Ταχυδρομικός Κωδικός»



Εικόνα 2.4: Ιεραρχία γενίκευσης γνωρίσματος «Φύλο»

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο	Ύψος	Μισθός
1	<35	[18000,18999]	*	1,80	500
2	<35	[18000,18999]	*	1,73	900
3	<35	[18000,18999]	*	1,77	700
4	[30,39]	[14000,14999]	*	1,67	900
5	[30,39]	[14000,14999]	*	1,55	800
6	[30,39]	[14000,14999]	*	1,67	800

Πίνακας 2.5: 3-ανωνυμοποίηση του Πίνακα 2.4

Όπως φαίνεται και στο παράδειγμα του Πίνακα 2.5 η *k*-ανωνυμία καλύπτει επαρκώς την ιδιωτικότητα σε επιθέσεις με ζητούμενο από τον επιτιθέμενο την επιβεβαίωση της ταυτότητας κάποιας εγγραφής.

2.2.2 Μοντέλα k-ανωνυμοποίησης

Στον τομέα της προστασίας της ιδιωτικότητας εξετάζεται η εφαρμογή της *k*-ανωνυμίας σε σύνολα δεδομένων που προέρχονται από διαφορετικά μοντέλα δεδομένων. Κατά την ανωνυμοποίηση εφαρμόζονται ανάλογες τεχνικές και μέθοδοι σε κάθε περίπτωση ώστε να ικανοποιούνται βέλτιστα οι απαιτήσεις για το κάθε σύνολο δεδομένων. Μέσα από αυτή τη διαδικασία έχουν προκύψει ήδη κάποια μοντέλα ανωνυμοποίησης. Σύμφωνα με [LDR05], τα

μέχρι τώρα προτεινόμενα μοντέλα *k*-ανωνυμοποίησης, μπορούν να κατηγοριοποιηθούν βάσει τριών κριτηρίων. Συγκεκριμένα μπορούν να διαχωριστούν αναλόγως αν εφαρμόζουν:

- *Γενίκευση (generalization) ή απόκρυψη (suppression):* Ο διαχωρισμός αυτός γίνεται ανάλογα με την τεχνική που χρησιμοποιούν για την επίτευξη της *k*-ανωνυμοποίησης, δηλαδή της μεθόδου με την οποία τροποποιούν τα δεδομένα ώστε να ικανοποιούν τελικά την *k*-ανωνυμία. Κάποια από τα προτεινόμενα μοντέλα χρησιμοποιούν μόνο την απόκρυψη δεδομένων, αποκρύπτοντας όσα από αυτά επηρεάζουν την *k*-ανωνυμία του συνόλου, ενώ κάποια άλλα γενικεύουν σημασιολογικά τις αρχικές τιμές των δεδομένων για τον ίδιο σκοπό. Η επιλογή της τεχνικής που θα χρησιμοποιηθεί για την *k*-ανωνυμοποίηση των δεδομένων μπορεί να είναι και ένας συνδυασμός αυτών των δύο τεχνικών.
- *Τοπική (local) ή καθολική (global) ανακωδικοποίηση*: Η διάκριση αυτή γίνεται βάσει του αν στο μοντέλο ανωνυμοποίησης όλες οι εμφανίσεις μιας τιμής του πεδίου τιμών σε γνωρίσματα του ψευδο-αναγνωριστικού ανακωδικοποιούνται με την ίδια πάντα γενικευμένη τιμή ή αν τροποποιούνται μόνο ορισμένες εμφανίσεις της τιμής στα δεδομένα, δηλαδή αν κάθε εμφάνιση της ίδια αρχικής τιμής μπορεί να ανακωδικοποιηθεί σε διάφορες γενικευμένες τιμές και όχι απαραίτητα στην ίδια. Ένα συνηθισμένο μοντέλο καθολικής ανακωδικοποίησης είναι η *γενίκευση πλήρους πεδίου (full domain generalization)*. Κατά το μοντέλο αυτό, το σύνολο των τιμών του πεδίου τιμών του γνωρίσματος αντιστοιχίζεται σε ένα πιο γενικό πεδίο μέσω της ιεραρχίας γενίκευσης του. Επομένως, στην περίπτωση αυτή κάθε εμφάνιση μιας τιμής του γνωρίσματος θα αντικαθίσταται από την ίδια πάντα γενικευμένη τιμή, όπως αυτή ορίζεται μέσα στο γενικευμένο πεδίο από την ιεραρχία γενίκευσης του πεδίου που χρησιμοποιείται. Στο παράδειγμα του Πίνακα 2.5 παρουσιάζεται ένα μοντέλο γενίκευσης πλήρους πεδίου μιας και κάθε αρχική τιμή, από τα κατώτερα επίπεδα των ιεραρχιών γενίκευσης αντικαθίστανται πάντα με την αμέσως πιο γενική τιμή της όπως δίνεται από τις ιεραρχίες γενίκευσης στις Εικόνες 2.3 και 2.4.

  Τα μοντέλα ανακωδικοποίησης που εφαρμόζουν τοπική ανακωδικοποίηση τείνουν να είναι πιο αποδοτικά, λόγω των περισσότερων επιλογών που προσφέρουν κατά την αντικατάσταση των αρχικών τιμών. Όπως και ο αλγόριθμος που παρουσιάζεται στην εργασία αυτή, δίνουν την δυνατότητα να επιλέγεται κατά περίπτωση η βέλτιστη γενικευμένη τιμή, κάτι που οδηγεί συνήθως σε μικρότερη απώλεια πληροφορίας από εκείνη που περιέχεται στα αρχικά δεδομένα.

  Τα μοντέλα καθολικής ανακωδικοποίησης διακρίνονται επιπλέον σε μοντέλα μονοδιάστατης ή πολυδιάστατης ανακωδικοποίησης.

Ένα μοντέλο ορίζεται ως *μοντέλο μονοδιάστατης ανακωδικοποίησης (single-dimension recoding)* όταν ανακωδικοποιεί κάθε γνώρισμα του ψευδο-αναγνωριστικού ξεχωριστά. Ο πλέον γνωστός αλγόριθμος που εφαρμόζει καθολική μονοδιάστατη ανακωδικοποίηση για την επίτευξη της *k*-ανωνυμοποίησης είναι ο αλγόριθμος Incognito, όπως παρουσιάζεται στο [LDR05]. Αντίστοιχα ορίζεται ως *μοντέλο ανακωδικοποίησης πολλαπλών διαστάσεων (multi-dimension recoding)* όταν τα πεδία των γνωρισμάτων του ψευδο-αναγνωριστικού αντιμετωπίζονται ως ένα διάνυσμα, όπου η ανακωδικοποίηση γίνεται στο διάνυσμα των τιμών κάθε πλειάδας, με παράδειγμα το μοντέλο ανακωδικοποίησης που εφαρμόζει ο αλγόριθμος Mondrian.

- *Βασισμένα σε ιεραρχία ή σε διαμέριση:* Ένα μοντέλο *βασισμένο σε ιεραρχία (hierarchy-based model)* χρησιμοποιεί κατά την ανακωδικοποίηση την δεδομένη ιεραρχία γενίκευσης τιμών του πεδίου. Οι αρχικές τιμές του πεδίου ανακωδικοποιούνται βάσει της ιεραρχίας γενίκευσης για το συγκεκριμένο γνώρισμα.

  Σε περίπτωση που το πεδίο τιμών είναι ένα πλήρως διατεταγμένο σύνολο, όπως οι φυσικοί αριθμοί, ορίζεται γενίκευση με *χρήση διαμέρισης (partition-based model)* του συνόλου σε ξένα μεταξύ τους διαστήματα. Η διάκριση αυτή γίνεται σε μοντέλα μονοδιάστατης ανακωδικοποίησης καθώς και σε μοντέλα ανακωδικοποίησης πολλαπλών διαστάσεων.

2.2.3 Αλγόριθμοι εύρεσης k-ανώνυμων πινάκων

Οι αλγόριθμοι υλοποίησης της *k*-ανωνυμοποίησης που μελετήθηκαν είναι ο Incognito και ο Mondrian. Οι δύο αυτοί αλγόριθμοι δέχονται ως είσοδο το σύνολο των αρχικών δεδομένων και επιστρέφουν την βέλτιστη προκύπτουσα από την διαδικασία που ακολουθούν γενίκευση του αρχικού συνόλου. Με τον όρο βέλτιστη γενίκευση θεωρείται η γενίκευση κατά την οποία παρέχεται προστασία της ιδιωτικότητας μέσω της εφαρμογής της *k*-ανωνυμίας και εξασφαλίζεται η μικρότερη απώλεια πληροφορίας. Και οι δύο αλγόριθμοι αντιπροσωπεύουν μοντέλα γενίκευσης, αφού αυτή την τεχνική χρησιμοποιούν. Η κύρια διαφορά τους βρίσκεται στο μοντέλο ανακωδικοποίησης που χρησιμοποιούν, ο Incognito είναι αλγόριθμος μονοδιάστατης καθολικής ανακωδικοποίησης, ενώ ο αλγόριθμος Mondrian εφαρμόζει πολυδιάστατη ανακωδικοποίηση στο επίπεδο της κλάσης ισοδυναμίας. Ο αλγόριθμος Incognito αντιμετωπίζει κάθε γνώρισμα χωριστά, ενώ ο αλγόριθμος Mondrian γενικεύει κάθε εγγραφή στο σύνολο των γνωρισμάτων της. Και οι δύο στοχεύουν στην εφαρμογή της *k*-ανωνυμίας στα δεδομένα, όμως παρέχουν διαφορετικής ποιότητας *k*-ανωνυμοποιήσεις λόγω των διαφορετικών μεθόδων που ακολουθούν.

2.2.3.1 Incognito

Ο αλγόριθμος Incognito, όπως αναλυτικά παρουσιάζεται από [LDR05], παρέχει την υλοποίηση του μοντέλου της *k*-ανωνυμοποίησης με γενίκευση πλήρους πεδίου. Αντιστοιχίζει δηλαδή κάθε εμφάνιση κάποιας τιμής του πεδίου του γνωρίσματος στην ίδια πάντα γενικευμένη τιμή για όλες τις τιμές του πεδίου, βασιζόμενος στην δοσμένη ιεραρχία γενίκευσης κάθε γνωρίσματος που ανήκει στο ψευδο-αναγνωριστικό.

Αρχικά, χρησιμοποιεί την προκαθορισμένη ιεραρχία γενίκευσης πεδίου για κάθε γνώρισμα ως μία *διάσταση*, με τη χρήση των οποίων φτιάχνει το πλέγμα γενίκευσης πολλαπλών γνωρισμάτων. Σε αυτό δίνονται όλοι οι δυνατοί συνδυασμοί μεταξύ των επιπέδων των ιεραρχιών γενίκευσης των γνωρισμάτων του ψευδο-αναγνωριστικού, όπου εκφράζονται ουσιαστικά όλες οι δυνατές γενικεύσεις των πλειάδων. Οι συνδυασμοί μετέπειτα ελέγχονται για την ικανοποίηση της *k*-ανωνυμίας. Στόχος του αλγορίθμου είναι η εύρεση της ελάχιστης γενίκευσης πλήρους πεδίου, η οποία ορίζεται με την βοήθεια του διανύσματος απόστασης.

- Ως *διάνυσμα απόστασης (distance vector)* μεταξύ δύο διανυσμάτων πεδίων τιμών $< D_{A_1}, \dots, D_{A_n} >$ και $< D_{B_1}, \dots, D_{B_n} >$ ορίζεται το διάνυσμα $DV = [d_1, \dots, d_n]$ όπου κάθε τιμή εκ των d_i ,$1 \leq i \leq n$ συμβολίζει το μήκος του μονοπατιού μεταξύ των πεδίων τιμών D_{A_i} και D_{B_i} από την ιεραρχία γενίκευσης πεδίου H_i .
- Η *ελάχιστη k-ανώνυμη γενίκευση πλήρους πεδίου (minimal k-anonymous full-domain generalization)* ορίζεται εκείνη που είναι *k*-ανώνυμη και έχει το μικρότερο ύψος γενίκευσης πολλαπλών γνωρισμάτων από όλες τις *k*-ανώνυμες αντίστοιχες γενικεύσεις.
- Ως *ύψος (height)* της γενίκευσης ορίζεται το άθροισμα των τιμών των αντίστοιχων διανυσμάτων απόστασης.

Ο αλγόριθμος ακολουθεί επαναληπτική διαδικασία για την επιλογή της ελάχιστης γενίκευσης πλήρους πεδίου.

Ξεκινώντας από κάθε γνώρισμα ελέγχει την ικανοποίηση της *k*-ανωνυμίας ως προς τις διαφορετικές δυνατές γενικεύσεις από την ιεραρχία γενίκευσης και προσθέτει σταδιακά τα υπόλοιπα γνωρίσματα του ψευδο-αναγνωριστικού μέχρι να εμφανιστεί *k*-ανώνυμος συνδυασμός γενικεύσεων του συνόλου των γνωρισμάτων του ψευδο-αναγνωριστικού. Για την εύρεση αυτού χρησιμοποιεί την *ιδιότητα του υποσυνόλου (subset property):* Αν μία σχέση-πίνακας $RT(A_1, A_2, \dots, A_n)$ είναι *k*-ανώνυμη ως προς το σύνολο γνωρισμάτων Q της σχέσης, τότε είναι *k*-ανώνυμη ως προς κάθε υποσύνολο γνωρισμάτων $P \subseteq Q$.

Για κάθε συνδυασμό γνωρισμάτων, υποσύνολο του ψευδο-αναγνωριστικού ακολουθεί την παρακάτω διαδικασία:

- Δημιουργεί το πλέγμα όλων των πιθανών συνδυασμών των διαφορετικών επιπέδων γενίκευσης των γνωρισμάτων που συμμετέχουν στο σύνολο που ελέγχει, βάσει των δοσμένων ιεραρχιών γενίκευσης.
- Ελέγχει αρχικά τα κατώτερα επίπεδα του πλέγματος, τα οποία αντιπροσωπεύουν τιμές των γνωρισμάτων πλησιέστερα στις αρχικές, ως προς την ικανοποίηση της *k*-ανωνυμίας των δεδομένων. Ελέγχει δηλαδή αν το σύνολο των γνωρισμάτων ικανοποιεί την *k*-ανωνυμία όταν οι αρχικές τιμές των γνωρισμάτων αντικατασταθούν με τις γενικευμένες τιμές του συνδυασμού των επιπέδων που εξετάζει.
- Αν ο συνδυασμός των γενικεύσεων των γνωρισμάτων είναι *k*-ανώνυμος, έπεται βάσει της ιδιότητας του υποσυνόλου ότι κάθε περαιτέρω γενίκευση τους είναι *k*-ανώνυμη.
- Αν ο συνδυασμός δεν είναι *k*-ανώνυμος, ανεβαίνει ένα επίπεδο στο πλέγμα των συνδυασμών των γενικεύσεων και εξετάζει εκεί την *k*-ανωνυμία.
- Αντιστρέφοντας την ιδιότητα του υποσυνόλου, έπεται πως αν ένα υποσύνολο γνωρισμάτων του ψευδο-αναγνωριστικού δεν ικανοποιεί την *k*-ανωνυμία, το ίδιο θα ισχύει και για κάθε σύνολο γνωρισμάτων που το περιέχει. Με χρήση αυτής δημιουργεί το πλέγμα όλων των *k*-ανώνυμων συνδυασμών των διαφορετικών επιπέδων γενίκευσης όλων των γνωρισμάτων του ψευδο-αναγνωριστικού απ'όπου επιλέγεται ο πιο αποδοτικός.

Ο Incognito είναι ένας ορθός και πλήρης αλγόριθμος ως προς την *k*-ανωνυμοποίηση που παράγει χρησιμοποιώντας γενίκευση πλήρους πεδίου.

Η πολυπλοκότητα του αλγορίθμου Incognito είναι τελικά εκθετική ως προς το μέγεθος του συνόλου των γνωρισμάτων του ψευδο-αναγνωριστικού.

2.2.3.2 Mondrian

Σε αρκετές περιπτώσεις, ο αλγόριθμος Incognito επιστρέφει μία πλήρους πεδίου γενίκευση η οποία υπεργενικεύει τα δεδομένα καθιστώντας τα άχρηστα για εκείνους στους οποίους απευθύνονται. Πράγματι, για αριθμητικά δεδομένα με γνωρίσματα από ένα κοινό πεδίο τιμών, μία γενίκευση πλήρους πεδίου σημαίνει αντικατάσταση όλων των αρχικών τιμών με σταθερά μη επικαλυπτόμενα μεταξύ τους διαστήματα ή πλήρη απόκρυψή τους. Ο αλγόριθμος Mondrian όπως ορίζεται και παρουσιάζεται από [LDR06], σε τέτοιες περιπτώσεις προσφέρει υψηλότερης ποιότητας ανωνυμοποίηση, με μεγαλύτερο βαθμό ελαστικότητας λόγω του πολυδιάστατου μοντέλου τοπικής ανακωδικοποίησης με το οποίο μπορεί να εφαρμοστεί.

Χρήσιμη για την παρουσίαση του αλγορίθμου είναι η χωρική αναπαράσταση της σχέσης $RT(A_1, A_2, \ldots, A_n)$ με ψευδο-αναγνωριστικό $QI_{RT} = (A_1, A_2, \ldots, A_d)$ με αντίστοιχα πεδία τιμών των γνωρισμάτων, $D_{A_1}, D_{A_2}, \ldots, D_{A_d}$. Θεωρώντας πως κάθε πεδίο των γνωρισμάτων του ψευδο-αναγνωριστικού έχει μία ολική διάταξη, οι προβολές του $RT(A_1, A_2, \ldots, A_n)$ στα $A_1, A_2, .., A_d$ μπορούν να αναπαρασταθούν ως ένα πολύ-σύνολο σημείων στον d-διάστατο χώρο.

Πρόκειται για έναν άπληστο αλγόριθμο, που χρησιμοποιεί την χωρική αναπαράσταση των πλειάδων και αναζητά πάνω σε αυτή την βέλτιστη πολυδιάστατη τομή σε κάθε στάδιο. Χρησιμοποιεί τις παρακάτω έννοιες:

- *Σύνολο συχνοτήτων (frequency set):* του γνωρίσματος A για την διαμέριση P ορίζεται από [LDR06], το σύνολο των μοναδικών τιμών του A στην P, με κάθε τιμή συνοδευόμενη από τον αριθμό εμφανίσεών της στην διαμέριση αυτή.
- *Αυστηρή πολυδιάστατη διαμέριση (strict multidimensional partitioning):* Ορίζεται ως το σύνολο των μη επικαλυπτόμενων περιοχών που καλύπτουν το d-διάστατο χώρο των πεδίων τιμών των γνωρισμάτων του ψευδο-αναγνωριστικού,

  $D_{A_1} \times D_{A_2} \times \ldots \times D_{A_d}$. Η συνάρτηση ανακωδικοποίησης φ αντιστοιχίζει κάθε πλειάδα $(x_1, x_2, \ldots, x_n) \in D_{A_1} \times D_{A_2} \times \ldots \times D_{A_d}$ σε ένα συναθροιστικό στατιστικό για την περιοχή στην οποία ανήκει.
- *Ελάχιστη αυστηρή πολυδιάστατη διαμέριση (minimal strict multidimensional partitioning):* Έστω $R_1, R_2, \ldots, R_n$ το σύνολο των περιοχών που προκύπτουν από μια αυστηρή πολυδιάστατη διαμέριση και κάθε περιοχή R_i να περιέχει το πολυσύνολο P_i αποτελούμενο από σημεία. Η πολυδιάστατη διαμέριση αυτή είναι *ελάχιστη* αν $\forall i, |P_i| \geq k$ και δεν υπάρχει άλλη επιτρεπόμενη πολυδιάστατη τομή για το σύνολο P_i.
- *Επιτρεπόμενη πολυδιάστατη τομή (allowable multidimensional cut):* Σε ένα πολυσύνολο σημείων P του d-διάστατου χώρου, μία τομή κάθετη στον άξονα A_i ορίζεται ως επιτρεπόμενη από [LDR06] στην τιμή a_i, αν και μόνο αν

  $Count(P.A_i > a_i) \geq k$ και $Count(P.A_i \leq a_i) \geq k$. Αν δηλαδή και στις δύο πλευρές της τομής υπάρχουν περισσότερα των k σημεία.

Για την περίπτωση της τοπικής πολυδιάστατης ανακωδικοποίησης η κύρια διαφορά εμφανίζεται στην επιλογή της διαμέρισης όπου επιτρέπονται επικαλυπτόμενα διαστήματα έτσι ώστε να επιτυγχάνεται η τοπική ανακωδικοποίηση. Για το λόγο αυτό ορίζεται:

- *Χαλαρή πολυδιάστατη διαμέριση (relaxed multidimensional partitioning):* Μια χαλαρή πολυδιάστατη διαμέριση για μία σχέση T ορίζει ένα σύνολο από (πιθανώς επικαλυπτόμενες) διακριτές πολυδιάστατες περιοχές οι οποίες καλύπτουν το d-διάστατο χώρο των πεδίων τιμών των γνωρισμάτων του ψευδο-αναγνωριστικού,

$D_{A_1} \times D_{A_2} \times ... \times D_{A_d}$. Η συνάρτηση ανακωδικοποίησης φ αντιστοιχίζει κάθε πλειάδα $(x_1, x_2, \ldots, x_n) \in D_{A_1} \times D_{A_2} \times ... \times D_{A_d}$ σε ένα συναθροιστικό στατιστικό για μία από τις περιοχές στην οποία ανήκει.

Ο αλγόριθμος ακολουθεί την διαδικασία:

- Αρχικά ορίζονται οι πολυδιάστατες περιοχές που καλύπτουν το χώρο των πεδίων του ψευδο-αναγνωριστικού $D_{A_1} \times D_{A_2} \times ... \times D_{A_d}$.
- Επιλέγεται η διάσταση κατά την οποία θα γίνει η διαμέριση. Υπάρχουν πολλοί τρόποι επιλογής, όμως κατά την υλοποίηση που χρησιμοποιείται στην εργασία επιλέγεται η διάσταση με το μεγαλύτερο εύρος τιμών .
- Υλοποιείται η διαμέριση κατά την παραπάνω επιλεγμένη διάσταση, βάσει του στατιστικού μέσου των τιμών του αντίστοιχου γνωρίσματος, έτσι ώστε οι τιμές που είναι μικρότερες ή ίσες με τον μέσο να βρίσκονται στην αριστερή κλάση και οι υπόλοιπες να βρίσκονται στην δεξιά κλάση ισοδυναμίας.
- Η διαδικασία επαναλαμβάνεται για κάθε μία από τις δύο προκύπτουσες κλάσεις ισοδυναμίας αναδρομικά μέχρι να μην υπάρχει άλλη επιτρεπόμενη πολυδιάστατη τομή για διαμέριση σε καμία διάσταση.
- Προκύπτει η βέλτιστη πολυδιάστατη διαμέριση και συνεπώς η κατάλληλη πολυδιάστατη γενίκευση που θα χρησιμοποιηθεί για την ανακωδικοποίηση των δεδομένων.

Με την διαδικασία αυτή, ο αλγόριθμος επιστρέφει την βέλτιστη πολυδιάστατη διαμέριση σε κάθε περιοχή της οποίας ανήκουν περισσότερες από k εγγραφές και συνεπώς ικανοποιείται η k-ανωνυμία.

Η πολυπλοκότητα του αλγορίθμου είναι $O(nlogn)$, όπου $n = |RT|$, ο αριθμός των εγγραφών που ανήκουν στην σχέση $RT(A_1, A_2, \ldots, A_n)$.

2.3 Αναγνώριση τιμής ευαίσθητων δεδομένων

2.3.1 Επιθέσεις κατά της k-ανωνυμίας

Η k-ανωνυμία παρότι αποτελεί βασική έννοια του τομέα της προστασίας της ιδιωτικότητας δεν εξασφαλίζει πλήρως την ιδιωτικότητα σε συγκεκριμένες επιθέσεις. Όπως έχει παρατηρηθεί από [MGK+06], η k-ανωνυμία δεν καλύπτει πλήρως σε επιθέσεις με ζητούμενο την ανακάλυψη της τιμής ευαίσθητων γνωρισμάτων των εγγραφών. Μετά την k-ανωνυμοποίηση των δεδομένων κάθε εγγραφή που ανήκει στον πίνακα δεν μπορεί να διακριθεί βάσει των τιμών που λαμβάνει στα γνωρίσματα του ψευδο-αναγνωριστικού,

ανάμεσα σε τουλάχιστον *k* εγγραφές του συνόλου. Όμως σε αρκετές περιπτώσεις όπως στατιστικές μελέτες ή έρευνες το ενδιαφέρον εστιάζεται στις τιμές του ευαίσθητου γνωρίσματος και απαιτείται αυτές να δημοσιεύονται ως έχουν. Στην περίπτωση που στα δεδομένα εμφανίζεται κάποιο ευαίσθητο γνώρισμα, οι τιμές του δεν επηρεάζονται από την *k*-ανωνυμοποίηση. Τότε οι τιμές του ευαίσθητου γνωρίσματος δημοσιεύονται με την αρχική τους μορφή, θεωρώντας πως εφόσον δεν μπορεί ο επιτιθέμενος να ταυτοποιήσει ένα άτομο με μία εγγραφή, δεν μπορεί να συμπεράνει με βεβαιότητα την τιμή που αυτό λαμβάνει στο ευαίσθητο γνώρισμα. Σε αυτές τις περιπτώσεις μπορεί ο επιτιθέμενος να μην έχει την δυνατότητα να συμπεράνει απευθείας την ταυτότητα κάποιας εγγραφής, μπορεί όμως να διεξάγει συμπεράσματα για την τιμή του ευαίσθητου γνωρίσματος μίας ή πολλών ομάδων εγγραφών. Ενέχει ο κίνδυνος να συνδυάσει κάποιες τιμές του ψευδο-αναγνωριστικού που γνωρίζει με κάποιες από αυτές που εμφανίζονται και αναλόγως των ιδιαιτεροτήτων που εμφανίζουν τα δεδομένα να συμπεράνει την κλάση ισοδυναμίας κάποιας εγγραφής και πιθανώς πληροφορίες σχετικά με την τιμή που λαμβάνει στο ευαίσθητο γνώρισμα.

Αναφορικά με την ανεπάρκεια της *k*-ανωνυμίας ως προς τη διαφύλαξη της ιδιωτικότητας, οι [MGK+06] επισημαίνουν δύο περιπτώσεις, τις επιθέσεις ομοιογένειας και τις επιθέσεις με πρότερη γνώση. Στις περιπτώσεις αυτές παρουσιάζονται δύο κίνδυνοι λόγω του ότι κατά την *k*-ανωνυμοποίηση το ευαίσθητο γνώρισμα δημοσιεύεται με τις αρχικές τιμές των εγγραφών. Εξαιτίας της κατανομής των ευαίσθητων τιμών του γνωρίσματος ή της γνώσης που κατέχει ο επιτιθέμενος σχετικά με κάποιες ευαίσθητες τιμές, ο επιτιθέμενος μπορεί να επιβεβαιώσει ή να αποκλείσει κάποιες από αυτές για το πρόσωπο που αναζητά. Με τον τρόπο αυτό συσχετίζει την εγγραφή με κάποιες τιμές και διεξάγει συμπεράσματα για την ευαίσθητη τιμή που λαμβάνει.

2.3.1.1 Επιθέσεις ομοιογένειας

Στο δημοσιευμένο σύνολο δεδομένων ενδέχεται σε κάποια κλάση ισοδυναμίας οι εγγραφές να εμφανίζουν ταυτόσημες τιμές στο ευαίσθητο γνώρισμα. Τότε οποιοσδήποτε έχει κάποια μερική γνώση, δηλαδή γνωρίζει μία ή περισσότερες τιμές για κάποια γνωρίσματα του ψευδο-αναγνωριστικού, έτσι ώστε να κατατάσσει το άτομο που αναζητά σε αυτή την κλάση ισοδυναμίας μπορεί με βεβαιότητα να συμπεράνει την τιμή που το άτομο παίρνει στο ευαίσθητο γνώρισμα χωρίς να χρειάζεται να το ταυτοποιήσει με κάποια εγγραφή από την κλάση ισοδυναμίας. Μια τέτοια περίπτωση παρουσιάζεται στο παράδειγμα του Πίνακα 2.7, ο οποίος προκύπτει από την 3-ανωνυμοποίηση του Πίνακα 2.6. Εκεί, αν ο επιτιθέμενος γνωρίζει πως η εγγραφή που αναζητά συμμετέχει στα δημοσιευμένα δεδομένα και έχει «Ηλικία» και «Ταχυδρομικό Κωδικό» που συμβαδίζει με τις τιμές της δεύτερης κλάσης

ισοδυναμίας των εγγραφών {4,5,6}, για παράδειγμα γνωρίζει πως το φυσικό πρόσωπο είναι μεγαλύτερο από 30 ετών όμως μικρότερο από 35, μπορεί να συμπεράνει με απόλυτη βεβαιότητα πως ο Μισθός του ατόμου αυτού θα είναι ίσος με 800.

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο	Μισθός
1	25	14540	Γυναίκα	500
2	27	14530	Άνδρας	1000
3	30	14550	Άνδρας	1000
4	31	14544	Άνδρας	800
5	37	14430	Γυναίκα	800
6	42	14600	Γυναίκα	800
7	40	14650	Άνδρας	900
8	35	14200	Γυναίκα	700
9	49	14660	Γυναίκα	700

Πίνακας 2.6

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο	Μισθός
1	≤30	[14500,14599]	*	500
2	≤30	[14500,14599]	*	1000
3	≤30	[14500,14599]	*	1000
4	>30	[14000,14999]	*	800
5	>30	[14000,14999]	*	800
6	>30	[14000,14999]	*	800
7	≥35	[14000,14999]	*	900
8	≥35	[14000,14999]	*	700
9	≥35	[14000,14999]	*	700

Πίνακας 2.7: 3-ανωνυμοποίηση του Πίνακα 2.6

Στην περίπτωση αυτή, ο επιτιθέμενος δε μπορεί να ταυτοποιήσει το πρόσωπο που ψάχνει με κάποια εγγραφή από το σύνολο, όμως έχοντας την κατάλληλη πληροφορία και βασιζόμενος στην ιδιαιτερότητα των τιμών του ευαίσθητου γνωρίσματος της δεύτερης κλάσης ισοδυναμίας, μπορεί να ανακαλύψει την τιμή αυτή για το συγκεκριμένο άτομο. Αυτό το γεγονός σε πολλές περιπτώσεις της καθημερινής ζωής ενδέχεται να παραβιάζει τον απόρρητο χαρακτήρα των προσωπικών δεδομένων. Τέτοια παραδείγματα μπορούν να ανεβρεθούν σε βάσεις δεδομένων που αφορούν εισοδήματα ή ακόμα και σε βάσεις δεδομένων με προσωπικά ιατρικά στοιχεία ατόμων τα οποία δημοσιεύονται σε ερευνητές ή φτάνουν στην κατοχή άλλων χωρίς την απαραίτητη δικαιοδοσία για κερδοσκοπικούς λόγους.

2.3.1.2 Επιθέσεις με πρότερη γνώση

Στο ίδιο μοντέλο δεδομένων, παρουσιάζεται από [MGK+06] μια ακόμα περίπτωση στην οποία εμφανίζεται η ανεπάρκεια της *k*-ανωνυμίας ως προς την προστασία της ιδιωτικότητας των δεδομένων. Στην περίπτωση αυτή ο επιτιθέμενος γνωρίζει με βεβαιότητα την συμμετοχή κάποιου ατόμου στο δημοσιευμένο σύνολο δεδομένων και έχει στην κατοχή του προηγούμενη ή γενική γνώση τέτοια ώστε να μπορεί να αποκλείσει κάποιες ευαίσθητες τιμές από την κλάση ισοδυναμίας στην οποία θεωρεί πως πιθανώς ανήκει το άτομο που αναζητά. Αποκλείοντας τις εγγραφές της κλάσης ισοδυναμίας που δεν επιβεβαιώνουν την πρότερη γνώση που μπορεί να έχει, του δίνεται η δυνατότητα να συμπεράνει την ισχύουσα τιμή του ευαίσθητου γνωρίσματος για το συγκεκριμένο άτομο. Ένα τέτοιο παράδειγμα μπορεί να προκύψει από τα δεδομένα του Πίνακα 2.7. Θεωρώντας τα ως το σύνολο δεδομένων που δημοσιεύει μια εταιρία για τους εργαζομένους της, ο επιτιθέμενος μπορεί να αναζητά τον μισθό που αντιστοιχεί σε έναν νέο συνάδελφό του που μόλις έχει προσληφθεί. Γνωρίζει πως αυτός θα εμφανίζεται στο σύνολο των δεδομένων, αφού εργάζονται μαζί και αντίστοιχα γνωρίζει πως η ηλικία του, ως νέος εργαζόμενος, πιθανώς είναι μικρότερη ή ίση των 30 ετών. Με τα δεδομένα αυτά μπορεί με μεγάλη βεβαιότητα να υποθέσει πως το ζητούμενο άτομο θα ανήκει στην πρώτη κλάση ισοδυναμίας του Πίνακα 2.7. Από τις τιμές του ευαίσθητου γνωρίσματος που δημοσιεύονται για αυτή την κλάση ισοδυναμίας μπορεί με βεβαιότητα να αποκλείσει την τιμή 1000 μιας και γνωρίζει πως αυτοί οι μισθοί αντιστοιχούν σε εργαζόμενους με αρκετή προϋπηρεσία όπως και ο ίδιος. Συνεπώς μπορεί να συμπεράνει ότι ο νέος συνάδελφός του λαμβάνει μισθό ίσο με 500, αφού έχει αποκλείσει όλες τις υπόλοιπες τιμές του γνωρίσματος «Μισθός» της κλάσης αυτής.

Στις δύο αυτές περιπτώσεις η *k*-ανωνυμία αφήνει δυνατότητες εκμετάλλευσης των δεδομένων και παραβίασης του προσωπικού απορρήτου ατόμων που συμμετέχουν σε δημοσιευμένες βάσεις δεδομένων. Με στόχο την κάλυψη τέτοιων περιπτώσεων προτάθηκε η έννοια της *l*-

διαφορετικότητας, η οποία μπορεί να εξασφαλίσει την ιδιωτικότητα στις προαναφερθείσες περιπτώσεις ελέγχοντας το πλήθος των διαφορετικών μεταξύ τους τιμών που λαμβάνει το ευαίσθητο γνώρισμα σε κάθε κλάση ισοδυναμίας.

2.3.2 l-Διαφορετικότητα

Οι δύο περιπτώσεις επίθεσης που αναφέρονται στην προηγούμενη παράγραφο, αντιπροσωπεύουν τις εγγυήσεις ιδιωτικότητας που αφορούν την ανακάλυψη της ευαίσθητης τιμής από τον επιτιθέμενο και ορίζονται από [MGK+06]. Συγκεκριμένα, δεδομένης της πρότερης γνώσης του επιτιθέμενου και της ιδιαιτερότητας της κατανομής των ευαίσθητων τιμών, ένας δημοσιευμένος πίνακας δεδομένων T^* μπορεί να επιτρέπει την διεξαγωγή πληροφορίας με δύο σημαντικούς τρόπους:

- *Θετική αποκάλυψη:* Η δημοσίευση του πίνακα T^* όπως αυτός προήλθε από τον αρχικό πίνακα δεδομένων T οδηγεί σε θετική αποκάλυψη αν ο επιτιθέμενος μπορεί ορθά να αναγνωρίσει την τιμή του ευαίσθητου γνωρίσματος με μεγάλη πιθανότητα. Το παράδειγμα της επίθεσης ομοιογένειας όπως εμφανίζεται στον Πίνακα 2.7 είναι ένα παράδειγμα θετικής αποκάλυψης, αφού ο επιτιθέμενος μπορεί με βεβαιότητα να συμπεράνει ότι η πραγματική ευαίσθητη τιμή του μισθού του ατόμου που αναζητά είναι 800, εφόσον αυτό ανήκει στην δεύτερη κλάση ισοδυναμίας.
- *Αρνητική αποκάλυψη:* Η δημοσίευση του πίνακα T^* όπως αυτός προήλθε από τον αρχικό πίνακα δεδομένων T οδηγεί σε αρνητική αποκάλυψη αν ο επιτιθέμενος μπορεί με βεβαιότητα να αποκλείσει ορισμένες από τις ευαίσθητες τιμές του γνωρίσματος. Ένα τέτοιο παράδειγμα εμφανίζεται από την επίθεση με πρότερη γνώση όπως έχει περιγραφεί προηγουμένως. Ο εργαζόμενος αναζητά τον μισθό ενός νέου συναδέλφου του από την πρώτη κλάση ισοδυναμίας του Πίνακα 2.7 και μπορεί με απόλυτη βεβαιότητα να αποκλείσει την ευαίσθητη τιμή 1000, η οποία γνωρίζει από πριν πως αντιστοιχεί σε εργαζόμενους με περισσότερα έτη προϋπηρεσίας και συνεπώς αναγνωρίζει ως ευαίσθητη τιμή του ατόμου που αναζητά την τιμή 500.

Όπως αναπτύσσεται στο [MGK+06], η *l*-διαφορετικότητα, επεκτείνοντας την *k*-ανωνυμία, επικεντρώνεται στην αποτροπή των επιθέσεων με στόχο την αναγνώριση της τιμής του ευαίσθητου γνωρίσματος κάποιας εγγραφής. Η *l*-διαφορετικότητα είναι μια εγγύηση ιδιωτικότητας η ικανοποίηση της οποίας από το σύνολο των δεδομένων μπορεί να αποτρέψει την αναγνώριση της ευαίσθητης τιμής κάποιας εγγραφής. Στην περίπτωση προστασίας από επιθέσεις που εφαρμόζεται η *l*-διαφορετικότητα, τα δεδομένα διαχωρίζονται σε ευαίσθητα και μη και θεωρείται ότι ο επιτιθέμενος γνωρίζει πως ο δημοσιευμένος πίνακας αποτελεί γενίκευση κάποιου αρχικού συνόλου δεδομένων, από όπου επιχειρεί να συμπεράνει την

ευαίσθητη τιμή κάποιας εγγραφής. Σε οποιαδήποτε βάση δεδομένων που αφορά προσωπικά δεδομένα, όπως για παράδειγμα το ιατρικό ιστορικό του συνόλου των ασθενών, ακόμα και έπειτα από την διαδικασία της *k*-ανωνυμοποίησής τους, υπάρχει η πιθανότητα παραβίασης του απορρήτου των ευαίσθητων δεδομένων που εμφανίζονται σε αυτήν από οποιονδήποτε επιτύχει πρόσβαση σε αυτά. Η *l*-διαφορετικότητα επιχειρεί να ορίσει το πλήθος των ευαίσθητων τιμών σε κάθε κλάση ισοδυναμίας έτσι ώστε να μην μπορεί κάποιος να συσχετίσει με απόλυτη βεβαιότητα μία εγγραφή με μία συγκεκριμένη ευαίσθητη τιμή. Με την ικανοποίηση της *l*-διαφορετικότητας ο επιτιθέμενος ακόμα και να γνωρίζει την κλάση ισοδυναμίας στην οποία υποθέτει ότι ανήκει η αναζητούμενη εγγραφή, δεν μπορεί να συμπεράνει την ευαίσθητη τιμή της αφού η εγγραφή θα συσχετίζεται με τουλάχιστον l ευαίσθητες τιμές που εμφανίζονται στην κλάση ισοδυναμίας.

Σύμφωνα με τον ορισμό της *l-διαφορετικότητας (l-diversity)* [MGK+06], ο αρχικός πίνακας δεδομένων $RT(A_1, A_2, \dots, A_n, S)$ περιέχει ένα ευαίσθητο γνώρισμα *S* και έχει ανωνυμοποιηθεί με τεχνικές γενίκευσης, από όπου προκύπτει ο γενικευμένος πίνακας $RT^*(A_1, A_2, \dots, A_n, S)$. Σε αυτόν οι εγγραφές χωρίζονται σε κλάσεις ισοδυναμίας (q^**-blocks*) ως προς τις τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού τους. Μια κλάση ισοδυναμίας ορίζεται ως *l-διαφορετική (l-diverse)* αν περιέχει τουλάχιστον $l \geq 2$ «καλώς ορισμένες τιμές» για το ευαίσθητο γνώρισμα. Αντίστοιχα, ένας πίνακας είναι *l-διαφορετικός* εάν κάθε κλάση ισοδυναμίας του είναι *l*-διαφορετική.

Στην πιο απλή της μορφή η *l*-διαφορετικότητα ορίζεται ως *διακριτή l-διαφορετικότητα (distinct l-diversity)* [LLV07], όπου ως «καλώς ορισμένες τιμές» θεωρούνται l διακριτές μεταξύ τους τιμές. Στην περίπτωση αυτή η συχνότητα εμφάνισης κάθε ευαίσθητης τιμής σε κάθε κλάση ισοδυναμίας δεν δέχεται περιορισμούς, με αποτέλεσμα την δυνατότητα διεξαγωγής συμπεράσματος από κάποιον με γνώση πάνω στην κατανομή των ευαίσθητων τιμών.

Για το λόγο αυτό η αρχή της *l*-διαφορετικότητας παρουσιάζεται με τέσσερις πιθανές υποστάσεις, ώστε να προσδιορίζονται κατάλληλα οι «καλώς ορισμένες τιμές»:

- *L-διαφορετικότητα με εντροπία (entropy l-diversity):* Θεωρώντας το ευαίσθητο γνώρισμα *S*, ένας πίνακας RT^* είναι *l-διαφορετικός με εντροπία* αν για κάθε κλάση ισοδυναμίας q^* όπως αυτή ορίζεται από τις εγγραφές με κοινές τιμές στα γνωρίσματα του ψευδο-αναγνωριστικού ισχύει η σχέση:

$$-\sum_{s \in S} p_{(q^*,s)} \cdot log p_{(q^*,s')} \geq log l$$

Όπου $p_{(q^*,s)} = \frac{n_{(q^*,s)}}{\sum_{s \in S} n_{(q^*,s')}}$ το κλάσμα των εγγραφών της κλάσης ισοδυναμίας q^* που έχουν ως ευαίσθητη τιμή την τιμή s και $n_{(q^*,s)}$ ο αριθμός των πλειάδων $t \in RT^*$ από

την κλάση ισοδυναμίας q^* με τιμή στο ευαίσθητο γνώρισμα ίση με s. Με τον τρόπο αυτό σε κάθε κλάση ισοδυναμίας εμφανίζονται τουλάχιστον l διάφορες μεταξύ τους τιμές για το ευαίσθητο γνώρισμα.

Για να είναι δυνατή η l-διαφορετικότητα με εντροπία σε έναν πίνακα η εντροπία του συνόλου του πίνακα πρέπει να είναι τουλάχιστον ίση με $logl$ κάτι που δεν ισχύει σε περιπτώσεις όπου μία τιμή του ευαίσθητου γνωρίσματος εμφανίζεται σε πολύ μεγαλύτερο ποσοστό από τις υπόλοιπες στο σύνολο των δεδομένων.

Λόγω της αυστηρής συνθήκης της, δεν μπορεί να εφαρμοστεί σε πολλά συχνά χρησιμοποιούμενα σύνολα δεδομένων, όπου τη λύση δίνει η αναδρομική (c, l)-διαφορετικότητα.

- *Αναδρομική (c,l)-διαφορετικότητα (recursive (c,l)- diversity)*: Έστω οι $s_1, s_2, \dots, s_m$ πιθανές τιμές του ευαίσθητου γνωρίσματος για μια κλάση ισοδυναμίας. Σε μια δοσμένη κλάση ισοδυναμίας q^* ορίζεται ως r_i ο αριθμός εμφανίσεων της i-οστής πιο συχνά εμφανιζόμενης τιμής μέσα στην κλάση, για το ευαίσθητο γνώρισμα S. Δοθείσης μιας σταθεράς c, η κλάση ισοδυναμίας q^* ικανοποιεί την αναδρομική (c, l) −διαφορετικότητα αν $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$. Ισοδύναμα η κλάση ισοδυναμίας ικανοποιεί την αναδρομική (c, l) −διαφορετικότητα αν αποκλείοντας μια από τις ευαίσθητες τιμές της, απομένει μία $(c, l-1)$-διαφορετική κλάση ισοδυναμίας, όπου και φαίνεται ο αναδρομικός χαρακτήρας του ορισμού. Ο πίνακας RT^* ικανοποιεί την αναδρομική (c, l) −διαφορετικότητα αν κάθε κλάση ισοδυναμίας του ικανοποιεί την αναδρομική (c, l) −διαφορετικότητα. Ο επιτιθέμενος με αυτόν τον ορισμό πρέπει να αποκλείσει τουλάχιστον $l - 1$ ευαίσθητες τιμές από την κλάση για μια επίθεση θετικής αποκάλυψης.

Για την περίπτωση κατά την οποία η θετική αποκάλυψη δεν αποτελεί σημαντική απειλή, όπως για παράδειγμα όταν κάποια ευαίσθητη τιμή είναι πολύ συχνή ή δεν αποτελεί κίνδυνο παραβίασης της ιδιωτικότητας, ορίζεται η αναδρομική (c, l)-διαφορετικότητα θετικής αποκάλυψης:

- *Αναδρομική (c,l)-διαφορετικότητα θετικής αποκάλυψης (positive disclosure-recursive (c,l)-diversity):* Έστω Y το σύνολο των τιμών του ευαίσθητου γνωρίσματος για τις οποίες επιτρέπεται η θετική αποκάλυψη. Έστω σε μια δοσμένη κλάση ισοδυναμίας q^* η πιο συχνά εμφανιζόμενη ευαίσθητη τιμή που δεν ανήκει στο σύνολο Y να είναι η y-οστή πιο συχνά εμφανιζόμενη τιμή και έστω r_i η συχνότητα εμφάνισης της i-οστής πιο συχνά εμφανιζόμενης τιμής μέσα στην κλάση ισοδυναμίας. Η κλάση αυτή ικανοποιεί την αναδρομική (c, l) −διαφορετικότητα θετικής αποκάλυψης αν ισχύει μία από τις παρακάτω συνθήκες:

- $y \leq l-1$ και $r_y < c\sum_{j=l}^{m} r_j$
- $y > l-1$ και $r_y < c\sum_{j=l-1}^{y-1} r_j + c\sum_{j=l+1}^{m} r_j$

Αντίστοιχα ορίζεται η αναδρομική (c_1, c_2, l)-διαφορετικότητα αρνητικής/θετικής αποκάλυψης κατά την οποία επιπλέον ελέγχονται οι ευαίσθητες τιμές για τις οποίες δεν επιτρέπεται η αρνητική αποκάλυψη:

- *Αναδρομική (c_1, c_2, l)-διαφορετικότητα αρνητικής/θετικής αποκάλυψης (negative/positive disclosure-recursive (c_1, c_2, l)-diversity):* Έστω W το σύνολο των ευαίσθητων τιμών για τις οποίες η αρνητική αποκάλυψη δεν είναι αποδεκτή. Ένας πίνακας ικανοποιεί την αναδρομική (c_1, c_2, l)-διαφορετικότητα αρνητικής/θετικής αποκάλυψης αν ικανοποιεί την αναδρομική (c, l) −διαφορετικότητα θετικής αποκάλυψης και κάθε τιμή $s \in W$ εμφανίζεται σε ποσοστό τουλάχιστον c_2 στο σύνολο των πλειάδων κάθε κλάσης ισοδυναμίας.

Επεκτείνοντας το παράδειγμα του Πίνακα 2.6, επιχειρείται η ικανοποίηση της διακριτής *l*-διαφορετικότητας σε ήδη *k*-ανωνυμοποιημένα δεδομένα προσαρμόζοντας τις προηγούμενες γενικεύσεις. Αντιμεταθέτουμε τις εγγραφές 3 και 4 ανάμεσα στις δύο πρώτες κλάσεις ισοδυναμίας και τις εγγραφές 6 και 8 μεταξύ των δύο τελευταίων κλάσεων ισοδυναμίας. Προκύπτει ο Πίνακας 2.8 που είναι 3-ανώνυμος και 3-διαφορετικός σε κάθε κλάση ισοδυναμίας του. Εκεί παρατηρείται ότι ενώ από τον αντίστοιχο 3-ανώνυμο Πίνακα 2.7, κάποιος μπορεί με απόλυτη βεβαιότητα να συμπεράνει την τιμή του μισθού για κάθε εγγραφή που μπορεί να ανήκει στην δεύτερη κλάση ισοδυναμίας ως την τιμή 800, η περίπτωση αυτή αποκλείεται εφόσον ισχύει η 3-διαφορετικότητα του Πίνακα 2.8. Η *l*-διαφορετικότητα επιβάλει τουλάχιστον 3 διαφορετικές μεταξύ τους ευαίσθητες τιμές σε κάθε κλάση ισοδυναμίας. Αντίστοιχα οι περιπτώσεις παραβίασης της ιδιωτικότητας κατά τις οποίες ο επιτιθέμενος έχει κάποια πρότερη γνώση, ελαχιστοποιούνται όσο μεγαλύτερη είναι η παράμετρος l καθώς τόσο μεγαλύτερη θα απαιτείται να είναι η προηγούμενη γνώση του επιτιθέμενου για να μπορεί να αποκλείσει τις υπόλοιπες $l-1$ ευαίσθητες τιμές και να βγάλει συμπεράσματα με βεβαιότητα. Στο συγκεκριμένο παράδειγμα λοιπόν ο εργαζόμενος που επιχειρεί να αναγνωρίσει τον μισθό του νέου συναδέλφου του θα πρέπει να γνωρίζει με ακρίβεια τις μισθοδοσίες για να μπορεί να απορρίψει τις δύο άλλες τιμές και να καταλήξει στον πραγματικό μισθό του συναδέλφου του.

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο	Μισθός
1	≤31	[14500,14599]	*	500
2	≤31	[14500,14599]	*	1000
4	≤31	[14500,14599]	*	800
3	[30,39]	[14000,14999]	*	1000
5	[30,39]	[14000,14999]	*	800
8	[30,39]	[14000,14999]	*	700
6	≥40	[14600,14699]	*	800
7	≥40	[14600,14699]	*	900
9	≥40	[14600,14699]	*	700

Πίνακας 2.8: 3-ανώνυμη και 3-διαφορετική έκδοση του Πίνακα 2.6

Η *l*-διαφορετικότητα υπερτερεί της *k*-ανωνυμίας, όταν το μοντέλο δεδομένων εμφανίζει ευαίσθητα γνωρίσματα επειδή:

- Δεν απαιτεί από τον κάτοχο των δεδομένων να έχει απαραίτητα την ίδια γνώση με τον επιτιθέμενο για να αποτρέψει την επίθεση, μιας και εφόσον υπάρχει ικανό πλήθος ευαίσθητων τιμών σε κάθε κλάση ισοδυναμίας καμία εγγραφή δεν μπορεί να συσχετιστεί μοναδικά με μία συγκεκριμένη ευαίσθητη τιμή.
- Αναλόγως του μεγέθους της παραμέτρου *l* εξασφαλίζει την ιδιωτικότητα απέναντι σε επιτιθέμενους με περισσότερη γνώση, αφού όσο μεγαλύτερο είναι το *l*, τόσο περισσότερη γνώση πρέπει να έχει ο επιτιθέμενος σχετικά με το ευαίσθητο γνώρισμα για να εξάγει συμπεράσματα.
- Καλύπτει τις περιπτώσεις επίθεσης προς μεμονωμένα πρόσωπα, αφού μία μόνο εγγραφή δεν μπορεί να συνδεθεί με μία τιμή του ευαίσθητου γνωρίσματος, παρά μόνο με τις ευαίσθητες τιμές της κλάσης ισοδυναμίας στην οποία ανήκει η εγγραφή.

2.3.3 t-Εγγύτητα

Παρά τα πλεονεκτήματα της *l*-διαφορετικότητας σε περιπτώσεις δημοσίευσης δεδομένων που οδηγούν στην αναγνώριση της τιμής ενός ευαίσθητου γνωρίσματος, η εφαρμογή της *l*-διαφορετικότητας σε κάποια σύνολα δεδομένων εμφανίζεται αναποτελεσματική [LLV07] . Συγκεκριμένα αποδεικνύεται πως δεν αποτελεί ικανή και αναγκαία συνθήκη για την

αποτροπή διεξαγωγής συμπεράσματος για το ευαίσθητο γνώρισμα. Η *l*-διαφορετικότητα περιορίζεται από την υπόθεση της γνώσης που κατέχει ο επιτιθέμενος, ο οποίος όμως στην πραγματικότητα μπορεί να αποκτήσει γνώση σχετικά με την τιμή του ευαίσθητου γνωρίσματος κάποιας εγγραφής λαμβάνοντας πληροφορία από την καθολική κατανομή του γνωρίσματος στο σύνολο των δεδομένων.

Η *l*-διαφορετικότητα παρουσιάζει κάποιους περιορισμούς στην εφαρμογή της σε πραγματικές βάσεις δεδομένων. Όπως έχει παρατηρηθεί σε πολλές περιπτώσεις είναι δύσκολο να ικανοποιηθεί, όπως για παράδειγμα όταν το ευαίσθητο γνώρισμα έχει ως δυνατές τιμές μόνο δύο εκ των οποίων η μία εμφανίζεται πολύ πιο συχνά από την άλλη. Στην περίπτωση αυτή, για να επιτευχθεί η *l*-διαφορετικότητα θα απαιτηθεί πολύ μεγαλύτερη γενίκευση, μιας και σε κάθε κλάση ισοδυναμίας θα πρέπει να εμφανίζονται και οι δύο ευαίσθητες τιμές. Συνεπώς θα σχηματιστούν πολύ μεγάλες κλάσεις ισοδυναμίας και άρα τα δεδομένα τους θα υπεργενικευτούν έτσι ώστε να παίρνουν τις ίδιες τιμές. Μια τέτοια περίπτωση μπορεί επίσης να οδηγήσει σε λανθασμένα συμπεράσματα, τα οποία όχι μόνο μπορεί να παραβιάσουν τα προσωπικά δεδομένα ενός ατόμου, αλλά και να δημοσιευθούν παραποιημένες πληροφορίες σχετικά με αυτό. Χαρακτηριστικό παράδειγμα τέτοιων περιπτώσεων βρίσκεται σε ιατρικές βάσεις δεδομένων όπου το ευαίσθητο γνώρισμα αναπαριστά το αποτέλεσμα κάποιου ιατρικού ελέγχου για μια σπάνια ασθένεια, στην οποία το μεγαλύτερο ποσοστό του πληθυσμού εμφανίζει αρνητική τιμή. Κατά την ικανοποίηση της *l*-διαφορετικότητας, θα πρέπει σε κάθε κλάση ισοδυναμίας να εμφανίζονται και οι δύο τιμές με αποτέλεσμα ακόμα και όσοι έχουν αρνητική τιμή να μπορούν να συσχετιστούν με την θετική τιμή απλά επειδή συμμετέχουν στο σύνολο των δεδομένων.

Η *l*-διαφορετικότητα εμφανίζεται ανεπαρκής σε κάποιες περιπτώσεις επιθέσεων όπου επιχειρείται η διεξαγωγή συμπεράσματος για την ευαίσθητη τιμή μιας εγγραφής. Οι περιπτώσεις που έχουν παρατηρηθεί [LLV07], κατατάσσονται συνήθως στις δύο κατηγορίες:

- *Επιθέσεις αλλοίωσης (skewness attacks):* Σε αυτήν την περίπτωση ανήκει το παράδειγμα που μόλις αναφέρθηκε. Σε αυτό, μπορεί να ικανοποιηθεί η *l*-διαφορετικότητα με υπεργενίκευση των δεδομένων ώστε να εμφανίζονται και οι δύο ευαίσθητες τιμές σε κάθε κλάση ισοδυναμίας, αλλά να συμπεραίνονται αλλοιωμένες πληροφορίες ως ακριβείς με σημαντική βεβαιότητα. Κάτι τέτοιο οδηγεί σε σημαντική παραβίαση της ιδιωτικότητας των ατόμων που συμμετέχουν στα δεδομένα μιας και τους προσδίδονται ιδιότητες που δεν έχουν στην πραγματικότητα.
- *Επιθέσεις ομοιότητας (similarity attacks):* Σε αυτή την κατηγορία κατατάσσονται οι περιπτώσεις κατά τις οποίες οι κλάσεις ισοδυναμίας είναι *l*-διαφορετικές, εμφανίζουν δηλαδή *l* διάφορες μεταξύ τους ευαίσθητες τιμές, όμως οι τιμές αυτές είναι σημασιολογικά σχετικές. Τότε η ιδιωτικότητα μπορεί να παραβιαστεί και ο

επιτιθέμενος μπορεί επίσης να συμπεράνει ευαίσθητες πληροφορίες. Παρόλο που δεν μπορεί να συσχετίσει κάποια εγγραφή με μία συγκεκριμένη ευαίσθητη τιμή, του δίνεται μία πληροφορία σχετικά με την πραγματική τιμή κάθε εγγραφής που συμμετέχει στην κλάση ισοδυναμίας. Μια τέτοια περίπτωση μπορεί να εμφανιστεί σε ιατρικές βάσεις δεδομένων, όπου οι ευαίσθητες τιμές του γνωρίσματος «Ασθένεια» σε μία κλάση ισοδυναμίας μπορεί να είναι σχετικές μεταξύ τους. Για παράδειγμα όλες οι εγγραφές μπορεί να έχουν ως τιμή κάποια διαφορετική μεταξύ τους ασθένεια συσχετιζόμενη με το ίδιο όργανο του σώματος και με τον τρόπο αυτό ο επιτιθέμενος να μπορεί να αναγνωρίσει μερικώς την ασθένεια του ατόμου που αναζητά. Άλλη μία περίπτωση μπορεί να εμφανιστεί σε βάσεις δεδομένων που αφορούν τη μισθοδοσία εργαζομένων όπου σε μία κλάση ισοδυναμίας, οι εγγραφές να εμφανίζουν κοντινές τιμές μεταξύ τους και να μπορεί έτσι να διεξαχθεί το συμπέρασμα ότι η κλάση αυτή αφορά τους υψηλόμισθους εργαζόμενους και να οριστεί ένα εύρος μέσα στο οποίο βρίσκονται οι μισθοί τους.

Για την αποτροπή τέτοιων περιπτώσεων, ορίστηκε η *t*–εγγύτητα, μια έννοια ιδιωτικότητας που μπορεί να αναπαραστήσει το γενικό γνωστικό υπόβαθρο του επιτιθέμενου πάνω στην κατανομή των τιμών του ευαίσθητου γνωρίσματος, όπως ορίζεται από [LLV07].

Η διαφορά της *t*-εγγύτητας με την *l*-διαφορετικότητα, αν και οι δύο λαμβάνουν υπόψη την πιθανή προηγούμενη γνώση του επιτιθέμενου, είναι πως η *l*-διαφορετικότητα επιχειρεί να περιορίσει την διαφοροποίηση της γνώσης του επιτιθέμενου μεταξύ της αρχικής του γνώσης για την ευαίσθητη τιμή και εκείνης που αποκτά έπειτα από την αναγνώριση της κλάσης ισοδυναμίας στην οποία συμμετέχει η εγγραφή. Εναλλακτικά, η *t*-εγγύτητα επιχειρεί να περιορίσει την διαφοροποίηση στην γνώση του επιτιθέμενου μεταξύ της γνώσης που αποκτά από το δημοσιευμένο σύνολο των δεδομένων αναφορικά με την κατανομή των τιμών του ευαίσθητου γνωρίσματος και της γνώσης που αποκτά για την κατανομή των ευαίσθητων τιμών στην κλάση ισοδυναμίας που βρίσκεται η εγγραφή που αναζητά. Μία κλάση ισοδυναμίας ικανοποιεί την *t-εγγύτητα (t-closeness)* αν η απόσταση της κατανομής των τιμών του ευαίσθητου γνωρίσματος μέσα στην κλάση ισοδυναμίας από την κατανομή των ευαίσθητων τιμών του γνωρίσματος στο σύνολο των δεδομένων δεν υπερβαίνει το άνω όριο *t*. Ένας πίνακας ικανοποιεί την *t*–εγγύτητα αν όλες οι κλάσεις ισοδυναμίας του την ικανοποιούν.

Το κύριο ερώτημα που προκύπτει σε συνέχεια του ορισμού της *t*-εγγύτητας είναι ο τρόπος μέτρησης της απόστασης μεταξύ των δύο κατανομών πιθανότητας των τιμών του ευαίσθητου γνωρίσματος. Αν και ορίζεται η *μεταβολική απόσταση (variational distance)* και η *απόσταση Kullback-Leibler,* η μέτρηση της απόστασης σκοπό έχει την αναπαράσταση της διαφοροποίησης των τιμών του ευαίσθητου γνωρίσματος με τρόπο ώστε μικρότερη απόσταση

να ερμηνεύεται ως λιγότερο διαφορετικές τιμές και συνεπώς μεγαλύτερος κίνδυνος εξόρυξης της προσωπικής πληροφορίας. Την καλύτερη αναπαράσταση της απόστασης, σύμφωνα με τους [LLV07], επιτυγχάνει η μετρική *Earth Mover's Distance*, η οποία ορίζεται από το πρόβλημα μετακίνησης. Βασίζεται στο ελάχιστο απαιτούμενο έργο για την μετατροπή της μίας κατανομής στην άλλη με την μαζική ανακατανομή ανάμεσά τους.

Έτσι, για δύο κατανομές, έστω $\boldsymbol{P} = [p_1, p_2, \dots, p_m]$ και $\boldsymbol{Q} = [q_1, q_2, \dots, q_m]$ με d_{ij} να αναπαριστά την σταθερή απόσταση μεταξύ των τιμών p_i και q_j, στο πρόβλημα μετακίνησης ζητείται η εύρεση ροής $F = [f_{ij}]$, η ροή της μάζας από το p_i στο q_j έτσι ώστε να ελαχιστοποιηθεί το συνολικό έργο το οποίο είναι ίσο με την μετρική *EMD* και δίνεται από τη σχέση:

$EMD[\boldsymbol{P}, \boldsymbol{Q}] = WORK(\boldsymbol{P}, \boldsymbol{Q}, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} \cdot f_{ij}$

Για το πρόβλημα αυτό προκύπτουν οι συνθήκες:

$f_{ij} \geq 0, 1 \leq i, j \leq m$

$p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i, 1 \leq i, j \leq m$

$\sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1$

Από τον ορισμό της *t*-εγγύτητας και με χρήση της μετρικής *EMD* προκύπτουν δύο ιδιότητες για τους πίνακες που ικανοποιούν την *t*-εγγύτητα:

- *Ιδιότητα γενίκευσης:* Έστω *T* ένας πίνακας δεδομένων και *A, B* δύο πίνακες-γενικεύσεις του *T* τέτοιες ώστε ο *A* να είναι πιο γενικευμένος από τον *B*. Αν ο *T* ικανοποιεί την *t*-εγγύτητα με χρήση του *B*, τότε ο *T* ικανοποιεί την *t*-εγγύτητα και με χρήση του πίνακα *A*.
- *Ιδιότητα υποσυνόλου:* Έστω *T* ένας πίνακας δεδομένων και *C* το σύνολο των γνωρισμάτων που συμμετέχουν στα δεδομένα του *T*. Αν ο *T* ικανοποιεί την *t*-εγγύτητα ως προς το σύνολο *C* τότε ικανοποιεί την *t*-εγγύτητα ως προς κάθε υποσύνολο γνωρισμάτων $D \subset C$.

Οι δύο ιδιότητες εγγυώνται τη δυνατότητα ένταξης της έννοιας της *t*-εγγύτητας με χρήση της μετρικής *EMD* στο γενικό πλαίσιο του αλγορίθμου Incognito.

Όπως και στις περισσότερες από τις εγγυήσεις ιδιωτικότητας, η παράμετρος *t* δίνει την δυνατότητα εναλλαγής μεταξύ της χρησιμότητας των δεδομένων και της ιδιωτικότητας που εξασφαλίζει.

2.3.4 Ανατομία

Κατά τη δημοσίευση των δεδομένων με τις παραπάνω τεχνικές χάνεται χρήσιμη πληροφορία, σε διαφορετικά ποσοστά, από το αρχικό σύνολο δεδομένων. Αυτό καθιστά το σύνολο των δημοσιευμένων δεδομένων χρηστικά μη αποδοτικό, μιας και χρήσιμες και ακριβείς πληροφορίες που θα μπορούσαν να εξαχθούν μέσα από την ανάλυση των δεδομένων αλλοιώνονται ή αποκρύπτονται πλήρως. Αυτό οφείλεται στις γενικεύσεις των αρχικών τιμών των γνωρισμάτων προκειμένου να ταυτίζονται οι τιμές των εγγραφών σε κάθε κλάση ισοδυναμίας. Παράλληλα δεν προστατεύεται πάντα η συσχέτιση κάθε εγγραφής με την ευαίσθητη τιμή της.

Σύμφωνα με τους [XT06], η έννοια της ανατομίας μπορεί να λύσει αυτό το πρόβλημα, δημοσιεύοντας με τον προτεινόμενο τρόπο, τις αρχικές τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού, αλλά και τις ακριβείς τιμές του ευαίσθητου γνωρίσματος, αποκρύπτοντας ουσιαστικά την συσχέτιση κάθε εγγραφής με την ευαίσθητη τιμή της.

Στο μοντέλο δεδομένων μέσα στο οποίο ορίζεται η ανατομία, σε αρμονία με τον ορισμό της *l*-διαφορετικότητας, στον πίνακα $RT(A_1, A_2, \dots, A_n, S)$ υπάρχει ένα ευαίσθητο δεδομένο S και το σύνολο των γνωρισμάτων του ψευδο-αναγνωριστικού $QI_{RT} = (A_1, A_2, \dots, A_d)$. Στόχος είναι ο επιτιθέμενος να μην μπορεί με σημαντική βεβαιότητα να συσχετίσει μοναδικά την τιμή του ευαίσθητου γνωρίσματος με κάποια από τις εγγραφές του πίνακα, ενώ παράλληλα τα δεδομένα να δημοσιεύονται με τις αρχικές τιμές τους. Όμως στην περίπτωση εφαρμογής της ανατομίας ο επιτιθέμενος που αναζητά να προσδιορίσει ένα φυσικό πρόσωπο ως μία εγγραφή του συνόλου ή ακόμα και να επιβεβαιώσει την συμμετοχή του σε αυτά, μπορεί να το επιτύχει μιας και οι τιμές που λαμβάνουν οι εγγραφές στα γνωρίσματα του ψευδο-αναγνωριστικού δημοσιεύονται με την αρχική τους μορφή.

Βάσει της λογικής της ανατομίας ορίζονται:

- Μια *διαμέριση (partition)* των εγγραφών του αρχικού πίνακα αποτελείται από m υποσύνολα των εγγραφών του, τις *ομάδες QI*, τέτοια ώστε κάθε εγγραφή να ανήκει ακριβώς σε ένα από αυτά.
- Μια *l-διαφορετική διαμέριση* ως μια διαμέριση όπου κάθε QI_j ομάδα, $1 \leq j \leq m$ είναι *l*-διαφορετική, δηλαδή:

  Αν *v* είναι η πιο συχνά εμφανιζόμενη τιμή του ευαίσθητου γνωρίσματος στην ομάδα QI_j και $c_j(v)$ ο αριθμός των εγγραφών της ομάδας που λαμβάνουν αυτήν την τιμή, ικανοποιείται η συνθήκη:

$$\frac{c_j(v)}{|QI_j|} \leq \frac{1}{l}$$

με $|QI_j|$ το πλήθος των εγγραφών στην ομάδα QI_j .

Δοθείσης μίας *l*-διαφορετικής διαμέρισης του πίνακα $RT(A_1, A_2, \ldots, A_n, S)$ σε *m* ομάδες QI η ανατομία παράγει δύο πίνακες:

- *Τον πίνακα του ψευδο-αναγνωριστικού QIT (quasi-identifier table)*: Σε αυτόν εμφανίζονται οι επακριβείς τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού κάθε εγγραφής και επιπλέον μία στήλη με τον αριθμό της ομάδας *QI* στην οποία ανήκει η εγγραφή. Ο διαχωρισμός των εγγραφών σε ομάδες γίνεται με μια δεδομένη στρατηγική.
- *Πίνακας ευαίσθητων τιμών ST (sensitive table)*: Στον δεύτερο πίνακα δημοσιεύονται τα στατιστικά σχετικά με τις τιμές του ευαίσθητου γνωρίσματος που εμφανίζονται σε κάθε ομάδα. Πιο συγκεκριμένα, σε κάθε εγγραφή του πίνακα αντιστοιχεί μία ευαίσθητη τιμή. Για την κάθε τιμή δημοσιεύεται η ομάδα στην οποία αυτή εμφανίζεται και ο αριθμός των πλειάδων της ομάδας που λαμβάνουν αυτήν την τιμή.

Με τον τρόπο αυτό η ανατομία εξασφαλίζει την ιδιωτικότητα σε μοντέλα δεδομένων που παρουσιάζουν ευαίσθητα γνωρίσματα ενάντια σε επιθέσεις με στόχο την αναγνώριση της ευαίσθητης τιμής της εγγραφής, αφού δεν συνδέει καμία ευαίσθητη τιμή με καμία ακολουθία τιμών του ψευδο-αναγνωριστικού. Κατά τον επίσημο ορισμό της ανατομίας χρησιμοποιείται η έννοια της *l*-διαφορετικότητας, έτσι ώστε ο επιτιθέμενος να μπορεί να ανακατασκευάσει κάθε πλειάδα του αρχικού συνόλου σωστά, με χρήση των δημοσιευμένων δεδομένων, με μέγιστη πιθανότητα $1/l$. Κατά την δημοσίευση δεδομένων βάσει της λογικής της ανατομίας επιτρέπεται στον κάτοχο των δεδομένων να γνωρίζει τις πιθανές πληροφορίες που μπορεί ο επιτιθέμενος να συμπεράνει, εφόσον γνωρίζει πολύ εύκολα τους δυνατούς συνδυασμούς που μπορεί να γίνουν μεταξύ των ευαίσθητων τιμών και των ακολουθιών των τιμών του ψευδο-αναγνωριστικού για κάθε κλάση ισοδυναμίας. Αν και η ανατομία ορίζεται μέσω μιας *l*-διαφορετικής διαμέρισης, παράγει ένα ζεύγος πινάκων και για οποιαδήποτε άλλη διαμέριση με αντίστοιχο τρόπο.

Για την βέλτιστη κατανόηση της ανατομίας θεωρείται ως παράδειγμα το σύνολο δεδομένων του Πίνακα 2.9, όπου ως γνωρίσματα του ψευδο-αναγνωριστικού λαμβάνεται το σύνολο {Ηλικία, Ταχυδρομικός κωδικός, Φύλο} και ως ευαίσθητο γνώρισμα ο «Μισθός». Τα δεδομένα στη συνέχεια διαχωρίζονται σε δύο κλάσεις ισοδυναμίας όπως φαίνεται στον Πίνακα 2.10, έτσι ώστε οι κλάσεις αυτές να ικανοποιούν την 4-ανωνυμία και την 3-διαφορετικότητα για τις τιμές του ευαίσθητου γνωρίσματος. Έπειτα σύμφωνα με τη λογική της ανατομίας, δημιουργείται ο Πίνακας 2.11, ο *QIT*, με τις τιμές του ψευδο-αναγνωριστικού. Στον πίνακα αυτό, για κάθε εγγραφή του αρχικού πίνακα εμφανίζονται οι αρχικές τιμές στα γνωρίσματα του ψευδο-αναγνωριστικού, χωρίς την παρουσία του ευαίσθητου γνωρίσματος,

μαζί με τον αριθμό της κλάσης ισοδυναμίας (QI group) στην οποία ανήκει η εγγραφή από τον Πίνακα 2.10.

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο	Μισθός
1	25	14540	Γυναίκα	500
2	27	14530	Άνδρας	1000
3	34	14550	Άνδρας	1000
4	31	14544	Άνδρας	800
5	37	17430	Γυναίκα	950
6	39	18600	Γυναίκα	900
7	40	17650	Άνδρας	900
8	38	18200	Γυναίκα	700

Πίνακας 2.9

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο	Μισθός
1	≤35	[14530,14550]	*	500
2	≤35	[14530,14550]	*	1000
3	≤35	[14530,14550]	*	1000
4	≤35	[14530,14550]	*	800
5	>35	[17430, 18600]	*	950
6	>35	[17430, 18600]	*	900
7	>35	[17430, 18600]	*	900
8	>35	[17430, 18600]	*	700

Πίνακας 2.10: 4-ανώνυμη και 3-διαφορετική έκδοση του Πίνακα 2.9

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο	Ομάδα QI
1	25	14540	Γυναίκα	1
2	27	14530	Άνδρας	1
3	34	14550	Άνδρας	1
4	31	14544	Άνδρας	1
5	37	17430	Γυναίκα	2
6	39	18600	Γυναίκα	2
7	40	17650	Άνδρας	2
8	38	18200	Γυναίκα	2

Πίνακας 2.11: Πίνακας ψευδο-αναγνωριστικού *QIT* του Πίνακα 2.9

Αντίστοιχα προκύπτει ο πίνακας με τις αρχικές τιμές του ευαίσθητου γνωρίσματος *ST*, που παρουσιάζεται στον Πίνακα 2.12, όπως υπολογίζεται από τον διαχωρισμό των κλάσεων ισοδυναμίας του Πίνακα 2.10. Εκεί, αναγράφεται η ευαίσθητη τιμή, ο αριθμός της ομάδας *QI* στην οποία αυτή εμφανίζεται καθώς και ο αριθμός των εμφανίσεών της μέσα στην ομάδα-κλάση ισοδυναμίας.

Ομάδα QI	Μισθός	Αριθμός Εμφανίσεων
1	500	1
1	1000	2
1	800	1
2	950	1
2	900	2
2	700	1

Πίνακας 2.12: Πίνακας ευαίσθητων τιμών *ST* του Πίνακα 2.9

Ο επιτιθέμενος θεωρείται πως έχει γνώση πάνω στις τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού κάποιου ατόμου και επιχειρεί να το ταυτοποιήσει με κάποια εγγραφή και συνεπώς να συμπεράνει την ευαίσθητη τιμή του. Με τη δημοσίευση των δύο τελευταίων πινάκων, μπορεί από τον Πίνακα 2.11 να προσδιορίσει αν όντως το άτομο που αναζητά

συμμετέχει στα δεδομένα, όμως δεν μπορεί να το συσχετίσει με απόλυτη βεβαιότητα με καμία από τις ευαίσθητες τιμές της QI ομάδας στην οποία ανήκει, αφού κάθε ομάδα ικανοποιεί την l-διαφορετικότητα. Για κάθε ευαίσθητη τιμή για την ομάδα αυτή θα υπάρχει πιθανότητα $1/l$ να αφορά την εγγραφή που αναζητά. Έτσι, οι αρχικές τιμές στο σύνολο των γνωρισμάτων μπορούν να δημοσιευθούν με την αρχική τους μορφή διασφαλίζοντας παράλληλα την ιδιωτικότητα των εγγραφών σε επιθέσεις με σκοπό την αναγνώριση της ευαίσθητης τιμής τους.

Η ανατομία συγκριτικά με τη χρήση της γενίκευσης:

- Στην περίπτωση κατά την οποία ο επιτιθέμενος γνωρίζει τις τιμές του ψευδο-αναγνωριστικού μιας εγγραφής-στόχου και είναι βέβαιος πως το ζητούμενο πρόσωπο εμφανίζεται στο σύνολο δεδομένων, η ανατομία παρέχει την ίδια προστασία με τη χρήση γενίκευσης, δηλαδή επιτυγχάνεται η παραβίαση της ιδιωτικότητας ενός προσώπου με μέγιστη πιθανότητα $1/l$, όπως με την εφαρμογή της l-διαφορετικότητας στα δεδομένα. Παράλληλα δεν αλλοιώνει την χρήσιμη πληροφορία που περιέχουν τα αρχικά δεδομένα, αφού αυτά δημοσιεύονται με την αρχική τους μορφή.
- Στην περίπτωση όπου ο επιτιθέμενος γνωρίζει την ακολουθία τιμών του ψευδο-αναγνωριστικού μιας εγγραφής-στόχου, αλλά δεν γνωρίζει με βεβαιότητα αν αυτή εμφανίζεται στο σύνολο των δεδομένων, με τη χρήση της ανατομίας, ο επιτιθέμενος μπορεί, συνδυάζοντας άλλα δημοσιευμένα δεδομένα να αναγνωρίσει την ύπαρξη ή μη του ατόμου αυτού μέσα στο σύνολο, λόγω του ότι οι τιμές του ψευδο-αναγνωριστικού δίνονται αυτούσιες στην αρχική τους μορφή. Συνεπώς στην περίπτωση αυτή είναι προτιμότερη η χρήση γενίκευσης.

2.4 Πολλαπλές δημοσιεύσεις

Οι έννοιες που μέχρι τώρα έχουν παρουσιαστεί είναι εφαρμόσιμες στην περίπτωση της απλής, μοναδικής δημοσίευσης μιας συλλογής δεδομένων. Όμως ένα συχνό φαινόμενο στις πραγματικές δημοσιεύσεις συλλογών προσωπικών δεδομένων είναι η ανανέωση των περιεχομένων τους. Στην περίπτωση επανέκδοσης μιας ήδη δημοσιευμένης βάσης δεδομένων με τροποποιημένα κάποια από τα δεδομένα της, δίνεται η δυνατότητα στον υποψήφιο επιτιθέμενο να συσχετίσει τις δύο εκδόσεις και να αποκομίσει πληροφορία σχετική με κάποια εγγραφή, οδηγώντας έτσι στην καταπάτηση του απορρήτου των προσωπικών δεδομένων του συγκεκριμένου ατόμου. Την περίπτωση αυτή επιχειρεί να καλύψει μια νέα εγγύηση ιδιωτικότητας, η m-αμεταβλητότητα, όπως ορίζεται από [XT07].

Σε αρμονία με τον χρησιμοποιούμενο ως τώρα συμβολισμό, θεωρείται ο αρχικός πίνακας $RT_j(A_1, A_2, \ldots, A_n, S)$ πριν την j δημοσίευση, με σύνολο γνωρισμάτων του ψευδο-αναγνωριστικού $QI_{RT} = (A_1, A_2, \ldots, A_d)$ ενώ η j κατά σειρά ανωνυμοποιημένη δημοσίευση του αρχικού πίνακα συμβολίζεται ως RT_j^*. Για τον ορισμό της m-αμεταβλητότητας ορίζονται:

- Μια *διαμέριση* του αρχικού πίνακα δεδομένων αποτελούμενη από ξένες μεταξύ τους ομάδες του ψευδο-αναγνωριστικού, των οποίων η ένωση ισοδυναμεί με τον αρχικό πίνακα. Ως *ομάδες του ψευδο-αναγνωριστικού (QI groups)* ορίζονται τα υποσύνολα των πλειάδων του αρχικού πίνακα RT_j.
- *Μια τροποποιημένη τεχνική γενίκευσης (counterfeited generalization)*: Η ανωνυμοποιημένη έκδοση RT_j^* υπολογίζεται βασισμένη σε μια διαμέριση του RT_j και έχει τις ιδιότητες:
  - Η έκδοση RT_j^* περιέχει μια στήλη A^g, όπου αναγράφεται ο αριθμός της ομάδας στην οποία ανήκει η πλειάδα.
  - Κάθε πλειάδα του $t \in RT_j$ έχει μια αντίστοιχη γενικευμένη πλειάδα στον πίνακα $t^* \in RT_j^*$ τέτοια ώστε το ευαίσθητο γνώρισμα να έχει την αρχική του τιμή, ενώ τα γνωρίσματα του ψευδο-αναγνωριστικού να έχουν γενικευθεί με την χρήση διαμέρισης του πεδίου τιμών.
  - Όλες οι πλειάδες του πίνακα RT_j^* με τον ίδιο αριθμό ομάδας στο γνώρισμα A^g έχουν ταυτόσημες τιμές σε κάθε γνώρισμα του ψευδο-αναγνωριστικού.
  - Για κάθε ομάδα του ψευδο-αναγνωριστικού του RT_j , ο προς δημοσίευση πίνακας RT_j^*, ενδέχεται να περιέχει οποιονδήποτε αριθμό *πλαστών πλειάδων, (counterfeit tuples)* t_c^*, τέτοιες ώστε να έχουν τιμή στο ευαίσθητο γνώρισμα από το πεδίο τιμών του ευαίσθητου γνωρίσματος και να έχουν τιμή στο γνώρισμα A^g τον αριθμό της *QI* ομάδας που ανήκουν με τιμές στα γνωρίσματα του ψευδο-αναγνωριστικού, αντίστοιχες της ομάδας στην οποία εισάγονται.

  Όπως γίνεται αντιληπτό η παραπάνω τεχνική γενίκευσης ταυτίζεται με την κλασσική προαναφερθείσα τεχνική γενίκευσης στην περίπτωση μη ύπαρξης πλαστών εγγραφών.
- *Μια βοηθητική σχέση-πίνακας (auxiliary relation):* Συμβολιζόμενη ως R_j, συνοδεύει τον δημοσιευμένο πίνακα RT_j στην εκάστοτε έκδοση. Έχει δύο στήλες, η μία αφορά τον αριθμό κάθε *QI* ομάδας που εμφανίζει πλαστές εγγραφές και η δεύτερη τον αριθμό των πλαστών εγγραφών τις οποίες περιέχει η αντίστοιχη *QI* ομάδα.

- *Ιστορική ένωση (historical union):* Για $n \geq 1$, η ιστορική ένωση $U(n)$ περιέχει όλες τις πλειάδες που βρίσκονται στον RT πριν από κάθε δημοσίευση $1,2,\dots,n$ κατ' ακολουθία.

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

- *Διάρκεια ζωής (lifespan):* Για κάθε πλειάδα $t \in U(n)$ ορίζεται ως διάρκεια ζωής το διάστημα $[x, y]$,με x τον μικρότερο ακέραιο j τέτοιος ώστε η πλειάδα να εμφανίζεται στον πίνακα RT_j και y τον αντίστοιχα μεγαλύτερο ακέραιο j για τον οποίο η πλειάδα εμφανίζεται στον πίνακα RT_j.
- Έστω QI^* μια ομάδα QI στον δημοσιευμένο πίνακα RT_j^* για κάποια δημοσίευση j. Ορίζεται ως *υπογραφή (signature)* της ομάδας QI^*, το σύνολο των διακριτών ευαίσθητων τιμών που εμφανίζονται σε αυτήν.
- Ένας γενικευμένος πίνακας RT_j^* ορίζεται ως *m-μοναδικός* αν κάθε ομάδα QI που ανήκει σε αυτόν περιέχει τουλάχιστον m εγγραφές και όλες οι εγγραφές της ομάδας έχουν διάφορες μεταξύ τους τιμές στο ευαίσθητο γνώρισμα.

Με χρήση των παραπάνω εννοιών ορίζεται η αρχή της m-αμεταβλητότητας:

Μια ακολουθία δημοσιευμένων πινάκων $RT_1^*, RT_2^*, \dots, RT_n^*$ ικανοποιεί την *m-αμεταβλητότητα* αν ισχύουν οι συνθήκες:

- Ο πίνακας RT_j^* είναι m-μοναδικός για όλα τα $j \in [1, \dots, n]$
- Για κάθε εγγραφή $t \in U(n)$ με διάρκεια ζωής $[x, y]$, οι ομάδες $QI^*(x), QI^*(x+1), \dots, QI^*(y)$ στις οποίες εμφανίζεται η εγγραφή t στην αντίστοιχη έκδοση έχουν την ίδια υπογραφή.

2.5 Ταυτοποίηση ύπαρξης

Όπως φαίνεται από τους προηγούμενους ορισμούς, το μεγαλύτερο ενδιαφέρον στον τομέα της προστασίας της ιδιωτικότητας στρέφεται προς την πιθανότητα ταυτοποίησης μιας εγγραφής από σύνολα προσωπικών δεδομένων με ένα φυσικό πρόσωπο. Κάθε σύνολο προσωπικών δεδομένων που δημοσιεύεται ικανοποιεί διαφορετικές συνθήκες ανωνυμίας με αποτέλεσμα να μην υπάρχει μία και μόνο σαφής ένδειξη του πόσο προστατεύεται η ιδιωτικότητα των εγγραφών του συνόλου. Οι [NAC07] εξέτασαν το πρόβλημα «πότε τα δεδομένα θεωρούνται επαρκώς ανώνυμα» μέσω της ανάλυσης του κινδύνου επιβεβαίωσης της συμμετοχής ή όχι, ενός φυσικού προσώπου στα ανωνυμοποιημένα δεδομένα. Όρισαν επισήμως το πρόβλημα της απόκρυψης της παρουσίας των ατόμων σε μια δοσμένη βάση

δεδομένων και απέδειξαν την ανεπάρκεια της εφαρμογής της *k*-ανωνυμίας σε περιπτώσεις δημοσίευσης των τιμών των ευαίσθητων γνωρισμάτων των εγγραφών. Όρισαν μια μετρική για την εκτίμηση του κινδύνου προσδιορισμού κάποιου ατόμου σε ένα σύνολο δεδομένων βασιζόμενη στη γενίκευση δημοσιευμένων δεδομένων.

Συγκεκριμένα, δοθέντος ενός δημοσιευμένου πίνακα P και ενός ιδιωτικού πίνακα T η *δ-παρουσία (δ-presence)* ισχύει για μία γενίκευση T^* του πίνακα T με $\delta = (\delta_{min}, \delta_{max})$ αν $\delta_{min} \leq P(t \in T \mid T^*) \leq \delta_{max}$, $\forall\, t \in P$.

Τότε κάθε εγγραφή από τον πίνακα P, $t \in P$ είναι *δ-παρούσα (δ-present)* στον πίνακα T ενώ το διάστημα τιμών $\delta = (\delta_{min}, \delta_{max})$ είναι το εύρος των αποδεκτών τιμών για την πιθανότητα $P(t \in T \mid T^*)$.

Με τον ορισμό αυτό κάθε γενίκευση ενός ιδιωτικού πίνακα T μπορεί να αξιολογηθεί ως προς τον κίνδυνο της επιβεβαίωσης της παρουσίας κάποιας εγγραφής δεδομένου ενός ήδη δημοσιευμένου πίνακα P. Ο υπολογισμός της *δ*-παρουσίας συνεπάγεται τον υπολογισμό του εύρους πιθανοτήτων $(\delta_{min}, \delta_{max})$. Για μία εγγραφή του δημοσιευμένου πίνακα P ο υπολογισμός της πιθανότητας αυτή να εμφανίζεται στον αρχικό πίνακα T δεδομένου του γενικευμένου πίνακα T^*, είναι ίση με το πλήθος των εγγραφών του πίνακα T^* που γενικεύονται στην ίδια κλάση ισοδυναμίας προς το πλήθος των εγγραφών του δημοσιευμένου πίνακα P που γενικεύονται στις ίδιες τιμές με εκείνες των γνωρισμάτων του ψευδο-αναγνωριστικού της κλάσης ισοδυναμίας του πίνακα T^*.

Για την καλύτερη κατανόηση του ορισμού αυτού θεωρείται ως δημοσιευμένος πίνακας P , ο Πίνακας 2.13, με ευαίσθητο γνώρισμα το γνώρισμα «Μισθός», ως ιδιωτικός πίνακας T, ο Πίνακας 2.14, και ως η γενίκευση του T^* που εξετάζεται ως προς την *δ*-παρουσία ο Πίνακας 2.15.

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο	Μισθός
1	25	18540	Γυναίκα	500
2	27	18530	Άνδρας	900
3	29	18050	Άνδρας	700
4	30	18500	Γυναίκα	800
5	31	14244	Άνδρας	900
6	37	14430	Γυναίκα	800
7	40	14558	Άνδρας	700

Πίνακας 2.13: Δημοσιευμένος πίνακας *P*

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο
1	25	18540	Γυναίκα
2	27	18530	Άνδρας
5	31	14244	Άνδρας
6	37	14430	Γυναίκα

Πίνακας 2.14: Ιδιωτικός πίνακας *T*

Α/Α	Ηλικία	Ταχυδρομικός κωδικός	Φύλο
1	≤30	[18000,18999]	*
2	≤30	[18000,18999]	*
5	>30	[14000,14999]	*
6	>30	[14000,14999]	*

Πίνακας 2.15: $\left(\frac{1}{2},\frac{2}{3}\right)$ −παρούσα γενίκευση του πίνακα *T*, T^*

Ο Πίνακας 2.15 παρουσιάζει μία $\left(\frac{1}{2},\frac{2}{3}\right)$ −παρούσα γενίκευση T^* του πίνακα *T* ως προς τον δημοσιευμένο πίνακα *P*. Για τη διεξαγωγή αυτού του συμπεράσματος υπολογίζεται η πιθανότητα $P(t \in T\,|T^*)$, $\forall\, t \in P$. Έτσι για την εγγραφή 1 η πιθανότητα είναι $P = \frac{|\{1,2\}|}{|\{1,2,3,4\}|} = \frac{1}{2}$, όπως και για τις εγγραφές {2,3,6}, ενώ για την εγγραφή 4 η πιθανότητα είναι $P = \frac{|\{5,6\}|}{|\{5,6,7\}|} = \frac{2}{3}$,όπως και για τις εγγραφές {5,7}.

Για την απόδοση των ιδιοτήτων της *δ*-παρουσίας χρήσιμος είναι ο ορισμός της μη επικαλυπτόμενης γενίκευσης.

Έστω *P* ένας δημοσιευμένος πίνακας, *T* ένας ιδιωτικός πίνακας και T^* μια γενίκευση του *T*. Η γενίκευση T^* αποτελεί μια *μη-επικαλυπτόμενη γενίκευση(non-overlapping generalization)* ως προς τους *P, T* αν και μόνο αν δεν υπάρχουν $p \in P$, $t_1^*, t_2^* \in T$ τέτοια ώστε $t_1^* \neq t_2^*$ και $t_1^* \in \psi(p)$ και $t_2^* \in \psi(p)$, όπου $\psi(p)$ η συνάρτηση γενίκευσης που επιστρέφει τις δυνατές γενικευμένες τιμές της εγγραφής p. Δηλαδή η γενίκευση T^* αποτελεί μια *μη-επικαλυπτόμενη γενίκευση* ως προς τους *P, T* αν κάθε αρχική εγγραφή του πίνακα *P* μπορεί να ταυτιστεί με το πολύ μία γενικευμένη εγγραφή από τον πίνακα-γενίκευση T^*.

Με χρήση αυτού του ορισμού ορίζονται οι ιδιότητες της *δ*-παρουσίας.

- *Θεώρημα*: Δοθέντος ενός δημοσιευμένου πίνακα δεδομένων *P*, ενός ιδιωτικού πίνακα *T*, μιας μη-επικαλυπτόμενης γενίκευσης T_1^* του *T* και μιας μη-επικαλυπτόμενης γενίκευσης T_2^* της T_1^*, αν η γενίκευση T_1^* είναι $(\delta_{min}, \delta_{max})$-παρούσα ως προς τους πίνακες *P, T*, τότε και η γενίκευση T_2^* είναι αντίστοιχα $(\delta_{min}, \delta_{max})$-παρούσα ως προς αυτούς.
- *Πόρισμα*: Αν η γενίκευση T_2^* δεν είναι *δ*-παρούσα ως προς τους πίνακες *P, T* τότε δεν είναι *δ*-παρούσα ως προς αυτούς και η γενίκευση T_1^*.

Οι παράμετροι παρουσίας $\delta_{min}, \delta_{max}$ καθορίζουν το επίπεδο συμβιβασμού μεταξύ χρηστικότητας των ανωνυμοποιημένων δεδομένων και προστασίας της ιδιωτικότητας των εγγραφών.

Όσο περισσότερο προσεγγίζουν οι τιμές των δύο παραμέτρων μεταξύ τους τόσο μεγαλύτερη προστασία παρέχεται μέσω της γενίκευσης, κάτι που συνήθως οδηγεί σε μικρότερη προσφερόμενη χρησιμότητα από τα ανωνυμοποιημένα δεδομένα. Όπως προκύπτει από την περίπτωση της γενίκευσης με όλα τα δεδομένα να αποκρύπτονται, κάθε απαίτηση παρουσίας εύρους $(\delta_{min}, \delta_{max})$,για ένα δημοσιευμένο πίνακα *P* και ένα ιδιωτικό πίνακα *T*, πρέπει να ικανοποιεί την σχέση $\delta_{min} \leq \frac{|T|}{|P|} \leq \delta_{max}$, μιας και στην περίπτωση που τα δεδομένα αποκρύπτονται πλήρως σε μια γενίκευση του *T*, T^*, η πιθανότητα θα είναι ίση με:

$$P(t \in T|T^*) = \left|\frac{T}{P}\right|, \forall t \in P.$$

Όπως παρατηρείται [NAC07], η *k*-ανωνυμία δεν επιλύει το πρόβλημα της επιβεβαίωσης ή μη της παρουσίας μιας εγγραφής λόγω του ότι δεν λαμβάνει υπόψη τον ήδη δημοσιευμένο πίνακα δεδομένων *P*. Είναι δυνατή η ικανοποίηση της *δ*-παρουσίας μέσω *k*-ανωνυμοποήσεων όμως δεν υπάρχει κάποια συσχέτιση μεταξύ των παραμέτρων *k* και *δ*.

3

Ορισμός προβλήματος

Πολλοί πλέον δημόσιοι οργανισμοί και επιχειρήσεις συλλέγουν προσωπικά δεδομένα με σκοπό την αξιοποίηση της πληροφορίας που αυτά περικλείουν, είτε για την βέλτιστη εξυπηρέτηση των ατόμων που απευθύνονται σε αυτούς, για ερευνητικούς σκοπούς ή με στόχο το κέρδος. Λόγω της ταχύτητας της επικοινωνίας και της ευκολίας της μεταφοράς, τροποποίησης ή δημοσίευσης μεγάλων συνόλων δεδομένων είναι σχεδόν αδύνατο να ελεγχθεί η νομιμότητα της κατοχής και πόσο μάλλον της διαχείρισης που μπορεί να έχει κάθε κάτοχος τέτοιων βάσεων δεδομένων, συνεπώς η ιδιωτικότητα των ατόμων που συμμετέχουν σε συλλογές προσωπικών δεδομένων σε κάθε περίπτωση δύναται να παραβιαστεί από τον οποιονδήποτε αποκτήσει πρόσβαση σε αυτά, νόμιμα ή μη. Μιας και πλέον σχεδόν όλοι βρισκόμαστε σε μία ή περισσότερες τέτοιες βάσεις δεδομένων, όπως για παράδειγμα το σύνολο των φορολογικών δεδομένων της Ελλάδας ή το πληροφοριακό πλέον σύστημα των ιατρικών δεδομένων, τα προσωπικά μας στοιχεία βρίσκονται στην διάθεση πολλών ατόμων εν αγνοία μας και πολλές ιδιωτικές πληροφορίες μπορούν να αποκαλυφθούν από τη συσχέτιση ήδη δημοσιευμένων δεδομένων με αυτά τα στοιχεία. Στα δεδομένα αυτά κατά τη δημοσίευσή τους, ακόμα και μετά την αφαίρεση των μοναδικών αναγνωριστικών γνωρισμάτων ενός ατόμου, όπως ο αριθμός «Αστυνομικής Ταυτότητας» ή ακόμα και ο «Αριθμός Μητρώου Κοινωνικής Ασφάλισης», υπάρχει η πιθανότητα να περιέχονται άλλα γνωρίσματα που σε συνδυασμό μεταξύ τους και με εξωτερική γνώση να μπορούν να προσδιορίσουν μοναδικά κάποιο φυσικό πρόσωπο.

Για την διασφάλιση του ιδιωτικού χαρακτήρα των δημοσιευμένων δεδομένων ο τομέας της προστασίας της ιδιωτικότητας αναπτύσσει συνεχώς αρχές και αλγορίθμους ανωνυμοποίησης, όπως αναλύεται στο Κεφάλαιο 2.

Λόγω των διαφορετικών μοντέλων δεδομένων αλλά και της πιθανής γνώσης που κατέχει ο επιτιθέμενος είναι δυνατή η διεξαγωγή διαφορετικής πληροφορίας και συνεπώς κάθε περίπτωση δεδομένων προς δημοσίευση απαιτεί διαφορετική επεξεργασία για την διατήρηση της ανωνυμίας των ατόμων που εμπεριέχουν. Στην παρούσα εργασία, εξετάζεται και επιχειρείται να επιλυθεί ένα πρόβλημα που δεν έχει ακόμα διερευνηθεί από τον τομέα της προστασίας της ιδιωτικότητας. Το ενδιαφέρον εστιάζεται σε βάσεις δεδομένων με γνωρίσματα από ένα κοινό πεδίο τιμών που αναπαριστούν το ίδιο είδος πληροφορίας και μπορούν να αποδώσουν έτσι συναθροιστική πληροφορία. Δεν συμμετέχει κάποιο ευαίσθητο γνώρισμα και επιδιώκεται η δημοσίευσή τους χωρίς την δυνατότητα ταυτοποίησης κάποιας εγγραφής με ένα φυσικό πρόσωπο μέσω της συναθροιστικής πληροφορίας.

Το μοντέλο δεδομένων του προβλήματος αντιπροσωπεύεται ήδη στην καθημερινή ζωή, για παράδειγμα στις βάσεις δεδομένων των τραπεζών όπου καταχωρούνται οι προσωπικές συναλλαγές κάθε καταθέτη. Εκεί, όλα τα γνωρίσματα αφορούν χρήματα και μπορούν να παρέχουν συναθροιστική πληροφορία όπως το ισοζύγιο των κινήσεων του ατόμου, την μέγιστη ή ελάχιστη κατάθεση και άλλα αντίστοιχα στατιστικά ποσά. Άλλο ένα χαρακτηριστικό παράδειγμα, πραγματικά δεδομένα της μορφής του οποίου χρησιμοποιήθηκαν για τη διεξαγωγή των πειραμάτων της εργασίας, αφορά την απογραφή των ετήσιων εισοδημάτων κάθε ατόμου. Τέτοιες βάσεις δεδομένων μπορούν να ανεβρεθούν από όποιον εργάζεται στο φορολογικό σύστημα της χώρας αλλά και δημοσίως μέσω της ανάρτησης των δεδομένων απογραφής προηγούμενων ετών για ερευνητικούς σκοπούς. Αν και η συλλογή των δεδομένων αυτών δεν είναι μεμπτή, η δημοσίευση, αναπαραγωγή ή τροποποίηση αυτών χωρίς την άδεια του προσώπου που αφορούν πρέπει να ελέγχεται έτσι ώστε προσωπικά στοιχεία να μη φτάνουν στην κατοχή ατόμων που δεν έχουν την απαραίτητη δικαιοδοσία.

Στις περιπτώσεις που αναφέρθηκαν, οποιοσδήποτε έχει πρόσβαση στο σύνολο των δεδομένων γνωρίζοντας μια συναθροιστική πληροφορία, όπως το συνολικό ετήσιο εισόδημα κάποιου ατόμου και χωρίς να γνωρίζει την πραγματική τιμή κάποιου γνωρίσματος, μπορεί υπολογίζοντας το συνολικό άθροισμα για κάθε εγγραφή στο σύνολο των δεδομένων των εισοδημάτων, να ταυτοποιήσει μοναδικά κάποια από αυτές με το φυσικό πρόσωπο που αναζητά και στη συνέχεια να ανακαλύψει τις επιμέρους τιμές των εισοδημάτων του.

Η τροποποίηση των δεδομένων έτσι ώστε να ικανοποιούν την *k*-ανωνυμία, όπως έχει οριστεί από [Swe02], μέσω κάποιου αλγορίθμου όπως ο Incognito και ο Mondrian, μπορεί αποκρύπτοντας μέρος της πληροφορίας που περιέχουν τα αρχικά δεδομένα μέσω της τεχνικής

της γενίκευσης να προστατέψει την ιδιωτικότητα των εγγραφών, καθώς ο επιτιθέμενος δε θα μπορεί να υπολογίσει την συναθροιστική τιμή με ακρίβεια για καμία εγγραφή. Θα μπορεί μόνο να υπολογίσει ένα εύρος τιμών μέσα στο οποίο θα βρίσκεται η συναθροιστική τιμή, κοινό για κάθε εγγραφή από την ίδια κλάση ισοδυναμίας. Εφόσον όμως ο επιτιθέμενος θεωρείται πως έχει μόνο συναθροιστική γνώση και δεν γνωρίζει την τιμή κάποιου επιμέρους γνωρίσματος για το άτομο που αναζητά, μερικές τιμές σε κάθε κλάση ισοδυναμίας για κάποια γνωρίσματα θα μπορούσαν να δημοσιευθούν με την αρχική τους μορφή, όσο διατηρείται κοινή η υπολογιζόμενη τιμή της συναθροιστικής συνάρτησης για τις εγγραφές της κλάσης.

Με τον τρόπο αυτό αποτρέπεται η παραβίαση της ιδιωτικότητας των εγγραφών από επιθέσεις αυτής της μορφής μιας και ο επιτιθέμενος δεν μπορεί να προσδιορίσει μοναδικά κάποια εγγραφή και κατά συνέπεια κάποια τιμή που αυτή λαμβάνει. Παράλληλα διατηρείται σημαντικά μεγαλύτερο ποσοστό χρήσιμης πληροφορίας στα δημοσιευμένα δεδομένα. Στη λογική αυτή βασίζεται ο αλγόριθμος που παρουσιάζεται με στόχο την επίλυση του προβλήματος που ορίζεται επίσημα στη συνέχεια.

3.1 Μοντέλο δεδομένων

Το προς συζήτηση πρόβλημα εστιάζει σε βάσεις δεδομένων με αριθμητικά γνωρίσματα, με το ενδιαφέρον στραμμένο στην περίπτωση της εφαρμογής μιας συναθροιστικής συνάρτησης πάνω στις τιμές των γνωρισμάτων κάθε εγγραφής, η οποία μπορεί να αναπαριστά κάποιο στατιστικό μέγεθος των τιμών αυτών. Το αντικείμενο της εργασίας μπορεί να επεκταθεί και σε βάσεις δεδομένων με κατηγορικά γνωρίσματα, όπου η συναθροιστική συνάρτηση θα ορίζεται με κατάλληλο τρόπο ώστε να αναπαριστά σημασιολογικές συσχετίσεις μεταξύ των γνωρισμάτων. Το εξεταζόμενο μοντέλο δεδομένων περιγράφεται από έναν πίνακα όπου κάθε γραμμή αντιστοιχεί σε μία εγγραφή και οι στήλες του στις τιμές που παίρνει η εγγραφή στα αντίστοιχα γνωρίσματα.

Ορίζεται ο πίνακας που περιέχει τα δεδομένα στην αρχική τους μορφή $RT(A_1, A_2, .., A_n)$ με ψευδο-αναγνωριστικό το σύνολο των γνωρισμάτων $QI_{RT} = (A_1, A_2, .., A_n)$ με κάθε γνώρισμα από το κοινό πεδίο τιμών, I υποσύνολο του συνόλου των πραγματικών αριθμών, $\mathbb{R}$. Με χρήση αυτών ορίζεται η συναθροιστική συνάρτηση με πεδίο ορισμού το καρτεσιανό γινόμενο των πεδίων τιμών των γνωρισμάτων του ψευδο-αναγνωριστικού, και πεδίο τιμών το $\mathbb{R}$, $f: I \times I \times ... \times I \rightarrow \mathbb{R}$. Η συνάρτηση αυτή μπορεί να αναπαριστά το συνολικό άθροισμα, τον μέσο όρο των τιμών κάθε εγγραφής ή οποιαδήποτε άλλη στατιστική συνάρτηση μπορεί να εφαρμοστεί αθροιστικά πάνω στις τιμές των γνωρισμάτων της κάθε εγγραφής και προσφέρει κάποια σημασιολογική χρησιμότητα.

Η χρήση βάσεων δεδομένων που αντιπροσωπεύουν το παραπάνω μοντέλο είναι πολύ συχνή στην καθημερινότητα. Εμφανίζεται για να περιγράψει τα ποσά από τις οικονομικές συναλλαγές των ατόμων που συμμετέχουν στα δεδομένα, τα φορολογικά συνολικά στοιχεία τους ή ακόμα και βάσεις δεδομένων με λεπτομερή στατιστικά στοιχεία από τις ιστοσελίδες τους Διαδικτύου, όπως για παράδειγμα την επισκεψιμότητά τους από κάθε χρήστη. Το παράδειγμα του Πίνακα 3.1 είναι μία περίπτωση αυτού του μοντέλου δεδομένων. Παρουσιάζει τα επιμέρους εισοδήματα, από διαφορετικές πηγές, ως τα γνωρίσματα της σχέσης. Χρησιμοποιείται το άθροισμα των γνωρισμάτων ως η προς εξέταση συναθροιστική συνάρτηση, που ισοδυναμεί με το συνολικό εισόδημα κάθε εγγραφής. Η τιμή αυτή στο παρόν πρόβλημα, δεν εμφανίζεται κατά την δημοσίευση και χρησιμοποιείται μόνο για υπολογιστικούς σκοπούς.

Α/Α	Μερικό Εισόδημα 1	Μερικό Εισόδημα 2	Μερικό Εισόδημα 3	Συνολικό Εισόδημα
1	100	190	210	500
2	190	150	250	590
3	210	400	310	920
4	210	490	200	900
5	250	280	410	940
6	280	250	450	980

Πίνακας 3.1

3.2 Απειλές κατά της ιδιωτικότητας

Στα σύνολα δεδομένων που αντιπροσωπεύουν το προαναφερθέν μοντέλο δεδομένων, μπορεί να επιχειρηθεί η παραβίαση της ιδιωτικότητας των ατόμων που εμφανίζονται μέσα σε αυτά, με στόχο την αναγνώριση της ταυτότητάς τους ή κάποιας από τις τιμές τους. Στο δεδομένο πρόβλημα, ο επιτιθέμενος δρα έχοντας πληροφορία για την τιμή της συνάρτησης που προκύπτει από τις τιμές των γνωρισμάτων που εμφανίζονται για κάποιο άτομο, είτε την τιμή δηλαδή της συναθροιστικής συνάρτησης ή κάποιο διάστημα μέσα στο οποίο αυτή μπορεί να βρίσκεται. Συνδυάζοντας τα στοιχεία που έχει και εκείνα που του παρέχονται από τη δημοσίευση του πίνακα προσπαθεί να ταυτοποιήσει κάποια συγκεκριμένη εγγραφή με το άτομο αυτό. Στο παράδειγμα του Πίνακα 3.1 ο επιτιθέμενος μπορεί να γνωρίζει ότι το συνολικό εισόδημα κάποιου συγκεκριμένου ατόμου είναι ακριβώς 590, και συνεπώς να το

αναγνωρίσει ως την εγγραφή 2, ή αν έχει ασαφή γνώση, μπορεί να γνωρίζει πως το συνολικό εισόδημα του ατόμου που αναζητά βρίσκεται μέσα στο διάστημα [900,1000), κάτι που τον οδηγεί στις εγγραφές 3,4,5,6.

Η βέβαιη ταυτοποίηση του ατόμου με κάποια εγγραφή είναι δυνατή όταν δεν υπάρχουν άλλες εγγραφές με την ίδια τιμή της δεδομένης συνάρτησης ή στην περίπτωση που ο επιτιθέμενος γνωρίζει με βεβαιότητα την τιμή κάποιου γνωρίσματος παράλληλα με μερική γνώση για την τιμή της συνάρτησης της εγγραφής. Αν για παράδειγμα ο επιτιθέμενος γνώριζε το προαναφερθέν διάστημα του συνολικού εισοδήματος του ατόμου [900,1000) και επιπλέον πως η τιμή του στο γνώρισμα "Μερικό Εισόδημα 1" είναι 250, θα μπορούσε με βεβαιότητα να αναγνωρίσει την εγγραφή 5 ως το πρόσωπο που αναζητά. Στις θεωρούμενες επιθέσεις ο επιτιθέμενος έχει γνώση μόνο για την τιμή της συναθροιστικής συνάρτησης του ατόμου που θέλει να ταυτοποιήσει. Γνωρίζει δηλαδή είτε την ακριβή τιμή της συναθροιστικής συνάρτησης της εγγραφής, για ευκολία θεωρούμε το άθροισμα όπως στο παράδειγμα, είτε ένα διάστημα τιμών μέσα στο οποίο αυτή εμφανίζεται. Έτσι ανατρέχοντας στη δημοσίευση του πίνακα με την αρχική της μορφή και υπολογίζοντας τα παρεχόμενα αθροίσματα μπορεί να ταυτοποιήσει κάποιο άτομο μέσα στον πίνακα και να ανακαλύψει όποιες από τις επιμέρους τιμές των γνωρισμάτων του, του δίνεται η δυνατότητα.

3.3 Μετρική κόστους απώλειας πληροφορίας

Σε κάθε περίπτωση ανωνυμοποίησης δεδομένων βασικό ρόλο κατά την εκτίμηση της αποδοτικότητας της μεθόδου που χρησιμοποιήθηκε έχει η απώλεια πληροφορίας που παρατηρείται στα δημοσιευμένα δεδομένα. Στην παρούσα εργασία ως εργαλείο σύγκρισης των αποτελεσμάτων των αναφερόμενων μεθόδων χρησιμοποιείται η *Κανονικοποιημένη Ποινή Βεβαιότητας (Normalized Certainty Penalty)* όπως ορίζεται από [XWP+06].

Συγκεκριμένα, σε έναν πίνακα με ψευδο-αναγνωριστικό το σύνολο $QI_{RT} = (A_1, A_2, \dots, A_n)$, όπου κάθε πλειάδα της μορφής $t = (x_1, x_2, \dots, x_n)$ γενικεύεται σε πλειάδα της μορφής $t = ([y_1, z_1], [y_2, z_2], \dots, [y_n, z_n])$ με $y_i \leq x_i \leq z_i$ για κάθε $1 \leq i \leq n$, για κάθε αριθμητικό γνώρισμα A_i ορίζεται η Κανονικοποιημένη Ποινή Βεβαιότητας ως:

$$NCP_{A_i}(t) = \frac{(z_i - y_i)}{|A_i|}$$

με $|A_i| = max_{t \in RT}\{t.A_i\} - min_{t \in RT}\{t.A_i\}$, δηλαδή το μέγιστο διάστημα των τιμών του γνωρίσματος A_i.

Η Κανονικοποιημένη Ποινή Βεβαιότητας υπολογίζεται κατά αντιστοιχία και για κατηγορικά γνωρίσματα. Έστω η πλειάδα t έχει την τιμή v στο κατηγορικό γνώρισμα A_i.

Κατά την ανωνυμοποίηση αυτή η τιμή θα αντικατασταθεί από ένα σύνολο τιμών $\{v_1, \dots, v_l\}$ τα οποία εμφανίζονται στην ίδια ομάδα γενίκευσης με την αρχική τιμή. Ως *u* συμβολίζεται ο *πλησιέστερος κοινός πρόγονος (closest common ancestor)* των τιμών του συνόλου, ο κόμβος της ιεραρχίας γενίκευσης του γνωρίσματος που είναι πρόγονος των τιμών-φύλλων στην ιεραρχία γενίκευσης και δεν υπάρχει άλλος κόμβος απόγονός του που είναι παράλληλα πρόγονος των τιμών αυτών. Τότε η Κανονικοποιημένη Ποινή Βεβαιότητας του γνωρίσματος ορίζεται :

$$NCP_{A_i}(t) = \frac{size(u)}{|A_i|},$$

όπου με $size(u)$ συμβολίζεται ο αριθμός των φύλλων-παιδιών του πλησιέστερου κοινού προγόνου των τιμών του συνόλου της ομάδας γενίκευσης της τιμής *v* και με $|A_i|$ συμβολίζεται ο αριθμός των διακριτών τιμών που λαμβάνει το γνώρισμα A_i .

Ο παραπάνω ορισμός δίνει την τιμή της μετρικής του A_i γνωρίσματος για την πλειάδα *t*, αν το γνώρισμα είναι αριθμητικό ή κατηγορικό. Η συνολική απώλεια πληροφορίας της πλειάδας *t*, ορίζεται από την *Σταθμισμένη Ποινή Βεβαιότητας (weighted certainty penalty),* αν κάθε γνώρισμα αντιστοιχίζεται με έναν αριθμό-βάρος w_i, που αντιπροσωπεύει την χρησιμότητά του στα ανωνυμοποιημένα δεδομένα και υπολογίζεται από:

$$NCP(t) = \sum_{i=1}^{n} w_i \cdot NCP_{A_i}(t) .$$

Βάσει αυτού, προκύπτει η Κανονικοποιημένη Ποινή Βεβαιότητας όλου του πίνακα από:

$$NCP(RT) = \sum_{t \in RT} NCP(t) .$$

Κατά την δημοσίευση των ανωνυμοποιημένων δεδομένων η διεξαγωγή της χρήσιμης πληροφορίας που σαν στόχο είχε η δημοσίευση έχει μειωθεί λόγω των γενικεύσεων που εκτελούνται ώστε να ικανοποιείται η επιλεγμένη εγγύηση ιδιωτικότητας. Η Κανονικοποιημένη Ποινή Βεβαιότητας αποτελεί ένα επαρκές εργαλείο σύγκρισης για την εκτίμηση της χρήσιμης πληροφορίας που χάνεται από τα αποτελέσματα των αλγορίθμων που χρησιμοποιούν την τεχνική της γενίκευσης με σκοπό την ανωνυμοποίηση των δεδομένων. Με χρήση αυτής μπορεί να επιλεγεί ο αλγόριθμος που βέλτιστα ικανοποιεί την ανωνυμία ενώ παράλληλα διατηρηθεί το μεγαλύτερο δυνατό ποσοστό χρήσιμης πληροφορίας από τα αρχικά δεδομένα.

3.4 Πιθανές λύσεις

Από το παραπάνω μοντέλο δεδομένων επιχειρείται η δημοσίευση του πίνακα με τις τιμές των εγγραφών στα συγκεκριμένα γνωρίσματα, όσο το δυνατόν πλησιέστερα στην αρχική τους

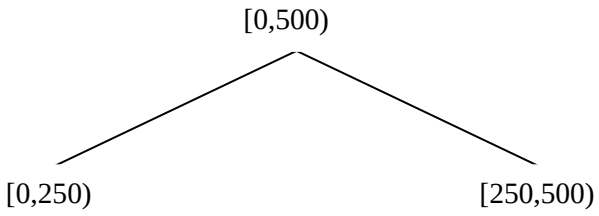
μορφή προϋποθέτοντας όμως την αδυναμία ταυτοποίησης μιας εγγραφής βάσει της τιμής της συναθροιστικής συνάρτησης.

Το ζητούμενο σύνολο όλων των πιθανών λύσεων του προβλήματος είναι το σύνολο όλων των δυνατών *k*-ανωνυμοποιήσεων του αρχικού πίνακα δεδομένων, όπου κάθε εγγραφή δε θα διακρίνεται μεταξύ τουλάχιστον *k* εγγραφών του πίνακα, ενώ όπως επισημαίνεται από την βιβλιογραφία, η εύρεση της βέλτιστης *k*-ανωνυμοποίησης υπό διαφορετικά μοντέλα, αποτελεί ένα NP-δύσκολο πρόβλημα. Σε κάθε μία από αυτές, το ενδιαφέρον εστιάζεται στο κατά πόσο εξασφαλίζει την ιδιωτικότητα των παρουσιαζόμενων προσωπικών δεδομένων αλλά και κατά πόσο διατηρεί την απαραίτητη πληροφορία που αυτά περιέχουν έπειτα από την εφαρμογή των αρχών γενίκευσης.

Όπως γίνεται αντιληπτό το πιο αποδοτικό μοντέλο ανακωδικοποίησης που μπορεί να χρησιμοποιηθεί για την επίλυση του προβλήματος αποφεύγει την απόκρυψη εγγραφών και εφαρμόζει τοπική γενίκευση.

3.4.1 Χρήση Incognito

Πράγματι, για το παράδειγμα του Πίνακα 3.1, για την περίπτωση της εφαρμογής καθολικής ανακωδικοποίησης, ορίζοντας μια υποθετική διαμέριση μη επικαλυπτόμενων διαστημάτων, κοινή για όλα τα γνωρίσματα, πάνω στο κοινό πεδίο τιμών [0, 500) της Εικόνας 3.1, με χρήση του αντιπροσωπευτικού αλγορίθμου μονοδιάστατης ανακωδικοποίησης Incognito, για $k = 2$, επιστρέφεται μια 2-ανωνυμοποίηση της μορφής του Πίνακα 3.2. Εκεί οι πλειάδες εμφανίζονται με την σειρά που επεξεργάστηκαν από τον αλγόριθμο, ενώ οι κλάσεις ισοδυναμίας που σχηματίζονται είναι τα σύνολα των εγγραφών {1,4}, {2,3}, {5,6}. Από τις γενικεύσεις αυτές προκύπτει το μέγιστο διάστημα μέσα στο οποίο θα ανήκει και το συνολικό άθροισμα κάθε εγγραφής, έτσι όπως μπορεί να το υπολογίσει ο επιτιθέμενος από την δημοσίευση των ανωνυμοποιημένων τιμών των γνωρισμάτων και παρουσιάζεται στην στήλη «Συνολικό Εισόδημα» του ίδιου πίνακα.



Εικόνα 3.1: Ιεραρχία γενίκευσης πεδίου τιμών των δεδομένων του Πίνακα 3.1

Στην περίπτωση αυτή ο επιτιθέμενος δε μπορεί να αναγνωρίσει με βεβαιότητα καμία εγγραφή βάσει της συναθροιστικής του γνώσης, ακόμα και αν γνωρίζει κάποια επιμέρους τιμή κάποιου γνωρίσματος της εγγραφής εκτός του αθροίσματος, αφού όλες οι εγγραφές στην ίδια κλάση ισοδυναμίας εμφανίζουν ακριβώς τις ίδιες γενικευμένες τιμές. Ο αλγόριθμος Incognito εξαιτίας της γενίκευσης πλήρους πεδίου που εφαρμόζει ικανοποιεί την *k*-ανωνυμία στο σύνολο των δεδομένων, όμως υπεργενικεύει τα δεδομένα. Αυτό έχει ως αποτέλεσμα την σημαντική απώλεια χρήσιμης πληροφορίας κατά την δημοσίευσή τους, χωρίς αυτό να είναι απαραίτητο, κυρίως στην περίπτωση που ο επιτιθέμενος έχει γνώση μόνο για την τιμή της συναθροιστικής συνάρτησης.

Α/Α	Μερικό Εισόδημα 1	Μερικό Εισόδημα 2	Μερικό Εισόδημα 3	Συνολικό Εισόδημα
1	[0,250)	[0,500)	[0,250)	[0,1000)
2	[0,250)	[0,500)	[250,500)	[250,1250)
3	[0,250)	[0,500)	[250,500)	[250,1250)
4	[0,250)	[0,500)	[0,250)	[0,1000)
5	[250,500)	[0,500)	[250,500)	[500,1500)
6	[250,500)	[0,500)	[250,500)	[500,1500)

Πίνακας 3.2: 2-ανωνυμοποίηση του Πίνακα 3.1 με χρήση Incognito

3.4.2 Χρήση Mondrian

Μια καλύτερη αντιμετώπιση για το συγκεκριμένο μοντέλο δεδομένων, όπως προκύπτει έπειτα από δοκιμές, παρουσιάζεται με την εφαρμογή μοντέλων πολυδιάστατης ανακωδικοποίησης, όπου οι τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού

αντιμετωπίζονται ως διάνυσμα αντί να γενικεύονται καθεμία χωριστά. Βέλτιστη απόδοση επίσης προκύπτει κατά την εφαρμογή της τοπικής γενίκευσης, αφού έτσι παρέχεται μεγαλύτερη ελαστικότητα. Με χρήση τοπικής γενίκευσης δεν επιβάλλεται κάθε εμφάνιση μιας αρχικής τιμής των δεδομένων να αντιστοιχίζεται στην ίδια πάντα γενικευμένη τιμή. Ο πιο αποδοτικός μέχρι στιγμής αλγόριθμος που αντιπροσωπεύει αυτό το μοντέλο ανακωδικοποίησης είναι ο αλγόριθμος Mondrian, με χρήση επικαλυπτόμενων διαστημάτων κατά την διαμέριση. Για το παράδειγμα του Πίνακα 3.1, ο αντίστοιχος 2-ανωνυμοποιημένος πίνακας των δεδομένων που προκύπτει από την εφαρμογή τοπικής γενίκευσης, με χρήση του Mondrian, εμφανίζεται στον Πίνακα 3.3. Στην περίπτωση αυτή οι εγγραφές διαχωρίζονται σε διαφορετικές κλάσεις ισοδυναμίας από πριν, ενώ παρουσιάζονται επικαλυπτόμενα διαστήματα πάνω στο κοινό πεδίο τιμών των γνωρισμάτων. Όπως παρατηρείται και στην περίπτωση του αλγορίθμου Mondrian τα δεδομένα υπεργενικεύονται καθώς δεν λαμβάνεται υπόψιν η συναθροιστική γνώση βάσει της οποίας μπορεί να γίνει η επίθεση. Εδώ, όλες οι τιμές αντικαθιστώνται με τις κατάλληλα γενικευμένες τιμές σύμφωνα με την βέλτιστη πολυδιάστατη διαμέριση που επέλεξε ο αλγόριθμος. Κάτι τέτοιο είναι σαφώς λιγότερο αποδοτικό μιας και από το αρχικό σύνολο δεδομένων φαίνεται πως υπάρχει η δυνατότητα δημοσίευσης κάποιων τιμών στην αρχική τους μορφή χωρίς να παραβιάζεται η ιδιωτικότητα των εγγραφών.

A/A	Μερικό Εισόδημα 1	Μερικό Εισόδημα 2	Μερικό Εισόδημα 3	Συνολικό Εισόδημα
1	[0,210]	[0,190]	[0,500]	[0,900]
2	[0,210]	[0,190]	[0,500]	[0,900]
3	[0,210]	[190,500]	[0,500]	[190,1210]
4	[0,210]	[190,500]	[0,500]	[190,1210]
5	[210,500]	[0,500]	[0,500]	[210,1500]
6	[210,500]	[0,500]	[0,500]	[210,1500]

Πίνακας 3.3: 2-ανωνυμοποίηση του Πίνακα 3.1 με χρήση Mondrian

3.4.3 Χρήση αλγορίθμου με συναθροιστική συνάρτηση

Στην περίπτωση της τοπικής γενίκευσης είναι δυνατή σημαντικά καλύτερη απόδοση με τον αλγόριθμο που εξετάζεται στην παρούσα εργασία. Η πιθανή προτεινόμενη λύση εφαρμόζει τοπική γενίκευση μέσα σε κάθε κλάση ισοδυναμίας με σκοπό την *k*-ανωνυμοποίηση των

δεδομένων. Η διαφορά με την προηγούμενη λύση, και κάθε άλλη υπάρχουσα μέχρι στιγμής, βρίσκεται στο ότι η κατάλληλη γενίκευση για κάθε γνώρισμα υπολογίζεται σε κάθε κλάση λαμβάνοντας υπόψη την τιμή της συναθροιστικής συνάρτησης των εγγραφών της κλάσης ισοδυναμίας. Η εφαρμογή του αλγορίθμου για το ίδιο σύνολο δεδομένων παρουσιάζει στον Πίνακα 3.4 το αντίστοιχο 2-ανώνυμο σύνολο, όπου οι εγγραφές εμφανίζονται διαχωρισμένες στις επιλεγμένες κλάσεις ισοδυναμίας. Ο αλγόριθμος είναι σαφώς πιο αποδοτικός αφού διαχωρίζει το σύνολο των δεδομένων σε κλάσεις ισοδυναμίας μία φορά και έπειτα επεξεργάζεται κάθε κλάση ισοδυναμίας χωριστά. Σε αντίθεση με αυτόν, οι δύο προηγούμενοι αλγόριθμοι εκτελούν πολλαπλές επαναλήψεις για την εύρεση της βέλτιστης διαμέρισης, διατρέχοντας τις αρχικές τιμές των γνωρισμάτων των εγγραφών του συνόλου. Όπως παρατηρείται από τον Πίνακα 3.4, εφόσον θεωρείται πως η γνώση του επιτιθέμενου βασίζεται μόνο στην συναθροιστική συνάρτηση πάνω στις τιμές της εγγραφής που αναζητά, είναι δυνατός ο διαχωρισμός των εγγραφών σε κλάσεις ισοδυναμίας με στόχο σε κάθε κλάση να εμφανίζονται οι κατάλληλα επιλεγμένες τιμές στα γνωρίσματα, γενικευμένες ή μη, έτσι ώστε ο επιτιθέμενος να υπολογίζει την ίδια τιμή της συναθροιστικής συνάρτησης για τις εγγραφές της κλάσης ισοδυναμίας. Με τον τρόπο αυτό, η απώλεια της πληροφορίας είναι μικρότερη μιας και δίνεται η δυνατότητα δημοσίευσης των αρχικών τιμών που δεν παραβιάζουν την *k*-ανωνυμία ως προς την συναθροιστική συνάρτηση.

Α/Α	Μερικό Εισόδημα 1	Μερικό Εισόδημα 2	Μερικό Εισόδημα 3	Συνολικό Εισόδημα
1	[100,190]	190	210	[500,590]
2	[100,190]	150	250	[500,590]
4	210	[400,490]	[200,310]	[810,1010]
3	210	[400,490]	[200,310]	[810,1010]
5	250	280	[410,450]	[940,980]
6	280	250	[410,450]	[940,980]

Πίνακας 3.4: 2-ανωνυμοποίηση του Πίνακα 3.1 με χρήση αλγόριθμου συναθροιστικής συνάρτησης

3.4.3.1 Χαλάρωση της εγγύησης της ανωνυμίας

Από τον αλγόριθμο της εργασίας δίνεται ακόμα η επιλογή της χαλάρωσης της εγγύησης της ανωνυμίας που προσφέρει κάτι που δεν εμφανίζεται στους υπόλοιπους αλγορίθμους. Η επιλογή αυτή γίνεται μέσω μίας παραμέτρου η οποία αναπαριστά την ασαφή γνώση που

μπορεί να έχει ο επιτιθέμενος πάνω στην τιμή της συναθροιστικής συνάρτησης της ζητούμενης από αυτόν εγγραφής. Θεωρούμε πως μπορεί να μην γνωρίζει την ακριβή τιμή, αλλά ένα διάστημα μέσα στο οποίο μπορεί αυτή να εμφανίζεται. Ο αλγόριθμος τότε μπορεί να τροποποιήσει τα δεδομένα μόνο τόσο όσο χρειάζεται ώστε να διασφαλίζει από απειλές με αυτή την γνώση. Ένα παράδειγμα αυτής της περίπτωσης αναφέρθηκε για τα στοιχεία του Πίνακα 3.1, όπου ο επιτιθέμενος μπορεί να γνωρίζει μόνο ότι το άθροισμα των εισοδημάτων της εγγραφής βρίσκεται μέσα στο διάστημα [900,1000) και έτσι να αμφιβάλλει μεταξύ των εγγραφών 3,4,5,6. Επειδή η περίπτωση ο επιτιθέμενος να μην γνωρίζει την ακριβή τιμή της συναθροιστικής συνάρτησης μπορεί να παρουσιαστεί σε επιθέσεις πάνω σε αυτό το μοντέλο δεδομένων, ο αλγόριθμος δίνει την δυνατότητα της χαλάρωσης της εγγύησης της ανωνυμίας, έτσι ώστε να μην τροποποιούνται τα αρχικά δεδομένα περισσότερο από όσο κρίνεται απαραίτητο από τον κάτοχο του συνόλου των δεδομένων. Σκοπός αυτής της επιλογής είναι η ελαχιστοποίηση της απώλειας της χρήσιμης πληροφορίας των πρωτότυπων δεδομένων σε περιπτώσεις δημοσίευσης προσωπικών δεδομένων που η ιδιωτικότητά τους δεν απειλείται στο μέγιστο βαθμό.

Οι τρεις προαναφερθείσες περιπτώσεις παρουσιάζουν διαφορές σχετικά με την διεξαγόμενη πληροφορία και συνεπώς την χρηστικότητα που διατηρούν στα δημοσιευμένα δεδομένα, οι οποίες περιγράφονται ικανοποιητικά από την Κανονικοποιημένη Ποινή Βεβαιότητας, βάσει της οποίας συγκρίνουμε την απώλεια πληροφορίας.

Η διαφορά μεταξύ του αλγορίθμου που λαμβάνει υπόψη την τιμή της συναθροιστικής συνάρτησης, εμφανίζεται στην απώλεια πληροφορίας στα δημοσιευμένα δεδομένα, αφού πολλές από τις εμφανιζόμενες τιμές του πίνακα παραμένουν στην αρχική τους μορφή. Η *k*-ανωνυμία ικανοποιείται κατά τις επιθέσεις όπου είναι γνωστή στον επιτιθέμενο η τιμή της συναθροιστικής συνάρτησης και εφόσον θεωρείται πως δεν έχει γνώση καμίας τιμής κάποιου επιμέρους γνωρίσματος, διασφαλίζεται η ιδιωτικότητα των ατόμων που εμφανίζονται στο σύνολο των δεδομένων.

4

Περιγραφή αλγορίθμου

4.1 Θεωρητικό Υπόβαθρο

Ο αλγόριθμος επιχειρεί την ικανοποίηση της *k*-ανωνυμίας ως προς την συναθροιστική συνάρτηση από τα προς δημοσίευση δεδομένα. Εφαρμόζει τοπική γενίκευση στις τιμές των γνωρισμάτων των εγγραφών, υπολογίζοντας την ζητούμενη διαμέριση κάθε γνωρίσματος σε κάθε κλάση ξεχωριστά.

Η *k-ανωνυμία,* όπως έχει προαναφερθεί, απαιτεί κάθε εγγραφή που εμφανίζεται στον πίνακα, να μην μπορεί να αναγνωρισθεί ανάμεσα από τουλάχιστον *k* εγγραφές του πίνακα. Αυτό επιτυγχάνεται αν *k* τουλάχιστον εγγραφές από το σύνολο, εμφανίζουν τις ίδιες τιμές ή αντίστοιχα τα ίδια διαστήματα τιμών σε όλα τους τα γνωρίσματα. Ο αλγόριθμος που παρουσιάζεται εξασφαλίζει την *k*-ανωνυμία ως προς την τιμή της συναθροιστικής συνάρτησης. Δηλαδή για κάθε τιμή της συνάρτησης που προκύπτει για κάποια εγγραφή του συνόλου, υπάρχουν τουλάχιστον άλλες *k* εγγραφές που λαμβάνουν την ίδια τιμή, ή το ίδιο διάστημα τιμών στο οποίο ανήκει η τιμή της συναθροιστικής συνάρτησης στα δημοσιευμένα δεδομένα. Εφαρμόζουμε *τοπική γενίκευση,* που όπως υπενθυμίζεται, κάθε εμφάνιση κάποιας τιμής του πεδίου τιμών των γνωρισμάτων αντικαθίσταται με την κατάλληλη για την κλάση γενικευμένη τιμή ενώ δεν αντιστοιχίζονται όλες οι εμφανίσεις της τιμής στον πίνακα στην ίδια γενικευμένη τιμή.

Το σύνολο των εγγραφών χωρίζεται σε κλάσεις ισοδυναμίας, που όπως έχει οριστεί στο Κεφάλαιο 2, είναι οι ομάδες που σχηματίζουν οι εγγραφές που εμφανίζουν τις ίδιες τιμές σε κάθε γνώρισμα και ικανοποιούν την k-ανωνυμία. Στον αλγόριθμο αυτό οι κλάσεις ισοδυναμίας δημιουργούνται με τέτοιο τρόπο ώστε οι εγγραφές μέσα σε μία κλάση ισοδυναμίας να εμφανίζουν την ίδια τιμή, γενικευμένη ή μη, κατά τον υπολογισμό της συναθροιστικής συνάρτησης. Η τοπική γενίκευση που πρέπει να εφαρμοστεί έτσι ώστε οι εγγραφές σε κάθε κλάση ισοδυναμίας να ικανοποιούν τον ορισμό, υπολογίζεται ξεχωριστά για κάθε κλάση.

Η είσοδος του αλγορίθμου είναι το αρχικό σύνολο δεδομένων $RT(A_1, A_2, .., A_n)$, η συναθροιστική συνάρτηση $f: I^n \rightarrow \mathbb{R}$, όπου ως I συμβολίζεται το κοινό πεδίο των γνωρισμάτων του ψευδο-αναγνωριστικού, στο οποίο ανήκουν n γνωρίσματα, η παράμετρος ανωνυμίας, που ισοδυναμεί με τον ελάχιστο αριθμό εγγραφών σε κάθε κλάση ισοδυναμίας, συμβολιζόμενη ως k, και η παράμετρος d, μια μεταβλητή που αντιπροσωπεύει την ασαφή πιθανή γνώση του επιτιθέμενου πάνω στην τιμή της συναθροιστικής συνάρτησης. Η μεταβλητή d χρησιμοποιείται για την χαλάρωση της εγγύησης της ανωνυμίας που προσφέρεται. Συγκεκριμένα η d ορίζεται ως το μέγιστο ποσοστό διαφοροποίησης που μπορεί να εμφανίζεται στις τιμές της συναθροιστικής συνάρτησης των εγγραφών κάθε κλάσης ισοδυναμίας, όπως αυτές υπολογίζονται από τα δημοσιευμένα δεδομένα. Με χρήση αυτής της παραμέτρου κατ'επιλογή ο εκδότης των δεδομένων μπορεί να χαλαρώσει την βασική συνθήκη του αλγορίθμου που απαιτεί οι εγγραφές σε κάθε κλάση να εμφανίζουν ακριβώς τις ίδιες τιμές κατά την εφαρμογή της συναθροιστικής συνάρτησης. Έτσι οι εγγραφές μπορούν να διαφοροποιούνται έως κατά αυτό το ποσοστό μέσα σε κάθε κλάση ισοδυναμίας.

Ο αλγόριθμος υλοποιεί την έννοια του δυναμικού προγραμματισμού καθώς διαιρεί το κύριο πρόβλημα της ανωνυμοποίησης του συνόλου των δεδομένων ως προς την τιμή της συναθροιστικής συνάρτησης, στα υπο-προβλήματα της ανωνυμοποίησης των κλάσεων ισοδυναμίας του αρχικού συνόλου. Το προς δημοσίευση k-ανώνυμο σύνολο των εγγραφών προκύπτει από την ένωση των ανωνυμοποιημένων κλάσεων ισοδυναμίας.

Σε κάθε κλάση ισοδυναμίας βρίσκονται και γενικεύονται τα κατάλληλα γνωρίσματα με προς-τα-πίσω αναδρομή, έτσι ώστε να ικανοποιείται η k-ανωνυμία στην κλάση για τις τιμές της συναθροιστικής συνάρτησης των εγγραφών της. Με τον τρόπο αυτό κάθε κλάση ισοδυναμίας διατηρεί περισσότερα μη γενικευμένα γνωρίσματα με τις αρχικές τιμές των εγγραφών, εκτελώντας παράλληλα μικρότερο αριθμό αναγκαίων πράξεων. Ο αριθμός των επιλεγμένων προς γενίκευση γνωρισμάτων μπορεί να διαφέρει μεταξύ των κλάσεων. Εξαρτάται από την συσχέτιση που εμφανίζουν οι τιμές των γνωρισμάτων των εγγραφών ως προς την συναθροιστική συνάρτηση, καθώς και από την τιμή της παραμέτρου χαλάρωσης d που έχει επιλεγεί, αφού όσο μικρότερο επιτρέπεται να είναι το ποσοστό διαφοροποίησης των τιμών

της συναθροιστικής συνάρτησης των εγγραφών σε κάθε κλάση ισοδυναμίας, τόσο περισσότερα γνωρίσματα είναι πιθανόν να γενικευθούν.

4.2 Υλοποίηση

Ο αλγόριθμος αρχικά υπολογίζει την τιμή της συναθροιστικής συνάρτησης για κάθε εγγραφή από το αντιπροσωπευτικό σύνολο δεδομένων του μοντέλου δεδομένων που παρουσιάστηκε στο Κεφάλαιο 3. Οι εγγραφές ταξινομούνται βάσει της τιμής που λαμβάνουν από την συναθροιστική συνάρτηση.

Πάνω στην διάταξη αυτή γίνεται η διαμέριση των εγγραφών σε κλάσεις ισοδυναμίας, μιας και λόγω της ταξινόμησης, κάθε εγγραφή θα γειτονεύει με εκείνες που έχουν την πλησιέστερη σε αυτήν τιμή που προκύπτει από την συναθροιστική συνάρτηση πάνω στις τιμές των γνωρισμάτων της. Η διαμέριση γίνεται κατά τη διάσχιση της διάταξης όπου οι εγγραφές χωρίζονται σε ομάδες μεγέθους $\geq k$ και $\leq 2k-1$. Η επιλογή του μεγέθους των κλάσεων ισοδυναμίας θεωρητικά είναι αυθαίρετη, όμως εντός του προαναφερθέντος διαστήματος, παρατηρείται ότι είναι προτιμότερο να είναι το μικρότερο δυνατόν ώστε να μην υπεργενικεύονται οι τιμές των εγγραφών.

Ο αλγόριθμος έπειτα, σε κάθε κλάση εκτελεί την ακόλουθη αναδρομική διαδικασία:

Ελέγχεται η τιμή της συνάρτησης πάνω στις τιμές των n γνωρισμάτων κάθε εγγραφής της κλάσης ισοδυναμίας. Αν όλες οι εγγραφές της κλάσης ικανοποιούν την συνθήκη που απαιτεί να εμφανίζουν την τιμή αυτή με ποσοστό διαφοροποίησης μεταξύ τους το πολύ d, τότε η κλάση δεν απαιτεί καμία γενίκευση και επιστρέφεται ως έχει.

Αλλιώς, εξετάζονται όλοι οι δυνατοί συνδυασμοί των $n-1$ γνωρισμάτων ως προς την τιμή της συναθροιστικής συνάρτησης των εγγραφών. Από αυτούς επιλέγεται ο συνδυασμός κατά τον οποίο οι εγγραφές εμφανίζουν όσο το δυνατόν πλησιέστερες τιμές της συναθροιστικής συνάρτησης και γενικεύεται το γνώρισμα που είχε εξαιρεθεί.

Η διαδικασία γενίκευσης που χρησιμοποιεί ο αλγόριθμος είναι η αντικατάσταση των αρχικών τιμών του γνωρίσματος από το εύρος των τιμών που αυτό λαμβάνει μέσα στην δεδομένη κλάση ισοδυναμίας.

Στη συνέχεια ελέγχονται τα εναπομείναντα $n-1$ γνωρίσματα του ψευδο-αναγνωριστικού ως προς τη χρησιμοποιούμενη συνάρτηση. Αν όλες οι εγγραφές έχουν τιμή στη συναθροιστική συνάρτηση πάνω στα $n-1$ γνωρίσματα που ικανοποιεί την βασική συνθήκη, τότε η κλάση ικανοποιεί την k-ανωνυμία ως προς την συνάρτηση και επιστρέφεται το ένα γενικευμένο γνώρισμα και τα υπόλοιπα $n-1$ με τις αρχικές τους τιμές.

Αλλιώς, αναδρομικά μειώνεται ο αριθμός των γνωρισμάτων σε $n-2$ και επαναλαμβάνεται η διαδικασία ελέγχου των συνδυασμών των $n-2$ γνωρισμάτων ως προς τις τιμές της συναθροιστικής συνάρτησης για τις εγγραφές της κλάσης ισοδυναμίας.

Ο αριθμός των συμμετεχόντων στον υπολογισμό της τιμής της συνάρτησης γνωρισμάτων μειώνεται έως ότου ικανοποιηθεί η συνθήκη ή γενικευθούν όλα τα γνωρίσματα της κλάσης ισοδυναμίας.

Ο αλγόριθμος με την μορφή ψευδο-κώδικα παρουσιάζεται παρακάτω:

Είσοδος: $RT(A_1, A_2, \dots, A_n)$ αρχικό σύνολο δεδομένων,

με $QI = (A_1, A_2, \dots, A_n)$ το σύνολο γνωρισμάτων του ψευδο-αναγνωριστικού

$f: I^n \to \mathbb{R}$, συναθροιστική συνάρτηση

k παράμετρος ανωνυμίας

d παράμετρος χαλάρωσης της εγγύησης της ανωνυμίας

Έξοδος: $RT^*(A_1, A_2, \dots, A_n)$ το σύνολο των ανωνυμοποιημένων δεδομένων

Βήματα αλγορίθμου

Για κάθε εγγραφή $t = (x_1, x_2, \dots, x_n) \in RT(A_1, A_2, \dots, A_n)$

Υπολόγισε την τιμή $f(x_1, x_2, \dots, x_n)$ της συναθροιστικής συνάρτησης

Ταξινόμησε τις εγγραφές βάσει της τιμής της συναθροιστικής συνάρτησης αυτής.

Διαχωρισμός ταξινομημένου συνόλου δεδομένων σε κλάσεις ισοδυναμίας μεγέθους $\geq k$ και $\leq 2k-1$.

Για κάθε κλάση ισοδυναμίας:

Αν όλες οι εγγραφές της κλάσης ισοδυναμίας έχουν τιμή συναθροιστικής συνάρτησης $f \in [f - f \cdot d, f + f \cdot d]$:

Πρόσθεσε την κλάση ισοδυναμίας στο σύνολο ανωνυμοποιημένων δεδομένων $RT^*(A_1, A_2, \dots, A_n)$.

Αλλιώς :

Όρισε το σύνολο γνωρισμάτων $QI = \{A_1, A_2, \dots, A_n\}$

$j = n - 1$

Όσο $j > 0$ {

Υπολόγισε όλους τους δυνατούς συνδυασμούς των j γνωρισμάτων από το σύνολο QI.

Για κάθε δυνατό συνδυασμό {

Υπολόγισε τις τιμές της f των εγγραφών της κλάσης ισοδυναμίας

}

Βρες τον συνδυασμό $QI_j \subset QI$ όπου οι εγγραφές εμφανίζουν τις πλησιέστερες τιμές της f μεταξύ τους.

Γενίκευσε το γνώρισμα $A_j = QI \setminus QI_j$, αντικαθιστώντας τις τιμές των εγγραφών με το εύρος του γνωρίσματος $\left[x_{min_{A_j}}, x_{max_{A_j}}\right]$.

$QI = QI \setminus A_j$

$j = j - 1$

Υπολόγισε τις τιμές fτων εγγραφών της κλάσης για το νέο σύνολο γνωρισμάτων QI

Αν όλες οι εγγραφές εμφανίζουν τιμή $f \in [f - f \cdot d, f + f \cdot d]${

Τερμάτισε το βρόχο

}

}

Πρόσθεσε την κλάση ισοδυναμίας στο σύνολο $RT^*(A_1, A_2, \dots, A_n)$

}

■

Η επιλογή της διαδικασίας για τον διαχωρισμό των εγγραφών σε κλάσεις ισοδυναμίας αφήνεται στην κρίση του κατόχου των δεδομένων, ώστε να κατανείμει όπως θεωρεί προτιμότερο τις εναπομένουσες εγγραφές από το σύνολο των δεδομένων κατά την διαίρεση του μεγέθους του με την παράμετρο k. Εδώ επιλέγεται η διαδικασία κατά την οποία, αφαιρούνται διαδοχικά από την ταξινόμηση των εγγραφών ομάδες k εγγραφών, ενώ οι υπολειπόμενες εγγραφές, που είναι λιγότερες από k, προσθέτονται στην τελευταία κλάση ισοδυναμίας.

Η επιλογή του συνδυασμού των γνωρισμάτων που αποδίδει τις πλησιέστερες τιμές της συναθροιστικής συνάρτησης σε κάθε βήμα, γίνεται με διαδικασία που εξαρτάται πάντα από την ίδια την συναθροιστική συνάρτηση. Στην υλοποίηση της παρούσας εργασίας επιλέγεται ο συνδυασμός που παρουσιάζει τις μικρότερες διαφορές μεταξύ των τιμών της συναθροιστικής συνάρτησης πάνω στις τιμές των εγγραφών της κλάσης ισοδυναμίας. Για την εύρεση αυτών αρχικά υπολογίζεται η μέγιστη διαφορά μεταξύ των τιμών της συναθροιστικής συνάρτησης

της κλάσης ισοδυναμίας για κάθε συνδυασμό n γνωρισμάτων και από αυτές επιλέγεται η μικρότερη. Έτσι εξασφαλίζεται πως όλες οι τιμές της συνάρτησης για αυτό το συνδυασμό γνωρισμάτων θα είναι οι πλησιέστερες δυνατές μεταξύ τους. Η σύγκριση θα μπορούσε να υλοποιηθεί και με χρήση του μέσου όρου των διαφορών που εμφανίζουν οι τιμές του αθροίσματος των τιμών κάθε εγγραφής για κάθε συνδυασμό και να επιλέγεται εκείνος με τον μικρότερο μέσο όρο διαφορών.

Για την καλύτερη κατανόηση της λειτουργίας του αλγορίθμου αναλύεται το παράδειγμα του Πίνακα 4.1, όπου επιχειρείται η 2-ανωνυμοποίηση των δεδομένων του, με χρήση της συνάρτησης του αθροίσματος και ποσοστό διαφοροποίησης $d = 0$. Σε συνέπεια με τα προηγούμενα παραδείγματα, τα δεδομένα αφορούν «Μερικά Εισοδήματα» ξεχωριστών ατόμων, και η συναθροιστική συνάρτηση το «Συνολικό Εισόδημα».

Α/Α	Μερικό Εισόδημα 1	Μερικό Εισόδημα 2	Μερικό Εισόδημα 3	Μερικό Εισόδημα 4	Συνολικό Εισόδημα
1	300	300	250	400	1250
2	400	300	100	650	1450
3	300	500	320	430	1550
4	150	350	200	350	1050

Πίνακας 4.1

Ακολουθώντας τα βήματα του αλγορίθμου, αφού έχει υπολογιστεί η τιμή της συναθροιστικής συνάρτησης, εδώ το συνολικό εισόδημα, για κάθε εγγραφή, ο πίνακας ταξινομείται σε αύξουσα διάταξη βάσει αυτής και διαχωρίζονται οι μετέπειτα κλάσεις ισοδυναμίας, με 2 εγγραφές σε κάθε ομάδα, όπως φαίνεται στον Πίνακα 4.2.

Α/Α	Μερικό Εισόδημα 1	Μερικό Εισόδημα 2	Μερικό Εισόδημα 3	Μερικό Εισόδημα 4	Συνολικό Εισόδημα
4	150	350	200	350	1050
1	300	300	250	400	1250
2	400	300	100	650	1450
3	300	500	320	430	1550

Πίνακας 4.2: Ταξινόμηση εγγραφών του Πίνακα 4.1 βάσει συναθροιστικής συνάρτησης και διαχωρισμός κλάσεων ισοδυναμίας

Επιλέγεται η πρώτη ομάδα των εγγραφών {4,1}.

Ελέγχονται οι τιμές των αθροισμάτων των δύο εγγραφών, οι οποίες διαφέρουν μεταξύ τους κατά ποσοστό μεγαλύτερο του $d = 0$.

Υπολογίζεται ο πίνακας των αθροισμάτων των εισοδημάτων σε κάθε εγγραφή, αφαιρώντας ένα γνώρισμα κάθε φορά με τη σειρά από το σύνολο του ψευδο-αναγνωριστικού όπως παρουσιάζεται στον Πίνακα 4.3:

A/A	{2,3,4}	{1,3,4}	{1,2,4}	{1,2,3}
4	900	700	850	700
1	950	950	1000	850

Πίνακας 4.3: Υπολογισμός συναθροιστικής συνάρτησης των δυνατών συνδυασμών των γνωρισμάτων ανά 3

Υπολογίζεται η διαφορά των τιμών των αθροισμάτων μεταξύ των εγγραφών σε κάθε στήλη και προκύπτει ότι από τα παραπάνω αθροίσματα τις πλησιέστερες τιμές στο «Συνολικό Εισόδημα» εμφανίζουν οι τιμές για το συνδυασμό των γνωρισμάτων {2,3,4}.

Επιλέγεται λοιπόν το πρώτο γνώρισμα προς γενίκευση αφού με την απουσία του από το άθροισμα οι τιμές των εγγραφών εμφανίζουν την μικρότερη διαφορά. Το «Μερικό Εισόδημα 1» γενικεύεται στο διάστημα τιμών [150,300], με αντικατάσταση των τιμών των εγγραφών σε αυτό το διάστημα, και αφαιρείται από το εξεταζόμενο σύνολο γνωρισμάτων του ψευδο-αναγνωριστικού. Αυτό φαίνεται στον Πίνακα 4.4, όπου το «Συνολικό Εισόδημα» υπολογίζεται από το άθροισμα των γνωρισμάτων που δεν έχουν γενικευτεί ακόμα, για κάθε εγγραφή.

A/A	Μερικό Εισόδημα 1	Μερικό Εισόδημα 2	Μερικό Εισόδημα 3	Μερικό Εισόδημα 4	Συνολικό Εισόδημα
4	[150,300]	350	200	350	900
1	[150,300]	300	250	400	950

Πίνακας 4.4: Αντικατάσταση του γενικευμένου γνωρίσματος στην κλάση ισοδυναμίας

Όπως αποτυπώνεται στον Πίνακα 4.4, η τιμή της συναθροιστικής συνάρτησης για τις τιμές των γνωρισμάτων που δεν έχουν γενικευτεί, για τις δύο εγγραφές είναι ίση με 900 και 950 αντίστοιχα. Εφόσον τα δύο αθροίσματα δεν είναι ίσα, γεγονός που επιβάλλεται από την απαίτηση της παραμέτρου $d = 0$, ο αλγόριθμος προχωράει στον υπολογισμό των

αθροισμάτων των τιμών των γνωρισμάτων για όλους τους δυνατούς συνδυασμούς των υπολειπόμενων τριών γνωρισμάτων ανά δύο, αφαιρώντας κάθε φορά ένα, όπως φαίνεται στον Πίνακα 4.5.

A/A	{3,4}	{2,4}	{2,3}
4	550	700	550
1	650	700	550

Πίνακας 4.5: Υπολογισμός συναθροιστικής συνάρτησης των δυνατών συνδυασμών των γνωρισμάτων ανά 2

Από εκεί επιλέγεται ο συνδυασμός που εμφανίζει τις μικρότερες διαφορές μεταξύ των αθροισμάτων των δύο εγγραφών, ο συνδυασμός των γνωρισμάτων {2,4}, αφού είναι ο πρώτος που συναντάται μεταξύ των {2,4}, {2,3} που δίνουν ίσο άθροισμα στις δύο εγγραφές. Η επιλογή αυτή οδηγεί στην γενίκευση του γνωρίσματος «Μερικό Εισόδημα 3» με την αντικατάσταση των τιμών του με το διάστημα τιμών [200,250].

Εφόσον τα αθροίσματα των τιμών των μη γενικευμένων γνωρισμάτων των δύο εγγραφών είναι ίσα, ο αλγόριθμος έχει ολοκληρώσει την επεξεργασία αυτής της κλάσης ισοδυναμίας. Την επιστρέφει, έχοντας γενικεύσει τα δύο γνωρίσματα «Μερικό Εισόδημα 1» και «Μερικό Εισόδημα 3» και συνεχίζει με την επεξεργασία της επόμενης κλάσης ισοδυναμίας. Στα δημοσιευμένα δεδομένα, ο καθένας μπορεί να υπολογίσει το ακριβές άθροισμα για κάθε εγγραφή αυτής της κλάσης, μόνο των γνωρισμάτων που δεν έχουν γενικευτεί, το οποίο θα είναι ακριβώς το ίδιο για τις δύο εγγραφές. Παράλληλα λόγω της γενίκευσης των άλλων δύο γνωρισμάτων, οι πιθανές τιμές των αθροισμάτων θα εμφανίζονται στο ίδιο διάστημα τιμών και για τις δύο εγγραφές.

Η μορφή της πρώτης κλάσης ισοδυναμίας που προκύπτει τελικά από τον αλγόριθμο εμφανίζεται στον Πίνακα 4.6.

Με την ίδια διαδικασία αντιμετωπίζεται η δεύτερη κλάση ισοδυναμίας των δεδομένων που αποτελείται από τις εγγραφές {2,3}. Ο αλγόριθμος εκεί επιλέγει να γενικεύσει τα δύο πρώτα γνωρίσματα ώστε να ικανοποιείται η ζητούμενη συνθήκη ισότητας των αθροισμάτων πάνω στις τιμές των εγγραφών της κλάσης. Τα αποτελέσματα για την κλάση αυτή εμφανίζονται επίσης στον Πίνακα 4.6, τον πίνακα των συνολικών ανωνυμοποιημένων δεδομένων που προκύπτουν από τον αλγόριθμο.

A/A	Μερικό Εισόδημα 1	Μερικό Εισόδημα 2	Μερικό Εισόδημα 3	Μερικό Εισόδημα 4	Συνολικό Εισόδημα
4	[150,300]	350	[200,250]	350	[1050,1250]
1	[150,300]	300	[200,250]	400	[1050,1250]
2	[300,400]	[300,500]	100	650	[1350,1650]
3	[300,400]	[300,500]	320	430	[1350,1650]

Πίνακας 4.6: 2-ανωνυμοποίηση του Πίνακα 4.1 με χρήση αλγόριθμου συναθροιστικής συνάρτησης

Ενδεικτικά περιγράφεται και ο τρόπος υπολογισμού της Κανονικοποιημένης Ποινής Βεβαιότητας όπως αυτή ορίστηκε στο Κεφάλαιο 3.

Για τα συγκεκριμένα δεδομένα, θεωρείται ότι όλα τα γνωρίσματα εμφανίζουν το κοινό πεδίο τιμών, αυτό το οποίο δημιουργείται από το διάστημα μεταξύ της μικρότερης και της μεγαλύτερης εμφανιζόμενης τιμής σε όλο το σύνολο δεδομένων.

Σε αρμονία με τον ορισμό, υπολογίζεται ο αριθμός των πιθανών τιμών από το πεδίο ορισμού των γνωρισμάτων:

650-100+1=551

Υπολογίζεται η Κανονικοποιημένη Ποινή Βεβαιότητας για κάθε τιμή που εμφανίζεται και στη συνέχεια το άθροισμα όλων διαιρείται με το πλήθος των τιμών των δεδομένων. Αυτή είναι μηδενική όταν η τιμή δεν έχει γενικευθεί, ή ίση με το εύρος του διαστήματος στο οποίο γενικεύεται, προς το πλήθος των πιθανών τιμών του γνωρίσματος. Επειδή σε κάθε κλάση ισοδυναμίας, όταν ένα γνώρισμα γενικεύεται, όλες οι εμφανιζόμενες τιμές του αντικαθίστανται με ένα κοινό διάστημα τιμών, έπεται πως αυτές οι τιμές θα έχουν και ίση Κανονικοποιημένη Ποινή Βεβαιότητας. Συνεπώς για το σύνολο των παραπάνω δεδομένων η συνολική απώλεια πληροφορίας θα είναι ο μέσος όρος της Κανονικοποιημένης Ποινής Βεβαιότητας όλων των τιμών:

$$NCP(RT) = \frac{2 \cdot (300 - 150) + 2 \cdot (250 - 200) + 2 \cdot (400 - 300) + 2 \cdot (500 - 300)}{4 \cdot 4 \cdot 551}$$

$$NCP(RT) = \frac{2 \cdot 150 + 2 \cdot 50 + 2 \cdot 100 + 2 \cdot 200}{8816}$$

$$NCP(RT) = 0.11343$$

5

Αξιολόγηση

5.1 Παράμετροι αξιολόγησης

Οι παράμετροι αξιολόγησης που χρησιμοποιήθηκαν για την διεξαγωγή των αποτελεσμάτων σχετικά με την χρηστικότητα και την αποδοτικότητα του αλγορίθμου είναι ο χρόνος και η μετρική απώλειας πληροφορίας.

Κύριο ρόλο έχει η μετρική απώλειας πληροφορίας που χρησιμοποιείται, βάσει της Κανονικοποιημένης Ποινής Βεβαιότητας όπως αυτή έχει οριστεί προηγουμένως. Λόγω της αριθμητικής φύσης των δεδομένων η μετρική που ορίζεται στο Κεφάλαιο 3 χρησιμοποιήθηκε με την ανταποκρίνουσα στο μοντέλο δεδομένων μορφή:

Έστω v η αρχική τιμή της εγγραφής στο γνώρισμα από το κοινό πεδίο τιμών των γνωρισμάτων I. Τότε η Κανονικοποιημένη Ποινή Βεβαιότητας υπολογίζεται από:

$$NCP(v) = \begin{cases} 0, & \text{αν η τιμή έχει γενικευθεί} \\ |max_v - min_v| \,/\, |A_i|, & \text{αλλιώς} \end{cases}$$

Όπου $|max_v - min_v|$ είναι το εύρος του γνωρίσματος στην κλάση ισοδυναμίας που εμφανίζεται η τιμή και το οποίο την αντικαθιστά κατά την γενίκευση από τον αλγόριθμο και $|A_i|$ η πληθικότητα του πεδίου τιμών του γνωρίσματος A_i στο οποίο ανήκει η τιμή v.

Τότε η συνολική απώλεια πληροφορίας του k-ανώνυμου συνόλου εγγραφών, από το αρχικό σύνολο RT, με πλήθος $|RT|$ και n γνωρίσματα είναι ο μέσος όρος για όλες τις εμφανιζόμενες τιμές, δηλαδή:

$$NCP(RT) = \frac{\sum_{i=0}^{|RT|}\left(\sum_{j=0}^{n} NCP(v_{i,j})\right)}{n \cdot |RT|}$$

Η μετρική αυτή υποδεικνύει την απώλεια πληροφορίας από τα αρχικά δεδομένα έπειτα από την τροποποίησή τους. Χρησιμοποιήθηκε για την αξιολόγηση των αποτελεσμάτων του αλγορίθμου που παρουσιάζεται στην εργασία, αλλά και για την σύγκριση αυτών με τα αντίστοιχα αποτελέσματα που παρέχει ο αλγόριθμος Mondrian για την ανωνυμοποίηση των δεδομένων.

Η δεύτερη παράμετρος αξιολόγησης που χρησιμοποιείται στην παρούσα εργασία είναι ο χρόνος εκτέλεσης του αλγορίθμου. Η σύγκριση μεταξύ των δύο αλγορίθμων ως προς τον χρόνο εκτέλεσης δεν είναι δυνατή μιας και ο υπολογισμός του χρόνου εξαρτάται σε μεγάλο βαθμό από την υλοποίηση. Ωστόσο είναι ένα χρήσιμο εργαλείο για την διεξαγωγή συμπερασμάτων ως προς την απόδοση του αλγορίθμου. Συγκεκριμένα μπορεί να παρατηρηθεί οποιαδήποτε διαφορά προκύπτει στον χρόνο εκτέλεσης αναφορικά με τα δεδομένα εισόδου, το μέγεθος του συνόλου δεδομένων, την τιμή της παραμέτρου ανωνυμίας k ή της παραμέτρου d. Από εκεί μπορεί να κριθεί η χρηστικότητα και η καταλληλότητα του αλγορίθμου σε πραγματικά σύνολα δεδομένων.

Οι τιμές που λαμβάνει η Κανονικοποιημένη Ποινή Βεβαιότητας αλλά και ο χρόνος εκτέλεσης σε κάθε εκτέλεση του αλγορίθμου επηρεάζονται άμεσα από τις τιμές που λαμβάνουν οι παράμετροι εισόδου του αλγορίθμου. Όπως θα παρουσιαστεί και στο Κεφάλαιο 5.3, ο χρόνος εκτέλεσης εξαρτάται από το μέγεθος του συνόλου των δεδομένων, τον αριθμό των εγγραφών που ανήκουν σε αυτά, καθώς και τον αριθμό των γνωρισμάτων που ανήκουν στο σύνολο του ψευδο-αναγνωριστικού. Αντίστοιχα, η Κανονικοποιημένη Ποινή Βεβαιότητας εξαρτάται και μεταβάλλεται σημαντικά για διαφορετικές τιμές της παραμέτρου ανωνυμίας k αλλά και της παραμέτρου χαλάρωσης της εγγύησης της ανωνυμίας d.

5.2 Οργάνωση πειραμάτων

Τα πειράματα που εκτελέστηκαν βασίστηκαν στην υλοποίηση του αλγορίθμου όπως περιγράφεται στο Κεφάλαιο 4, με χρήση C++. Κατά την υλοποίηση, ως συναθροιστική συνάρτηση χρησιμοποιήθηκε σε όλες τις εκτελέσεις το άθροισμα των τιμών των γνωρισμάτων κάθε εγγραφής, επιλογή η οποία μπορεί να αντικατασταθεί από τον μέσο όρο των τιμών ή κάποια άλλη συναθροιστική συνάρτηση.

5.2.1 Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν και παρήγαγαν τα αποτελέσματα που παρουσιάζονται στην παρούσα εργασία αναπαριστούν πραγματικά δεδομένα, προερχόμενα από απογραφές προσωπικών φορολογικών δεδομένων των κατοίκων των περιοχών Los Angeles και Long Beach, όπως αυτά δημοσιεύονται στην ιστοσελίδα UCI Machine Learning Repository [1].

Αν και δοκιμάστηκε η εκτέλεση του αλγορίθμου με συνθετικά δεδομένα, προτιμήθηκαν τα προαναφερθέντα δεδομένα για την ανάδειξη της λειτουργίας και της χρησιμότητας του αλγορίθμου σε πραγματικές συνθήκες.

Από το σύνολο των δεδομένων και αναλόγως των δυνατοτήτων του χρησιμοποιούμενου προσωπικού υπολογιστή επιλέχθηκε το αρχείο που αφορά την χρονική περίοδο του 1980, στο οποίο εμφανίζονται 74954 εγγραφές. Στο αρχικό σύνολο δεδομένων υπήρχαν 61 γνωρίσματα από τα οποία απομονώθηκαν τα 7 που αντιπροσώπευαν αποκλειστικά οικονομικά στοιχεία των ατόμων, όπως μισθοί και γενικότερα εισοδήματα από διαφορετικές πηγές.

5.2.2 Διαδικασία πειραμάτων

Από το αρχείο δεδομένων με τα οικονομικά πλέον γνωρίσματα, με 74954 εγγραφές και 7 γνωρίσματα, δημιουργήθηκαν υποσύνολα δεδομένων μεγέθους 50000, 25000,10000, 5000, 1000 προερχόμενα το καθένα από το αμέσως μεγαλύτερό του, διαδοχικά. Τα επιλεγμένα γνωρίσματα, θεωρήθηκαν πως προέρχονται από ένα κοινό πεδίο τιμών, μιας και όλα αφορούν οικονομικά δεδομένα ξεχωριστών ατόμων. Ο αριθμός των πιθανών τιμών του πεδίου τιμών για κάθε γνώρισμα, απαραίτητος για τον υπολογισμό της Κανονικοποιημένης Ποινής Βεβαιότητας, υπολογίσθηκε για κάθε αρχείο διαφορετικού μεγέθους πριν από κάθε εκτέλεση ως την διαφορά της μέγιστης και της ελάχιστης εμφανιζόμενης τιμής στο σύνολο των αρχικών δεδομένων συν ένα.

Η παρούσα υλοποίηση του αλγορίθμου επιβάλλει μη αρνητικό πεδίο τιμών των γνωρισμάτων, και για το λόγο αυτό τα δεδομένα που χρησιμοποιήθηκαν μετατοπίστηκαν κατά την μικρότερη αρνητική τιμή που εμφανιζόταν σε αυτά.

Με σκοπό την σύγκριση των αποτελεσμάτων του αλγορίθμου ως προς την κλασσική *k*-ανωνυμία, χρησιμοποιήθηκε ο καταλληλότερος αντιπροσωπευτικός αλγόριθμος Mondrian. Ο αλγόριθμος έχει υλοποιηθεί από το Πανεπιστήμιο του Dallas [2] και συγκεκριμένα από το UT Dallas Data Security and Privacy Lab, με χρήση java και xml.

Αφού τα δεδομένα τροποποιήθηκαν κατάλληλα, εκτελέστηκαν επαναλήψεις της υλοποίησης του αλγορίθμου για κάθε ένα από τα υποσύνολα δεδομένων και για τις τιμές των παραμέτρων k=(5,10,20,50) και d=(0,1,10,20), όπου υπολογίσθηκε η Κανονικοποιημένη Ποινή

Βεβαιότητας και ο χρόνος εκτέλεσης. Αντίστοιχες εκτελέσεις του αλγορίθμου Mondrian έγιναν για τα ίδια υποσύνολα δεδομένων και τις ίδιες τιμές της παραμέτρου *k*, ενώ δεν εξετάστηκε κάποια αντιστοιχία για την παράμετρο *d*, αφού ο αλγόριθμος Mondrian δεν εμπεριέχει κάποια επιλογή χαλάρωσης της εγγύησης της ανωνυμίας που παρέχει. Τα ανωνυμοποιημένα δεδομένα που προκύπτουν από τους δύο αλγορίθμους συγκρίθηκαν ως προς την απώλεια πληροφορίας που εμφανίζουν με χρήση της Κανονικοποιημένης Ποινής Βεβαιότητας. Ο υπολογισμός της μετρικής αυτής έγινε με την ανάπτυξη αντίστοιχων εργαλείων και στις δύο περιπτώσεις.

Το σύνολο των πειραμάτων έγινε με τη χρήση προσωπικού υπολογιστή, με επεξεργαστή Intel Core i5-2430M, με CPU 2,40GHz, με RAM 6,00 GB και λειτουργικό σύστημα Windows 7. Η υλοποίηση του αλγορίθμου έγινε ως εφαρμογή τερματικού παραθύρου, με σκοπό την δυνατότητα εκτέλεσης των πειραμάτων ανεξαρτήτως του λειτουργικού συστήματος του υπολογιστή. Για την διεξαγωγή των πειραμάτων στα Windows 7 χρησιμοποιήθηκε το MinGW, σύνολο προγραμματιστικών εργαλείων που διευκόλυνε τη φορητότητα μεταξύ Windows και Ubuntu, όπου αυτό ήταν απαραίτητο.

5.3 Αποτελέσματα

Το ενδιαφέρον εστιάστηκε στις δύο παραμέτρους αξιολόγησης όπως αναφέρονται στην παράγραφο 5.1. Εξετάστηκε η απόδοση του αλγορίθμου σε σχέση με τις διαφορετικές τιμές των παραμέτρων εισόδου του αλγορίθμου, ενώ όπου ήταν εφικτό και χρήσιμο τα αποτελέσματά του συγκρίθηκαν με τα αντίστοιχα αποτελέσματα του αλγορίθμου Mondrian.

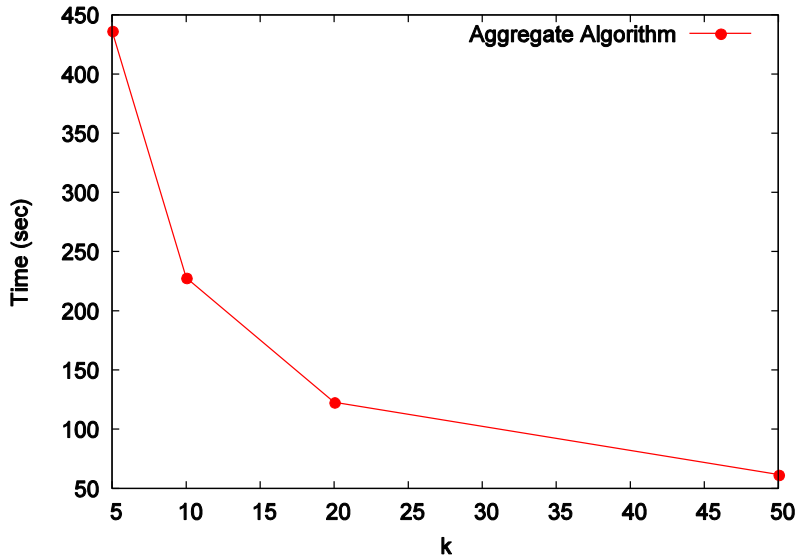
Συγκεκριμένα οι δύο αλγόριθμοι συγκρίθηκαν βάσει της Κανονικοποιημένης Ποινής Βεβαιότητας. Ο αλγόριθμος που παρουσιάζεται στην εργασία ήταν πιο γρήγορος στο σύνολο των πειραμάτων, όμως η σύγκριση βάσει του χρόνου εκτέλεσης δεν κρίθηκε σκόπιμη μιας και πρόκειται για διαφορετικές υλοποιήσεις σε διαφορετικές πλατφόρμες και γλώσσες προγραμματισμού.

5.3.1 Χρόνος εκτέλεσης

Αρχικά παρουσιάζονται τα αποτελέσματα από τις διαφορετικές εκτελέσεις του αλγορίθμου αναφορικά με τον χρόνο εκτέλεσής του. Αν και ο χρόνος εκτέλεσης που υπολογίζεται έχει άμεση εξάρτηση από τον τρόπο υλοποίησης του αλγορίθμου, ο συγκεκριμένος αλγόριθμος χάρη στην εξέταση κάθε κλάσης ισοδυναμίας μεμονωμένα από τις υπόλοιπες εγγραφές μπορεί να προσφέρει εξαιρετικά μικρούς χρόνους εκτέλεσης, κυρίως αν κατά την υλοποίησή

του παραλληλοποιηθεί η επαναληπτική διαδικασία. Με τον τρόπο αυτό θα μπορούσε να εξετάζει ταυτόχρονα περισσότερες από μία κλάσεις ισοδυναμίας.

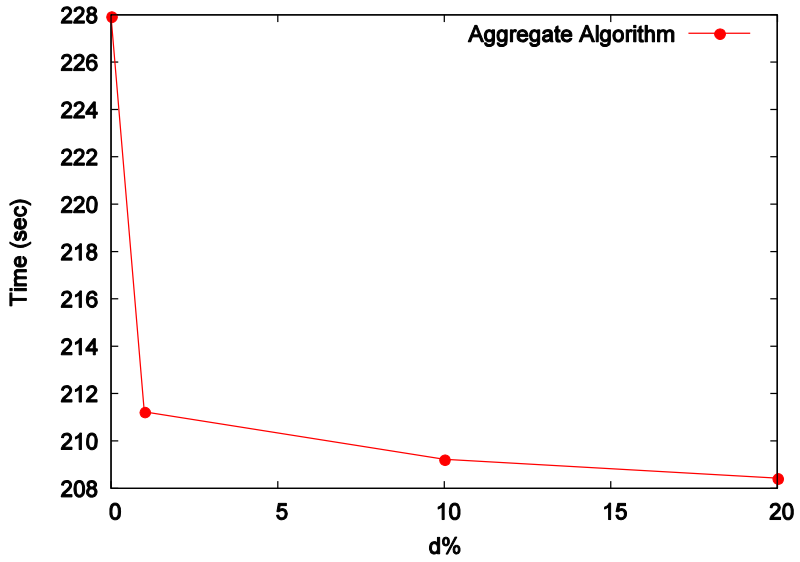
Στην Εικόνα 5.1 αποτυπώνεται ο χρόνος εκτέλεσης του αλγορίθμου, με είσοδό του το μεγαλύτερο εξεταζόμενο σύνολο δεδομένων 74954 εγγραφών και τιμές της παραμέτρου k από το σύνολο {5,10,20,50}, ενώ η παράμετρος d θεωρείται σταθερή και ίση με μηδέν. Όπως παρατηρείται ο χρόνος εκτέλεσης αυξάνεται σημαντικά όταν η παράμετρος ανωνυμίας k μειώνεται. Κάτι τέτοιο είναι αναμενόμενο μιας και η πολυπλοκότητα του αλγορίθμου εξαρτάται κατεξοχήν από τον αριθμό των κλάσεων ισοδυναμίας που θα δημιουργήσει και στη συνέχεια θα εξετάσει. Ο αριθμός των κλάσεων ισοδυναμίας, θεωρώντας το μέγεθος του συνόλου των εγγραφών σταθερό, αυξάνεται όσο η παράμετρος k μειώνεται αφού ο διαχωρισμός τους στο πειραματικό μέρος έχει επιλεχθεί να γίνεται λαμβάνοντας υπόψη την παράμετρο k.



Εικόνα 5.1: Χρόνος εκτέλεσης αλγορίθμου- παράμετρος ανωνυμίας k

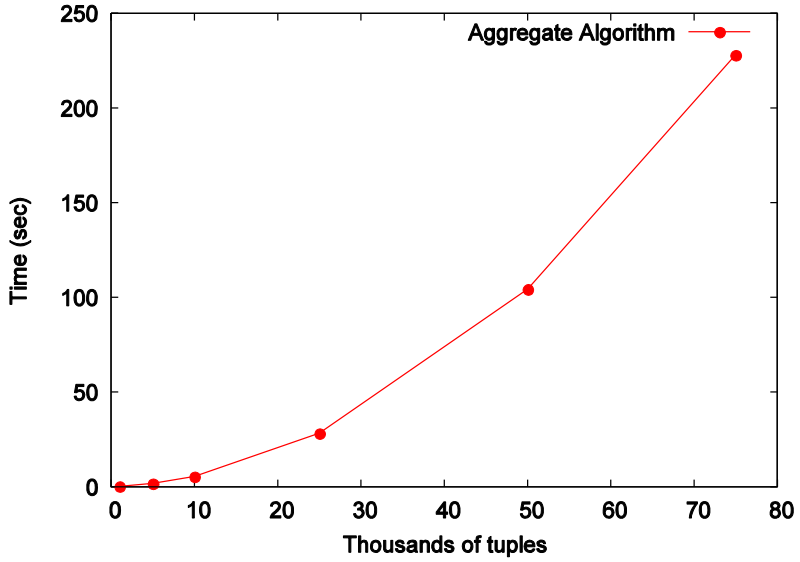
Στην Εικόνα 5.2 παρουσιάζονται οι χρόνοι εκτέλεσης του αλγορίθμου επίσης για το μεγαλύτερο σύνολο εγγραφών που χρησιμοποιήθηκε, θεωρώντας σταθερή την τιμή της παραμέτρου ανωνυμίας $k = 10$ και μεταβάλλοντας την παράμετρο d στις τιμές του συνόλου {0,1,10,20}. Όπως είναι αναμενόμενο ο χρόνος εκτέλεσης εμφανίζει μόνο μικρές διακυμάνσεις, μιας και η παράμετρος d δεν επηρεάζει σε τόσο μεγάλο βαθμό τον αριθμό των πράξεων που εκτελούνται. Η παράμετρος αυτή εκφράζει το μέγιστο επιτρεπόμενο ποσοστό διαφοροποίησης που μπορούν να εμφανίζουν οι τιμές της συναθροιστικής συνάρτησης μέσα σε κάθε κλάση ισοδυναμίας. Επηρεάζει συνεπώς τον αριθμό των γνωρισμάτων που μπορεί να χρειαστούν τροποποίηση μέσα σε μια κλάση ισοδυναμίας και γιαυτό απαιτεί λίγες πράξεις περισσότερες σε κάθε κλάση ισοδυναμίας όταν η τιμή της μειώνεται. Η μεταβολή αυτή

θεωρείται μικρή σε σχέση με την διαφοροποίηση που παρατηρείται στον χρόνο εκτέλεσης για τα διαφορετικά μεγέθη των συνόλων δεδομένων, όπως παρουσιάζεται στην Εικόνα 5.3.



Εικόνα 5.2: Χρόνος εκτέλεσης αλγορίθμου- παράμετρος χαλάρωσης της εγγύησης ανωνυμίας d

Στην Εικόνα 5.3 καταγράφονται οι χρόνοι εκτέλεσης του αλγορίθμου για τα σύνολα δεδομένων διαφορετικού μεγέθους, για την τιμή $k = 10$ και $d = 0$. Ο χρόνος εκτέλεσης αυξάνει σημαντικά καθώς αυξάνεται το μέγεθος του συνόλου των εγγραφών, κάτι που μπορεί να καταστήσει την συγκεκριμένη υλοποίηση ακατάλληλη για πολύ μεγάλα σύνολα δεδομένων. Η υλοποίηση θα μπορούσε να επεκταθεί έτσι ώστε να διαχειρίζεται παράλληλα διαφορετικές κλάσεις ισοδυναμίας και να οδηγήσει σε βέλτιστα αποτελέσματα καταμέτρησης του χρόνου εκτέλεσης.

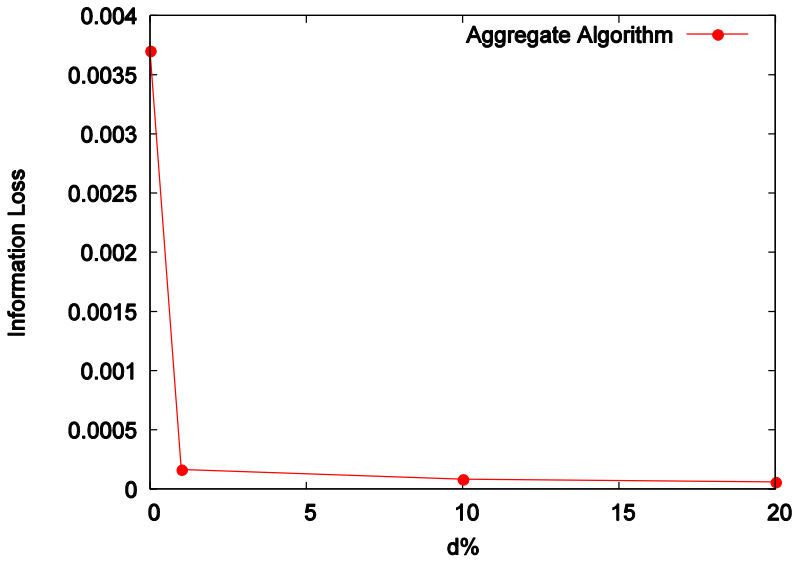


Εικόνα 5.3: Χρόνος εκτέλεσης αλγορίθμου- πλήθος εγγραφών συνόλου δεδομένων ανά 1000

5.3.2 Κανονικοποιημένη Ποινή Βεβαιότητας

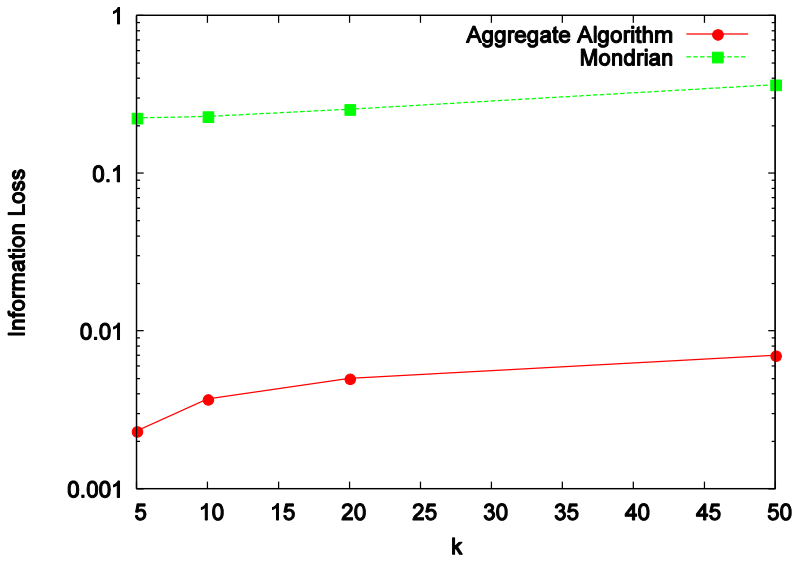
Στη συνέχεια μελετήθηκε η Κανονικοποιημένη Ποινή Βεβαιότητας για τις επαναλήψεις του αλγορίθμου και συγκρίθηκε με εκείνη του αλγορίθμου Mondrian για τα αντίστοιχα σύνολα δεδομένων, για την περίπτωση που η τιμή της παραμέτρου χαλάρωσης της εγγύησης της ανωνυμίας d είναι μηδέν, μιας και ο αλγόριθμος Mondrian δεν εμφανίζει κάποια ανάλογη συνθήκη. Όπως έχει ήδη προαναφερθεί, η Κανονικοποιημένη Ποινή Βεβαιότητας θεωρήθηκε κατάλληλη για τη σύγκριση των δύο αλγορίθμων ως προς την απώλεια πληροφορίας που εμφανίζουν τα ανωνυμοποιημένα δεδομένα που προκύπτουν από τους δύο αλγορίθμους.

Στην Εικόνα 5.4 αποτυπώνεται η απώλεια πληροφορίας με χρήση της Κανονικοποιημένης Ποινής Βεβαιότητας, όπως υπολογίστηκε για τις διαφορετικές τιμές της παραμέτρου d από το σύνολο $\{0,1,10,20\}$, σε επαναλήψεις του αλγορίθμου για το σύνολο των 74954 εγγραφών, με $k = 10$. Για μικρότερες τιμές της παραμέτρου χαλάρωσης d παρατηρείται σημαντική αύξηση της απώλειας πληροφορίας κατά την ανωνυμοποίηση των δεδομένων, όπως είναι αναμενόμενο, αφού οι τιμές της συναθροιστικής συνάρτησης που υπολογίζονται από τα δημοσιευμένα δεδομένα σε κάθε κλάση ισοδυναμίας επιβάλλεται να έχουν την ελάχιστη διαφορά ή ακόμα και να είναι ακριβώς ίσες. Από την απαίτηση αυτή συνεπάγεται πιθανώς η γενίκευση περισσότερων γνωρισμάτων σε κάθε κλάση ισοδυναμίας. Εντύπωση προκαλεί στα συγκεκριμένα αποτελέσματα η μεγάλη διαφορά στην απώλεια πληροφορίας που προκύπτει μεταξύ δύο διαφορετικών εκτελέσεων εκ των οποίων στη πρώτη το μέγιστο επιτρεπόμενο ποσοστό διαφοροποίησης μεταξύ των τιμών της συναθροιστικής συνάρτησης είναι μηδενικό ($d = 0$), και στην δεύτερη είναι ίσο με 1% , ($d = 1$). Συνεπώς μια τόσο μικρή χαλάρωση της εγγύησης της ανωνυμίας μπορεί να ωφελήσει στην διατήρηση περισσότερης πληροφορίας στα δημοσιευμένα δεδομένα. Από την τιμή της παραμέτρου $d = 1$ και για κάθε μεγαλύτερη τιμή, παρατηρείται μικρή διαφοροποίηση στην απώλεια πληροφορίας, καθώς όσο αυξάνει η τιμή της τόσο μεγαλώνει η πιθανότητα οι τιμές της συναθροιστικής συνάρτησης να ανήκουν στο κοινό εύρος τιμών που προκύπτει από την παράμετρο d.



Εικόνα 5.4: Μετρική απώλειας πληροφορίας – παράμετρος χαλάρωσης της εγγύησης της ανωνυμίας d

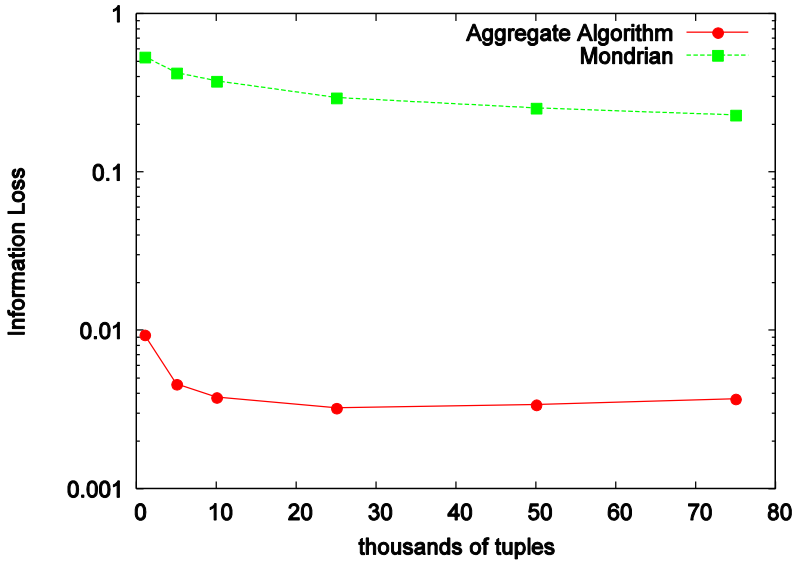
Στην Εικόνα 5.5, παρουσιάζεται η απώλεια πληροφορίας που προκύπτει έπειτα από την εκτέλεση και των δύο αλγορίθμων πάνω στο ίδιο σύνολο δεδομένων 74954 εγγραφών. Η τιμή της παραμέτρου *d* για τον αλγόριθμο της παρούσας εργασίας διατηρήθηκε σταθερή και ίση με μηδέν σε κάθε επανάληψη, ώστε να συμβαδίζει με τα δεδομένα εισόδου του αλγορίθμου Mondrian. Για την καλύτερη ανάδειξη των αποτελεσμάτων προτιμήθηκε η απόδοση της Κανονικοποιημένης Ποινής Βεβαιότητας σε λογαριθμική κλίμακα, μιας και οι δύο αλγόριθμοι εμφανίζουν μεγάλη διαφορά ως προς την απώλεια πληροφορίας η οποία δεν αποτυπώνεται με τον καλύτερο τρόπο κατά την δεκαδική κλίμακα. Από την Εικόνα 5.5 διεξάγεται το συμπέρασμα της υπεροχής του εν λόγω αλγορίθμου σε σύγκριση με τον αλγόριθμο Mondrian αναφορικά με την απώλεια πληροφορίας που εμφανίζουν τα ανωνυμοποιημένα δεδομένα. Παρατηρείται σημαντική διαφορά στην απώλεια πληροφορίας μεταξύ των δύο αλγορίθμων για κάθε συνδυασμό των παραμέτρων εισόδου, γεγονός που οφείλεται κατά κύριο λόγο στην χρήση της συναθροιστικής συνάρτησης από τον αλγόριθμο για την κατάλληλη ανωνυμοποίηση των δεδομένων, κάτι που ο αλγόριθμος Mondrian αγνοεί.



Εικόνα 5.5: Μετρική απώλειας πληροφορίας - παράμετρος ανωνυμίας k

Στην Εικόνα 5.6 παρουσιάζεται η απώλεια πληροφορίας που εμφανίζεται στα ανωνυμοποιημένα δεδομένα κατά την εκτέλεση των δύο αλγορίθμων για σύνολα δεδομένων διαφορετικού μεγέθους. Η παράμετρος d θεωρείται μηδέν και η παράμετρος ανωνυμίας διατηρεί την τιμή $k = 10$.

Οι δύο αλγόριθμοι για κάθε πλήθος εγγραφών διατηρούν μεγάλη διαφορά στην απώλεια πληροφορίας των ανωνυμοποιημένων δεδομένων που παρέχουν και για το λόγο αυτό τα αποτελέσματα παρουσιάζονται επίσης σε λογαριθμική κλίμακα. Και στους δύο παρατηρείται μικρή αύξηση της απώλειας πληροφορίας για τα μικρότερα σύνολα δεδομένων, γεγονός που τεκμηριώνεται λογικά, μιας και όσο λιγότερες είναι οι εγγραφές, τόσο μικρότερη είναι η πιθανότητα εμφάνισης κοινών τιμών στις εγγραφές και συνεπώς απαιτείται μεγαλύτερη γενίκευση.



Εικόνα 5.6: Μετρική απώλειας πληροφορίας - πλήθος εγγραφών συνόλου δεδομένων ανά 1000

5.4 Σύνοψη συμπερασμάτων αξιολόγησης

Σύμφωνα με τα παραπάνω αποτελέσματα, ο αλγόριθμος που παρουσιάζεται στην παρούσα εργασία προσφέρει μια αποδοτικότερη λύση ως προς την διατήρηση της χρήσιμης πληροφορίας στα προς δημοσίευση δεδομένα, για το πρόβλημα της ανωνυμοποίησης των δεδομένων παρέχοντας προστασία κατά των επιθέσεων με ύπαρξη συναθροιστικής γνώσης πάνω στις τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού.

Ο αλγόριθμος, όπως υλοποιήθηκε εμφανίζει αποδεκτούς χρόνους εκτέλεσης κάτι που μπορεί να ελαχιστοποιηθεί με διαφορετική υλοποίηση μέσω της δυνατότητας της παράλληλης εξέτασης των κλάσεων ισοδυναμίας που προσφέρει.

Στην περίπτωση που ο επιτιθέμενος κατέχει συναθροιστική πληροφορία για τα γνωρίσματα του ψευδο-αναγνωριστικού η *k*-ανωνυμοποίηση με τον αλγόριθμο αυτό φαίνεται πολύ πιο αποδοτική από εκείνη του αλγορίθμου Mondrian ως προς την διατήρηση της πληροφορίας που έχουν τα δεδομένα. Προσφέρει επαρκή προστασία απέναντι στην ανακάλυψη της ταυτότητας των εγγραφών βάσει της συναθροιστικής συνάρτησης αφού γενικεύει τα δεδομένα μόνο όσο είναι απαραίτητο ώστε σε κάθε κλάση ισοδυναμίας οι εγγραφές να εμφανίζουν την ίδια τιμή κατά τον υπολογισμό της συναθροιστικής συνάρτησης. Με τον τρόπο αυτό διαφυλάσσει την ταυτότητα των εγγραφών στις συγκεκριμένες επιθέσεις και εξοικονομεί χρήσιμη πληροφορία κατά την ανωνυμοποίηση.

6

Τεχνικές λεπτομέρειες

Στο Κεφάλαιο αυτό παρουσιάζονται όλες οι τεχνικές λεπτομέρειες αναφορικά με την υλοποίηση του αλγορίθμου σε εφαρμογή τερματικού παραθύρου αλλά και σε εφαρμογή γραφικού περιβάλλοντος, τη διεξαγωγή των πειραμάτων και τις επιπλέον απαιτούμενες ενέργειες που έγιναν.

6.1 Λεπτομέρειες υλοποίησης

Η υλοποίηση του αλγορίθμου με τη μορφή εφαρμογής τερματικού παραθύρου έγινε με χρήση της γλώσσας C++. Κατά τη διάρκεια της ανάπτυξης της εφαρμογής αλλά και για την διεξαγωγή των πειραμάτων χρησιμοποιήθηκε το σύνολο προγραμματιστικών εργαλείων MinGW, με τον μεταγλωττιστή g++. Στην συγκεκριμένη υλοποίηση του αλγορίθμου ενσωματώθηκε ως συναθροιστική συνάρτηση το άθροισμα των τιμών κάθε εγγραφής σε όλα τα εμφανιζόμενα γνωρίσματα και συνεπώς η αλλαγή της επιλογής αυτής απαιτεί την κατάλληλη τροποποίηση του κώδικα της εφαρμογής.

6.1.1 Χαρακτηριστικά υλοποίησης

6.1.1.1 Μορφή δεδομένων εισόδου - εξόδου

Στην παρούσα υλοποίηση το αρχικό σύνολο των δεδομένων εισάγεται με τη μορφή αρχείου απλού κειμένου. Κάθε εγγραφή αφορά μία γραμμή κειμένου, ενώ οι τιμές της εγγραφής αυτής για τα αντίστοιχα γνωρίσματα διαχωρίζονται μεταξύ τους με το χαρακτήρα «Tab».

Ο αλγόριθμος μετά την εκτέλεσή του παράγει ένα αρχείο κειμένου με τα ανωνυμοποιημένα δεδομένα, αντίστοιχα των τιμών των παραμέτρων που εισήγαγε ο χρήστης. Στο αρχείο αυτό καταγράφεται επίσης η τιμή της Κανονικοποιημένης Ποινής Βεβαιότητας που προκύπτει από τα ανωνυμοποιημένα δεδομένα συναρτήσει των αρχικών δεδομένων, καθώς και ο χρόνος εκτέλεσης του αλγορίθμου.

6.1.1.2 Κλήση της εφαρμογής

Η εφαρμογή καλείται από το παράθυρο του τερματικού, αφού πρώτα έχει μεταγλωττιστεί, με την παρακάτω ακολουθία για τα ορίσματα εισόδου της:

- *Όνομα αρχείου δεδομένων:* Το πλήρες όνομα του αρχείου στο οποίο βρίσκονται τα αρχικά δεδομένα, μαζί με την επέκταση τύπου του αρχείου, για παράδειγμα microdata.dat
- *Τιμή παραμέτρου ανωνυμίας k:* Η τιμή της παραμέτρου *k* κατά την οποία απαιτούμε τα ανωνυμοποιημένα δεδομένα να ικανοποιούν την *k*-ανωνυμία ως προς την συναθροιστική συνάρτηση. Όπως είναι λογικό, η τιμή της παραμέτρου *k* ορίζεται ως ακέραια και μεγαλύτερη του μηδενός.
- *Τιμή παραμέτρου χαλάρωσης της εγγύησης της ανωνυμίας d:* Το μέγιστο ποσοστό διαφοροποίησης που επιτρέπεται να εμφανίζουν μεταξύ τους οι τιμές της συναθροιστικής συνάρτησης των εγγραφών σε κάθε κλάση ισοδυναμίας, στα ανωνυμοποιημένα δεδομένα. Η τιμή εισάγεται σε μορφή ακεραίου από το διάστημα τιμών [0,100].

6.1.1.3 Δομές δεδομένων

Κατά την ανάπτυξη του αλγορίθμου με χρήση της γλώσσας C++ κρίθηκαν καταλληλότερες για την αποθήκευση και την επεξεργασία των δεδομένων οι προσφερόμενες δομές δεδομένων, τα διανύσματα τύπου *vector*, και η ταξινομημένη δομή *multimap*, από το σύνολο C++ Standard Template Library.

Τα διανύσματα vector αποθηκεύουν τα δεδομένα τους κατ' ακολουθία, όπως ακριβώς και τα arrays, με την σημαντική διαφορά της δυναμικής τροποποίησης του μεγέθους τους. Συνεπώς προτιμήθηκαν για την εξοικονόμηση της απαιτούμενης μνήμης, Χρησιμοποιήθηκαν για τον ορισμό μονοδιάστατων διανυσμάτων αλλά και για την υλοποίηση πινάκων δύο διαστάσεων με τη μορφή ενός διανύσματος με διανύσματα ως στοιχεία του (<vector <vector> >).

Η δομή δεδομένων multimap<key, value> είναι μια δομή που ουσιαστικά συσχετίζει δεδομένα της μορφής ζεύγους. Κάθε ζεύγος αποτελείται από μία *τιμή-κλειδί (key)* και την αντίστοιχη *τιμή (value*) της. Η δομή multimap ταξινομεί τις εισαγόμενες σε αυτήν τιμές βάσει της τιμής-κλειδί που έχουν, με χρήση της δοσμένης συνάρτησης ταξινόμησης, ενώ επιτρέπει την εισαγωγή δύο εγγραφών με την ίδια τιμή-κλειδί. Στην παρούσα εφαρμογή, η δομή multimap χρησιμοποιήθηκε κατά την εκτέλεση της ταξινόμησης των εγγραφών βάσει της τιμής της συναθροιστικής συνάρτησης, όπου ως τιμή-κλειδί χρησιμοποιήθηκε η τιμή του αθροίσματος των επιμέρους τιμών των εγγραφών και αυτή αντιστοιχίστηκε με το διάνυσμα των τιμών κάθε εγγραφής.

Στο σύνολο της υλοποίησης προτιμήθηκε η αναπαράσταση των τιμών των δεδομένων από αριθμούς κινητής υποδιαστολής διπλής ακρίβειας, τύπου *double*. Αν και τα δεδομένα που εξετάστηκαν είχαν την μορφή ακεραίων, η επιλογή αυτή προσφέρει μεγαλύτερη ακρίβεια στις πράξεις που εκτελούνται ενώ παράλληλα επιτρέπει τη χρήση μεγαλύτερων αριθμών.

6.1.1.4 Συναρτήσεις

Η εφαρμογή αποτελείται από ένα αρχείο γλώσσας C++. Σε αυτό αναπτύχθηκαν οι συναρτήσεις βασικών πράξεων και μετατροπών μεταξύ δομών και τύπων που χρειάζονται και οι συναρτήσεις που υλοποιούν τον αλγόριθμο. Οι συναρτήσεις αυτές είναι οι:

- *loadFile:* Συνάρτηση που ανοίγει το αρχείο με το όνομα που εισάγεται ως αλφαριθμητικό.
- *mapFromFile:* Συνάρτηση που δημιουργεί τη δομή multimap με το σύνολο των δεδομένων από το αρχικό αρχείο. Τα δεδομένα ταξινομούνται σε αύξουσα σειρά βάσει του αθροίσματος των τιμών που εμφανίζονται σε κάθε γραμμή του αρχείου.
- *getclass:* Συνάρτηση που δέχεται ως όρισμα την τιμή της παραμέτρου *k* και το σύνολο των δεδομένων που δεν έχει ακόμη εξεταστεί. Διαχωρίζει και επιστρέφει την επόμενη κατάλληλη κλάση ισοδυναμίας.
- *findDiff:* Συνάρτηση που υπολογίζει τις διαφορές των στοιχείων ανά δύο από το μονοδιάστατο διάνυσμα-όρισμα εισόδου και τις επιστρέφει σε ένα μονοδιάστατο διάνυσμα.

- *findPerDiff:* Συνάρτηση που υπολογίζει τις ποσοστιαίες διαφορές των στοιχείων του διανύσματος εισόδου ανά δύο και τις επιστρέφει σε ένα νέο διάνυσμα.
- *samesum:* Δέχεται ως είσοδο ένα μονοδιάστατο διάνυσμα και την τιμή της παραμέτρου *d* και υπολογίζει τις ποσοστιαίες διαφορές των στοιχείων του διανύσματος ανά δύο μέσω της findPerDiff. Επιστρέφει *true* αν όλες είναι μικρότερες ή ίσες από την τιμή της παραμέτρου *d*.
- *sameval:* Επιστρέφει *true* αν όλα τα στοιχεία του διανύσματος που δέχεται ως όρισμα έχουν την ίδια τιμή. Χρησιμοποιείται για την εύρεση τετριμμένων περιπτώσεων όπου ένα γνώρισμα εμφανίζει την ίδια τιμή για όλες τις εγγραφές σε μια κλάση ισοδυναμίας και προφανώς δεν απαιτείται γενίκευση.
- *findPositiveMin:* Εντοπίζει και επιστρέφει την μικρότερη τιμή που εμφανίζεται μέσα στο διάνυσμα εισόδου. Χρησιμοποιείται κατά την επιλογή του γνωρίσματος προς γενίκευση.
- *findposition:* Εντοπίζει και επιστρέφει την θέση του στοιχείου στο διάνυσμα εισόδου με τιμή το δεύτερο όρισμα εισόδου.
- *checkanonymity:* Αποτελεί βασική μέθοδο της υλοποίησης και η λειτουργία της αναλύεται στην επόμενη παράγραφο. Δέχεται ως είσοδο την εξεταζόμενη κλάση ισοδυναμίας και την τιμή της παραμέτρου *d* και επιστρέφει την κλάση ισοδυναμίας υποδεικνύοντας το γνώρισμα που έχει επιλέξει να γενικεύσει ο αλγόριθμος σε κάθε βήμα.
- *recursion:* Συνάρτηση για την υλοποίηση της αναδρομής του αλγορίθμου, η λειτουργία της οποίας αναλύεται επίσης παρακάτω. Δέχεται ως είσοδο την εξεταζόμενη κλάση ισοδυναμίας και την τιμή της παραμέτρου *d* και επιστρέφει την κλάση ισοδυναμίας υποδεικνύοντας τα προς γενίκευση γνωρίσματα, αφού ο αλγόριθμος ολοκληρωθεί για αυτήν την κλάση.
- *replaceAttToString:* Συνάρτηση που αντικαθιστά στην κλάση ισοδυναμίας όπως αυτή προκύπτει από την συνάρτηση recursion τα οριζόμενα από τον αλγόριθμο γνωρίσματα με τις αντίστοιχες γενικεύσεις τους. Επιστρέφει την ανωνυμοποιημένη κλάση ισοδυναμίας στην τελική της μορφή, ως πίνακα αλφαριθμητικών.
- *replaceClassToString:* Αντικαθιστά την ανωνυμοποιημένη κλάση ισοδυναμίας στην αρχική της θέση στο σύνολο των δεδομένων, έτσι ώστε αφού ελεγχθούν όλες οι κλάσεις ισοδυναμίας να προκύπτει το ανωνυμοποιημένο σύνολο δεδομένων.
- *lossmetric:* Υπολογίζει τον παράγοντα της κλάσης ισοδυναμίας που συμμετέχει στην συνολική Κανονικοποιημένη Ποινή Βεβαιότητας των δεδομένων.

- *Main-βασική συνάρτηση:* Συνδέει όλες τις προαναφερθείσες συναρτήσεις με τρόπο ώστε να εκτελείται ο θεωρητικός αλγόριθμος. Η λειτουργία της αναλύεται στην επόμενη παράγραφο.

6.1.2 Ανάλυση βασικών μεθόδων κώδικα

6.1.2.1 Βασική Συνάρτηση

Υλοποιεί το κύριο μέρος του αλγορίθμου με τα παρακάτω βήματα.

- Διαβάζει τις επιλογές του χρήστη για το αρχείο δεδομένων, την τιμή της παραμέτρου ανωνυμίας *k* και την τιμή της παραμέτρου χαλάρωσης της εγγύησης της ανωνυμίας *d*.
- Ανοίγει το αρχείο των αρχικών δεδομένων και διαβάζοντάς το γραμμή-γραμμή περνάει τις εγγραφές μέσα στη δομή multimap με χρήση της συνάρτησης mapFromFile(). Με τον τρόπο αυτό γίνεται παράλληλα η ταξινόμηση των εγγραφών-γραμμών βάσει του αθροίσματος των τιμών που εμφανίζουν στα γνωρίσματα, σε αύξουσα σειρά.
- Τροποποιεί την δομή που χρησιμοποιεί για το σύνολο των ταξινομημένων πλέον δεδομένων σε πίνακα με τη χρήση της κατάλληλης συνάρτησης, για την διευκόλυνση των πράξεων που ακολουθούν.
- Υπολογίζει το πλήθος των πιθανών τιμών του πεδίου τιμών που εμφανίζουν τα γνωρίσματα στο σύνολο των δεδομένων. Αυτό γίνεται βρίσκοντας την μικρότερη και την μεγαλύτερη εμφανιζόμενη τιμή και υπολογίζοντας την τιμή $max - min + 1$.
- Ξεκινάει την επαναληπτική διαδικασία:
    - Αρχικά διαχωρίζει και αποθηκεύει σε ξεχωριστό πίνακα την κλάση ισοδυναμίας που πρόκειται να εξετάσει, με την κλήση της συνάρτησης getclass.
    - Για την δεδομένη κλάση ισοδυναμίας καλεί την συνάρτηση recursion όπου εφαρμόζει την αναδρομική διαδικασία που επιβάλει ο αλγόριθμος και η λειτουργία της αναλύεται παρακάτω. Η συνάρτηση αυτή επιστρέφει την κλάση ισοδυναμίας έχοντας «σημειώσει» τα γνωρίσματα που υποδεικνύει ο αλγόριθμος πως πρέπει να γενικευθούν με σκοπό την ικανοποίηση της *k*-ανωνυμίας.
    - Αντικαθιστά τα γνωρίσματα αυτά με τα διαστήματα τιμών που εμφανίζουν στην δεδομένη κλάση ισοδυναμίας, με κλήση της συνάρτησης replaceAttToString, όπου υπολογίζονται και τα ζητούμενα διαστήματα τιμών.

- Αντικαθιστά τις εγγραφές της ανωνυμοποιημένης κλάσης ισοδυναμίας στις θέσεις τους στο αρχικό σύνολο των ταξινομημένων εγγραφών με τη συνάρτηση replaceClassToString.
- Υπολογίζει από τα ανωνυμοποιημένα δεδομένα της κλάσης ισοδυναμίας τον παράγοντα κατά τον οποίο συμμετέχει η κλάση ισοδυναμίας που εξετάστηκε στην Κανονικοποιημένη Ποινή Βεβαιότητας μέσω της συνάρτησης lossmetric.

Ο βρόχος επαναλαμβάνεται έως ότου να μην υπάρχει καμία εγγραφή προς εξέταση στο αρχικό σύνολο δεδομένων.

- Έπειτα υπολογίζεται ο μέσος όρος της Κανονικοποιημένης Ποινής Βεβαιότητας για το σύνολο των ανωνυμοποιημένων δεδομένων.

6.1.2.2 Συνάρτηση διαχωρισμού κλάσεων

Η συνάρτηση getclass καλείται μία φορά για κάθε κλάση ισοδυναμίας. Αποσπάει σε κάθε επανάληψή της την επόμενη κλάση ισοδυναμίας από το σύνολο των προς εξέταση υπολειπόμενων εγγραφών. Τα ορίσματα εισόδου της είναι το τρέχον σύνολο εγγραφών και η τιμή της παραμέτρου *k*. Ελέγχει από το σύνολο αυτό αν οι εγγραφές που δεν έχουν εξεταστεί είναι περισσότερες από *2k*. Αυτός ο έλεγχος πραγματοποιείται με την διαίρεση του πλήθους των υπολειπόμενων προς εξέταση εγγραφών με την τιμή της *k*. Αν το πηλίκο είναι μεγαλύτερο της μονάδας, έπεται ότι απομένουν περισσότερες των *2k* εγγραφές, οπότε αποθηκεύει σε έναν νέο πίνακα τις πρώτες *k* εγγραφές από το σύνολο και τις διαγράφει από το αρχικό σύνολο. Αν όμως το πηλίκο είναι ίσο ή μικρότερο της μονάδας, έπεται πως απομένουν λιγότερες των *2k* εγγραφές προς εξέταση. Εφόσον όλες οι κλάσεις ισοδυναμίας προκύπτουν με τον τρόπο που περιγράφεται εδώ, το πλήθος των υπολειπόμενων εγγραφών κατά την περίπτωση που συναντάει λιγότερες από *2k* εγγραφές πάντα θα είναι ίσο με το $k + |RT| mod\ k$, όπου $|RT|$ το αρχικό πλήθος των εγγραφών που εισήχθησαν στον αλγόριθμο. Με αυτή τη διαδικασία, στην πρώτη περίπτωση εντάσσονται όλες οι κλάσεις ισοδυναμίας πλην της τελευταίας που δημιουργείται από το σύνολο, η οποία αναλόγως του πλήθους του αρχικού συνόλου εγγραφών είναι μεγέθους k ή $k + |RT| mod\ k$.

6.1.2.3 Συνάρτηση αναδρομής

Η συνάρτηση recursion μαζί με την συνάρτηση checkanonymity υλοποιούν την αναδρομή που εμφανίζει ο αλγόριθμος. Η συνάρτηση δέχεται ως είσοδο την εξεταζόμενη κλάση

ισοδυναμίας και επιστρέφει την κλάση ισοδυναμίας με σημειωμένα όλα τα γνωρίσματα που πρέπει να γενικευθούν όπως αυτό υπολογίζεται από τον αλγόριθμο.

Αρχικά ελέγχει τις εγγραφές για τετριμμένες περιπτώσεις στις τιμές του κάθε γνωρίσματος. Ελέγχει δηλαδή αν υπάρχουν γνωρίσματα που εμφανίζουν την ίδια ακριβώς τιμή σε όλες τις εγγραφές της κλάσης ισοδυναμίας, τα οποία και εξαιρεί πριν ξεκινήσει η εκτέλεση του αλγορίθμου για την κλάση αυτή. Τα γνωρίσματα αυτά αντικαθίστανται με τις αρχικές τιμές τους, κατά την ολοκλήρωση του αλγορίθμου, αφού διατηρούν την *k*-ανωνυμία όταν αυτή ικανοποιείται. Στην υλοποίηση τα γνωρίσματα αυτά σημειώνονται όπως εκείνα που γενικεύονται, με σκοπό να μην συμμετάσχουν στην διαδικασία που ακολουθεί, ενώ όταν αντικαθίστανται οι τιμές των γνωρισμάτων προς γενίκευση, αυτά παίρνουν την ίδια τιμή που είχαν στα αρχικά δεδομένα.

Στη συνέχεια, ξεκινάει ο βρόχος μέσα στον οποίο ελέγχονται οι τιμές της συναθροιστικής συνάρτησης για τα μη γενικευμένα γνωρίσματα σε κάθε εγγραφή. Όταν αυτές δεν ικανοποιούν τη συνθήκη ισότητας με μέγιστη διαφοροποίηση ανά δύο ποσοστό ίσο με την παράμετρο *d*, καλείται η συνάρτηση checkanonymity. Η διαδικασία επαναλαμβάνεται έως ότου τεθεί ψευδής η συνθήκη του βρόχου, δηλαδή δεν υπάρχουν άλλα γνωρίσματα προς γενίκευση ή ικανοποιηθεί η ζητούμενη συνθήκη *k*-ανωνυμίας ως προς την συναθροιστική συνάρτηση για κάποιο συνδυασμό γνωρισμάτων.

6.1.2.4 Συνάρτηση ανωνυμίας

Η συνάρτηση checkanonymity χρησιμοποιείται σε κάθε βήμα για τον έλεγχο των τιμών της συναθροιστικής συνάρτησης των εγγραφών της κλάσης ισοδυναμίας και τελικά την εύρεση του επόμενου γνωρίσματος που πρέπει να γενικευθεί με στόχο την ικανοποίηση της *k*-ανωνυμίας. Δέχεται ως όρισμα την εξεταζόμενη κλάση ισοδυναμίας και την τιμή της παραμέτρου *d* και επιστρέφει την κλάση ισοδυναμίας με «σημειωμένο» έως ένα γνώρισμα προς γενίκευση. Η συνάρτηση υλοποιεί ένα βήμα του αλγορίθμου σε κάθε της εκτέλεση για την συγκεκριμένη κλάση ισοδυναμίας.

Αρχικά υπολογίζει για την κλάση ισοδυναμίας την τιμή της συναθροιστικής συνάρτησης κάθε εγγραφής των γνωρισμάτων που δεν έχουν γενικευθεί, έστω *n*. Ελέγχει αν αυτές ικανοποιούν την συνθήκη ισότητας που έχει οριστεί. Αν την ικανοποιούν επιστρέφει την κλάση ως έχει. Αν δεν την ικανοποιούν, δημιουργεί ένα νέο διάνυσμα δύο διαστάσεων όπου αποθηκεύει τις τιμές της συναθροιστικής συνάρτησης για τους διαφορετικούς συνδυασμούς των $n-1$ γνωρισμάτων, όπως υποδεικνύει ο αλγόριθμος. Στον πίνακα αυτό, σε κάθε στήλη υπολογίζεται η τιμή της συναθροιστικής συνάρτησης έτσι όπως προκύπτει από τα

γνωρίσματα του συνόλου των μη γενικευμένων γνωρισμάτων, πλην εκείνου που βρίσκεται στον αρχικό πίνακα δεδομένων στην στήλη αυτή.

Στην συνέχεια υπολογίζει για κάθε στήλη τις διαφορές ανά δύο στις τιμές του πίνακα των τιμών της συναθροιστικής συνάρτησης που δημιούργησε, με χρήση της συνάρτησης findDiff. Από αυτές, υπολογίζει την μέγιστη διαφορά που εμφανίζει κάθε στήλη με χρήση της συνάρτησης findMax και την εισάγει στο διάνυσμα που αποθηκεύει τις μέγιστες διαφορές που συναντάει. Επαναλαμβάνει την διαδικασία με σκοπό την εύρεση της μέγιστης εμφανιζόμενης διαφοράς για κάθε στήλη στον πίνακα των τιμών της συναθροιστικής συνάρτησης για τους διαφορετικούς συνδυασμούς των γνωρισμάτων. Στη συνέχεια επιλέγει από το διάνυσμα μεγίστων διαφορών την μικρότερη τιμή. Επιλέγοντας την μικρότερη από τις μέγιστες εμφανιζόμενες διαφορές στα αθροίσματα των συνδυασμών των γνωρισμάτων εξασφαλίζουμε πως όλες οι διαφορές για τον συνδυασμό που επιλέγουμε θα είναι οι μικρότερες όλων στον πίνακα. Η επιλογή της συνθήκης αυτής θα μπορούσε να είναι και ο μέσος όρος των διαφορών έτσι ώστε να επιλέγεται ο συνδυασμός των γνωρισμάτων που εμφανίζουν τον μικρότερο μέσο όρο στις διαφορές της συναθροιστικής συνάρτησης.

Ο συνδυασμός που επιλέγεται υποδεικνύει το γνώρισμα που πρόκειται να γενικευθεί το οποίο βρίσκεται στην αντίστοιχη στήλη στην αρχική μορφή της κλάσης ισοδυναμίας, με χρήση της συνάρτησης findposition. Οι τιμές του γνωρίσματος της κλάσης ισοδυναμίας που πρόκειται να γενικευθεί αντικαθίστανται με την τιμή-δείκτη «-1». Με τον τρόπο αυτό δεν λαμβάνονται υπόψη στις επόμενες επαναλήψεις της συνάρτησης, ούτε κατά τον υπολογισμό του πίνακα των τιμών της συναθροιστικής συνάρτησης, όπου και εκεί αντικαθίστανται με την τιμή «-1», ενώ στην συνάρτηση main αντικαθίστανται από τις γενικευμένες τιμές μέσω της συνάρτησης replaceAttToString.

6.1.2.5 Συνάρτηση υπολογισμού μετρικής

Η συνάρτηση lossmetric υλοποιήθηκε με σκοπό τον υπολογισμό σε κάθε κλάση ισοδυναμίας του παράγοντα που αυτή συνεισφέρει στην συνολική απώλεια πληροφορίας, όπως περιγράφεται από την Κανονικοποιημένη Ποινή Βεβαιότητας. Ως είσοδο δέχεται την θέση της πρώτης εγγραφής της κλάσης ισοδυναμίας στον ταξινομημένο πίνακα δεδομένων, το πλήθος των πιθανών τιμών του κοινού πεδίου τιμών των γνωρισμάτων, το αρχικό σύνολο δεδομένων και την κλάση ισοδυναμίας μετά την εφαρμογή του αλγορίθμου, με σημειωμένα τα προς γενίκευση γνωρίσματα με την τιμή «-1». Το αποτέλεσμα της συνάρτησης είναι ένα διάνυσμα με τον παράγοντα συμμετοχής στην Κανονικοποιημένη Ποινή Βεβαιότητας για κάθε γνώρισμα.

Σκοπός της συνάρτησης είναι ο υπολογισμός της πράξης

$$NCP(v) = \begin{cases} 0, & \text{αν η τιμή έχει γενικευθεί} \\ |max_v - min_v| \,/\, |A_i|, & \text{αλλιώς} \end{cases}$$

για κάθε τιμή που εμφανίζεται στα ανωνυμοποιημένα πλέον δεδομένα.

Για να το πετύχει αυτό, για κάθε στήλη-γνώρισμα στην κλάση ισοδυναμίας που έχει γενικευθεί υπολογίζει την μέγιστη και την ελάχιστη τιμή που εμφανίζει καθώς και το εύρος που έχουν. Η $NCP(v)$ για κάθε τιμή που εμφανίζεται βάσει του παραπάνω τύπου είναι ίση με το εύρος αυτό προς το πλήθος των πιθανών τιμών του πεδίου τιμών του γνωρίσματος. Εφόσον οι τιμές που εμφανίζονται στο γνώρισμα γενικεύονται όλες στο ίδιο διάστημα τιμών, έπεται πως θα έχουν όλες την ίδια $NCP(v)$. Συνεπώς ο παράγοντας που επιστρέφει για κάθε γνώρισμα είναι ίσος με την τιμή $NCP(v)$ που υπολόγισε επί τον αριθμό των εγγραφών της κλάσης ισοδυναμίας. Ο παράγοντας που προκύπτει για κάθε γνώρισμα αποθηκεύεται σε ένα διάνυσμα, το οποίο επιστρέφεται από την συνάρτηση.

Στη συνέχεια στην βασική συνάρτηση main αθροίζονται όλα τα στοιχεία του διανύσματος και διαιρούνται με το σύνολο των τιμών των αρχικών δεδομένων, ακολουθώντας τον τύπο

$$NCP(RT) = \frac{\sum_{i=0}^{|RT|}\left(\sum_{j=0}^{n} NCP(v_{i,j})\right)}{n \cdot |RT|}$$

6.1.3 Ανάπτυξη χρήσιμων εργαλείων

Για την κατάλληλη τροποποίηση του συνόλου των αρχικών δεδομένων με σκοπό την συμβατότητά τους με τους δύο αλγορίθμους αναπτύχθηκαν τα παρακάτω εργαλεία με χρήση της C++.

6.1.3.1 Αρχείο μετακίνησης δεδομένων

Κατά την παρούσα υλοποίηση του αλγορίθμου επιλέχθηκε η δυνατότητα εισαγωγής μόνο μη αρνητικών τιμών με σκοπό τη διευκόλυνση της επικοινωνίας των προαναφερόμενων συναρτήσεων, με την αντικατάσταση των τιμών των προς γενίκευση γνωρισμάτων με «-1», ώστε να μην λαμβάνουν μέρος στις πράξεις. Κατά τη διαδικασία προσαρμογής του κώδικα ώστε να δέχεται και αρνητικές τιμές στο σύνολο των δεδομένων αποδείχθηκε πως προτιμότερη είναι η προσαύξηση όλων των τιμών του συνόλου των δεδομένων κατά την απόλυτη τιμή του μικρότερου εμφανιζόμενου αρνητικού αριθμού. Για το λόγο αυτό αναπτύχθηκε το κατάλληλο εργαλείο μετακίνησης δεδομένων, το οποίο καλείται πριν από την κύρια εφαρμογή πάνω στα δεδομένα και τα τροποποιεί όταν περιέχουν αρνητικές τιμές.

Με αυτό, αρχικά υπολογίζεται η μέγιστη και η ελάχιστη εμφανιζόμενη τιμή στο σύνολο των δεδομένων, αφού κατά το μοντέλο δεδομένων της παρούσας εργασίας όλα τα γνωρίσματα

έχουν το ίδιο πεδίο τιμών. Όταν η ελάχιστη εμφανιζόμενη τιμή είναι αρνητική, τότε όλες οι τιμές των δεδομένων εγγράφονται στο νέο δοσμένο από τον χρήστη αρχείο, προσαυξημένες με την απόλυτη ελάχιστη τιμή.

6.1.3.2 Αρχείο επιλογής γνωρισμάτων

Το εργαλείο αυτό υλοποιήθηκε με σκοπό να απομονώνει τα δοσμένα γνωρίσματα από το αρχικό σύνολο δεδομένων. Η ανάπτυξή του κρίθηκε απαραίτητη ώστε να απομονωθούν από το αρχείο των προσωπικών φορολογικών δεδομένων που χρησιμοποιήθηκαν, μόνο εκείνα που αναφέρονται σε οικονομικά γνωρίσματα, ώστε να αναδεικνύεται βέλτιστα η χρηστικότητα του αλγορίθμου. Το εν λόγω αρχείο δέχεται ως όρισμα το αρχικό αρχείο δεδομένων και το όνομα του αρχείου στο οποίο θα αντιγράψει μόνο τα επιλεγμένα γνωρίσματα.

6.1.3.3 Αρχείο δειγματοληψίας

Για την καλύτερη συσχέτιση των αρχείων δεδομένων διαφορετικού μεγέθους και την ορθότερη διεξαγωγή συμπερασμάτων από την εκτέλεση του αλγορίθμου σε αυτά, προτιμήθηκε η διαδοχική δημιουργία υποσυνόλων δεδομένων από το αρχικό σύνολο. Για την διαδικασία αυτή αναπτύχθηκε το κατάλληλο αρχείο κώδικα. Δέχεται ως ορίσματα το όνομα του αρχικού αρχείου δεδομένων, το όνομα του νέου αρχείου-υποσυνόλου και τον αριθμό των εγγραφών που ζητείται να περιέχει το δεύτερο. Επιλέγονται τυχαία εγγραφές από το πρώτο και εγγράφονται στο δεύτερο μέχρι να καλυφθεί το ζητούμενο πλήθος. Επιστρέφει ένα υποσύνολο δεδομένων και η διαδικασία επαναλαμβάνεται με την εκτέλεσή του, με αρχικό πλέον αρχείο δεδομένων το τελευταίο υποσύνολο που δημιουργήθηκε, μέσω των κατάλληλων σεναρίων φλοιού που αναπτύχθηκαν.

6.1.3.4 Αρχείο υπολογισμού μετρικής Mondrian

Η εφαρμογή αυτή αναπτύχθηκε για την δυνατότητα σύγκρισης των αποτελεσμάτων του παρουσιαζόμενου αλγορίθμου σε σχέση με τα αντίστοιχα αποτελέσματα του αλγορίθμου Mondrian αναφορικά με την απώλεια χρήσιμης πληροφορίας που παρουσιάζουν. Δημιουργήθηκε ξεχωριστή εφαρμογή για τον υπολογισμό της Κανονικοποιημένης Ποινής Βεβαιότητας από τα ανωνυμοποιημένα δεδομένα του Mondrian λόγω της διαφορετικότητας της μορφής τους σε σχέση με τον αλγόριθμο της εργασίας. Η μετρική αυτή υπολογίζεται με τη χρήση του κώδικα του αρχείου υπολογισμού μετρικής Mondrian, χρησιμοποιώντας την

ίδια διαδικασία όπως στον αλγόριθμο που αναπτύσσει η παρούσα εργασία. Η κύρια διαφορά πέραν της μορφής με την οποία καταγράφονται τα ανωνυμοποιημένα δεδομένα, όσον αφορά την μετρική, είναι πως ο αλγόριθμος Mondrian επιλέγοντας την κατάλληλη γενίκευση σε κάθε κλάση ισοδυναμίας για κάθε γνώρισμα, δύναται να επιλέξει την αντικατάσταση των αρχικών τιμών με κλειστά και ανοιχτά διαστήματα τιμών. Η εφαρμογή αυτή δέχεται ως ορίσματα το αρχείο των ανωνυμοποιημένων δεδομένων όπως αυτό προκύπτει από την εκτέλεση του αλγορίθμου Mondrian με την υλοποίηση από το Πανεπιστήμιο του Dallas, τον αριθμό των εγγραφών που ανήκουν στο σύνολο δεδομένων, τον αριθμό των γνωρισμάτων τους και το πλήθος των δυνατών τιμών του κοινού πεδίου τιμών των γνωρισμάτων. Επιστρέφει την Κανονικοποιημένη Ποινή Βεβαιότητας για τα συγκεκριμένα δεδομένα, η οποία υπολογίζεται όπως έχει περιγραφεί παραπάνω, ενώ λαμβάνει υπόψη την ύπαρξη ανοιχτών και κλειστών διαστημάτων κατά τον υπολογισμό του εύρους των γενικευμένων τιμών – διαστημάτων.

6.1.4 Λεπτομέρειες λειτουργίας αλγορίθμου Mondrian

Ο αλγόριθμος Mondrian, όπως έχει ήδη αναφερθεί χρησιμοποιήθηκε με την υλοποίηση σε Java από το Πανεπιστήμιο του Dallas. Η υλοποίηση αυτή έχει διαμορφωθεί έτσι ώστε να εκτελεί και άλλους αλγορίθμους εκτός του Mondrian, οι οποίοι δεν εξετάζονται από αυτή την εργασία. Η εφαρμογή δέχεται τα προς ανωνυμοποίηση δεδομένα σε μορφή απλού κειμένου, θεωρώντας αντίστοιχα με τον αλγόριθμο της εργασίας κάθε γραμμή ως μία ξεχωριστή εγγραφή. Οι παράμετροι που ζητούνται κατά την εκτέλεση της εφαρμογής συγκεντρώνονται στο αρχείο παραμετροποίησης και με χρήση αυτού εισέρχονται στην εφαρμογή. Στο αρχείο αυτό καθορίζονται τα γνωρίσματα του ψευδο-αναγνωριστικού και τα ευαίσθητα γνωρίσματα. Οι παράμετροι που απαιτούνται για την εκτέλεση της υλοποίησης είναι ο αλγόριθμος που ζητείται να εκτελεστεί, η τιμή της παραμέτρου ανωνυμίας *k*, το αρχείο με τα αρχικά δεδομένα, το αρχείο στο οποίο θα εγγραφούν τα ανωνυμοποιημένα δεδομένα και κατ' επιλογή, αναλόγως τον αλγόριθμο που χρησιμοποιείται, ζητείται η ιεραρχία γενίκευσης του πεδίου τιμών κάθε γνωρίσματος που θέλουμε να χρησιμοποιηθεί.

Αφού οριστεί το αρχείο παραμετροποίησης με τις ζητούμενες τιμές των παραμέτρων, ο αλγόριθμος εκτελείται από το τερματικό παράθυρο μέσω του κατάλληλου σεναρίου φλοιού.

Από το αρχείο των αποτελεσμάτων, των ανωνυμοποιημένων δεδομένων γίνεται ο υπολογισμός της Κανονικοποιημένης Ποινής Βεβαιότητας με χρήση της εφαρμογής υπολογισμού μετρικής Mondrian.

6.2 Πλατφόρμες και προγραμματιστικά εργαλεία

6.2.1 Εφαρμογή γραφικού περιβάλλοντος

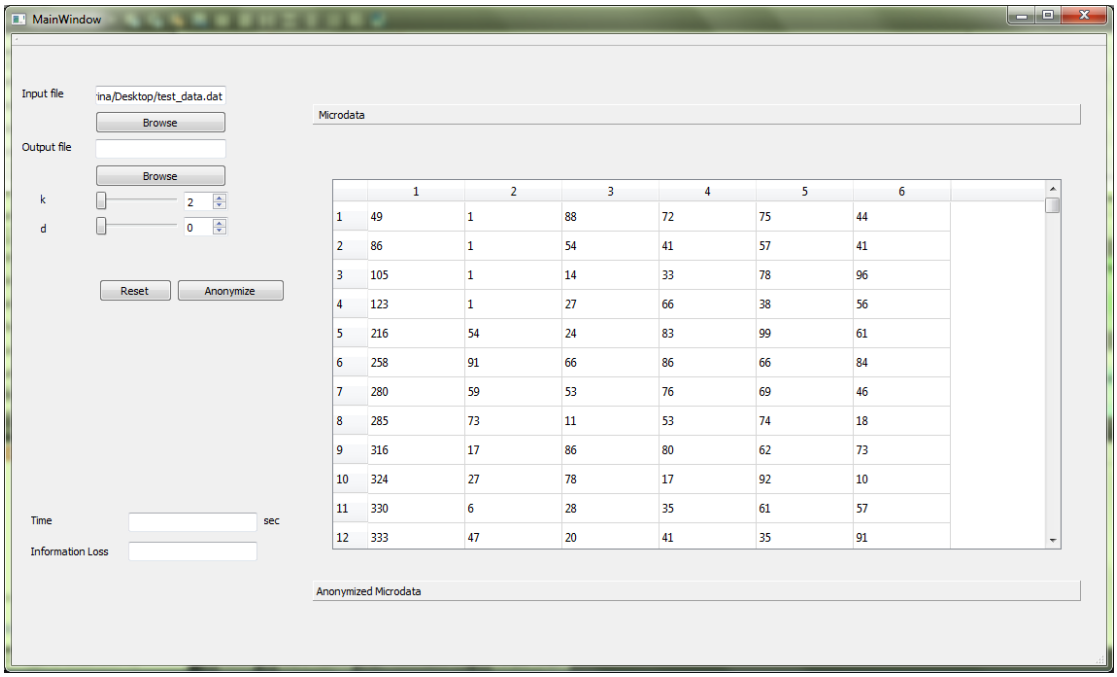
Η εφαρμογή τερματικού παραθύρου αναπτύχθηκε με στόχο την διεξαγωγή των πειραμάτων ενώ παράλληλα αναπτύχθηκε εφαρμογή γραφικού περιβάλλοντος που υλοποιεί τον αλγόριθμο. Η εφαρμογή γραφικού περιβάλλοντος δημιουργήθηκε ως εκτελέσιμο αρχείο λειτουργικού συστήματος Windows και μπορεί να χρησιμοποιηθεί από οποιονδήποτε με εισαγωγή των δεδομένων και των τιμών των παραμέτρων που ζητούνται με πιο εύχρηστο τρόπο.

Για την ανάπτυξη της εφαρμογής γραφικού περιβάλλοντος χρησιμοποιήθηκε το ανεξαρτήτου πλατφόρμας ολοκληρωμένο περιβάλλον ανάπτυξης της γλώσσας C++ (cross-platform C++ integrated development environment), Qt Creator, τμήμα του πλαισίου εφαρμογών (application framework) Qt SDK. Το εργαλείο αυτό χρησιμοποιεί τα βασικά στοιχεία της γλώσσας C++, προσφέροντας περισσότερες δυνατότητες για την ανάπτυξη εφαρμογών γραφικού περιβάλλοντος (GUI). Για την ορθή λειτουργία του αλγορίθμου χρησιμοποιήθηκε ο κώδικας της εφαρμογής τερματικού ο οποίος εμπλουτίστηκε με τα γραφικά στοιχεία.

Τα γραφικά στοιχεία που εμφανίζονται στην γραφική εφαρμογή αναπαριστούν όλες τις επιλογές που μπορεί να εισάγει ο χρήστης για την εκτέλεση του αλγορίθμου και παρουσιάζονται στο στιγμιότυπο της Εικόνας 6.1. Για την επιλογή και την εμφάνιση των αρχείων εισόδου-εξόδου χρησιμοποιήθηκαν στοιχεία της μορφής *γραμμής κειμένου* (line edit) και αντίστοιχα σε αυτά *κουμπιά περιήγησης*. Τα κουμπιά περιήγησης ανοίγουν *παράθυρα διαλόγου* για την περιήγηση του χρήστη στους καταλόγους αρχείων, από όπου επιλέγει τα αρχεία δεδομένων που θέλει να ανωνυμοποιήσει. Για την επιλογή των τιμών των παραμέτρων ανωνυμίας k και της παραμέτρου χαλάρωσης της εγγύησης της ανωνυμίας d, χρησιμοποιήθηκαν *μπάρες ολίσθησης* (slider bar) και *πλαίσια αυξομείωσης τιμών*(spin box), οι λειτουργίες των οποίων είναι συνδεδεμένες μεταξύ τους για την διευκόλυνση του χρήστη. Το διάστημα τιμών που εμφανίζεται για την τιμή της παραμέτρου k είναι [2, 99], επειδή για $k < 2$ η έννοια της k-ανωνυμίας είναι τετριμμένη, ενώ θεωρείται απίθανη η επιλογή τιμών $k > 99$ μιας και πιθανότατα τα δεδομένα θα έχουν ήδη υπεργενικευτεί.

Το διάστημα τιμών που εμφανίζεται για την παράμετρο της χαλάρωσης της εγγύησης της ανωνυμίας d είναι [0,99]. Κάτι τέτοιο αναπαριστά την μορφή του ποσοστού που έχει η παράμετρος d, ενώ η τιμή $d = 100$ δεν εισάγεται μιας και σε μια τέτοια περίπτωση οι τιμές της συναθροιστικής συνάρτησης σε κάθε κλάση ισοδυναμίας θα μπορούν να έχουν την μέγιστη δυνατή διαφορά μεταξύ τους οπότε καμία λειτουργία του αλγορίθμου δεν θα είναι απαραίτητη. Κάτι τέτοιο ενδέχεται να ισχύει και για πολύ μικρότερες τιμές της παραμέτρου

d, όμως παρέχεται η δυνατότητα αυτών στον χρήστη γιατί κάτι τέτοιο εξαρτάται πάντα από τα δεδομένα που εισάγονται.



Εικόνα 6.1: Στιγμιότυπο εισαγωγής δεδομένων προς ανωνυμοποίηση στην εφαρμογή γραφικού περιβάλλοντος

Για την αναπαράσταση των αρχικών και ανωνυμοποιημένων δεδομένων χρησιμοποιήθηκαν γραφικά στοιχεία με τη μορφή *πίνακα* (table widgets) ενώ για την παρουσίαση των τιμών του χρόνου εκτέλεσης και της μετρικής της απώλειας πληροφορίας χρησιμοποιήθηκαν επίσης αντικείμενα της μορφής γραμμής κειμένου.

Σε αρμονία με την εφαρμογή τερματικού παραθύρου, ο χρήστης πατώντας το κουμπί περιήγησης της επιλογής «Input file» , διαλέγει το αρχείο των αρχικών δεδομένων μέσα από το παράθυρο διαλόγου περιήγησης που εμφανίζεται. Ο χρήστης εισάγει τα δεδομένα σε μορφή αρχείου κειμένου, όπου κάθε γραμμή αναπαριστά μια εγγραφή και τις τιμές που αυτή λαμβάνει σε κάθε γνώρισμα. Με το άνοιγμα του αρχείου, τα δεδομένα εμφανίζονται με τη μορφή πίνακα υπό τον τίτλο «Microdata», στο αντικείμενο που χρησιμοποιείται για την αναπαράσταση των αρχικών δεδομένων.

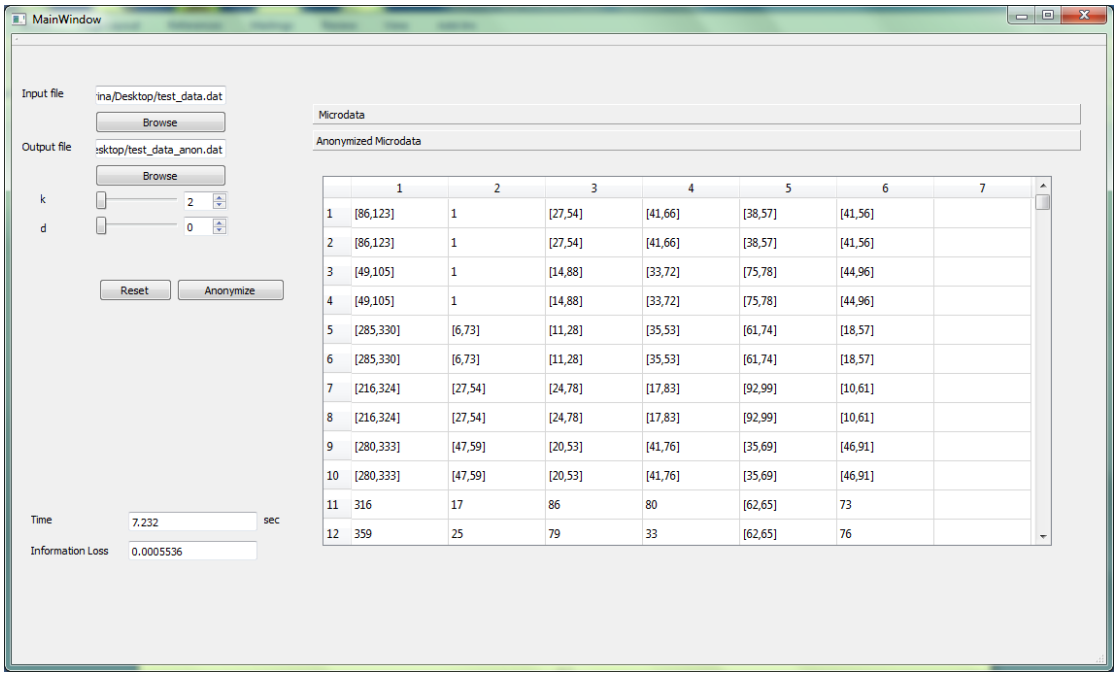
Κατά αντιστοιχία επιλέγει το αρχείο στο οποίο επιθυμεί να αποθηκευτούν τα ανωνυμοποιημένα δεδομένα από το παράθυρο διαλόγου περιήγησης που ανοίγει πατώντας το κουμπί «Browse» του «Output file». Στο μενού αυτό έχει την επιλογή να δημιουργήσει ένα νέο αρχείο για την αποθήκευση απλά εισάγοντας το όνομα που επιθυμεί στην γραμμή κειμένου με τίτλο «Output file». Στη συνέχεια επιλέγει τις τιμές των παραμέτρων *k* και *d* με

χρήση της μπάρας ολίσθησης ή του αντίστοιχου πλαισίου αυξομείωσης τιμών, στο οποίο και εμφανίζεται η τιμή που επιλέγει για κάθε παράμετρο.

Ο χρήστης οδηγείται σε δύο κουμπιά με δύο διαφορετικές επιλογές. Πατώντας το κουμπί «Anonymize», εκτελείται ο αλγόριθμος της εργασίας για τα δεδομένα που εισήγαγε ο χρήστης και για τις τιμές των παραμέτρων που επέλεξε μέσα από το γραφικό περιβάλλον. Τα ανωνυμοποιημένα δεδομένα εμφανίζονται στο γραφικό στοιχείο με μορφή πίνακα με τίτλο «Anonymized Microdata» και ο χρήστης μπορεί να τα δει πατώντας πάνω στην ετικέτα αυτή, όπως παρουσιάζεται στην Εικόνα 6.2. Τα δεδομένα στον πίνακα εμφανίζονται διαχωρισμένα σε κλάσεις ισοδυναμίας έτσι ώστε να είναι εμφανής η ορθή λειτουργία του αλγορίθμου και να εντοπίζονται πιο εύκολα τυχόν λάθη.

Με την εκτέλεση του αλγορίθμου, στο πλαίσιο γραμμής κειμένου με τίτλο «Time» παρουσιάζεται ο χρόνος εκτέλεσης του αλγορίθμου σε δευτερόλεπτα, ενώ στο πλαίσιο με τίτλο «Information Loss» παρουσιάζεται η Κανονικοποιημένη Ποινή Βεβαιότητας η οποία υπολογίζεται με τον ίδιο ακριβώς τρόπο όπως και στην εφαρμογή τερματικού παραθύρου.

Η δεύτερη επιλογή του χρήστη αφορά το κουμπί αναίρεσης «Reset». Με αυτό αναιρούνται όλες οι επιλογές που έχει κάνει σχετικά με τα αρχεία εισόδου-εξόδου και τις τιμές των παραμέτρων, καθώς απομακρύνονται από το γραφικό περιβάλλον τα δεδομένα που έχουν εισαχθεί, τα ανωνυμοποιημένα δεδομένα και οι τιμές του χρόνου και της μετρικής της απώλειας πληροφορίας.



	1	2	3	4	5	6	7
1	[86,123]	1	[27,54]	[41,66]	[38,57]	[41,56]	
2	[86,123]	1	[27,54]	[41,66]	[38,57]	[41,56]	
3	[49,105]	1	[14,88]	[33,72]	[75,78]	[44,96]	
4	[49,105]	1	[14,88]	[33,72]	[75,78]	[44,96]	
5	[285,330]	[6,73]	[11,28]	[35,53]	[61,74]	[18,57]	
6	[285,330]	[6,73]	[11,28]	[35,53]	[61,74]	[18,57]	
7	[216,324]	[27,54]	[24,78]	[17,83]	[92,99]	[10,61]	
8	[216,324]	[27,54]	[24,78]	[17,83]	[92,99]	[10,61]	
9	[280,333]	[47,59]	[20,53]	[41,76]	[35,69]	[46,91]	
10	[280,333]	[47,59]	[20,53]	[41,76]	[35,69]	[46,91]	
11	316	17	86	80	[62,65]	73	
12	359	25	79	33	[62,65]	76	

Εικόνα 6.2: Στιγμιότυπο προβολής ανωνυμοποιημένων αποτελεσμάτων εφαρμογής γραφικού περιβάλλοντος

7

Επίλογος

7.1 Σύνοψη και συμπεράσματα

Η εργασία αυτή ασχολήθηκε με το πρόβλημα της διασφάλισης της ιδιωτικότητας των εγγραφών σε βάσεις δεδομένων υπό συνθήκες που μέχρι τώρα δεν έχουν ερευνηθεί. Σε αντίθεση με αντίστοιχες εργασίες, θεωρήθηκε η περίπτωση επίθεσης κατά την οποία ο επιτιθέμενος έχει συναθροιστική γνώση πάνω στο σύνολο των γνωρισμάτων του ψευδο-αναγνωριστικού. Τα γνωρίσματα του ψευδο-αναγνωριστικού προέρχονται από το ίδιο πεδίο τιμών, έτσι ώστε οι τιμές τους να εμφανίζουν κάποια συσχέτιση που παρέχει την συναθροιστική πληροφορία. Με στόχο την εγγύηση της ανωνυμίας των εγγραφών, επεκτάθηκε η έννοια της *k*-ανωνυμίας έτσι ώστε να λαμβάνει υπόψη την συναθροιστική συνάρτηση που αναπαριστά την γνώση του επιτιθέμενου. Αναπτύχθηκε αναδρομικός αλγόριθμος που εγγυάται την ικανοποίηση της *k*-ανωνυμίας αυτής της μορφής και επιστρέφει την *k*-ανωνυμοποίηση του συνόλου των δεδομένων. Ο αλγόριθμος βασίζεται στην χρήση της τοπικής γενίκευσης ενώ εξετάζει κάθε κλάση ισοδυναμίας μεμονωμένα. Ο αλγόριθμος υλοποιήθηκε ως εφαρμογή με την οποία εκτελέστηκαν πειράματα σε πραγματικά δεδομένα. Η απόδοσή του εξετάστηκε σε σύγκριση με τον αλγόριθμο *k*-ανωνυμοποίησης πολυδιάστατης διαμέρισης με χρήση τοπικής γενίκευσης, Mondrian. Η απόδοση των δύο αλγορίθμων πάνω στα ίδια σύνολα δεδομένων αξιολογήθηκε με χρήση της Κανονικοποιημένης Ποινής Βεβαιότητας στην κατάλληλη μορφή της. Όπως φαίνεται και από τα αποτελέσματα, η υπεροχή του αλγορίθμου που παρουσιάζεται για το πρόβλημα αυτό είναι ευδιάκριτη καθώς

επιτυγχάνει πολύ μικρότερη απώλεια πληροφορίας από τον αλγόριθμο Mondrian για κάθε συνδυασμό των τιμών των παραμέτρων. Ακόμα, υπερτερεί κατά την δυνατότητα ελαχιστοποίησης του χρόνου εκτέλεσής του μιας και με την κατάλληλη υλοποίηση μπορεί να παραλληλοποιηθεί έτσι ώστε να εξετάζει περισσότερες από μία κλάσεις ισοδυναμίας του συνόλου δεδομένων ταυτόχρονα. Αυτό οφείλεται στον τρόπο λειτουργίας του όπου εξετάζει τις εγγραφές κάθε κλάσης ισοδυναμίας χωριστά, χωρίς να λαμβάνει υπόψη τις υπόλοιπες εγγραφές του συνόλου.

Συμπεραίνεται πως για το πρόβλημα της προστασίας της ιδιωτικότητας σε σύνολα δεδομένων της παραπάνω μορφής από επιθέσεις με συναθροιστική γνώση στις τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού κάποιας εγγραφής, ο αλγόριθμος που αναπτύχθηκε επιτυγχάνει μικρότερη απώλεια πληροφορίας και ελαχιστοποίηση του χρόνου εκτέλεσης συγκριτικά με τον αλγόριθμο Mondrian.

7.2 Μελλοντικές επεκτάσεις

Ο αλγόριθμος που παρουσιάστηκε προστατεύει από επιθέσεις με συναθροιστική γνώση πάνω στις τιμές των γνωρισμάτων, όμως λόγω της πρακτικής χρησιμότητας που παρουσιάζει μπορεί να επεκταθεί και σε διαφορετικά μοντέλα επιθέσεων. Χρήσιμη είναι η επέκτασή του έτσι ώστε να προστατεύει τα δεδομένα από τη διεξαγωγή συμπερασμάτων μέσω της παράλληλης χρήσης περισσοτέρων από μία συναθροιστικών συναρτήσεων. Στην περίπτωση αυτή, η ιδιωτικότητα των ατόμων που συμμετέχουν θα μπορεί να διασφαλιστεί από παραβιάσεις όπου ο επιτιθέμενος έχει περισσότερη γνώση πάνω στα δεδομένα είτε από διαφορετικές παράλληλες επιθέσεις της μορφής της αρχικής επίθεσης. Ένα παράδειγμα στο οποίο πιθανώς να μπορεί να δώσει λύση κάποια επέκταση του αλγορίθμου θα ήταν σε σύνολα δεδομένων προσωπικών εισοδημάτων, όπου θα προστατεύει από την πιθανή γνώση του συνολικού αθροίσματος των τιμών μιας εγγραφής ταυτόχρονα με την προστασία από επίθεση με γνώση του μέγιστου εισοδήματός της.

Για μια ακόμη πιθανή επέκταση του αλγορίθμου θα μπορούσε να θεωρηθεί διαφορετικό μοντέλο δεδομένων τέτοιο ώστε ένα ή περισσότερα γνωρίσματα του ψευδο-αναγνωριστικού να είναι ευαίσθητα και συνεπώς να απαιτείται η προστασία από επιθέσεις με στόχο την αναγνώριση των μεμονωμένων τιμών της εγγραφής. Σε αυτήν την περίπτωση μπορεί να επιχειρηθεί η ικανοποίηση της *l*-διαφορετικότητας για τα ευαίσθητα γνωρίσματα σε συνδυασμό με την *k*-ανωνυμία.

Τέλος ο αλγόριθμος μπορεί να επεκταθεί ώστε να καλύπτει σύνολα δεδομένων προερχόμενα από μοντέλο δεδομένων που έχει τα γνωρίσματα του ψευδο-αναγνωριστικού από το ίδιο

πεδίο τιμών και ένα ευαίσθητο γνώρισμα που αφορά κάποια άλλη μορφή πληροφορίας και όχι αντίστοιχη των υπολοίπων γνωρισμάτων. Στην περίπτωση αυτή, ο αλγόριθμος μπορεί να λειτουργήσει ώστε να ικανοποιεί την *k*-ανωνυμία ως προς την επιλεγμένη συναθροιστική συνάρτηση για τα γνωρίσματα του ψευδο-αναγνωριστικού και να ικανοποιεί την *l*-διαφορετικότητα για τις τιμές του ευαίσθητου γνωρίσματος.

8

Βιβλιογραφία

[Swe02] L. Sweeney. *k*-Anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Volume 10, no. 5, 2002

[MGK+06] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam. *l*-Diversity: Privacy Beyond *k*-Anonymity, In Proc. Intl. Conference on Data Engineering, 2006

[NC06] M. Nergiz, C. Clifton. Thought on *k*-Anonymization, In Proc. Intl. Conference on Data Engineering Workshops, 2006

[LLV07] N. Li, T. Li, S. Venkatasubramanian. *t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity, In Proc. Intl. Conference on Data Engineering, 2007

[XT07] X. Xiao, Y. Tao. *m*-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets, In Proc. Special Interest Group on Management of Data , 2007

[NAC07] M.E. Nergiz, M. Atzori, C. Clifton. Hiding the Presence of Individuals from Shared Databases, In Proc. Special Interest Group on Management of Data, 2007

[LDR05] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Incognito: Efficient Full-Domain *k*-Anonymity, In Proc. Special Interest Group on Management of Data, 2005

[LDR06] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Mondrian Multidimensional *k*-Anonymity, In Proc. Intl. Conference on Data Engineering, 2006

[XT06] X. Xiao, Y. Tao. Anatomy: Simple and Effective Privacy Preservation, In Proc. Very Large Data Bases, 2006

[TM08] M. Terrovitis, N. Mamoulis, Privacy-preserving Anonymization of Set-valued Data, Very Large Data Bases,2008

[XWP+06] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A. Fu, Utility-Based Anonymization Using Local Recoding, KDD, 2006

[1] "Uci Repository",

http://archive.ics.uci.edu/ml/datasets/IPUMS+Census+Database .

[2] "UTD Anonymization Toolbox", http://cs.utdallas.edu/dspl/cgi-bin/toolbox/.