ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΊΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ (ΠΑΤΡΑ)

ΠΡΟΣΤΑΣΙΑ ΙΔΙΩΤΙΚΟΤΗΤΑΣ ΕΥΑΙΣΘΗΤΩΝ ΔΕΔΟΜΕΝΩΝ: ΤΟ ΜΟΝΤΕΛΟ ΤΗΣ κ-ΑΝΩΝΥΜΙΑΣ

PRIVACY PROTECTION OF SENSITIVE DATA: THE κ-ANONYMITY MODEL

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΤΩΝ:
ΛΑΓΟΥΔΑΚΗ ΕΛΙΣΣΑΒΕΤ (ΑΜ 12688)
ΛΩΛΟΥ ΙΩΑΝΝΑ (ΑΜ 12764)

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΔΡ. ΗΛΙΑΣ Κ. ΣΤΑΥΡΟΠΟΥΛΟΣ

ΠΑΤΡΑ, ΙΟΥΝΙΟΣ 2015

ΕΥΧΑΡΙΣΤΙΕΣ

Θα θέλαμε να ευχαριστήσουμε τον επιβλέποντα καθηγητή κ. Ηλία Σταυρόπουλο για την ευκαιρία και την εμπιστοσύνη που μας έδειξε καθώς και την δυνατότητα που μας έδωσε να ασχοληθούμε με ένα τόσο πρωτοποριακό και ενδιαφέρον θέμα στα πλαίσια της πτυχιακής μας εργασίας αλλά και να αποκομίσουμε ουσιαστικά προσόντα μέσα από αυτήν.

ΠΕΡΙΛΗΨΗ

Πολλοί οργανισμοί ιδιωτικοί ή δημόσιοι φορείς συλλέγουν και διαχειρίζονται προσωπικά δεδομένα, διαφόρων χρηστών, τα οποία μπορούν να δημοσιευτούν για ερευνητικούς σκοπούς και στατιστικές μελέτες. Ο σκοπός της εργασίας είναι η χρήση και η ανάπτυξη του προστατευτικού μοντέλου k-ανωνυμία για την διασφάλιση της ιδιωτικότητας των προσωπικών δεδομένων.

Με τον τρόπο δημοσίευσης και την χρήση των δεδομένων ενέχει ο κίνδυνος ευαίσθητων προσωπικών στοιχείων ή και αναγνώρισης κάποιων ατόμων που εμφανίζονται στις συλλογές δεδομένων.

Για την προστασία της ιδιωτικής ζωής, έχουν προταθεί διάφορες τεχνικές οι οποίες υλοποιούν αλγορίθμους ανωνυμοποίησης στα δεδομένα. Μέσω της εφαρμογής, προσφέρεται η δυνατότητα χρήσης αρκετών από τους πλέον διαδεδομένους αλγορίθμους ανωνυμοποίησης δεδομένων, με στόχο την προστασία των προσωπικών πληροφοριών των ατόμων σε προς δημοσίευση σύνολα δεδομένων. Ταυτόχρονα, παρέχονται πληροφορίες για το ποσοστό γενίκευσης τιμών, τον χρόνο εκτέλεσης του αλγορίθμου και την απώλεια πληροφορίας που προκύπτει λόγω της ανωνυμοποίησης, δίνοντας έτσι την δυνατότητα σύγκρισης και επιλογής του κατάλληλου αλγορίθμου για την κάθε περίσταση.

Λέξεις κλειδιά: Ανωνυμία δεδομένων, Ιδιωτικότητα

ABSTRACT

Many private or public organizations collect and manage private data of various users that can be published for scientific research and statistical studies. The purpose of this paper is to present the use and the development of the k-anonymity protecting model in privacy protection.

The way that these data could be published poses the risk of disclosing sensitive personal data or identifying the persons who are related to the data-collection.

There have been proposed many techniques, for the privacy protection that implement algorithms specific for data anonymity. Via this application it is provided the ability of using common algorithms that brings privacy protection. That aims to protect person-specific information from been published in datasets. At the same time there are provided information for the percentage of the generalization of the values, the time that the algorithms are carried out, and the information loss that comes from the anonymization. That grants the ability of comparing and choosing the most appropriate algorithm for each occasion.

Keywords: Data anonymity, Privacy

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

1 ΕΙΣΑΓΩΓΗ

1.1 ΙΔΙΩΤΙΚΟΤΗΤΑ ΚΑΙ ΑΝΩΝΥΜΙΑ ΔΗΜΟΣΙΕΥΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ

Τα τελευταία χρόνια, έχει σημειωθεί εκθετική αύξηση του όγκου των πληροφοριών που είναι διαθέσιμες στο ευρύ κοινό, καθώς η ιλιγγιώδες ταχύτητα του Διαδικτύου αλλά και ο αποθηκευτικός χώρος γίνονται ολοένα και πιο διαθέσιμα. Το γεγονός αυτό, σε συνδυασμό με την προσωπική φύση του μεγαλύτερου όγκου αυτής της πληροφορίας, έχουν οδηγήσει τόσο στην ανάπτυξη νομοθεσίας, όσο και τεχνικών "ανωνυμοποίησης" των δεδομένων προς δημοσίευση με σκοπό την προστασία της ταυτότητας του ατόμου, αλλά και πιθανών "ευαίσθητων" γνωρισμάτων.

Η διαθεσιμότητα όμως των αναλυτικών προσωπικών δεδομένων θέτει σημαντικούς κινδύνους στην ιδιωτικότητα του κάθε ατόμου. Ακόμη και με την απόκρυψη στοιχείων που προσδιορίζουν μοναδικά ένα άτομο όπως είναι για παράδειγμα το ονοματεπώνυμο ή ο Αριθμός Φορολογικού Μητρώου (ΑΦΜ), είτε ο συνδυασμός άλλων στοιχείων όπως ο ταχυδρομικός κώδικας, το φύλο και η ηλικία του, θα μπορούσαν να λειτουργήσουν σαν ψευδό-αναγνωριστικά και να οδηγήσουν στην ταυτοποίηση του ατόμου αποκαλύπτοντας ευαίσθητα προσωπικά δεδομένα.

Το χαρακτηριστικότερο παράδειγμα όπου περιγράφει η καθηγήτρια του πανεπιστημίου του Harvard, Sweeney, είναι η σύνδεση του κυβερνήτη της Μασαχουσέτης William Weld, που ζούσε εκείνη την εποχή στο Cambridge με τα ιατρικά του στοιχεία, χρησιμοποιώντας την λίστα 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 των υποψηφίων σε συνδυασμό με τις φαινομενικά ανωνυμοποιημένες πληροφορίες του οργανισμού GIC ο οποίος είναι υπεύθυνος για την ιατρική ασφάλιση των εργαζομένων της πολιτείας, δύο δημοσιευμένα και φαινομενικά ανεξάρτητα σύνολα δεδομένων.

Προκειμένου να γίνει πιο κατανοητή η μελέτη του προβλήματος της προστασίας της ιδιωτικότητας κρίνεται σκόπιμο να περιγράψουμε ένα ακόμα παράδειγμα όπως φαίνεται στον πίνακα 1. Θεωρούμε ότι υπάρχει ένας ιατρικός οργανισμός ο οποίος δημοσιεύει τα ιατρικά δεδομένα των ασθενών του και ο οποίος αποκρύπτει χαρακτηριστικά του ατόμου τα οποία τον προσδιορίζουν άμεσα. Επιπροσθέτως, υπάρχει και ένας άλλος πίνακας με τα εκλογικά δεδομένα μιας περιοχής που αποτελείται από ένα σύνολο κοινών γνωρισμάτων τα οποία είναι ημερομηνία γεννήσεως, φύλο, ταχυδρομικός κώδικας. Από την σύνδεση των δυο πινάκων μπορεί εύκολα να προσδιοριστεί ότι ο Παύλος εισήχθηκε στο νοσοκομείο με

κάταγμα χεριού. Ο λόγος που προκύπτει αυτό το συμπέρασμα είναι γιατί υπάρχει μόνο μια εγγραφή με ημερομηνία γεννήσεως 1978, ταχυδρομικό κώδικα 26331 και φύλο άνδρας.

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.

ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ			
Ημερομηνία Γεννήσεως	Φύλο	Ταχυδρομικός Κώδικας	Ασθένεια
1978	Άνδρας	26331	Κάταγμα χεριού
1978	Άνδρας	26331	Πυρετός
1991	Άνδρας	22810	Γρίπη
1991	Γυναίκα	30002	Πόνος στη πλάτη
1985	Γυναίκα	26332	Εμετός
1988	Άνδρας	22810	Γρίπη

ΕΚΛΟΓΙΚΑ ΔΕΔΟΜΕΝΑ			
Όνομα	Ημερομηνία Γεννήσεως	Φύλο	Ταχυδρομικός Κώδικας
Παύλος	1978	Άνδρας	26331
Γιάννης	1986	Άνδρας	26332
Κώστας	1990	Άνδρας	22810
Ειρήνη	1980	Γυναίκα	30002
Πόπη	1993	Γυναίκα	30002
Ηρώ	1989	Γυναίκα	26331

Πίνακας 1: Στοιχεία Ιατρικού Οργανισμού και Εκλογικού Καταλόγου

Έτσι για την αποφυγή της παράνομης επεξεργασίας ή της δημοσίευσης προσωπικών αλλά και ευαίσθητων δεδομένων της ανθρώπινης δραστηριότητας αναπτύχθηκαν διάφοροι αλγόριθμοι και τεχνικές ανωνυμοποίησης, οι οποίοι αφορούν αναγνωριστική πληροφορία από το σύνολο δεδομένων που είναι διαθέσιμο έτσι ώστε να μην μπορεί ο επιτιθέμενος να προσδιορίσει ένα άτομο.

Μία τεχνική που χρησιμοποιείται κατά κόρον στα μοντέλα ανωνυμοποίησης είναι αυτή της γενίκευσης (generalization) που ορίζεται ως η διαδικασία κατά την οποία η αρχική τιμή που εμφανίζεται στο πίνακα, αντικαθίστανται με μία πιο γενική τιμή. Το σύνολο των ενδεχόμενων γενικεύσεων των τιμών κάθε ψευδό-αναγνωριστικού γνωρίσματος σχηματίζουν την ιεραρχία γενίκευσης. Ο στόχος της τεχνικής της γενίκευσης είναι να αποκρύψει ένα μέρος της χρήσιμης πληροφορίας που εμφανίζεται στα αρχικά δεδομένα. Για το λόγο αυτό, αναζητούνται εγγυήσεις ιδιωτικότητας οι οποίες δεν επιτρέπουν τη εξαγωγή προσωπικών δεδομένων και παράλληλα αφαιρούν όσο το δυνατό λιγότερη πληροφορία από τα αρχικά δεδομένα διατηρώντας την χρηστικότητα τους για τα άτομα που θέλουν να τα αξιοποιήσουν. Το ισχυρότερο μοντέλο ιδιωτικότητας που αποτρέπει σε ικανοποιητικό βαθμό την αναγνώριση κάποιου ατόμου μοναδικά είναι το μοντέλο της k-ανωνυμίας. Σκοπός του μοντέλου αυτού είναι να καταστήσει κάθε εγγραφή αδιάκριτη ανάμεσα σε άλλες k-1 εγγραφές.

Ωστόσο, επειδή υπάρχει πιθανότητα η πληροφορία που έχει στη διάθεση του ένας κακόβουλος τρίτος να έχει πολλαπλές μορφές και να προέρχεται από διαφορετικές δημοσιεύσεις σε πηγές, έχουν αναπτυχθεί διάφορες παραλλαγές της k-ανωνυμίας που εκμεταλλεύονται κάθε φορά τη μορφή της γνώσης του επιτιθέμενου. Μία από αυτές είναι και η k^m-ανωνυμία, η οποία εγγυάται ότι η βάση είναι k-ανώνυμη ακόμα και αν ο επιτιθέμενος έχει ως γνωστικό υπόβαθρο ένα σύνολο m τιμών από το σύνολο του ψευδό-αναγνωριστικού μιας εγγραφής, και προσπαθεί να εντοπίσει την εγγραφή αυτή στα δημοσιευόμενα δεδομένα.

1.2 ΑΝΤΙΚΕΙΜΕΝΟ ΠΤΥΧΙΑΚΗΣ

Η παρούσα εργασία ασχολείται με το πρόβλημα της προστασίας της ιδιωτικότητας των ατόμων που εμφανίζονται σε δημοσιευμένες βάσης δεδομένων. Κατά την δημοσίευση των δεδομένων, ο συνδυασμός των τιμών κάποιων γνωρισμάτων μπορεί να λειτουργήσει ως ψευδό-αναγνωριστικό και να έχει ως αποτέλεσμα την αναγνώριση της ταυτότητας κάποιου ατόμου.

Γενικά, αναλύεται το σενάριο οπού ο επιτιθέμενος έχει κάποιες πληροφορίες στη κατοχή του και προσπαθεί με τη χρήση αυτών των γνώσεων να αναγνωρίσει την ταυτότητα κάποιας εγγραφής που συμμετέχει στα δεδομένα. Παράλληλα όμως, επιχειρείται κάθε απαραίτητη τροποποίηση στα δεδομένα με στόχο την αποτροπή της ταυτοποίησης της εγγραφής.

Για την επίλυση αυτού του προβλήματος, προτάθηκε από την L.Sweeney το μοντέλο της k-ανωνυμίας. Με τη χρήση αυτής της μεθόδου εξασφαλίζεται ότι ο επιτιθέμενος δεν θα μπορεί να προσδιορίσει μοναδικά μια εγγραφή στη βάση δεδομένων, αφού υπάρχουν άλλες k-1 εγγραφές με τις ίδιες τιμές. Εντούτοις, θα πρέπει να σημειωθεί ότι το μοντέλο αυτό έχει ένα αρνητικό στοιχείο, δηλαδή προκαλεί πολλές φορές υπεργενίκευση στα δεδομένα με αποτέλεσμα να προκύπτει μεγάλη απώλεια χρήσιμης πληροφορίας.

Τελικά, μετά από μελέτη που έγινε, η καλύτερη προτεινόμενη λύση για την διασφάλιση της ιδιωτικότητας των εγγραφών στο πρόβλημα αυτό, είναι η k^m ανωνυμοποίηση των δεδομένων, με χρήση του αλγορίθμου Mondrian. Η τροποποίηση των δεδομένων χρησιμοποιώντας αυτό το μοντέλο αποτρέπει τον επιτιθέμενο να προσδιορίσει μοναδικά κάποια εγγραφή, για οποιοδήποτε σύνολο μερικής γνώσης που κατέχει πάνω στα δημοσιευμένα δεδομένα, όπου αυτό συνεπάγεται και λιγότερη απώλεια πληροφορίας.

Συνοπτικά μπορούμε να πούμε ότι ο αλγόριθμος Datafly γενικεύει συνεχώς τα ψευδό-αναγνωριστικά (QI) δημιουργώντας buckets παρόμοιας συχνότητας. Με αυτό τον τρόπο οι τιμές γίνονται όλο και πιο γενικές με την ολοκλήρωση κάθε επανάληψης.

1.3 ΟΡΓΑΝΩΣΗ ΚΕΙΜΕΝΟΥ

Η παρούσα πτυχιακή εργασία παρουσιάζεται σύμφωνα με τα παρακάτω κεφάλαια:

Σε αυτό το **πρώτο** κεφάλαιο, περιγράψαμε το πρόβλημα που αναφέρεται στην εργασία αυτή, αναφέραμε τη σημασία του προβλήματος, ενώ ταυτόχρονα κάναμε και μια πρώτη περιγραφή του τρόπου με τον οποίο επιλύσαμε το πρόβλημα.

Στο **δεύτερο** κεφάλαιο αναλύεται η βιβλιογραφία που μελετήθηκε και αφορά το μοντέλο της προστασίας της ιδιωτικότητας σε σχεσιακές βάσεις δεδομένων προσωπικής πληροφορίας και τους σχετικούς ορισμούς για την καλύτερη κατανόηση και μελέτη του προβλήματος της προστασίας των ευαίσθητων προσωπικών δεδομένων καθώς και τους αλγόριθμους εύρεσης ανωνυμοποιημένων πινάκων.

Στο **τρίτο** κεφάλαιο ορίζονται οι τεχνικές ανωνυμοποίησης πινάκων που στόχο έχουν την προστασία ιδιωτικότητας από επιθέσεις με μερική γνώση σε συλλογές με συνεχή γνωρίσματα, και αναλύονται οι πιθανές λύσεις που προτείνονται για την επίλυση του προβλήματος.

Στο **τέταρτο** κεφάλαιο περιγράφονται και αναλύονται οι επιθέσεις ενάντια στην k-ανωνυμία, όπου ένας κακόβουλος τρίτος προσπαθεί να ανακαλύψει προσωπικές πληροφορίες ενός ατόμου από έναν δημοσιευμένο πίνακα, ο οποίος ικανοποιεί την k-ανωνυμία, χρησιμοποιώντας τις τεχνικές ανωνυμοποίησης πινάκων.

Στο **πέμπτο** κεφάλαιο αναλύονται οι μέθοδοι αντιμετώπισης επιθέσεων όπου μια από τις μεθόδους είναι επέκταση της k –ανωνυμίας, έχουν στόχο την αποτροπή των επιθέσεων αναγνώρισης της τιμής του ευαίσθητου γνωρίσματος αλλά και την διασφάλιση ότι ο επιτιθέμενος δεν θα μπορέσει να βρει προσωπικά στοιχεία για ένα φυσικό πρόσωπο.

Στο **έκτο** κεφάλαιο μελετώνται και αναλύονται αλγόριθμοι οι οποίοι επιχειρούν την ικανοποίηση της k-ανωνυμίας δημοσιευμένων συλλογών δεδομένων καθώς και συστήματα πραγματικού κόσμου όπου επιδιώκουν να παρέχουν προστασία της k-ανωνυμίας χρησιμοποιώντας γενίκευση και συμπίεση. Εφαρμόζουν τοπική γενίκευση στις τιμές των γνωρισμάτων των εγγραφών, υπολογίζοντας την ζητούμενη διαμέριση κάθε γνωρίσματος ξεχωριστά.

Στο **έβδομο** κεφάλαιο συνοψίζονται τα αποτελέσματα της πτυχιακής εργασίας σχετικά με την βέλτιστη λύση του προβλήματος της ιδιωτικότητας από επιτιθέμενους με μερική γνώση, σε συλλογές δεδομένων με συνεχή γνωρίσματα και προτείνονται μελλοντικές επεκτάσεις.

2 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Το πρόβλημα της προστασίας ιδιωτικότητας ευαίσθητων δεδομένων, είναι ένα ανοικτό ζήτημα στο χώρο της επιστήμης των υπολογιστών. Διαρκώς όλο και περισσότερα δεδομένα, τα οποία αφορούν ευαίσθητες πληροφορίες, δημοσιεύονται για στατιστικούς και ερευνητικούς σκοπούς. Συνεπώς, είτε αυτές οι προσωπικές πληροφορίες δημοσιεύονται σε ιστοσελίδες, είτε χρησιμοποιούνται για δημογραφικές έρευνες και στατιστικές αναλύσεις, το ζήτημα σχετικά με την προστασία της ιδιωτικότητας αποτελεί ενδιαφέρον για πολλούς επιστήμονες.

Η παρούσα εργασία, 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 τον 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.

2.1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΠΡΟΣΤΑΣΙΑ ΤΗΣ ΙΔΙΩΤΙΚΟΤΗΤΑΣ

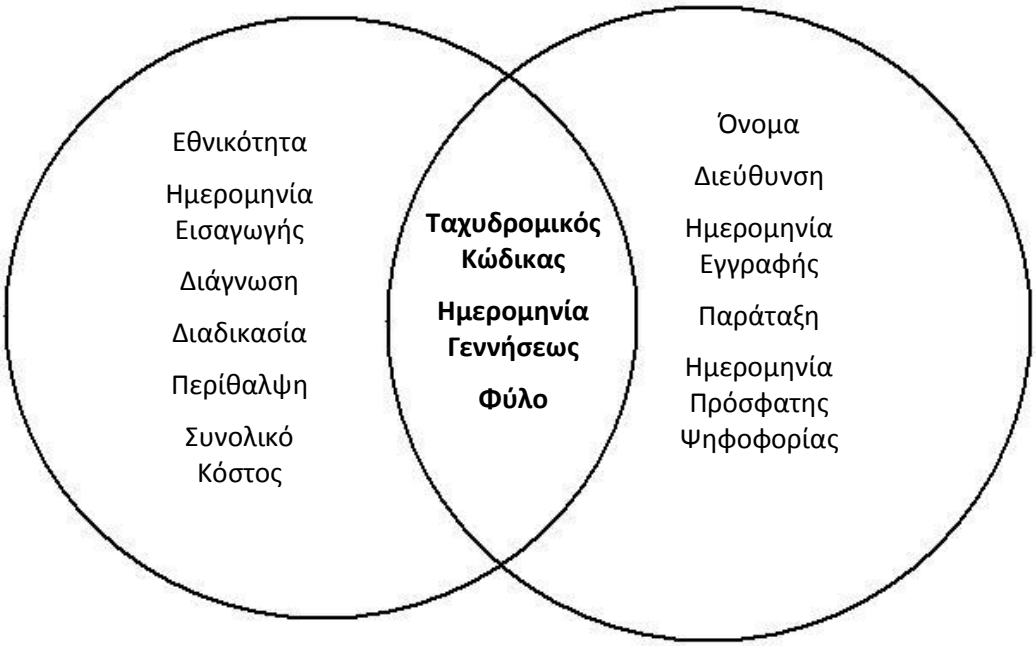
Η εκθετική αύξηση του αριθμού και της ποικιλίας των δεδομένων, που βιώνει η κοινωνία, διατηρούνται και αποθηκεύονται σε διάφορες βάσεις δεδομένων που αφορούν συγκεκριμένα άτομα είναι πλέον γεγονός. Καθημερινά, ιδιωτικοί και δημόσιοι φορείς συλλέγουν όλο και περισσότερες πληροφορίες για διάφορους λόγους. Η τεχνολογία των υπολογιστών σε συνδυασμό με την συνδεσιμότητα των δικτύων τον φθηνό και προσιτό αποθηκευτικό χώρο διευκολύνουν ακόμα περισσότερο την οργάνωση όλης αυτής της πληροφορίας σε διάφορα ψηφιακά μέσα. Δεν είναι λίγες οι εφαρμογές που εξυπηρετούν διάφορες κοινωνικές ανάγκες και έμμεσα διαδραματίζεται ακόμα πιο έντονα η συλλογή πληροφορίας που αφορούν ατομικές συνήθειες, ενέργειες ή ακόμα και χαρακτηριστικά. Όλες οι προαναφερθείσες ενέργειες πραγματοποιούνται στο πλαίσιο της προστασίας των προσωπικών δεδομένων και της ιδιωτικότητας, αλλά πολλές φορές κακόβουλοι χρήστες μπορούν να εκμεταλλευτούν αυτή την πληθώρα πηλοφοριών για την ταυτοποίηση των ατόμων.

Έτσι, μια κοινή πρακτική για τους οργανισμούς είναι να εκδίδουν δεδομένα που αφορούν συγκεκριμένα άτομα χωρίς την παρουσία χαρακτηριστικών αναγνωριστικών όπως

όνομα, διεύθυνση και τηλεφωνικός αριθμός. Αυτή η ενέργεια εξασφαλίζει την ανωνυμία των προσώπων που εμπλέκονται. Παρ' όλα αυτά στις περισσότερες περιπτώσεις είναι δυνατή η επανα-αναγνώριση μέσω διαφόρων τεχνικών από τα εναπομείναντα δεδομένα. Τέτοιες τεχνικές μπορεί να είναι η σύνδεση ή η αντιστοίχηση των δεδομένων με άλλα δεδομένα που έχουν ήδη εκδοθεί.

Σε μια έρευνα, (Sweeney, 2002) διεξήχθηκαν πειράματα χρησιμοποιώντας περιληπτικά δεδομένα της απογραφής των Η.Π.Α για το έτος 1990 προκειμένου να προσδιοριστεί πόσα άτομα μέσα σε γεωγραφικά εγκατεστημένους πληθυσμούς είχαν συνδυασμούς από δημογραφικές μετρήσεις που εμφανίζονταν σπάνια. Συνδυασμοί μερικών χαρακτηριστικών συχνά συνδέονται με πληθυσμούς για να αναγνωρίζονται μοναδικά ή σχεδόν μοναδικά μερικά άτομα. Για παράδειγμα ένα αποτέλεσμα της έρευνας ήταν πως το 87% (216 εκατ. από τα 248 εκατ.) του πληθυσμού των Η.Π.Α είχαν αναφέρει χαρακτηριστικά που πιθανώς τους έκαναν μοναδικούς βασισμένα μόνο στον πενταψήφιο Τ.Κ., το φύλλο και την ημερομηνία γέννησης. Αυτό μας οδηγεί στο συμπέρασμα πως εκδόσεις δεδομένων που περιέχουν τέτοιες πληροφορίες δεν θα πρέπει να θεωρούνται ανώνυμες. Παρ' όλα αυτά, ιατρικές πληροφορίες και άλλα δεδομένα που αφορούν σε συγκεκριμένα άτομα είναι συχνά διαθέσιμα μη τηρώντας τον προηγούμενο περιορισμό.

Στη Μασαχουσέτη, η επιτροπή κοινωνικής ασφάλισης συνέλλεξε χιλιάδες δεδομένα που αφορούν συγκεκριμένους ασθενείς και τις οικογένειες τους αφού οι ασθενείς πίστεψαν ότι τα δεδομένα ήταν ανώνυμα. Η επιτροπή ωστόσο έδωσε ένα αντίγραφο των δεδομένων στους ερευνητές και πούλησε ένα αντίγραφο σε μια βιομηχανική εταιρία. Παράλληλα, για 20 δολάρια αγοράστηκε η λίστα των ψηφοφόρων της Μασαχουσέτης για το Cambridge η οποία στάλθηκε σε 20 δισκέτες. Η λίστα περιλάμβανε το όνομα, τη διεύθυνση, το φύλο, τον Τ.Κ και ημερομηνία γέννησης κάθε ψηφοφόρου. Αυτή η λίστα όμως μπορεί να συνδεθεί με την λίστα των ιατρικών πληροφοριών μέσω Τ.Κ., ημερομηνία γέννησης και φύλου και άρα με στοιχεία που αφορούν διαγνώσεις, διαδικασίες και συνταγές φαρμάκων που αφορούν σε συγκεκριμένα άτομα. Για παράδειγμα, ο William Weld ήταν κυβερνήτης της Μασαχουσέτης εκείνη την περίοδο και τα ιατρικά του αρχεία ήταν μέσα στη λίστα της επιτροπής κοινωνικής ασφάλισης. Ο κυβερνήτης, ζούσε εκείνη την εποχή στο Cambridge της Μασαχουσέτης. Σύμφωνα με την λίστα των ψηφοφόρων 6 άτομα μοιράζονταν την ίδια ημερομηνία γέννησης με την δικά του εκ των οποίων μόλις τρείς ήταν άντρες κανείς από τους οποίους δεν μοιράζονταν τον ίδιο πενταψήφιο Τ.Κ.



Εικόνα 1: Συνδυασμός Δεδομένων από Διαφορετικά Σύνολα Εγγραφών

2.1.1 ΧΡΗΣΙΜΟΙ ΟΡΙΣΜΟΙ ΠΡΟΣΤΑΣΙΑΣ ΙΔΙΩΤΙΚΟΤΗΤΑΣ

Προκειμένου να γίνει πιο κατανοητή η μελέτη του προβλήματος της προστασίας ιδιωτικότητας κρίνεται σκόπιμο σε αυτό το σημείο να αναλυθούν κάποιοι ορισμοί που χρησιμοποιούνται συχνά στην επιστημονική κοινότητα (Jing , και συν., 2012) και αφορούν την σχεσιακή βάση δεδομένων.

1. **Ευαίσθητα γνωρίσματα (sensitive attributes)**

Ένα γνώρισμα κάποιου πίνακα οντοτήτων θα χαρακτηρίζεται ως ευαίσθητο αν δεν πρέπει να επιτραπεί σε έναν επιτιθέμενο να ανακαλύψει την τιμή του για οποιοδήποτε άτομο στο σύνολο των δεδομένων. Συνήθως, το ευαίσθητο γνώρισμα είναι μια προσωπική πληροφορία που θα πρέπει να μείνει κρυφή απέναντι σε κάθε κακόβουλο τρίτο που θα προσπαθήσει να την αποσπάσει. Η επιλογή των γνωρισμάτων που θα θεωρηθούν ευαίσθητα εξαρτάται από τις συνθήκες ιδιωτικότητας που πρέπει να διασφαλίζουν τα δημοσιευμένα δεδομένα αλλά και από την κρίση του κατόχου των δεδομένων. Έτσι, κάποιες φορές επιλέγονται κάποια γνωρίσματα για την κατηγορία αυτή τα οποία στην πραγματικότητα δεν είναι και τόσο ευαίσθητα αλλά θεώρησε ο ιδιοκτήτης της βάσης δεδομένων ότι μπορεί να οδηγήσουν στην αποκάλυψη πληροφορίας την οποία πρέπει να προστατέψει.

Στο παρακάτω παράδειγμα το ευαίσθητο γνώρισμα είναι οι {θρησκεύτηκες πεποιθήσεις}.

Ημερομηνία Γεννήσεως	Φύλο	Χώρα	Θρησκευτικές Πεποιθήσεις
1978	Άνδρας	Ελλάδα	Χριστιανισμός
1978	Άνδρας	Ρωσία	Αθρησκεία
1991	Άνδρας	Ινδία	Ινδουισμός
1991	Γυναίκα	Γαλλία	Αθρησκεία
1985	Γυναίκα	Ελλάδα	Χριστιανισμός
1988	Άνδρας	Ινδία	Ινδουισμός

Πίνακας 2: Παράδειγμα για ευαίσθητα δεδομένα

2. Προσωπικά Δεδομένα

Το σύνολο των προσωπικών πληροφοριών ενός ατόμου τα οποία το καθορίζουν. Τα δεδομένα αυτά συνήθως συγκεντρώνονται σε βάσεις δεδομένων, για να μπορεί να είναι πιο εύκολη η μαζική επεξεργασία και μεταφορά τους.

Παραδείγματα προσωπικών δεδομένων αποτελούν τα παρακάτω:

- το φύλο
- το επάγγελμά
- η ηλικία
- ο μισθός

3. Πίνακας (Table)

Τα δεδομένα που βρίσκονται αποθηκευμένα σε μια βάση, είναι οργανωμένα σε μορφή πίνακα RT (A1, A2,..,An) σχεσιακής βάσης δεδομένων όπου τα A1, A2,..,An είναι οι στήλες-γνωρίσματα του.

Στο παρακάτω παράδειγμα τα δεδομένα αφορούν ένα πίνακα με σύνολο γνωρισμάτων (Ηλικία, Φύλο, Ταχυδρομικός Κώδικας, Μισθός).

Ηλικία	Φύλο	Ταχυδρομικός Κώδικας	Μισθός
30	Άνδρας	26331	700
50	Άνδρας	26331	1000
24	Άνδρας	22810	750
21	Γυναίκα	30002	750
32	Γυναίκα	26332	900
44	Άνδρας	22810	1500

Πίνακας 3: Παράδειγμα πίνακα με τέσσερα γνωρίσματα

4. Πλειάδα (Tuple)

Ένα σύνολο τιμών στον πίνακα σχεσιακής βάσης δεδομένων. Πρόκειται για μια εγγραφή η οποία αφορά ένα άτομο και τις τιμές του στα αντίστοιχα πεδία πληροφορίας.

Για παράδειγμα έστω η γραμμή ενός πίνακα που έχει ένα σύνολο τιμών (1976,Γυναίκα,53712,πόνος στη πλάτη) αναφέρεται σε ένα συγκεκριμένο άτομο.

ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ			
Ημερομηνία Γεννήσεως	Φύλο	Ταχυδρομικός Κώδικας	Ασθένεια
1978	Άνδρας	26331	Κάταγμα χεριού
1978	Άνδρας	26331	Πυρετός
1991	Άνδρας	22810	Γρίπη
1976	Γυναίκα	53712	Πόνος στη πλάτη
1985	Γυναίκα	26332	Εμετός
1988	Άνδρας	22810	Γρίπη

Πίνακας 4: Παράδειγμα πλειάδας

5. Στήλη – γνώρισμα (Attribute)

Κάθε στήλη του πίνακα σχεσιακής βάσης δεδομένων αναφέρεται σε ένα ξεχωριστό γνώρισμα, που αντιπροσωπεύει μια κατηγορία πληροφορίας και έχει ένα σύνολο πιθανών τιμών.

Για παράδειγμα η στήλη-γνώρισμα {τηλέφωνα} έχει πεδίο τιμών τους αριθμούς των τηλεφώνων της περιοχής.

Ημερομηνία Γεννήσεως	Φύλο	Ταχυδρομικός Κώδικας	Τηλέφωνα
1978	Άνδρας	26331	6982529186
1978	Άνδρας	26331	6982525333
1991	Άνδρας	22810	6976565200
1991	Γυναίκα	30002	6970001147
1985	Γυναίκα	26332	6980007891
1988	Άνδρας	22810	6974566211

Πίνακας 5: Παράδειγμα για στήλη- γνώρισμα

6. Ιδιότητες κλειδιά(Key Attribute)

Γνωρίσματα κάποιου πίνακα οντοτήτων τα οποία δεν πρέπει να δημοσιευθούν γιατί καθορίζουν άμεσα κάποιο φυσικό πρόσωπο.

Για παράδειγμα ο Αριθμός Δελτίου Ταυτότητας ή ο Αριθμός Μητρώου Κοινωνικής Ασφάλισης.

7. Ψευδό- αναγνωριστικό (quasi-identifier)

Ως ψευδό-αναγνωριστικό (quasi-identifier) ορίζεται το ελάχιστο σύνολο γνωρισμάτων $QI=A_1,A_2,....A_d$ με το οποίο ένας πίνακας PT μπορεί να διασταυρωθεί με κάποιες εξωτερικές πληροφορίες και να αναγνωριστεί άμεσα η ταυτότητα κάποιας εγγραφής.

Έστω ένας πληθυσμός οντοτήτων U, ένας πίνακας οντοτήτων $T(A_1,\ldots,A_n)$, μία συνάρτηση f_c: $U \rightarrow T$ και μία συνάρτηση f_g: $T \rightarrow U'$, με $U \subseteq U'$. Ένα ψευδο-αναγνωριστικό του T, συμβολίζεται με Q_T και είναι ένα σύνολο γνωρισμάτων $\{A_i,\ldots,A_j\} \subseteq \{A_1,\ldots,A_n\}$ για το οποίο ισχύει ότι: $\exists p_i \in U : f_g(f_c(p_i)[Q_T]) = p_i$.

Το σύνολο των γνωρισμάτων που αποτελούν το ψευδό-αναγνωριστικό καθορίζεται από τον κάτοχο των δεδομένων, με αναζήτηση σε εξωτερικούς καταλόγους στους οποίους εμφανίζονται γνωρίσματα που περιλαμβάνονται στα σύνολα δεδομένων που έχουν δημοσιευθεί. Τα γνωρίσματα του πίνακα που θα χρησιμοποιηθούν στην πράξη ως ψευδό-αναγνωριστικά, επιλέγονται 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.

Συνήθως, τα γνωρίσματα αυτά δεν είναι τόσο επιβλαβή για το άτομο που περιγράφουν και μπορούν να γίνουν γνωστά από εξωτερικές πηγές δημόσιες ή ακόμα και ιδιωτικές. Ωστόσο, όταν συνδυαστούν με άλλα ψευδό-αναγνωριστικά μπορούν να οδηγήσουν σε ταυτοποίηση του ατόμου. Παράδειγμα ψευδό-αναγνωριστικού μπορεί να θεωρηθεί ο Ταχυδρομικός Κώδικας.

2.2 ΜΟΝΤΕΛΟ ΠΡΟΣΤΑΣΙΑΣ Κ-ΑΝΩΝΥΜΙΑΣ

Οι συλλογές δεδομένων, συχνά περιέχουν πληροφορίες που δεν πρέπει να δημοσιευτούν. Οι χρήστες των βάσεων δεδομένων έχουν την ευθύνη να δημοσιεύουν πληροφορίες χωρίς να θέτουν σε κίνδυνο την προστασία της ιδιωτικότητας, την εμπιστευτικότητα ή ακόμα και γνώσεις εθνικών συμφερόντων. Λειτουργώντας αυτόνομα και συχνά χωρίς ιδιαίτερες γνώσεις, είναι δύσκολο να το καταφέρουν.

Σε πολλές περιπτώσεις, η ύπαρξη της προστασίας της βάσης δεδομένων εξαρτάται από την ικανότητα των χρηστών να δημοσιεύουν ανώνυμα δεδομένα, καθώς οφείλουν να παρέχουν τέτοιο όγκο πληροφοριών που να διατηρεί την χρησιμότητά τους. Σε αντίθετη περίπτωση, μια αποτυχία προστασίας των δεδομένων μπορεί να επιφέρει συνθήκες που βλάπτουν το κοινό.

Η προστασίας ιδιωτικότητας θέτει ως στόχο να μειώσει την πιθανότητα να προσδιοριστεί μοναδικά μια συγκεκριμένη οντότητα, ακόμα και με τη διασταύρωση δημοσιευμένων εγγραφών που μπορεί να είναι ανωνυμοποιημένες. Για να γίνει αυτό, πρέπει να υπάρχει μια μέγιστη πιθανότητα το πολύ k, ώστε ένας επιτιθέμενος να μπορεί να ανακαλύψει με κάποια σύνδεση πινάκων, σε ποιο άτομο ανήκει μια εγγραφή.

Βάσει του ορισμού που είχε δοθεί από τη L. Sweeney το 2002, ένας προς δημοσίευση πίνακας *RT* $(A_1,A_2,\ldots,A_n)$ με ψευδό-αναγνωριστικό *QIRT* =(*Ai*,…, *Aj*) το σύνολο των γνωρισμάτων $A_1,A_2,\ldots,A_n$, ικανοποιεί την *k-ανωνυμία (k-anonymity)* αν κάθε ακολουθία τιμών στον πίνακα RT[*QIRT*] του ψευδό-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 ψευδό-αναγνωριστικού θα υπάρχουν το λιγότερο *k* εγγραφές που θα τον περιέχουν.

Τέλος, ο στόχος της ανωνυμοποίησης ενός πίνακα Τ είναι η παραγωγή μιας όψης V του πίνακα Τ, η οποία θα μετασχηματίζει τα δεδομένα του Τ, έτσι ώστε η V να είναι k-ανώνυμη με βάση το ψευδό-αναγνωριστικό. Επιπλέον, ο σκοπός της k-ανωνυμίας είναι να καταστήσει κάθε μια εγγραφή αδιάκριτη ανάμεσα σε άλλες k-1.

2.2.1 K-ANONYMITY

Έστω ένα σύνολο δεδομένων D. Τα δεδομένα είναι ορισμένα κατά γραμμές και στήλες έτσι ώστε κάθε γραμμή αφορά διαφορετικό άτομο και κάθε στήλη αφορά διαφορετική ιδιότητα του ατόμου. Όπως αναφέρεται και από την μελέτη του (Γιαννακόπουλος, 2012), το σύνολο των δεδομένων D το αποκαλούμε και πίνακα δεδομένων και κάθε γραμμή του πίνακα θα την αποκαλούμε στο εξής πλειάδα (tuple). Για κάθε πίνακα D ορίζουμε:

Ορισμός 1: Έστω D(A1, A2, ..., An) πίνακας δεδομένων με πεπερασμένο αριθμό από πλειάδα. Το πεπερασμένο σύνολο {A1, A2, ..., An} είναι το σύνολο των χαρακτηριστικών του πίνακα D. Για παράδειγμα, έστω ο πίνακας δεδομένων που φαίνεται στον Πίνακα 2., στον οποίο απεικονίζονται ιατρικά δεδομένα κάποιων ασθενών. Σε αυτόν τον πίνακα, με βάση τον παραπάνω ορισμό το σύνολο των χαρακτηριστικών είναι το σύνολο {Ηλικία, Φύλο, Τ.Κ., Ασθένεια} ενώ οι πλειάδες είναι οι γραμμές του πίνακα. Όπως είναι σαφές από τον παραπάνω ορισμό και τον παραπάνω πίνακα, κάθε χαρακτηριστικό του πίνακα είναι μοναδικό, αφού αποδίδει μια σημασιολογική ιδιότητα σε κάθε πλειάδα του πίνακα. Πέρα από την ιδιότητα, κάθε χαρακτηριστικό συνδέεται άμεσα και με ένα πεδίο τιμών. Για παράδειγμα, το χαρακτηριστικό {Φύλο} του πίνακα 2, έχει ως πεδίο τιμών το σύνολο {Άνδρας, Γυναίκα}. Αντίθετα, κάθε πλειάδα στον πίνακα δεν είναι υποχρεωτικά μοναδικό. Ορίζουμε τώρα την έννοια της προβολής πίνακα σε άλλον πίνακα που περιέχει τις ίδιες πλειάδες αλλά ένα υποσύνολο από χαρακτηριστικά του αρχικού πίνακα.

Ηλικία	Φύλο	Τ.Κ.	Ασθένεια
25	Άνδρας	30002	Γρίπη
25	Γυναίκα	30002	Βρογχίτιδα
26	Άνδρας	30500	Πυρετός
27	Άνδρας	84100	Δύσπνοια
27	Γυναίκα	30500	Υπέρταση
28	Άνδρας	84100	Γρίπη

Πίνακας 6: Ένας απλός πίνακας δεδομένων D

Ορισμός 2: Έστω πίνακας δεδομένων D με σύνολο χαρακτηριστικών A και έστω A' ⊆ A. Ο πίνακας D' ο οποίος περιέχει τις ίδιες πλειάδες με τον D, με σύνολο χαρακτηριστικών το A' θα ονομάζεται προβολή του πίνακα D ως προς A' και θα συμβολίζεται με D[A']. Για παράδειγμα, η προβολή του πίνακα 2 ως προς το σύνολο A' = {Ηλικία, Τ.Κ.} συμβολίζεται ως D {Ηλικία, Τ.Κ.} και είναι ο πίνακας που ακολουθεί. Βλέπουμε ότι οι πίνακες περιέχουν ακριβώς τις ίδιες πλειάδες με τη διαφορά ότι στον πίνακα προβολή υπάρχουν λιγότερα χαρακτηριστικά.

Ηλικία	Τ.Κ.
25	30002
25	30002
26	30500
27	84100
27	30500
28	84100

Πίνακας 7: Προβολή του πίνακα D ως προς το QI_D

Συνεχίζοντας, ορίζουμε την έννοια του ψευδό-αναγνωριστικού που παίζει πολύ σημαντικό ρόλο στο μοντέλο k-ανωνυμία.

Ορισμός 3: Έστω πίνακας D με σύνολο χαρακτηριστικών A. Ορίζουμε ως ψευδό-αναγνωριστικό (quasi-identifier) ή QI και συμβολίζουμε με QI_D ένα ελάχιστο υποσύνολο του A, τέτοιο ώστε ο πίνακας D{QI_D} αν συνενωθεί με άλλα δεδομένα να μπορεί να οδηγήσει σε κατάργηση της ανωνυμίας ενός ή περισσότερων ατόμων. Είναι σαφές ότι το QI δεν είναι μοναδικό και μπορούμε να πούμε ότι η επιλογή των χαρακτηριστικών που δημιουργούν το QI σε κάθε περίπτωση εξαρτάται από το είδος των εξωτερικών δεδομένων που έχουμε διαθέσιμα για σύνδεση. Αν για παράδειγμα, ενώσουμε τον πίνακα D με QI_D = {Ηλικία, Τ.Κ.}

με τον πίνακα που ακολουθεί (έστω V) και αφορά δεδομένα που περιέχονται σε εκλογικό κατάλογο τότε μπορούμε να βρούμε ότι ο Γιώργος πάσχει από ίωση (Γρίπη).

Όνομα	Ηλικία	Φύλο	Τ.Κ.
Γιώργος	25	Άνδρας	30002
Αλέξης	28	Άνδρας	30002
Νίκη	31	Γυναίκα	30500
Πέτρος	19	Άνδρας	84100
Εβελίνα	40	Γυναίκα	30500

Πίνακας 8: Εκλογικός κατάλογος V

Ηλικία	Τ.Κ.
25	30002
28	30002
31	30500
19	84100
40	30500

Πίνακας 9: Προβολή του πίνακα V ως προς το QI_D

Αυτό μπορούμε να το βρούμε αν συνενώσουμε τους πίνακες D{QI_D} και V{QI_D} (πίνακας 6 και πίνακας 8 αντίστοιχα). Από τους δυο πίνακες βρίσκουμε ότι η πλειάδα (25, 30002) είναι κοινή. Από τις πληροφορίες που μας δίνουν οι πίνακες D και V προκύπτει ότι το άτομο που αναπαριστά την πλειάδα είναι {Γιώργος, πάσχει από ίωση (Γρίπη) και είναι Άντρας}. Βλέπουμε λοιπόν ότι μπορούμε με κατάλληλη επιλογή του QI να άρουμε την ανωνυμία με τη βοήθεια της σύνδεσης.

2.2.2 Μεθοδοι Για Την Επιτευξη Της κ-Ανωνυμιας

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.

Η ανωνυμοποίηση είναι μια απαραίτητη διαδικασία που πρέπει να προηγηθεί πριν την δημοσίευση δεδομένων. Αν συνυπολογίσουμε μάλιστα και το γεγονός ότι ο όγκος των δεδομένων συνεχώς αυξάνεται και ότι διαρκώς γεννιούνται νέα δεδομένα, γίνεται σαφές ότι η ανωνυμοποίηση πρέπει να γίνεται γρήγορα. Για αυτό το λόγο επιδιώκουμε να εκτελέσουμε την ανωνυμοποίηση με κατανεμημένο τρόπο.

Το μοντέλο της k-ανωνυμίας χρησιμοποιεί, δυο βασικές τεχνικές, έτσι ώστε να ανωνυμοποιεί τους πίνακες της βάσης δεδομένων που δημοσιεύονται με σκοπό να καταστήσει κάθε μια εγγραφή αδιάκριτη ανάμεσα σε άλλες k-1.

Μια από τις βασικές τεχνικές είναι η γενίκευση και ορίζεται ως η διαδικασία κατά την οποία η αρχική τιμή που εμφανίζεται στα δεδομένα αντικαθίστανται με μια πιο γενική τιμή ή με ένα ευρύτερο πλαίσιο τιμών που σημασιολογικά περιέχει την αρχική τιμή. Στόχος είναι η διατήρηση μέρους της πληροφορίας που περιέχει η αρχική τιμή χωρίς αυτή να αλλοιώνεται πλήρως. Η τεχνική της γενίκευσης χρησιμοποιείται σε αριθμητικά αλλά και κατηγορικά πεδία τιμών. Στην περίπτωση ενός αριθμητικού πεδίου τιμών, μια αρχική τιμή γενικεύεται σε ένα διάστημα τιμών και στην περίπτωση των κατηγορικών δεδομένων η γενίκευση εφαρμόζεται βάσει της σημασιολογίας των αρχικών τιμών.

Η δεύτερη βασική τεχνική που χρησιμοποιείται για την ανωνυμία είναι η μέθοδος συμπίεσης ή απόκρυψη τιμών. Στόχος είναι η ελαχιστοποίηση του επιπέδου γενίκευσης και η μείωση απώλειας της πληροφορίας στα δεδομένα, προκειμένου να επιτευχθεί κάτι τέτοιο αφαιρούνται δεδομένα από το σύνολο εγγραφών, τα οποία παραβιάζουν την k-Ανωνυμία.

Τις τεχνικές ανωνυμοποίησης πινάκων για την προστασία της ιδιωτικότητας σε μια βάση ευαίσθητων δεδομένων θα αναφέρουμε και θα αναλύσουμε με παραδείγματα στο επόμενο κεφάλαιο.

2.2.3 ΑΛΓΟΡΙΘΜΟΙ ΕΥΡΕΣΗΣ Κ-ΑΝΩΝΥΜΩΝ ΠΙΝΑΚΩΝ

Οι πιο ευρέως διαδεδομένοι αλγόριθμοι υλοποίησης της *k*-ανωνυμοποίησης είναι ο Apriori, Incognito και ο Mondrian, οι οποίοι χρησιμοποιούν την τεχνική της γενίκευσης. Αυτοί οι αλγόριθμοι δέχονται ως είσοδο το σύνολο των αρχικών δεδομένων και επιστρέφουν κάθε φορά την αντίστοιχη βέλτιστη γενίκευση. Λέγοντας βέλτιστη γενίκευση νοείται η γενίκευση κατά την οποία διασφαλίζεται η προστασία της ιδιωτικότητας με χρήση του μοντέλου *k*-ανωνυμίας και εξασφαλίζεται όσο το δυνατόν μικρότερη απώλεια πληροφορίας γίνεται.

Η κύρια διαφορά των δύο αλγορίθμων είναι ότι ο Incognito είναι αλγόριθμος μονοδιάστατης καθολικής ανακωδικοποίησης, ενώ ο Mondrian είναι αλγόριθμος πολυδιάστατης ανακωδικοποίησης στο επίπεδο της κλάσης ισοδυναμίας. Επίσης, ο αλγόριθμος Incognito γενικεύει ένα-ένα τα γνωρίσματα, ενώ ο αλγόριθμος Mondrian γενικεύει κάθε εγγραφή στο σύνολο των γνωρισμάτων της. Στόχος και των δύο αλγορίθμων είναι να χρησιμοποιήσουν το μοντέλο της *k*-ανωνυμίας στα δεδομένα. Παρόλα αυτά, προσφέρουν διαφορετική ποιότητα *k*-ανωνυμοποίησης εξαιτίας των διαφορετικών μεθόδων που ακολουθούν.

2.2.3.1 Apriori

Ένας αποδοτικός ευριστικός αλγόριθμος που εφαρμόζει k^m- ανωνυμία σε σύνολα δεδομένων είναι αυτός που εκμεταλλεύεται την αρχή της apriori ιδιότητας. Σύμφωνα με την ιδιότητα αυτή εάν ένα σύνολο J παραβιάζει την ιδιωτικότητα της βάσης, τότε το ίδιο θα συμβαίνει και για οποιοδήποτε υπερσύνολο του J. Ο αλγόριθμος ξεκινά και ελέγχει για παραβιάσεις ιδιωτικότητας υποθέτοντας ότι ο επιτιθέμενος γνωρίζει μόνο μια τιμή από το σύνολο του ψευδο-αναγνωριστικού, στη συνέχεια

επαναλαμβάνει τον έλεγχο για 2 τιμές και συνεχίζει μέχρι να ελέγξει για m τιμές. Το πλεονέκτημα του συγκεκριμένου αλγόριθμου είναι ότι εκμεταλλεύεται τις γενικεύσεις που έγιναν στο βήμα i με αποτέλεσμα να μειώνεται ο αριθμός των γενικεύσεων στο βήμα i+1.

Ο αλγόριθμος εφαρμόζει τη διαδικασία γενίκευσης σε όλους τους συνδυασμούς τιμών μεγέθους $i = \{1,2, \ldots ,m\}$. Σε κάθε βήμα επανάληψης i, ο apriori αλγόριθμος καταγράφει σε ένα δέντρο συχνοτήτων count-tree τις εμφανίσεις του κάθε συνδυασμού τιμών μεγέθους i της βάσης δεδομένων. Στη συνέχεια εντοπίζει στο δέντρο συχνοτήτων τις τιμές στους κόμβους- φύλλα που παρουσιάζουν συχνότητα εμφάνισης μικρότερη από *k*. Για κάθε μια από αυτές τις τιμές ανατρέχει στο δέντρο ιεραρχίας γενίκευσης, και αντικαθιστά τις προβληματικές τιμές με πιο γενικευμένες με στόχο να αυξήσει τη συχνότητα εμφάνισης της κάθε τιμής σε πλήθος μεγαλύτερο από k. Ο αλγόριθμος επαναλαμβάνει τη διαδικασία για όλες τις προβληματικές τιμές του δέντρου συχνοτήτων.

Ακολουθεί ο ψευδοκώδικας του αλγόριθμου που βασίζεται στην apriori ιδιότητα:

Apriori Αλγόριθμος Ανωνυμοποίησης

AA *(D, I, k, m)*

1: αρχικοποίηση δέντρου ιεραρχίας γενίκευσης

2: ***για*** *ι:=1 μέχρι m ⇒ για* ***όλα*** *τα μεγέθη εγγραφών*

3: δημιούργησε νέο δέντρο συχνοτήτων count-tree

4: ***για όλες*** *τις εγγραφές t ∈ D*

5: ενημέρωσε το count-tree με όλους τους συνδυασμούς μεγέθους i της εγγραφής

6: ***για όλα*** *τα φύλλα v του δέντρου συχνοτήτων*

7: ***εάν*** *το support(v) < k*

8: βρες γενίκευση στο δέντρο ιεραρχίας τέτοια ώστε support(v)>=k

9: ανωνυμοποίησε τα δεδομένα και ενημέρωσε δέντρο συχνοτήτων

2.2.3.2 INCOGNITO

Ο αλγόριθμος Incognito όπως παρουσιάζεται και αναλύεται με παραδείγματα από τους (Ριζομυλιώτης, 2012) και (Αγγέλη, 2014), παρέχει την υλοποίηση του μοντέλου της *k*-ανωνυμοποίησης με γενίκευση πλήρους πεδίου, δηλαδή αντιστοιχίζει κάθε τιμή ενός γνωρίσματος, με την ίδια γενικευμένη τιμή σε όλες τις τιμές του πίνακα. Αρχικά, χρησιμοποιεί την προκαθορισμένη ιεραρχία γενίκευσης πεδίου δημιουργώντας το πλέγμα γενίκευσης πολλαπλών γνωρισμάτων. Στο πλέγμα 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. Μετέπειτα, ελέγχεται εάν κάθε ένας από αυτούς τους συνδυασμούς ικανοποιεί την *k*-ανωνυμία. Στόχος του αλγορίθμου Incognito είναι η αναζήτηση και τελικά η εύρεση της ελάχιστης γενίκευσης πλήρους πεδίου, η οποία ορίζεται με την βοήθεια του διανύσματος απόστασης και με αυτό τον τρόπο μειώνεται το φαινόμενο της απώλειας της πληροφορίας.

Ο αλγόριθμος Incognito εφαρμόζει την *ιδιότητα του υποσυνόλου (subset property),* σύμφωνα με την οποία, αν ένας πίνακας RT(A_1,A_2,….A_n) είναι k-ανώνυμος ως προς ένα σύνολο γνωρισμάτων Q του συνόλου δεδομένων, τότε είναι k-ανώνυμος και ως προς οποιοδήποτε υποσύνολο γνωρισμάτων P ⊆ Q. Ακόμα, εφαρμόζει την ιδιότητα της γενίκευσης δηλαδή: έστω ένας πίνακας RT και έστω P και Q δύο σύνολα από γνωρίσματα στο RT τέτοια ώστε $D_P < {}_DD_Q$. Αν ο πίνακας RT ικανοποιεί την k-ανωνυμία με βάση το P τότε ο RT είναι k-ανώνυμος με βάση το Q.

Ο αλγόριθμος Incognito ακολουθεί τα εξής βήματα:

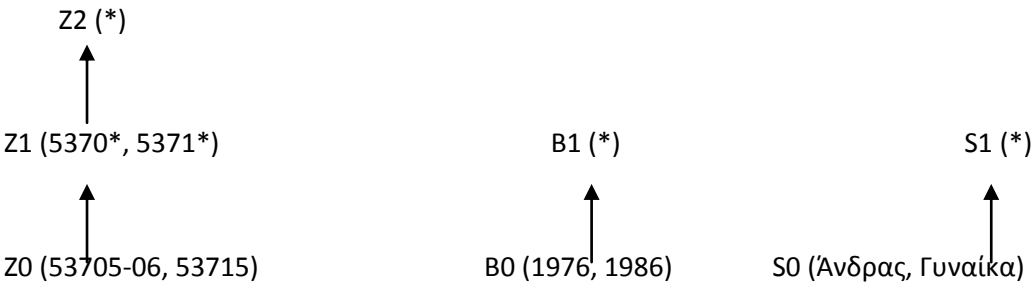
Είσοδος: *Ένας πίνακας RT προς k-ανωνυμοποίηση, ένα σύνολο Q του ψευδό-αναγνωριστικού και μία ιεραρχία για κάθε γνώρισμα του ψευδό-αναγνωριστικού.*
Έξοδος: *Ένα σύνολο από k-ανώνυμων γενικεύσεων γενικού συνόλου.*
C_1 *= Κόμβοι της ιεραρχίας γενίκευσης συνόλου τιμών των γνωρισμάτων του Q.*
E_1 *= Ακμές της ιεραρχίας γενίκευσης συνόλου τιμών των γνωρισμάτων του Q.*
Queue *= μία άδεια ουρά*
For i=1 to d
//C_i και E_i ορίζουν ένα γράφο γενικεύσεων
S_i = ένα αντίγραφο του C_i
Roots = όλοι οι κόμβοι στο C_i χωρίς κάποια ακμή στο E_i κατευθυνόμενοι προς αυτούς
Εισαγωγή των roots στην queue, διατηρώντας την queue ταξινομημένοι ως προς το ύψος.
While queue δεν είναι άδειο do
Node = αφαίρεση του πρώτου στοιχείου της queue
If node δεν είναι μαρκαρισμένος then
If το RT είναι k-ανώνυμο με βάση τα γνωρίσματα του κόσμου then
Μάρκαρε όλες τις άμεσες γενικεύσεις του κόμβου
Else
Διέγραψε τον κόμβο από το S_i
Εισήγαγε τις άμεσες γενικεύσεις του κόμβου στην ουρά,
κρατώντας την ουρά ταξινομημένη ως προς το ύψος.
End if
End if
End while
C_{i-1}, E_{i+1} = ΚατασκευήΓράφου(S_i, E_i)
End for
Προβολή όλων των γνωρισμάτων του S_n στο RT και στις ιεραρχίες.

Παρακάτω, θα αναλύσουμε την διαδικασία και τα βήματα που ακολουθεί ο αλγόριθμος Incognito μέσα από ένα παράδειγμα. Έστω ότι έχουμε ένα πίνακα με σύνολο γνωρισμάτων (Ημερομηνία Γεννήσεως, Φύλο, Ταχυδρομικός Κώδικας, Ασθένεια).

ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ			
Ημερομηνία Γεννήσεως	Φύλο	Ταχυδρομικός Κώδικας	Ασθένεια
1976	Άνδρας	53715	Γρίπη
1986	Γυναίκα	53715	Ηπατίτιδα
1976	Άνδρας	53703	Βρογχίτιδα
1976	Άνδρας	53703	Κάταγμα Χεριού
1986	Γυναίκα	53706	Πυρετός
1976	Άνδρας	53706	Διάστρεμμα

Πίνακας 10: Ιατρικά δεδομένα οργανισμού

Για i=1, ο αλγόριθμος incognito ελέγχει αν ο πίνακας RT ικανοποιεί την k-ανωνυμία για γενικεύσεις ενός συνόλου γνωρισμάτων με μέγεθος i=1. Αρχικά, απομακρύνει τα πεδία <Φύλο> και <Ταχυδρομικός Κώδικας> και κρατάει το πεδίο <Ημερομηνία Γεννήσεως>. Με αυτό τον τρόπο ελέγχει εάν ο πίνακας είναι k-ανώνυμος με βάση το υποσύνολο, και έτσι εάν ισχύει αυτό συνεπάγεται ότι είναι k-ανώνυμος για όλες τις γενικευμένες τιμές που ορίζονται από την ιδιότητα γενίκευσης. Η ίδια διαδικασία επαναλαμβάνεται για τα υπόλοιπα γνωρίσματα του ψευδό-αναγνωριστικού.



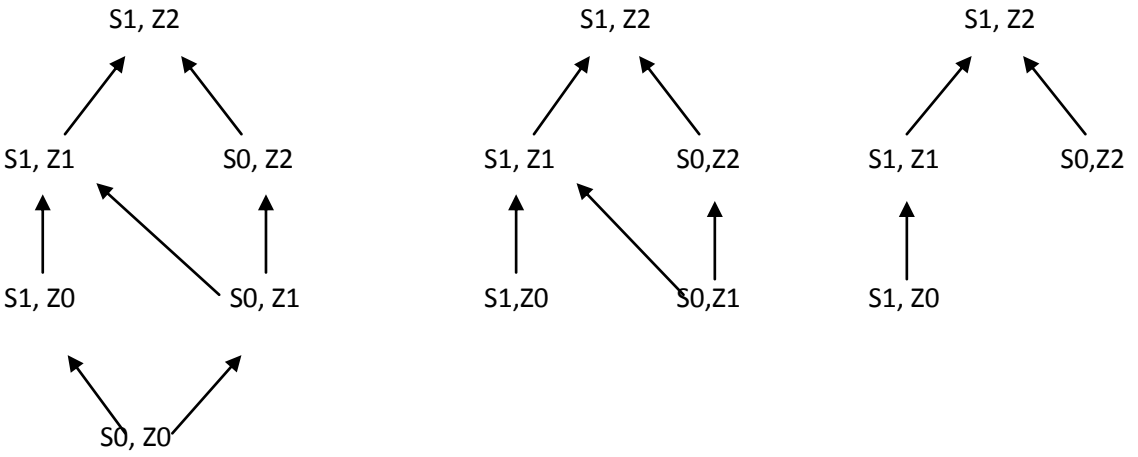
Εικόνα 2: Ιεραρχίες γενίκευσης πεδίων {Ημερομηνία Γεννήσεως, Ταχυδρομικός Κώδικας, Φύλο}

Για i=2, ο αλγόριθμος ελέγχει εάν ικανοποιείται η k-ανωνυμία για τα υποσύνολα γνωρισμάτων με μέγεθος i=2 που είναι τα εξής: <Ημερομηνία Γεννήσεως, Φύλο>, <Ημερομηνία Γεννήσεως, Ταχυδρομικός Κώδικας> και <Φύλο, Ταχυδρομικός Κώδικας>.

Έτσι στο παράδειγμα που απεικονίζεται παραπάνω, ο αλγόριθμος ξεκινάει ελέγχοντας:

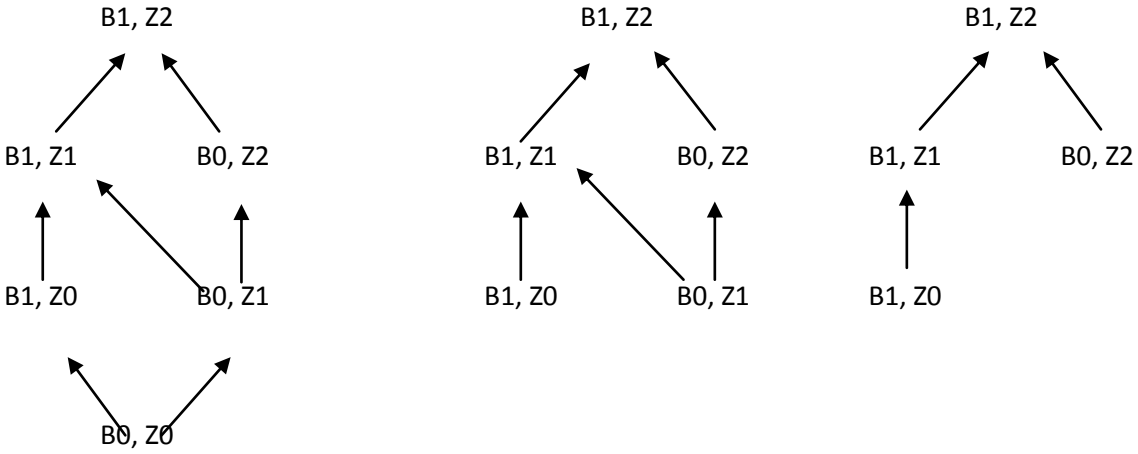
- **Το frequency set του συνόλου <S0, Z0>, το οποίο δεν ικανοποιεί την k-ανωνυμία.**

- **Έτσι, συνεχίζει ελέγχοντας το <S1, Z0> και το <S0, Z1>. Ο πίνακας ικανοποιεί την k-ανωνυμία με βάση το <S1, Z0> και για όλες τις γενικεύσεις του. Μετά ελέγχει το υποσύνολο <S0, Z1>,το οποίο δεν ικανοποιεί την k-ανωνυμία.**
- **Το <S1, Z1> δεν ελέγχεται γιατί είναι γενίκευση του <S1,Z0>.**
- **Στη συνέχεια ελέγχεται το σύνολο <S0, Z2> με βάση το οποίο ικανοποιείται η k-ανωνυμία. Σε αυτό το σημείο σταματάει και ο έλεγχος.**

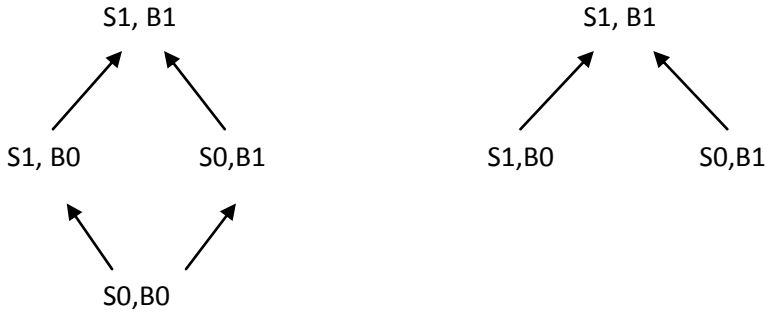


Εικόνα 3: Απόρριψη κόμβων στο υποσύνολο γνωρισμάτων «Φύλο, Ταχυδρομικός Κώδικας»

Στο τελευταίο βήμα του αλγορίθμου αντιστρέφεται η ιδιότητα του υποσυνόλου. Δηλαδή, εάν ένα υποσύνολο γνωρισμάτων του ψευδό-αναγνωριστικού δεν ικανοποιεί την k-ανωνυμία τότε το ίδιο θα ισχύει και για κάθε σύνολο γνωρισμάτων που το περιέχει. Με βάση αυτή την ιδιότητα, δημιουργείται το τελικό πλέγμα όλων των συνδυασμών που ικανοποιούν την k-ανωνυμία.

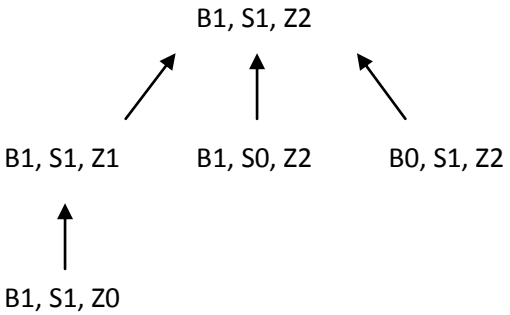


Εικόνα 4:Απόρριψη κόμβων στο υποσύνολο γνωρισμάτων «Φύλο, Ημερομηνία Γεννήσεως»



Εικόνα 5: Απόρριψη κόμβων στο υποσύνολο γνωρισμάτων

<Ταχυδρομικός Κώδικας, Ημερομηνία Γεννήσεως>



Εικόνα 6: Τελικό πλέγμα γενίκευσης

Ένα άλλο πρόβλημα του αλγορίθμου είναι η υλοποίηση του γράφου, η οποία γίνεται σε τρεις φάσεις. Στην αρχή έχουμε μια φάση join μετά μία prune φάση ώστε να υλοποιηθεί το σύνολο των υποψήφιων κόμβων C_i τα οποία θα μπορούσαν να οδηγήσουν σε k-anonymity. Στην τρίτη και στη τελευταία φάση κατασκευάζεται ο πίνακας με τις ακμές, δηλαδή όλα τα χαρακτηριστικά μες τις γενικευμένες σχέσεις. Η φάση join: κατασκευάζεται ένα υπερσύνολο του C_i με βάση το S_i. Να σημειωθεί ότι απαιτείται για να υλοποιηθεί το join τα στοιχεία να έχουν κάποια μορφή ταξινόμησης. Το join είναι λοιπόν της παρακάτω μορφής:

Αλγόριθμος 2 Join Phase

INSERT *INTO* C_i *(dim1; index1,…,dimi, indexi, parent1, parent2)*

SELECT *p.dim1, p.index1,…,p.*$dimi_{i-1}$*,* $p.indexi_{i-1}$*, q.*$dimi_{i-1}$*,* $q.indexi_{i-1}$*, p.ID, q.ID*

FROM *Si-1p, Si-1q*
WHERE *p.dim1 = q.dim1 ^p.index1 = q.index1 ^…^ p.dimi* $_{i-2}$ *= q.dimi* $_{i-2}$ *^p.indexi*$_{i-2}$ *= q.indexi*$_{i-2}$ *^p.dimi*$_{i-1}$ *< q.dimi*$_{i-1}$

Εν συνεχεία εφαρμόζουμε την prune φάση, στην οποία αφαιρούμε κόμβους οι οποίοι δεν χρειάζονται. Η διαδικασία είναι απλή. Ένα σύνολο ιδιοτήτων s με πληθάριθμο n θα περιέχεται S_i στο αν και μόνο στο S_{i-1} περιέχονται όλα τα δυνατά υποσύνολα του s με πληθάριθμο n-1. Βασισμένοι σε αυτήν την ιδιότητα μπορούμε να αφαιρέσουμε όλα τα επιπλέον σύνολα που παράχθηκαν στη φάση του join.

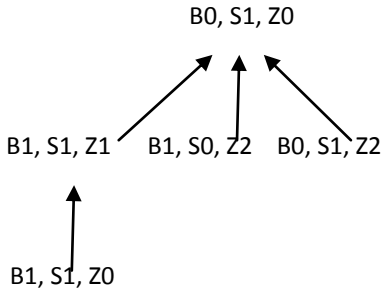
Ο λόγος που αυτά τα δυο παραπάνω μας παράγουν το C_i είναι προφανής. Επί της ουσίας είναι ο apriori αλγόριθμος, και έχουμε εκμεταλλευτεί το subset και generalization property. Γνωρίζουμε ότι ο πίνακας δεν είναι ανωνυμοποιημένος με βάση ένα σύνολο P, τότε δεν μπορεί να είναι ανωνυμοποιημένο με βάση ένα σύνολο Q αν Q $\leq$ P. Πως εκμεταλλευόμαστε όμως αυτό στην υλοποίηση του C_i; Στην επανάληψη i-1 έχουμε Q μέγεθος i-1, έστω το $(q_1,…q_{i-1})$ το οποίο δεν υπάρχει στο S_{i-1}. Κάθε σύνολο της μορφής $(q_1,…,q_{i-1}, t)$ όπου t είναι η αξία μιας νέας διάστασης, δεν μπορεί να οδηγηθεί σε k-ανωνυμία. Άρα αν έχουμε τα στοιχεία ταξινομημένα με βάση το ύψος τότε μπορούμε με το join σε πρώτη φάση και σε δεύτερη το prune να κρατήσουμε μόνο όσους κόμβους έχουν όντως πιθανότητα να μας οδηγήσουν σε k-ανωνυμία.

Στην τρίτη φάση πρέπει να υλοποιηθούν όλες οι ακμές. Αυτό στηρίζεται στην εξής παρατήρηση, ότι ένας κόμβος A είναι γενίκευση ενός άλλου B όταν συμβαίνει ένα από τα δυο:

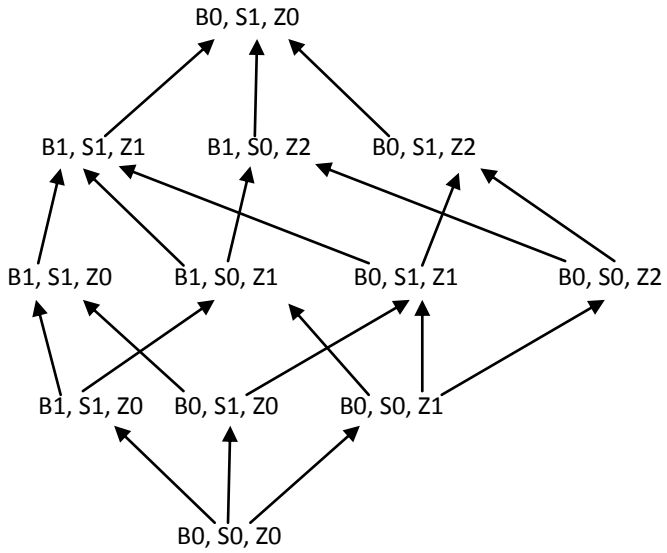
- **Οι γονείς του B είναι γενικεύσεις στους γονείς του A.**
- **Ένας από τους δυο γονείς του B είναι γενίκευση του αντίστοιχου γονέα του A και οι άλλοι δυο είναι ίσοι. Όλες οι implied generalization σχέσεις αφαιρούνται στο τέλος.**

Δίνεται η διαδικασία εφαρμοσμένη σε SQL:

Algorithm 3 Prune phase
INSERT INTO *Ei (start, end)*
WITH CandidateEdges (start, end) AS (
SELECT *p.ID, q.ID*
FROM *Ci p, Ci q, Ei-1 e, Ei-1 f*
WHERE *(e.start = p.parent1^e:end = q:parent1^f:start = p:parent2^f:end = q:parent2) V (e:start = p.parent1 ^ e.end = q.parent1 ^ p.parent2 = q.parent2) V (e.start = p.parent2 ^ e.end = q.parent2 ^ p.parent1 = q.parent1))*
SELECT *D.start, D.end*
FROM *CandidateEdges D*
EXCEPT
SELECT *D1, start, D2.end*
FROM *CandidateEdges D1, CandidateEdges D2*
WHERE *D1.end = D2.start*



(α΄) Ο γράφος που κατασκευάζει ο αλγόριθμος



(β΄) Όλος ο γράφος

Εικόνα 7: Αφαίρεση κόμβων μέσω των join και prune φάσεων

Χάρη στον apriori αλγόριθμο μειώνουμε ένα σημαντικό πλήθος έλεγχων. Για παράδειγμα στο παραπάνω σχήμα, στην πρώτη περίπτωση δίνεται το δίκτυο το οποίο θα εξεταστεί στην επόμενη επανάληψη με βάση την συγκεκριμένη υλοποίηση και το δίκτυο το οποίο θα εξεταζόταν αν δεν είχαμε αυτές τις τρείς φάσεις.

2.2.3.3 ΣΧΟΛΙΑ ΑΛΓΟΡΙΘΜΟΥ

Ο αλγόριθμος Incognito έχει το πλεονέκτημα ότι υπολογίζει όλες τις δυνατές k-ανωνυμοποιημένες όψεις ενός πίνακα βασισμένος σε μια ιεραρχία γενικεύσεων σε αρκετά ικανοποιητικό χρόνο.

Στην πραγματικότητα τα μειονεκτήματα απαιτούν μια k-ανωνυμοποιημένη όψη, αλλά ο υπολογισμός όλων των δυνατών k-ανωνυμοποιημένων όψεων είναι χρονοβόρος.

Το άλλο μειονέκτημα του αλγορίθμου είναι ότι δεν μας δίνει καμία ένδειξη για την απώλεια πληροφορίας. Όπως στον Incognito αλλά και σε άλλους αλγορίθμους, δεν μας ενδιαφέρει απλά να βρούμε μια k-ανωνυμοποιημένη όψη αλλά να βρούμε μια τέτοια όψη η οποία να έχει όσον το δυνατόν λιγότερη απώλεια πληροφορίας. Έτσι για να επιλέξουμε μια από τις δυνατές λύσεις που μας παρέχει ο αλγόριθμος πρέπει να ελέγξουμε όλες τις λύσεις για να βρούμε την βέλτιστη, εισάγοντάς μας έτσι ακόμα μεγαλύτερη χρονική πολυπλοκότητα.

Ο αλγόριθμος προϋποθέτει ότι υπάρχει ήδη μια ιεραρχία που ορίζει τις γενικεύσεις. Δεν τον ενδιαφέρει πως έχει προκύψει αυτή και με ποιο υπολογιστικό κόστος. Αυτό μπορεί να είναι ιδιαίτερα αρνητικό για αριθμητικά δεδομένα. Στα αριθμητικά δεδομένα συνήθως επιτυγχάνουμε την γενίκευση κάνοντας μια τιμή range και γενικότερα ένα range γενικεύεται επεκτείνοντας τα άκρα του. Είναι προφανές ότι τα άκρα του range δεν είναι απαραίτητο να είναι μεγαλύτερα και μικρότερα αντίστοιχα από ότι την μέγιστη και ελάχιστη τιμή αντίστοιχα. Μια ιεραρχία δεδομένων η οποία έχει οριστεί, πριν αναζητήσουμε γενικεύσεις δεν μπορεί να το πετύχει αυτό.

2.2.3.4 MONDRIAN

Ο αλγόριθμος Mondrian, όπως αναφέραμε και στην εισαγωγή του κεφαλαίου είναι ένας αλγόριθμος πολυδιάστατης ανακωδικοποίησης ο οποίος εφαρμόζεται όταν αντιμετωπίζουμε υπεργενίκευση των δεδομένων. Λέγοντας υπεργενίκευση των δεδομένων και με βάση τον (Αγγέλη, 2014), εννοούμε ότι τα δεδομένα μετατρέπονται σε άχρηστη πληροφορία διότι εφαρμόστηκε σε αυτά γενίκευση πλήρους πεδίου με αποτέλεσμα να αντικατασταθούν όλες οι αρχικές τιμές με σταθερά μη επικαλυπτόμενα μεταξύ τους διαστήματα ή ακόμα μπορεί να αποκρύφτηκαν εντελώς.

Συνεπώς, ο αλγόριθμος Mondrian προσφέροντας υψηλότερης ποιότητας ανωνυμοποίηση από τον αλγόριθμο Incognito μπορεί να λύσει τέτοιου είδους προβλήματα. Με βάση αυτόν τον αλγόριθμο, ορίζεται ένας χώρος μ-διαστάσεων, όπου μ το πλήθος των ψευδο-αναγνωριστικών (quasi-identifiers). Χωρίζοντας αυτό το χώρο σε διαμερίσεις (partitions), αναζητείται μια k-ανώνυμη λύση.

Ο αλγόριθμος Modrian περιγράφεται από τα εξής βήματα:

1. **Επιλέγει την διάσταση σύμφωνα με την οποία θα γίνει η διαμέριση του χώρου.**
2. **Υλοποιεί τη διαμέριση βάσει της πιο πάνω διάστασης, από την οποία προκύπτουν δύο υποχώροι R1 και R2.**
3. **Για κάθε ένα από τους δύο υποχώρους R1 και R2, επαναλαμβάνεται η διαδικασία μέχρι να μην υπάρχει άλλη επιτρεπόμενη τομή για διαμέριση σε καμία διάσταση.**
4. **Προκύπτει η βέλτιστη πολυδιάστατη διαμέριση και συνεπώς η κατάλληλη πολυδιάστατη γενίκευση που θα χρησιμοποιηθεί.**

Ψευδοκώδικας του αλγορίθμου Mondrian:

```
K// αριθμός k-anonymity
2 Mondrian (p a r t i t i o n)
3     if (|p a r t i t i o n| < 2k)
```

```
4       return
5     else
6       dim = chooseDimension (p a r t i t i o n)
7       s p l i t V a l u e = findMedian (p a r t i t i o n , dim)
8       l P a r t = {t E p a r t i t i o n  t . dim _ s p l i t V a l u e}
9       rPart = {t E  p a r t i t i o n : t . dim > s p l i t V a l u e}
10      return Mondrian (l P a r t) [ Mondrian (rPart)]
```

Οι γραμμές 6-10 του ψευδοκώδικα αποτελούν τη συνάρτηση split όπου δέχεται ως παράμετρο το αρχικό σύνολο και επιστρέφει δυο σύνολα (part1, part2) που η ένωση τους είναι το αρχικό σύνολο ενώ παράλληλα τα part1, part2 είναι ξένα μεταξύ τους. Στην αρχή του ψευδοκώδικα επιλέγεται μια διάσταση του πίνακα. Η διάσταση αυτή ισοδυναμεί με κάποιο χαρακτηριστικό των πλειάδων και ανήκει στο ψευδο-αναγνωριστικό QID βάση του οποίου εκτελείται η ανωνυμοποίηση. Στη συνέχεια, εξετάζουμε όλες τις πλειάδες των χαρακτηριστικών που επιλέχθηκε και επιλέγουμε την τιμή του ενδιάμεσου χαρακτηριστικού (splitValue). Για παράδειγμα, αν οι πλειάδες στο χαρακτηριστικό που είναι προς εξέταση έχουν τις τιμές {10, 11, 9, 10, 2}, η ενδιάμεση τιμή είναι το 10^2. Στη συνέχεια, χωρίζουμε τις πλειάδες με βάση την τιμή που έχουν στο υποεξέταση χαρακτηριστικό και το splitValue. Οι πλειάδες που είναι μεγαλύτερες από το ενδιάμεσο στοιχείο στην επιλεγμένη διάσταση, τοποθετούνται στο rPart ενώ όλα τα υπόλοιπα τοποθετούνται στο lPart. Τέλος, η διαδικασία συνεχίζεται αναδρομικά, μέχρι όλα τα υποσύνολα να παραμείνουν με λιγότερα από 2k στοιχεία (έτσι ώστε να μην διαμερίζονται).

Ένα πολύ σημαντικό τμήμα του αλγορίθμου αφορά τον τρόπο με τον οποίο επιλέγεται η διάσταση ως προς την οποία γίνεται η διαμέριση. Γενικά, μπορούν να χρησιμοποιηθούν πολλοί ευριστικοί τρόποι για την επιλογή της διάστασης. Ένας απλός και αποδοτικός τρόπος που επιλέχθηκε στην παρούσα εργασία, είναι η επιλογή της διάστασης που έχει το μέγιστο κανονικοποιημένο εύρος τιμών. Για κάθε διάσταση (ή αλλιώς, για κάθε χαρακτηριστικό που ανήκει στο ψευδο-αναγνωριστικό QID) υπολογίζεται το εύρος τιμών του εξεταζόμενου partition και στη συνέχεια, το εύρος αυτό διαιρείται με τη μέγιστη τιμή που εμφανίζει το χαρακτηριστικό σε όλες τις πλειάδες του πίνακα δεδομένων (έτσι τα εύρη τιμών όλων των διαστάσεων κανονικοποιούνται και έχουν μέγιστη τιμή ίση με 1). Από όλα τα χαρακτηριστικά επιλέγεται αυτό που έχει το μεγαλύτερο κανονικοποιημένο εύρος. Σε περίπτωση που όλα τα χαρακτηριστικά έχουν το ίδιο εύρος (π.χ., στο πρώτο βήμα εκτέλεσης του αλγορίθμου) τότε επιλέγεται το χαρακτηριστικό που έχει το μικρότερο απόλυτο εύρος.

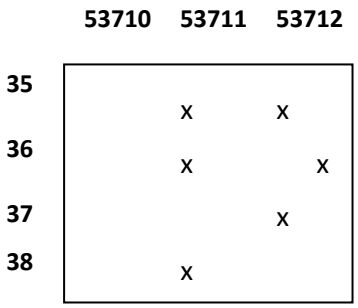
Μετά την επιλογή της διάστασης ακολουθεί η εύρεση του ενδιάμεσου στοιχείου και στη συνέχεια γίνεται η διαμέριση των πλειάδων με βάση το στοιχείο αυτό. Γενικά, ένας αλγόριθμος που στηρίζεται στη διαμέριση όπως ο Mondrian μπορεί να ακολουθήσει ορισμένες πολιτικές κατά τη διαμέριση ενός συνόλου. Μια πιθανή πολιτική είναι η διαμέριση ενός συνόλου με στόχο τη δημιουργία δυο συνόλων που περιέχουν περίπου τον ίδιο αριθμό στοιχείων, ακόμη και αν αυτό σημαίνει ότι πολλά σημεία που βρίσκονται εκατέρωθεν του ενδιάμεσου στοιχείου τοποθετούνται στο ίδιο υποσύνολο. Μια δεύτερη πιθανή πολιτική είναι η ακριβώς αντίθετη της πρώτης: κατά τη διαμέριση τα στοιχεία τοποθετούνται "αυστηρά" εκεί που θα έπρεπε να ανήκουν με βάση την θέση τους ως προς το ενδιάμεσο στοιχείο ακόμα και αν αυτό σημαίνει ότι θα δημιουργήσουν δυο σύνολα που θα έχουν μεγάλη διαφορά στον αριθμό των στοιχείων τους. Η πρώτη πολιτική διαμέρισης είναι γνωστή ως relaxed partitioning και η δεύτερη είναι γνωστή ως strict partitioning.

Έτσι, μέσα από αυτή τη διαδικασία ο αλγόριθμος επιτυγχάνει να βρει τη βέλτιστη πολυδιάστατη διαμέριση, σε κάθε περιοχή της οποίας ανήκουν περισσότερες από k-εγγραφές και συνεπώς ικανοποιείται η k-ανωνυμία.

Το πρόβλημά εύρεσης της βέλτιστης διαμέρισης είναι NP-hard. Ο αλγόριθμος Mondrian δεν δίνει την βέλτιστη λύση. Ωστόσο, έχει πολυπλοκότητα O(*nlogn*), όπου n ο αριθμός των εγγραφών του πίνακα, η οποία είναι σχετικά μια ικανοποιητική προσέγγιση. Τέλος, ένα ζήτημα που παραμένει ανοιχτό είναι το γεγονός ότι ο αλγόριθμος Modrian δεν μπορεί να υπολογίσει την απώλεια πληροφορίας.

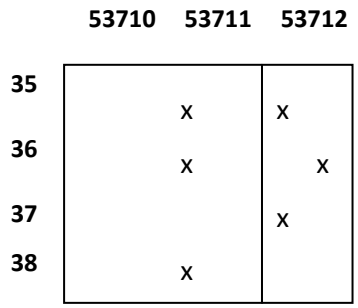
Παρακάτω ορίζονται τα μοντέλα μονοδιάστατης και πολυδιάστατης ανωνυμοποίησης. Στη μονοδιάστατη ανωνυμοποίηση, σε κάθε επίπεδο της ιεραρχίας γενίκευσης βρίσκονται τιμές από μη επικαλυπτόμενα διαστήματα ενώ στην πολυδιάστατη ανωνυμοποίηση επιτρέπονται τα επικαλυπτόμενα διαστήματα. Επιπροσθέτως, στο μοντέλο της μονοδιάστατης ανωνυμοποίησης, ο χώρος μοιράζεται σχηματίζοντας παράλληλες γραμμές ως προς τους άξονες οι οποίες διασχίζουν όλο το χώρο. Αντιθέτως, όταν πρόκειται για το πολυδιάστατο μοντέλο ανωνυμοποίησης υπάρχουν δύο υποχώροι, στους οποίους ορίζονται άλλοι υποχώροι. Τέλος, για να ικανοποιείται η k-ανωνυμία πρέπει σε κάθε υποχώρο να υπάρχουν τουλάχιστον k εγγραφές, οι οποίες καταγράφονται σαν ένα σημείο στο χώρο.

ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ			
Ημερομηνία Γεννήσεως	Φύλο	Ταχυδρομικός Κώδικας	Ασθένεια
35	Άνδρας	53711	Γρίπη
35	Γυναίκα	53712	Εμετός
36	Άνδρας	53711	Βρογχίτιδα
37	Άνδρας	53710	Κάταγμα Χεριού
37	Γυναίκα	53712	Πυρετός
38	Άνδρας	53711	Πόνος στη πλάτη



Πίνακας 11: Αρχικός πίνακας

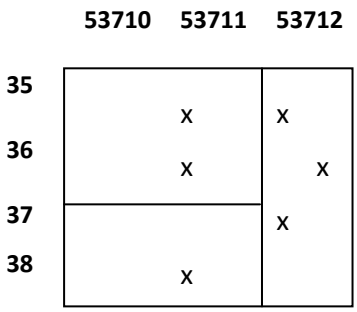
ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ			
Ημερομηνία Γεννήσεως	Φύλο	Ταχυδρομικός Κώδικας	Ασθένεια
[35-38]	Άνδρας	[53710-53711]	Γρίπη
[35-38]	Γυναίκα	53712	Εμετός
[35-38]	Άνδρας	[52710-53711]	Βρογχίτιδα
[35-38]	Άνδρας	[53710-53711]	Κάταγμα Χεριού
[35-38]	Γυναίκα	53712	Πυρετός
[35-38]	Άνδρας	[53710-53711]	Πόνος στη πλάτη



Πίνακας 12: Μονοδιάστατη ανωνυμοποίηση

ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ			
Ημερομηνία Γεννήσεως	Φύλο	Ταχυδρομικός Κώδικας	Ασθένεια

[35-36]	Άνδρας	53711	Γρίπη
[35-37]	Γυναίκα	53712	Εμετός
[35-36]	Άνδρας	53711	Βρογχίτιδα
[37-38]	Άνδρας	[53710-53711]	Κάταγμα Χεριού
[35-37]	Γυναίκα	53712	Πυρετός
[37-38]	Άνδρας	[53710-53711]	Πόνος στη πλάτη



Πίνακας 13: Πολυδιάστατη ανωνυμοποίηση

2.2.3.5 ΣΧΟΛΙΑ ΑΛΓΟΡΙΘΜΟΥ

Ο αλγόριθμος Mondrian είναι ένας ικανοποιητικά γρήγορος αλγόριθμος ο οποίος δουλεύει ¨καλά¨ για αριθμητικά δεδομένα. Το βασικό μειονέκτημά του είναι ότι δεν μπορεί να επεξεργαστεί κατηγορικά δεδομένα, εκτός αν υπάρχει κάποια διάταξη για αυτά. Ακόμα όμως και να υπάρχει διάταξη δεν μπορεί να ικανοποιήσει το μέγιστο όριο των partitions. Παράλληλα, ο αλγόριθμος δεν μας παρέχει κανένα μέτρο για την απώλεια της πληροφορίας. Το μέγεθος του partition αποτελεί κάποια μορφή ελέγχου αλλά δεν είναι αρκετή. Είναι δυνατόν ο χώρος να χωριστεί σε διάφορα partitions, το θέμα είναι ποιος από όλους αυτούς τους χωρισμούς προσφέρει την καλύτερη πληροφορία.

Τέλος, θα μπορούσε κανείς να αναρωτηθεί γιατί δεν εφαρμόζουμε πάντα local recoding έναντι του global recoding αφού το πρώτο χωρίζει σίγουρα σε μικρότερα partitions. Ο βασικός λόγος αναφέρθηκε παραπάνω και είναι η απώλεια πληροφορίας. Δεν μας εγγυάται κανείς ότι το local recoding είναι βέλτιστο ως προς αυτό. Μην ξεχνάμε ότι το k-anonymity απαιτεί να υπάρχουν τουλάχιστον k ίδιες εγγραφές, εν συνεχεία αυτό που μας νοιάζει δεν είναι κατά το πόσο ξεπερνάμε το k αλλά κατά πόσο διατηρούμε την πληροφορία. Θα μπορούσε εύκολα κανείς να φανταστεί περιπτώσεις όπου αν επιλέξουμε επικαλυπτόμενες περιοχές και τις γενικεύσουμε τότε είναι πιθανό να έχουμε μεγαλύτερη απώλεια πληροφορίας (ισχύει και το αντίστροφο).

3 ΤΕΧΝΙΚΕΣ ΑΝΩΝΥΜΟΠΟΙΗΣΗΣ

Σε αυτό το κεφάλαιο θα εξετάσουμε δύο ειδών τεχνικές στον τομέα της προστασίας ιδιωτικότητας, οι οποίες αποτρέπουν επιθέσεις αναγνώρισης ταυτότητας και επιθέσεις αναγνώρισης τιμής ευαίσθητων δεδομένων. Στην ακόλουθη ενότητα θα μελετήσουμε και θα αναλύσουμε την πρώτη τεχνική ανωνυμοποίησης πινάκων.

3.1 ΤΕΧΝΙΚΕΣ ΓΕΝΙΚΕΥΣΗΣ

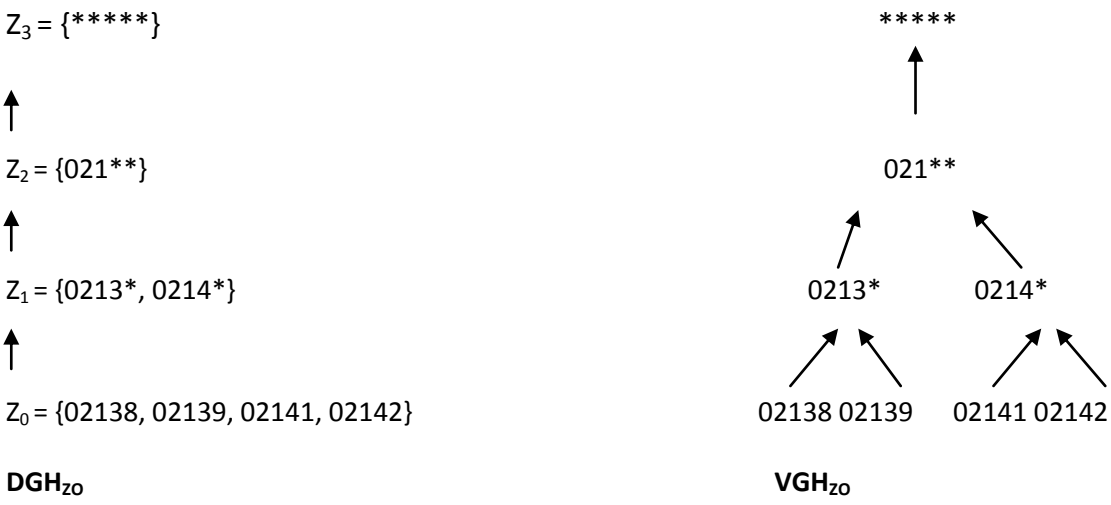
Σύμφωνα με το άρθρο της (Sweeney, 2002), η ιδέα της γενίκευσης ενος χαρακτηριστικού έχει απλή λογική. Μια τιμή αντικαθίσταται από μια λιγότερο συγκεκριμένη, πιο γενική τιμή που είναι πιστή στην αρχική. Σε μια βάση δεδομένων τα πεδία χρησιμοποιούνται για να περιγράψουν το σύνολο των τιμών που παριστάνουν τα χαρακτηριστικά γνωρίσματα. Για παράδειγμα μπορεί να υπάρχει ένα πεδίο Τ.Κ., ένα πεδίο αριθμός και ένα πεδίο αλφαριθμητικό. Στην αρχική βάση δεδομένων όπου κάθε τιμή είναι όσο πιο καθορισμένη γίνεται, κάθε χαρακτηριστικό γνώρισμα βρίσκεται στο κύριο πεδίο. Για παράδειγμα, η τιμή 02139 βρίσκεται στο κύριο πεδίο Τ.Κ.

Για την επίτευξη της k-ανωνυμίας μπορούμε να τροποποιήσουμε τις τιμές του πεδίου ώστε να παρέχει λιγότερη πληροφορία. Η τροποποίησή του μπορεί να γίνει ως εξής: υπάρχει ένα πιο γενικό, λιγότερο συγκεκριμένο πεδίο που μπορεί να χρησιμοποιηθεί για την περιγραφή του Τ.Κ., όπου το τελευταίο ψηφίο έχει αντικατασταθεί από το *. Έτσι με βάση τα παραπάνω, η πληροφορία του πεδίου Τ.Κ. γίνεται 02139$\longrightarrow$ 0213*. Αυτή η σχεδίαση μεταξύ των πεδίων ξεκίνησε με σκοπό την γενίκευση των δεδομένων. Ωστόσο είναι απαραίτητο να ικανοποιούνται οι παρακάτω συνθήκες:

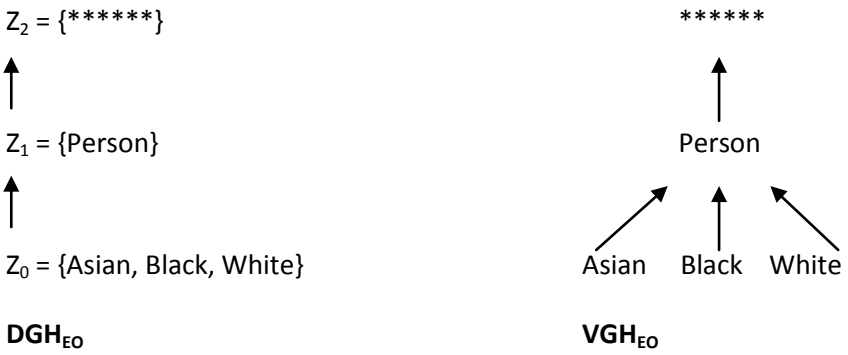
- Κάθε πεδίο έχει το πολύ ένα άμεσο γενικευμένο πεδίο
- Όλα τα μέγιστα πεδία είναι μοναδικά.

Η γενίκευση υποθέτει την ύπαρξη για κάθε πεδίο D ϵ Dom από μια ιεραρχία που ονομάζεται ιεραρχία γενικευμένου πεδίου (DGH_A). Εφόσον, οι γενικευμένες τιμές μπορούν να χρησιμοποιηθούν άλλων πιο συγκεκριμένων, είναι σημαντικό όλα τα πεδία στην ιεραρχία να είναι συμβατά. Η συμβατότητα μπορεί να διασφαλιστεί με την χρήση της ίδιας φόρμας αποθήκευσης για όλα τα πεδία σε μια γενικευμένη ιεραρχία.

Στη συνέχεια ακολουθεί ένα παράδειγμα των πεδίων και την ιεραρχία που ακολουθούν οι γενικευμένες τιμές για τα εξής πεδία: Τ.Κ., Άτομο, Φυλή.



Εικόνα 8: Πεδίο Τ.Κ. Συμπεριλαμβανομένης της Συμπίεσης



Εικόνα 9: Πεδίο Φυλή Συμπεριλαμβανομένης της Συμπίεσης

3.1.1 Έννοια Ελαχιστης Γενικευσης

Δεδομένου ενός αρχικού πίνακα PT, η γενίκευση μπορεί να είναι αποτελεσματική στην παραγωγή ενός νέου πίνακα RT, βάσει του αρχικού πίνακα που τηρεί την k-ανωνυμία όπως παρουσιάζεται από τους (Ryan, και συν., 2007). Οι τιμές των χαρακτηριστικών γνωρισμάτων που είναι αποθηκευμένες στον πίνακα PT μπορούν να υποκατασταθούν, κατά την έκδοση, από τις γενικευμένες τιμές. Εφόσον, οι πολλαπλές τιμές μπορούν να οδηγήσουν σε μία μόνο γενικευμένη τιμή, η γενίκευση μπορεί να μειώσει τον αριθμό των ξεχωριστών πλειάδων και με αυτό τον τρόπο πιθανώς να αυξηθεί το μέγεθος των συστοιχιών που περιέχουν πλειάδες με τις ίδιες τιμές.

Μια συνάρτηση γενίκευσης στην πλειάδα t σε σχέση με το $A_1,...,A_n$ είναι μια συνάρτηση f_t για $A_1x,...,xA_n$ τέτοια ώστε: $f_t(A_1,…,A_n) = (f_{t1}(A_1),…,f_{tn}(A_n))$ όπου για κάθε i: 1,...,n, f_{ti} είναι μια γενίκευση της τιμή t $[A_i]$. Η συνάρτηση f_t είναι ένα σύνολο συναρτήσεων.

Η συνάρτηση f_t παράγεται από το f_{ti}. . Δίνονται η συνάρτηση f, A_1, ..., A_n, ένας πίνακας T (A_1, ..., A_n) και μια πλειάδα tE T, i.e., t(a_1,…,a_n), δηλαδή,

$$g(T) = \{k * f(t) : t \in T \text{ and } |f^{-1}(f(t))| = k \}$$

Η συνάρτηση g είναι συνάρτηση πολλών συνόλων και η f^1 είναι η αντίστροφη συνάρτηση της f. Η g παράγεται από την f και από το fi της. Επιπλέον, η συνάρτηση g (t) είναι μια γενίκευση του πίνακα T.

Γενίκευση χαρακτηριστικού γνωρίσματος σημαίνει υποκατάσταση των τιμών με αντίστοιχες τιμές από ένα πιο γενικό πεδίο. Η γενίκευση σε ένα επίπεδο χαρακτηριστικού γνωρίσματος διασφαλίζει ότι όλες οι τιμές ενός χαρακτηριστικού γνωρίσματος ανήκουν στο ίδιο πεδίο. Παρ' όλα αυτά, ένα αποτέλεσμα της διαδικασίας της γενίκευσης είναι πως το πεδίο ενός χαρακτηριστικού γνωρίσματος μπορεί να αλλάξει, γι' αυτό και πρέπει να υπάρχει μια ισορροπία μεταξύ της γενίκευσης που θέλουμε να επιτύχουμε και του βαθμού απώλεια πληροφορίας που συνεπάγεται η γενίκευση. Γι' αυτό το λόγο ορίζουμε την ελάχιστη γενίκευση, η οποία είναι η γενίκευση μέχρι το επίπεδο που διασφαλίζεται η ιδιότητα της k-ανωνυμίας και δεν θίγεται ο απαιτούμενος βαθμός αξία στης πληροφορίας και γνώσης που μπορεί να εξαχθεί.

Αν DGHi είναι οι ιεραρχίες γενίκευσης πεδίου για τα χαρακτηριστικά A_{li} όπου i = 1, ...,An. Έστω ότι δυο πίνακες T_l [A_{L1}, ...,A_{lAn}] και T_m [A_{m1}, ..., A_{mAn}] είναι δύο πίνακες τέτοιοι ώστε για κάθε i : 1,…,n, A_{li}, A_{mi} E DGH_i. Στη συνέχεια, ο πίνακας T_m είναι μια γενίκευση του πίνακα T, έχοντας $T_l \leq T_m$, αν και μόνο αν υπάρχει μια λειτουργία γενίκευσης g τέτοια ώστε g [T_l] = T_m και να παράγεται από fi, όπου: t_1 E T_1, $a_{li} \leq f_i(a_{li}) = a_{mi}$, και κάθε fi να βρίσκεται στην DGHi του χαρακτηριστικού A_{li}.

3.1.2 Υπολογισμος Καταλληλης Γενικευσης

Έστω ότι έχουμε PT[QI] ενός πίνακα PT με ψευδο-αναγνωριστικό το QI. Αρχικά, εμφανίζονται όλες οι διακριτές πλειάδες μαζί με το πλήθος εμφάνισης τους. Στη συνέχεια υπολογίζεται η απόσταση κάθε εγγραφής και της ακεραίας τιμής. Έπειτα, κατασκευάζεται ένας κατευθυνόμενος γράφος με κόμβους όλα τα διανύσματα. Υπάρχει ένα τόξο από κάθε διάνυσμα προς όλα τα μικρότερα κυριαρχώντας έτσι στο σύνολο. Ο αλγόριθμος αποφασίζει αν μια γενίκευση είναι τοπικά ελάχιστη απλά ελέγχοντας τον συνδυασμό των εμφανίσεων των εγγραφών, χωρίς να εκτελεί πρακτικά τη γενίκευση. Το κόστος υπολογισμού αυτού του αλγορίθμου μειώνεται από τα εξής τρία χαρακτηριστικά:

- Ο υπολογισμός των διανυσμάτων μεταξύ των πλειάδων μειώνουν τον αριθμό των γενικεύσεων που εξετάζονται.
- Οι γενικεύσεις δεν υπολογίζονται πραγματικά αλλά προβλέπονται, παρατηρώντας πως οι εμφανίσεις των εγγραφών θα συνδυαστούν.
- Ο αλγόριθμος παρακολουθεί τις εκτιμημένες γενικεύσεις έτσι σταματά τον υπολογισμό σε ένα μονοπάτι, όταν διασχίζει ένα μονοπάτι που έχει ήδη υπολογίσει.

3.2 ΤΕΧΝΙΚΗ ΣΥΜΠΙΕΣΗΣ

Η γενίκευση έχει το πλεονέκτημα της έκδοσης όλων των μονών πλειάδων του πίνακα, σε μια πιο γενική μορφή, όπως παρουσιάζεται από την (Σπίνου, Σεπτέμβριος 2011). Η δεύτερη μέθοδος είναι μια συμπληρωματική προσέγγιση για την επίτευξη της k-ανωνυμίας, που είναι η συμπίεση (suppression). Η συμπίεση δεδομένων ως τεχνική ελέγχου προστασίας των ατομικών πληροφοριών δεν είναι καινούρια και σημαίνει την απόκρυψη εγγραφών από τον πίνακα έτσι ώστε να μην είναι δυνατή η ταυτοποίηση φυσικών προσώπων.

Εφαρμόζουμε την συμπίεση σε επίπεδο εγγραφής, δηλαδή, επηρεάζονται μεμονωμένες εγγραφές καθώς αυτές αποκρύπτονται από την βάση δεδομένων. Η συμπίεση χρησιμοποιείται για να μετριάσει την διαδικασία της γενίκευσης, όταν υπάρχουν εγγραφές με λιγότερες από k εμφανίσεις.

Με άλλα λόγια, επιλέγονται κάποιες τιμές των γνωρισμάτων οι οποίες αποκρύπτονται από τον δημοσιευμένο πίνακα έτσι ώστε να ικανοποιούν την k-ανωνυμία. Διαφορετικά αν αυτά τα δεδομένα παραμείνουν στον πίνακα τα αντικαθιστούμε από ένα γενικότερο αναγνωριστικό.

Αυτή η τεχνική ανωνυμοποίησης αφορά κυρίως την πλήρη απόκρυψη μιας τιμής στον πίνακα και υπάρχουν τρείς εκδοχές εφαρμογής της:

- Απαλοιφή πλειάδας ή εγγραφής όπου καταστέλλεται ολόκληρη η εγγραφή.
- Απαλοιφή τιμής που οδηγεί σε καταστολή μιας συγκεκριμένης τιμής σε ολόκληρο τον πίνακα.
- Απαλοιφή κελιού που καταστέλλει μόνο μερικά κελιά του πίνακα.

Για την καλύτερη κατανόηση παρουσιάζεται η προβολή του πίνακα, όπως φαίνεται πιο κάτω. Έστω ότι θέλουμε να διατηρείται η k-ανωνυμία με k=2. Το χαρακτηριστικό γνώρισμα Ημερομηνία Γέννησης έχει ένα πεδίο με τις ακόλουθες γενικεύσεις: από την συγκεκριμένη ημερομηνία, στο μήνα, στο χρόνο, σε ένα πενταετές διάστημα, σε ένα δεκαετές διάστημα, εικοσιπενταετές διάστημα κ.ο.κ. Είναι εύκολο να παρατηρήσουμε πως η ύπαρξη της τελευταίας πλειάδας απαιτεί πολλά βήματα γενίκευσης, ένα βήμα γενίκευσης στην Ημερομηνία Γέννησης, ένα βήμα γενίκευσης στον Ταχυδρομικό Κώδικα και ένα βήμα ακόμα γενίκευσης στο Φύλο. Πρακτικά, και στις τρείς περιπτώσεις, σχεδόν όλα τα χαρακτηριστικά γνωρίσματα πρέπει να γενικευτούν.

Εθνικότητα	Ημερομηνία Γέννησης	Φύλο	Ταχυδρομικός Κώδικας
Ασιάτης	09/27/1964	Γυναίκα	30002
Ασιάτης	02/30/1964	Γυναίκα	30002
Ασιάτης	04/18/1969	Άνδρας	84100
Ασιάτης	04/15/1968	Άνδρας	30500
Έγχρωμος	10/13/1968	Γυναίκα	84100
Έγχρωμος	07/18/1969	Άνδρας	30500
Λευκός	09/15/1969	Γυναίκα	84100

Πίνακας 14: Η προβολή του πίνακα

Εθνικότητα	Ημερομηνία Γέννησης	Φύλο	Ταχυδρομικός Κώδικας
Άτομο	1964	Γυναίκα	30002
Άτομο	1964	Γυναίκα	30002
Άτομο	1969	Άνδρας	84100
Άτομο	1968	Άνδρας	30500
Άτομο	1968	Γυναίκα	84100
Άτομο	1969	Άνδρας	30500
Άτομο	1969	Γυναίκα	84100

Πίνακας 15: Πιθανές ελάχιστες γενικεύσεις

Παρατηρούμε πως αν δεν υπήρχε η τελευταία εγγραφή, η k-ανωνυμία θα είχε επιτευχθεί εύκολα με ένα βήμα γενίκευσης στο χαρακτηριστικό γνώρισμα Ημερομηνία Γέννησης, όπως παρουσιάζεται στον παρακάτω πίνακα. Η απομάκρυνση της συγκεκριμένης εγγραφής επιτρέπει την εφαρμογή λιγότερης γενίκευσης.

Εθνικότητα	Ημερομηνία Γέννησης	Φύλο	Ταχυδρομικός Κώδικας
Ασιάτης	1964	Γυναίκα	30002
Ασιάτης	1964	Γυναίκα	30002
Ασιάτης	1969	Άνδρας	84100
Ασιάτης	1968	Άνδρας	30500
Έγχρωμος	1968	Γυναίκα	84100
Έγχρωμος	1969	Άνδρας	30500

Πίνακας 16: Παράδειγμα με συμπίεση εγγραφής

Η γενίκευση και η συμπίεση δεδομένων είναι δυο διαφορετικές προσεγγίσεις, για την απόκτηση ενός πίνακα (από έναν άλλον, αρχικό πίνακα) που ικανοποιεί την k-ανωνυμία. Με την γενίκευση έχουμε λιγότερη ακρίβεια ενώ με την συμπίεση λιγότερη πληρότητα. Από παρατηρήσεις σε πραγματικές εφαρμογές και απαιτήσεις προτιμάται η συμπίεση παρά η γενίκευση. Ο λόγος είναι πως η συμπίεση επηρεάζει συγκεκριμένες εγγραφές ενώ η γενίκευση τροποποιεί όλες τις τιμές σε ένα πίνακα, επηρεάζοντας έτσι όλες τις πλειάδες. Παρ' όλα αυτά η γενίκευση παραμένει μια ισχυρή τεχνική που χρησιμοποιείται ευρέως.

3.2.1 ΕΛΑΧΙΣΤΗ ΑΠΟΚΡΥΨΗ

Σκοπός της δημοσίευσης μιας βάσης δεδομένων είναι η εκμετάλλευση της χρήσιμης πληροφορίας που περιέχουν τα δεδομένα, για στατιστικούς ή ερευνητικούς σκοπούς. Προκειμένου να εξασφαλισθεί η ιδιωτικότητα των εγγραφών πρέπει τα δεδομένα να τροποποιηθούν. Οι αρχικές τιμές αντικαθιστώνται με πιο γενικές τιμές, ώστε να διαφυλαχθεί η προσωπική πληροφορία και να μην δημοσιεύεται.

Μια τέτοια γενίκευση έχει σαν αποτέλεσμα, την ελάχιστη απόκρυψη χρήσιμης πληροφορίας που έχουν τα αρχικά δεδομένα, με αποτέλεσμα τα συμπεράσματα που μπορεί να βγουν από μια στατιστική μελέτη των δεδομένων να μην είναι τόσο ακριβή όσο θα ήταν αν δημοσιεύονταν οι αρχικές τιμές.

Η έννοια της ελάχιστης απόκρυψης χρησιμοποιείται για την διαδικασία της γενίκευσης όταν υπάρχει περιορισμένος αριθμός των πλειάδων με λιγότερο από k εμφανίσεις. Ένα χαρακτηριστικό παράδειγμα για την καλύτερη κατανόηση της ελάχιστης απόκρυψης επισημαίνεται παρακάτω.

Έστω ένας πίνακας Tj όπου είναι μια γενίκευση του Ti ο οποίος ικανοποιεί k-ανωνυμία. Ο Tj επιβάλει την ελάχιστη καταστολή αν υπάρχει Tz τέτοιο ώστε:

- Ti $\leq$ Tz
- DVi,z = DVi,j
- Sizeof (Tj) < sizeof (Tz)
- Tz να ικανοποιεί k-ανωνυμία

Οι εγγραφές των πινάκων όπου αναγράφονται με έντονα γράμματα είναι το σύνολο των πλειάδων που πρέπει να κατασταλούν για να επιτύχουν k-ανωνυμία του 2. Κάθε υπερσύνολο της συμπίεσης δεν ικανοποιεί την ελάχιστη απαιτούμενη καταστολή.

Εθνικότητα	Ταχυδρομικός Κώδικας:
Ασιάτης	02138
Ασιάτης	02138
Ασιάτης	02142
Ασιάτης	02142
Έγχρωμος	**02138**
Έγχρωμος	**02141**
Έγχρωμος	**02142**
Λευκός	**02138**

Εθνικότητα	Ταχυδρομικός Κώδικας:
Άτομο	02138
Άτομο	02138
Άτομο	02142
Άτομο	02142
Άτομο	02138
Άτομο	**02141**
Άτομο	02142
Άτομο	02138

Πίνακας 17: Πίνακες PT και GT[1,0]

Στον πίνακα PT αλλά και στους υπόλοιπους πίνακες που ακολουθούν το σύνολο των χαρακτηριστικών είναι {Εθνικότητα, Ταχυδρομικός Κώδικας} μια βάσης δεδομένων.

Εθνικότητα	Ταχυδρομικός Κώδικας:
Ασιάτης	02130
Ασιάτης	02130
Ασιάτης	02140
Ασιάτης	02140
Έγχρωμος	**02130**
Έγχρωμος	02140
Έγχρωμος	02140
Λευκός	**02130**

Εθνικότητα	Ταχυδρομικός Κώδικας:
Ασιάτης	02100
Ασιάτης	02100
Ασιάτης	02100
Ασιάτης	02100
Έγχρωμος	02100
Έγχρωμος	02100
Έγχρωμος	02100
Λευκός	**02100**

Εθνικότητα	Ταχυδρομικός Κώδικας:
Άτομο	02130
Άτομο	02130
Άτομο	02140
Άτομο	02140
Άτομο	02130
Άτομο	02140
Άτομο	02140
Άτομο	02130

Πίνακας 18: Πίνακες GT[0,1], GT[0,2], GT[1,1]

3.2.2 ΕΛΑΧΙΣΤΗ ΓΕΝΙΚΕΥΣΗ ΜΕ ΣΥΜΠΙΕΣΗ

Σύμφωνα με (Aggarwal, και συν., 2005) η τεχνική της γενίκευσης σε συνδυασμό με την τεχνική της συμπίεσης ικανοποιούν την k-ανωνυμία. Η μέθοδος της γενίκευσης είναι αντιστρόφως ανάλογη της συμπίεσης και η δεύτερη είναι αποδεκτή μέχρι το όριο MaxSup, δηλαδή εντός του ορίου αυτού, η συμπίεση θεωρείται ότι είναι καλύτερη. Αυτό συμβαίνει γιατί η γενίκευση επηρεάζει όλες τις πλειάδες ενώ η συμπίεση επηρεάζει μόνο μία πλειάδα.

Ο ορισμός της k-ελάχιστης γενίκευσης με συμπίεση είναι: Έστω ότι Ti (A1, ..., An) και Tj (A1, ..., An) είναι δύο πίνακες, έτσι ώστε Ti ≤ Tj και MaxSup είναι το όριο της συμπίεσης. Ακόμα, Tj είναι η k-ελάχιστη γενίκευση του Ti και:

- Tj ικανοποιεί την K-ανωνυμία
- Sizeof (Ti) - sizeof (Tj) ≤ MaxSup
- Δεν υπάρχει Tz για το οποίο: Ti ≤ Tz, Tz ικανοποιεί συνθήκες 1 και 2 και DVi,z<DVi,j.

Έστω ανωνυμοποιημένος πίνακας με χρήση της μεθόδου της γενίκευσης και της συμπίεσης: Ti (A1, ..., An) και Tj (A1, ..., An) είναι δύο πίνακες με ίδια γνωρίσματα. Ο πίνακας Tj θεωρούμε ότι είναι μια γενίκευση του Ti αν Tj ≤ Ti. Για όλα τα z = 1, ..., n: dom (AZ, Ti) ≤ dom (AZ, Ti). Υπάρχει Ti και Tj που συνδέει τις πλειάδες ti (σε Ti) και tj (σε TJ) έτσι ώστε ti [az] ≤ tj [az].

Με βάση όλα τα προαναφερθείσα, στη συνέχεια αναφέρουμε ένα παράδειγμα για την καλύτερη κατανόηση της ελάχιστης γενίκευσης σε συνδυασμό με την συμπίεση.

Έστω οι παρακάτω πίνακες όπου ικανοποιούν την k-ανωνυμία:

- **MaxSup** = 0: GT [1,1]

(οι πίνακες GT [1,0], GT [0,1]), ή GT [0,2], αποκρύπτουν περισσότερες πλειάδες από ότι επιτρέπεται, ο GT [1,2] δεν είναι ελάχιστος λόγω της GT [1,1].

- **MaxSup** = 1: G T [1,0] και GT [0,2]

(ο πίνακας GT [0,1] αποκρύπτει περισσότερες πλειάδες από ότι επιτρέπεται, GT [1,1] δεν είναι ελάχιστος λόγω του GT [1,0], και ο GT [1,2] δεν είναι ελάχιστος λόγω των GT[1,0] και GT [0,2].

Εθνικότητα	Ταχυδρομικός Κώδικας:
Ασιάτης	02138
Ασιάτης	02138
Ασιάτης	02142
Ασιάτης	02142
Έγχρωμος	02138
Έγχρωμος	02141
Έγχρωμος	02142
Λευκός	02138

Εθνικότητα	Ταχυδρομικός Κώδικας:
Άτομο	02138
Άτομο	02138
Άτομο	02142
Άτομο	02142
Άτομο	02138
Άτομο	02142
Άτομο	02138

Εθνικότητα	Ταχυδρομικός Κώδικας:
Ασιάτης	02130
Ασιάτης	02130
Ασιάτης	02140
Ασιάτης	02140
Έγχρωμος	02140
Έγχρωμος	02140

Πίνακας 19: Πίνακες PT, GT [1,0], GT [0,1]

- **MaxSup ≥ 2**: GT [1,0] ΚΑΙ GT [0,1]

(ο GT [0,2] δεν είναι ελάχιστος λόγω του πίνακα GT [0,1], οι πίνακες GT [1,1] και GT[1,2] δεν είναι ελάχιστοι λόγω των πινάκων GT [1,0] και GT [0,1]).

Εθνικότητα	Ταχυδρομικός Κώδικας:
Ασιάτης	02100
Ασιάτης	02100
Ασιάτης	02100
Ασιάτης	02100
Έγχρωμος	02100
Έγχρωμος	02100
Έγχρωμος	02100

Εθνικότητα	Ταχυδρομικός Κώδικας:
Άτομο	02130
Άτομο	02130
Άτομο	02140
Άτομο	02140
Άτομο	02130
Άτομο	02140
Άτομο	02140
Άτομο	02130

Πίνακας 20: Πίνακες GT [0,2], GT [1,1]

4 ΕΠΙΘΕΣΕΙΣ ΕΝΑΝΤΙΑ ΣΤΗΝ K-ΑΝΩΝΥΜΙΑ

Σε αυτή την ενότητα θα αναφερθούμε στο πως ένας κακόβουλος τρίτος, ο οποίος κατέχει γνώση για τα δεδομένα, μπορεί να ανακαλύψει τις προσωπικές πληροφορίες ενός ατόμου από τον δημοσιευμένο πίνακα, ακόμα και εάν ο πίνακας αυτός ικανοποιεί την k-ανωνυμία. Με άλλα λόγια, παρόλο που ο κάτοχος της βάσης δεδομένων έχει χρησιμοποιήσει μία από τις τεχνικές που περιγράψαμε παραπάνω με σκοπό ο πίνακας που θα δημοσιευθεί να ικανοποιεί την k-ανωνυμία, δηλαδή κάθε εγγραφή να είναι αδιάκριτη ανάμεσα σε άλλες k-1, αποδεικνύεται πώς δεν είναι αρκετό για να εμποδίσει έναν κακόβουλο τρίτο να συλλέξει πληροφορίες που αφορούν τα ιδιωτικά δεδομένα κάθε προσώπου.

Έτσι, γεγονός είναι ότι στις περιπτώσεις των επιθέσεων που θα μελετήσουμε παρακάτω, ο επιτιθέμενος προσπαθεί άμεσα ή έμμεσα να συνδέσει κάποια εγγραφή της βάσης δεδομένων με κάποιο φυσικό πρόσωπο. Ο επιτιθέμενος επιδιώκει να συσχετίσει τις πληροφορίες που έχει στην κατοχή του για κάποιο άτομο με εκείνες που εμφανίζονται στις δημοσιευμένες συλλογές δεδομένων, έτσι ώστε να συμπεράνει με ακρίβεια την ταυτότητα κάποιου ατόμου που εμφανίζεται στον δημοσιευμένο πίνακα. Ακόμα και εάν έχει εφαρμοστεί η μέθοδος της συμπίεσης ή της γενίκευσης σε έναν πίνακα προτού δημοσιευθεί έτσι ώστε κάθε εγγραφή να είναι ταυτόσημη με τουλάχιστον k-1 άλλες εγγραφές, ένας κακόβουλος τρίτος έχει την δυνατότητα να συναγάγει συμπεράσματα για την τιμή του ευαίσθητου γνωρίσματος. Επίσης, συνδυάζοντας τις τιμές του ψευδό-αναγνωριστικού που γνωρίζει με κάποιες από αυτές που εμφανίζονται στον δημοσιευμένο πίνακα, ενέχει ο κίνδυνος να ανακαλύψει πληροφορίες ως προς τα άτομα.

Επιπροσθέτως, η προσπάθεια κάποιου ατόμου, που έχει πρόσβαση στα δημοσιευμένα δεδομένα, να ανακαλύψει περισσότερα προσωπικά δεδομένα από εκείνα που γνωρίζει για κάποιο άτομο μέσω της δημοσίευσης, παραβιάζει τον απόρρητο χαρακτήρα των προσωπικών δεδομένων και καλείται επίθεση στην ιδιωτικότητα του ατόμου. Υπάρχουν πολλά είδη επιθέσεων ενάντια στην k-ανωνυμία, όπως η επίθεση αναγνώρισης πεδίου, η επίθεση αναγνώρισης τοποθεσίας, η επίθεση αναγνώρισης ταυτότητας, η επίθεση αναγνώρισης τιμής ευαίσθητων δεδομένων και άλλες. Ωστόσο, εμείς θα μελετήσουμε τις δύο τελευταίες κατηγορίες επιθέσεων στις οποίες κάνουν αναφορά η (Sweeney, 2002) και οι (Abou-el-ela, και συν., 2013), (Dalenius, 1986), (Fung, et al., 2010).

Για να γίνουν, όμως, κατανοητά τα σενάρια των επιθέσεων που θα μελετήσουμε θα πρέπει πρώτα να γίνει μία σύντομη περιγραφή των ρόλων που συμμετέχουν σε ένα τυπικό σενάριο επίθεσης. Οι ρόλοι είναι οι εξής:

Ο κάτοχος των δεδομένων/εκδότης: είναι ο οργανισμός ή το άτομο που έχει τα δεδομένα προς ανωνυμοποίηση με σκοπό την προστασία της ιδιωτικότητας.

Οι κάτοχοι των εγγραφών: είναι οντότητες που κατέχουν μια ή περισσότερες εγγραφές στο σύνολο δεδομένων που έχει επιλεγεί για δημοσίευση.

Ο αποδέκτης δεδομένων: είναι οποιοδήποτε άτομο έχει πρόσβαση στον ανωνυμοποιημένο πίνακα.

Ο επιτιθέμενος: είναι ένας κακόβουλος τρίτος που επιδιώκει να κερδίσει επιπρόσθετη γνώση από αυτή που έχει σε σχέση με τα ευαίσθητα δεδομένα ενός ατόμου.

4.1 ΕΠΙΘΕΣΗ ΑΝΑΓΝΩΡΙΣΗΣ ΤΑΥΤΟΤΗΤΑΣ

Στην πρώτη κατηγορία επιθέσεων που εξετάζεται, ο επιτιθέμενος προσπαθεί να συνδέσει μια εγγραφή του πίνακα που έχει στη διάθεση του με κάποιο φυσικό πρόσωπο. Ο επιτιθέμενος, χρησιμοποιώντας την γνώση που έχει στην κατοχή του ή συνδυάζοντας δημοσιευμένες βάσεις δεδομένων, μπορεί να ανακαλύψει την ταυτότητα κάποιου προσώπου που εμφανίζεται στον δημοσιευμένο πίνακα δεδομένων με αντιπαραβολή των τιμών του ψευδό-αναγνωριστικού του. Με άλλα λόγια, ο επιτιθέμενος προσπαθεί να ταυτοποιήσει μια εγγραφή των δεδομένων με ένα συγκεκριμένο πρόσωπο χρησιμοποιώντας τα ψευδό-αναγνωριστικά του.

Στον πίνακα 21, θεωρούμε ότι ο επιτιθέμενος γνωρίζει κάποιες τιμές του ψευδο-αναγνωριστικού, για παράδειγμα το σύνολο {1990, Άνδρας} και έτσι συμπεραίνει αμέσως ότι ο συγκεκριμένος υπάλληλος παίρνει μισθό ίσο με 500 ευρώ το μήνα.

Ημερομηνία Γεννήσεως	Ταχυδρομικός Κώδικας	Φύλο	Μισθός
199*	11855	Άνδρας	500
1993	8410*	Γυναίκα	1000
198*	26223	Γυναίκα	1500
199*	11855	Άνδρας	500
198*	26223	Γυναίκα	700
1993	8410*	Γυναίκα	750

Πίνακας 21: Στοιχεία εργαζομένων μιας εταιρείας

Εφόσον τέτοιες επιθέσεις είναι συχνές όταν το σύνολο των δεδομένων δημοσιεύεται με την αρχική του μορφή, ο τομέας της προστασίας της ιδιωτικότητας ευαίσθητων δεδομένων ασχολήθηκε αρχικά με αυτές, διερευνώντας τρόπους ώστε να εμποδίζονται. Για την αποτροπή των επιθέσεων αναγνώρισης της ταυτότητας των ατόμων, απαιτείται τροποποίηση των αρχικών δεδομένων έτσι ώστε τα δεδομένα που τελικά δημοσιεύονται να περιέχουν όση περισσότερη πληροφορία είναι δυνατό και ταυτόχρονα να προστατεύουν την ευαίσθητη πληροφορία των συμμετεχόντων. Ακόμα, για την προστασία από τέτοιου είδους επιθέσεις, τα αρχικά δεδομένα θα πρέπει να ανωνυμοποιηθούν έτσι ώστε να πληρούν ορισμένες προϋποθέσεις. Η πιο διαδεδομένη τεχνική είναι η k-ανωνυμία (k-anonymity) που εγγυάται ότι κάθε εγγραφή είναι αδιάκριτη ανάμεσα σε άλλες k-1 εγγραφές, με βάση τα ψευδό-αναγνωριστικά, που σημαίνει ότι κάθε συνδυασμός ψευδό-αναγνωριστικών θα πρέπει να εμφανίζεται μηδέν ή περισσότερες από k φορές στον ανωνυμοποιημένο πίνακα. Το σύνολο των εγγραφών με τα ίδια QI καλείται κλάση ισοδυναμίας (equivalence class). Στην περίπτωση όπου ο επιτιθέμενος γνωρίζει τα QI, η πιθανότητα να ταυτοποιήσει επιτυχώς το άτομο που αναζητά δεν είναι ποτέ μεγαλύτερη από 1/k.

Υπάρχουν οι ακόλουθες παραλλαγές όσον αφορά την τεχνική της k-ανωνυμίας:

(1, k)-ανωνυμία ((1, k)-anonymity): Όταν ο επιτιθέμενος γνωρίζει τις δημόσιες πληροφορίες του ατόμου που αναζητά, αντί να εφαρμόσει k-ανωνυμοποίηση, μπορεί να χρησιμοποιήσει την μέθοδο της γενίκευσης για να γενικεύσει τις εγγραφές του πίνακα έτσι ώστε κάθε δημόσια πληροφορία να είναι συμβατή με τουλάχιστον k εγγραφές του προς δημοσίευση πίνακα T. Σε αυτό το σημείο πρέπει να

τονιστεί ότι κάθε k-ανώνυμος πίνακας είναι και (1, k)-ανωνυμοποιημένος, χωρίς να ισχύει απαραίτητα και το αντίστροφο.

(k, 1)-ανωνυμία ((k, 1)-anonymity): Έχουμε στην περίπτωση οπού κάθε εγγραφή είναι συνεπής με τουλάχιστον k εγγραφές του αρχικού πίνακα T.

(k, k)- ανωνυμία ((k, k)- anonymity): Οι δύο παραπάνω παραλλαγές που μελετήσαμε προσφέρουν λιγότερο ισχυρή προστασία της ιδιωτικότητας σε σχέση με την k-ανωνυμία. Για αυτό το λόγο θα πρέπει να χρησιμοποιούνται σε συνδυασμό. Ένας ανώνυμος πίνακας που ικανοποιεί την (k, 1)-ανωνυμία και την (1, k)-ανωνυμία είναι ένας (k, k)-ανώνυμος πίνακας. Όταν το σενάριο επίθεσης είναι ένας επιτιθέμενος που έχει πλήρη γνώση για κάποια από τα άτομα του πίνακα, τότε η προστασία που προσφέρει ένας (k, k)-ανώνυμος πίνακας μοιάζει με την προστασία που προσφέρει η k-ανωνυμία. Η μοναδική διαφορά όταν χρησιμοποιούμαι (k, k)-ανωνυμία σε σχέση με την k-ανωνυμία είναι ότι ο κάτοχος των δεδομένων μπορεί να δει μεγαλύτερη χρησιμότητά ως προς τις πληροφορίες που δίνουν τα δεδομένα.

Μια ακόμα παραλλαγή της k-ανωνυμίας που μετασχηματίζει τα αρχικά δεδομένα σε μικρότερο βαθμό και ταυτοχρόνως μειώνει την αλλοίωση της πληροφορίας είναι η k^m-ανωνυμία. Η προϋπόθεση αυτής της τεχνικής είναι ότι κάθε συνδυασμός έως και m QIs να εμφανίζεται το λιγότερο k φορές στον δημοσιευμένο πίνακα. Η ιδέα που απεικονίζεται στην k^m -ανωνυμία είναι ότι όταν ο κακόβουλος γνωρίζει σχεδόν όλα τα πεδία μιας εγγραφής μπορεί να επιτευχθεί ελάχιστη ιδιωτικότητα και για να γίνει αυτό θα πρέπει να χαθεί μεγάλο μέρος της πληροφορίας.

Τέλος, γνωστές είναι και οι μέθοδοι που προσφέρουν αμοιβαία χωρική k-ανωνυμία (reciprocal spatial k-anonymity) για την καταπολέμηση επιθέσεων ελαχιστοποίησης (minimality attacks). Το σενάριο αυτής της επίθεσης είναι ότι υπάρχει ένας επιτιθέμενος ο οποίος συγκρίνει τα CRs όλων των k συμμετεχόντων που περιλαμβάνονται σε ένα συγκεκριμένο CR ώστε να εντοπίσει τις διαφορές τους και έτσι να προσδιορίσει μοναδικά τους χρήστες που περιλαμβάνονται σε αυτό. Τα CR(cloaking region-περιοχή απόκρυψης) είναι μία μεθοδολογία κατά την οποία αντικαθιστάται η ακριβής περιοχή του ατόμου με μία ευρύτερη περιοχή. Ωστόσο, αν και μόνο αν το ίδιο CR έχει παραχθεί για καθέναν από τους k χρήστες τότε μόνο ο επιτιθέμενος δεν μπορεί να προσδιορίσει τα άτομα.

4.2 Επίθεση Αναγνώρισης Τιμής Ευαίσθητων Δεδομένων

Στη δεύτερη κατηγορία επιθέσεων, ο επιτιθέμενος προσπαθεί να εντοπίσει τις τιμές των ευαίσθητων γνωρισμάτων κάποιου πίνακα οντοτήτων. Για παράδειγμα, θεωρούμε ότι ένας κακόβουλος τρίτος γνωρίζει ότι το άτομο που αναζητά συμμετέχει στον δημοσιευμένο πίνακα, καθώς και κάποιες από τις τιμές που λαμβάνει σε κάποια από τα γνωρίσματα του συνόλου δεδομένων. Οι πληροφορίες αυτές είναι αρκετές για να μπορέσει ο επιτιθέμενες να συνδέσει το άτομο με κάποια εγγραφή και ταυτόχρονα να διεξάγει συμπεράσματα για τις τιμές των γνωρισμάτων που αναφέρονται σε απόρρητη αλλά και εξολοκλήρου προσωπική πληροφορία.

Τέλος, οι επιθέσεις που αφορούν την αναγνώριση τιμής ευαίσθητων δεδομένων είναι δύσκολο να εμποδιστούν ή τουλάχιστον να αντιμετωπιστούν με τη χρήση της k-ανωνυμίας, όσο καλά και αν επιλεγούν τα ψευδό-αναγνωριστικά.

4.2.1 ΕΠΙΘΕΣΗ ΟΜΟΙΟΓΕΝΕΙΑΣ

Η επίθεση της ομοιογένειας εκμεταλλεύεται το γεγονός ότι στον πίνακα που δίνεται προς δημοσίευση υπάρχει περίπτωση να υπάρχουν εγγραφές με ταυτόσημες τιμές στο ευαίσθητο γνώρισμα. Έτσι, όταν ένα άτομο γνωρίζει μία ή περισσότερες τιμές του ψευδό-αναγνωριστικού, μπορεί εύκολα να συμπεράνει την τιμή του ευαίσθητου γνωρίσματος χωρίς να προβεί σε ενέργειες σύγκρισης με παραπλήσιες εγγραφές.

Μια τέτοια περίπτωση παρουσιάζεται στο παράδειγμα του Πίνακα 22, ο οποίος προκύπτει από την 3-ανωνυμοποίηση του Πίνακα 21. Αν ο επιτιθέμενος γνωρίζει πως το άτομο που αναζητά συμμετέχει στα δημοσιευμένα δεδομένα και για παράδειγμα γνωρίζει πως έχει ηλικία μεγαλύτερη ή ίση από 35 ετών, μπορεί να συμπεράνει με βεβαιότητα πως ο Μισθός του ατόμου αυτού θα είναι ίσος με 1200 ευρώ το μήνα.

Ημερομηνία Γεννήσεως	Ταχυδρομικός Κώδικας	Φύλο	Μισθός
25	26331	Άνδρας	500
32	26332	Γυναίκα	1000
33	26331	Γυναίκα	1500
40	26335	Άνδρας	1200
28	26334	Γυναίκα	500
51	26334	Γυναίκα	1200

Πίνακας 22: Στοιχεία εργαζομένων μιας εταιρείας

Ημερομηνία Γεννήσεως	Ταχυδρομικός Κώδικας	Φύλο	Μισθός
<= 35	[26331,26335]	*	500
<= 35	[26331,26335]	*	1000
<= 35	[26331,26335]	*	1500
>=35	[26331,26335]	*	1200
<= 35	[26331,26335]	*	500
>=35	[26331,26335]	*	1200

Πίνακας 23: Αποτελεί την 3-ανωνυμοποίηση του Πίνακα 3

Έτσι συμπεραίνουμε για ακόμη μία φορά ότι το μοντέλο της k-ανωνυμίας είναι ανεπαρκή και για το λόγο αυτό παραβιάζεται το απόρρητο των προσωπικών δεδομένων κατά την διάρκεια της επίθεσης. Επιπλέον, πρέπει να σημειωθεί ότι η επίθεση της ομοιογένειας είναι συχνό κρούσμα σε βάσεις δεδομένων που αφορούν εισοδήματα ή ακόμα και σε συλλογές δεδομένων με προσωπικά ιατρικά δεδομένα ατόμων τα οποία δημοσιεύονται σε ερευνητές, σε ιστοσελίδες ή φτάνουν στην κατοχή άλλων χωρίς την απαραίτητη δικαιοδοσία για κερδοσκοπικούς λόγους.

4.2.2 Επιθεση Με Προτερη Γνωση

Μια ακόμα περίπτωση στην οποία εμφανίζεται η ανεπάρκεια της k-ανωνυμίας ως προς την προστασία της ιδιωτικότητας των δεδομένων είναι η επίθεση με πρότερη γνώση. Σε αυτή την περίπτωση, ο επιτιθέμενος εκτός από τις πληροφορίες που έχει στην κατοχή του, γνωρίζει με βεβαιότητα την συμμετοχή κάποιου ατόμου στον δημοσιευμένο πίνακα και έτσι μπορεί να αποκλείσει κάποιες ευαίσθητες τιμές από την κλάση ισοδυναμίας στην οποία θεωρεί πως πιθανώς ανήκει το άτομο που αναζητά. Με αυτό τον τρόπο του δίνεται η δυνατότητα να συμπεράνει την ισχύουσα τιμή του ευαίσθητου γνωρίσματος για το συγκεκριμένο άτομο.

Ένα τέτοιο παράδειγμα προκύπτει από τα δεδομένα του Πίνακα 18. Θεωρώντας τα ως το σύνολο δεδομένων που δημοσιεύει μια εταιρία για τους υπαλλήλους της, ο επιτιθέμενος μπορεί να αναζητά τον μισθό που αντιστοιχεί σε έναν νέο συνάδελφό του. Γνωρίζει πως αυτός θα εμφανίζεται στον δημοσιευμένο πίνακα, αφού εργάζονται στην ίδια εταιρεία και αντίστοιχα γνωρίζει πως η ηλικία του, ως νέος εργαζόμενος, είναι μικρότερη ή ίση των 35 ετών. Επίσης, ο επιτιθέμενος μπορεί με βεβαιότητα να αποκλείσει την τιμή 1000 και 1500 μιας και γνωρίζει πως αυτοί οι μισθοί αντιστοιχούν σε εργαζόμενους με επαγγελματική εμπειρία. Έτσι προκύπτει ότι ο νέος συνάδελφός του λαμβάνει μισθό ίσο με 500, αφού έχει αποκλείσει όλες τις υπόλοιπες τιμές του γνωρίσματος «Μισθός».

Τελικά, προκύπτει ότι στην επίθεση της ομοιογένειας καθώς και στην επίθεση με πρότερη γνώση η k-ανωνυμία όχι μόνο δεν εξασφαλίζει την προστασία της ιδιωτικότητας των ευαίσθητων δεδομένων αλλά αφήνει και δυνατότητες εκμετάλλευσης των δεδομένων καθώς και παραβίασης της ιδιωτικής ζωής των ατόμων που συμμετέχουν σε δημοσιευμένες συλλογές δεδομένων. Επίσης, εξαιτίας της κατανομής των ευαίσθητων τιμών του γνωρίσματος και της γνώσης που έχει ο επιτιθέμενος σχετικά με τις ευαίσθητες τιμές, υπάρχει ο κίνδυνος να μπορεί να επιβεβαιώσει ή να αποκλείσει κάποιες από αυτές για το πρόσωπο που αναζητά. Με τον τρόπο αυτό επιτυγχάνει την συσχέτιση μεταξύ των εγγραφών και διεξάγει συμπεράσματα για την τιμή του ευαίσθητου γνωρίσματος.

Με στόχο την κάλυψη τέτοιων περιπτώσεων προτάθηκε η έννοια της l-διαφορετικότητας (l-diversity), η οποία μπορεί να κατοχυρώσει και να εγγυηθεί την ιδιωτικότητα στις παραπάνω περιπτώσεις ελέγχοντας το πλήθος των διαφορετικών μεταξύ τους τιμών που λαμβάνει το ευαίσθητο γνώρισμα σε κάθε κλάση ισοδυναμίας. Με άλλα λόγια, η l-διαφορετικότητα είναι επέκταση της k-ανωνυμίας και στόχο έχει να ορίσει το πλήθος των ευαίσθητων τιμών σε κάθε κλάση ισοδυναμίας έτσι ώστε να μην μπορεί ένας κακόβουλος να συνδέσει μία εγγραφή με μία ευαίσθητη τιμή.

4.2.3 Επιθεση Μη Ταξινομημενης Αντιστοιχισης

Αυτή η επίθεση είναι βασισμένη στη σειρά στην οποία οι πλειάδες εμφανίζονται στον ανωνυμοποιημένο πίνακα. Στο παράδειγμα που ακολουθεί οι πίνακες GT1 και GT2 προκύπτουν από την ανωνυμοποίηση του αρχικού πίνακα PT χρησιμοποιώντας την τεχνική της γενίκευσης. Οι πίνακες GT1 και GT2 ικανοποιούν k-Ανωνυμία για k=2. Η σειρά εμφάνισης των πλειάδων του πίνακα PT, GT1 και GT2 είναι η ίδια. Εάν δημοσιευθεί πρώτα ο πίνακας GT1 και μετά ο πίνακας GT2, είναι δυνατό με μια σύνδεση να κατασκευαστούν όλες οι εγγραφές του αρχικού πίνακα και έτσι να

αποκαλυφθεί η ευαίσθητη πληροφορία. Αυτό το πρόβλημα δεν θα υπήρχε εάν είχε γίνει τυχαία ταξινόμηση των πλειάδων πριν από την δημοσίευση του πίνακα. (Σπίνου, Σεπτέμβριος 2011)

PT	
Φυλή	**Τ.Κ.**
Ασιάτης	02138
Ασιάτης	02139
Ασιάτης	02141
Ασιάτης	02142
Έγχρωμος	02138
Έγχρωμος	02139
Έγχρωμος	02141
Έγχρωμος	02142
Λευκός	02138
Λευκός	02139
Λευκός	02141
Λευκός	02142

GT1	
Φυλή	**Τ.Κ.**
Άτομο	02138
Άτομο	02139
Άτομο	02141
Άτομο	02142
Άτομο	02138
Άτομο	02139
Άτομο	02141
Άτομο	02142
Άτομο	02138
Άτομο	02139
Άτομο	02141
Άτομο	02142

GT2	
Φυλή	**Τ.Κ.**
Ασιάτης	02130
Ασιάτης	02130
Ασιάτης	02140
Ασιάτης	02140
Έγχρωμος	02130
Έγχρωμος	02130
Έγχρωμος	02140
Έγχρωμος	02140
Λευκός	02130
Λευκός	02130
Λευκός	02140
Λευκός	02140

Πίνακας 24: Παράδειγμα Επίθεσης μη Ταξινομημένης Αντιστοίχησης

4.2.4 Επιθεση Συμπληρωματικης Έκδοσης

Έστω ότι έχουμε έναν αρχικό πίνακα PT από τον οποίο προκύπτουν οι πίνακες GT1 και GT2 οι οποίοι ικανοποιούν την k-ανωνυμία για k=2. Εάν δημοσιευθεί πρώτα ο πίνακας GT1 και ύστερα ο πίνακας GT2 η προστασία της k-ανωνυμίας δεν μπορεί να εγγυηθεί ακόμη και αν η σειρά των πλειάδων είναι διαφορετική.

PT				
Φυλή	**Ημερομηνία Γεννήσεως**	**Φύλο**	**Τ.Κ.**	**Ασθένεια**
Έγχρωμος	9/20/1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	2/14/1965	Άνδρας	02141	Πόνος στο στήθος
Έγχρωμος	10/23/1965	Γυναίκα	02138	Πόνος στο μάτι
Έγχρωμος	8/24/1965	Γυναίκα	02138	Βρογχιολίτιδα
Έγχρωμος	11/7/1964	Γυναίκα	02138	Παχυσαρκία
Έγχρωμος	12/1/1964	Γυναίκα	02138	Πόνος στο στήθος
Λευκός	10/23/1964	Άνδρας	02138	Δύσπνοια
Λευκός	3/15/1965	Γυναίκα	02139	Υπέρταση
Λευκός	8/13/1964	Άνδρας	02139	Παχυσαρκία
Λευκός	5/5/1964	Άνδρας	02139	Πυρετός
Λευκός	2/13/1967	Άνδρας	02138	Εμετός
Λευκός	3/21/1967	Άνδρας	02138	Πόνος στην πλάτη

LT				
Φυλή	**Ημερομηνία Γεννήσεως**	**Φύλο**	**Τ.Κ.**	**Ασθένεια**
Έγχρωμος	1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	1965	Άνδρας	02141	Πόνος στο στήθος
Έγχρωμος	1965	Γυναίκα	02138	Πόνος στο μάτι
Έγχρωμος	1965	Γυναίκα	02138	Βρογχιολίτιδα
Έγχρωμος	1964	Γυναίκα	02138	Παχυσαρκία
Έγχρωμος	1964	Γυναίκα	02138	Πόνος στο στήθος
Λευκός	1964	Άνδρας	02138	Δύσπνοια
Λευκός	1965	Γυναίκα	02139	Υπέρταση
Λευκός	1964	Άνδρας	02139	Παχυσαρκία
Λευκός	1964	Άνδρας	02139	Πυρετός
Λευκός	1967	Άνδρας	02138	Εμετός
Λευκός	1967	Άνδρας	02138	Πόνος στην πλάτη

Πίνακας 25: Αρχικός Πίνακας PT και Πίνακας LT

Αυτό συμβαίνει γιατί αν συνδυαστούν οι πίνακες GT1 και GT2 με βάση το γνώρισμα {Ασθένεια}, μπορεί να κατασκευαστεί ο πίνακας LT (Linked Table) ο οποίος δεν ικανοποιεί k-ανωνυμία αφού τα γνωρίσματα [Λευκός, 1964, Άνδρας, 02138, Δύσπνοια] και [Λευκός, 1965, Γυναίκα, 02139, Υπέρταση] είναι μοναδικά. Αυτό το πρόβλημα δεν θα υπήρχε εάν ο πίνακας GT2 είχε κατασκευαστεί με ψευδό-αναγνωριστικό το σύνολο QI∪{Ασθένεια}, ή αν ο GT2 είχε σαν βάση τον GT1.

GT1				
Φυλή	**Ημερομηνία Γεννήσεως**	**Φύλο**	**Τ.Κ.**	**Ασθένεια**
Έγχρωμος	1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	1965	Άνδρας	02141	Πόνος στο στήθος
Άτομο	1965	Γυναίκα	0213*	Πόνος στο μάτι
Άτομο	1965	Γυναίκα	0213*	Βρογχιολίτιδα
Έγχρωμος	1964	Γυναίκα	02138	Παχυσαρκία
Έγχρωμος	1964	Γυναίκα	02138	Πόνος στο στήθος
Λευκός	1964	Άνδρας	0213*	Δύσπνοια
Άτομο	1965	Γυναίκα	0213*	Υπέρταση
Λευκός	1964	Άνδρας	0213*	Παχυσαρκία
Λευκός	1964	Άνδρας	0213*	Πυρετός
Λευκός	1967	Άνδρας	02138	Εμετός
Λευκός	1967	Άνδρας	02138	Πόνος στην πλάτη

GT2				
Φυλή	**Ημερομηνία Γεννήσεως**	**Φύλο**	**Τ.Κ.**	**Ασθένεια**
Έγχρωμος	1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	1965	Άνδρας	02141	Πόνος στο στήθος
Έγχρωμος	1965	Γυναίκα	02138	Πόνος στο μάτι
Έγχρωμος	1965	Γυναίκα	02138	Βρογχιολίτιδα
Έγχρωμος	1964	Γυναίκα	02138	Παχυσαρκία
Έγχρωμος	1964	Γυναίκα	02138	Πόνος στο στήθος
Λευκός	1960-69	Άνδρας	02138	Δύσπνοια
Λευκός	1960-69	Άνθρωπος	02139	Υπέρταση
Λευκός	1960-69	Άνθρωπος	02139	Παχυσαρκία
Λευκός	1960-69	Άνθρωπος	02139	Πυρετός
Λευκός	1960-69	Άνδρας	02138	Εμετός
Λευκός	1960-69	Άνδρας	02138	Πόνος στην πλάτη

Πίνακας 26: Πίνακας GT1 και GT2 που ικανοποιούν την k-ανωνυμία για k=2

4.2.5 ΧΡΟΝΙΚΗ ΕΠΙΘΕΣΗ

Αυτή η επίθεση αφορά την δυναμική αλλαγή των στοιχείων του πίνακα, δηλαδή μπορεί οποιαδήποτε χρονική στιγμή να προστεθεί, να αφαιρεθεί ή ακόμα και να αλλάξει μια πλειάδα. Στο παρακάτω παράδειγμα υπάρχει ένας αρχικός πίνακας PT. Μετά από ανωνυμοποίηση του PT προκύπτει ο πίνακας RT που ικανοποιεί την k-ανωνυμία.

PT				
Φυλή	**Ημερομηνία Γεννήσεως**	**Φύλο**	**Τ.Κ.**	**Ασθένεια**
Έγχρωμος	9/20/1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	2/14/1965	Άνδρας	02141	Πόνος στο στήθος
Έγχρωμος	10/23/1965	Γυναίκα	02138	Πόνος στο μάτι
Έγχρωμος	8/24/1965	Γυναίκα	02138	Βρογχιολίτιδα
Έγχρωμος	11/7/1964	Γυναίκα	02138	Παχυσαρκία
Έγχρωμος	12/1/1964	Γυναίκα	02138	Πόνος στο στήθος
Λευκός	10/23/1964	Άνδρας	02138	Δύσπνοια
Λευκός	3/15/1965	Γυναίκα	02139	Υπέρταση
Λευκός	8/13/1964	Άνδρας	02139	Παχυσαρκία
Λευκός	5/5/1964	Άνδρας	02139	Πυρετός
Λευκός	2/13/1967	Άνδρας	02138	Εμετός
Λευκός	3/21/1967	Άνδρας	02138	Πόνος στην πλάτη

Πίνακας 27: Αρχικός πίνακας PT

Έστω ότι σε χρόνο t προσθέτω στον αρχικό πίνακα PT δύο εγγραφές και προκύπτει ένας νέος πίνακας PTt= PT∪{Λευκός, 10/10/67, Άνδρας, 02139, Πυρετός} {Λευκός, 10/21/67, Άνδρας, 02141, Εμετός}. Με ανωνυμοποίηση του PTt προκύπτει ο RTt που ικανοποιεί k-ανωνυμία.

RT				
Φυλή	**Ημερομηνία Γεννήσεως**	**Φύλο**	**Τ.Κ.**	**Ασθένεια**
Έγχρωμος	1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	1965	Άνδρας	02141	Πόνος στο στήθος
Άτομο	1965	Γυναίκα	0213*	Πόνος στο μάτι
Άτομο	1965	Γυναίκα	0213*	Βρογχιολίτιδα
Έγχρωμος	1964	Γυναίκα	02138	Παχυσαρκία
Έγχρωμος	1964	Γυναίκα	02138	Πόνος στο στήθος
Λευκός	1964	Άνδρας	0213*	Δύσπνοια
Άτομο	1965	Γυναίκα	0213*	Υπέρταση
Λευκός	1964	Άνδρας	0213*	Παχυσαρκία
Λευκός	1964	Άνδρας	0213*	Πυρετός
Λευκός	1967	Άνδρας	02138	Εμετός
Λευκός	1967	Άνδρας	02138	Πόνος στην πλάτη

Πίνακας 28: Πίνακας RT που ικανοποιεί την k-ανωνυμία

Λόγω του ότι δεν υπάρχει καμία εγγύηση η οποία να εξασφαλίζει ότι ο πίνακας RTt έχει σαν βάση τον RT, τότε η σύνδεση των δύο πινάκων μπορεί να μην ακολουθεί την k-ανωνυμία. Τέλος, αυτό το πρόβλημα μπορεί να επιλυθεί εάν οι λύσεις της k-ανωνυμίας θεωρηθούν ως ένωση εξωτερικών πληροφοριών δηλαδή: RT∪(PTt-PT).

PTt				
Φυλή	**Ημερομηνία Γεννήσεως**	**Φύλο**	**Τ.Κ.**	**Ασθένεια**
Έγχρωμος	9/20/1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	2/14/1965	Άνδρας	02141	Πόνος στο στήθος
Έγχρωμος	10/23/1965	Γυναίκα	02138	Πόνος στο μάτι
Έγχρωμος	8/24/1965	Γυναίκα	02138	Βρογχιολίτιδα
Έγχρωμος	11/7/1964	Γυναίκα	02138	Παχυσαρκία
Έγχρωμος	12/1/1964	Γυναίκα	02138	Πόνος στο στήθος
Λευκός	10/23/1964	Άνδρας	02138	Δύσπνοια
Λευκός	3/15/1965	Γυναίκα	02139	Υπέρταση
Λευκός	8/13/1964	Άνδρας	02139	Παχυσαρκία
Λευκός	5/5/1964	Άνδρας	02139	Πυρετός
Λευκός	2/13/1967	Άνδρας	02138	Εμετός
Λευκός	3/21/1967	Άνδρας	02138	Πόνος στην πλάτη
Λευκός	10/10/1967	Άνδρας	02139	Πυρετός
Λευκός	10/21/1967	Άνδρας	02141	Εμετός

RTts				
Φυλή	**Ημερομηνία Γεννήσεως**	**Φύλο**	**Τ.Κ.**	**Ασθένεια**
Έγχρωμος	1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	1965	Άνδρας	02141	Πόνος στο στήθος
Έγχρωμος	1965	Γυναίκα	02138	Πόνος στο μάτι
Έγχρωμος	1965	Γυναίκα	02138	Βρογχιολίτιδα
Έγχρωμος	1964	Γυναίκα	02138	Παχυσαρκία
Έγχρωμος	1964	Γυναίκα	02138	Πόνος στο στήθος
Λευκός	1960-69	Άνδρας	02138	Δύσπνοια
Λευκός	1960-69	Άνθρωπος	02139	Υπέρταση
Λευκός	1960-69	Άνθρωπος	02139	Παχυσαρκία
Λευκός	1960-69	Άνθρωπος	02139	Πυρετός
Λευκός	1960-69	Άνδρας	02138	Εμετός
Λευκός	1960-69	Άνδρας	02138	Πόνος στην πλάτη
Λευκός	1960-69	Άνδρας	02139	Πυρετός
Λευκός	1960-69	Άνδρας	02141	Εμετός

Πίνακας 29: Πίνακας PTt προκύπτει από την εισαγωγή δυο νέων εγγραφών στον PT και Πίνακας RTt προκύπτει από την ανωνυμοποίηση του πίνακα PTt

5 ΜΕΘΟΔΟΙ ΑΝΤΙΜΕΤΩΠΙΣΗΣ ΕΠΙΘΕΣΕΩΝ

5.1 L-ΔΙΑΦΟΡΕΤΙΚΟΤΗΤΑ (L-DIVERSITY)

Η επίθεση ομοιογένειας και η επίθεση με πρότερη γνώση που αναφέρθηκαν στο προηγούμενο κεφάλαιο αντιπροσωπεύουν τις εγγυήσεις ιδιωτικότητας που αφορούν την ασφάλεια της ταυτότητας μιας εγγραφής και τη διασφάλιση ότι ο επιτιθέμενος δεν θα μπορέσει να ανακαλύψει τις ευαίσθητες τιμές ενός φυσικού προσώπου, όπως αναφέρονται και παρουσιάζονται από την (Λεπενιώτη, 2013). Έτσι, εάν ο επιτιθέμενος κατέχει γνώση για τα δεδομένα και την κατανομή των ευαίσθητων τιμών, τότε ένας δημοσιευμένος πίνακας δεδομένων μπορεί να επιτρέπει την διεξαγωγή πληροφορίας με τους εξής τρόπους:

- Θετική αποκάλυψη: Είναι η κατάσταση όπου ένας κακόβουλος τρίτος μπορεί να αναγνωρίσει σωστά την τιμή του ευαίσθητου γνωρίσματος με μεγάλη πιθανότητα.

- Αρνητική αποκάλυψη: Είναι η κατάσταση όπου ένας επιτιθέμενος μπορεί με σιγουριά να αποκλείσει κάποιες από τις ευαίσθητες τιμές του γνωρίσματος.

Η l-διαφορετικότητα αποτελεί επέκταση της k-ανωνυμίας. Η μέθοδος αυτή, έχει ως στόχο την αποτροπή των επιθέσεων αναγνώρισης της τιμής του ευαίσθητου γνωρίσματος αλλά και την διασφάλιση ότι ο επιτιθέμενος δεν θα μπορέσει να βρει προσωπικά στοιχεία για ένα φυσικό πρόσωπο. Κατά την εφαρμογή της l-διαφορετικότητας, τα δεδομένα διαχωρίζονται σε ευαίσθητα και μη και ο επιτιθέμενος πιστεύει πως ο δημοσιευμένος πίνακας αποτελεί γενίκευση κάποιου αρχικού πίνακα, από όπου προσπαθεί να ανακαλύψει την ευαίσθητη τιμή κάποιας εγγραφής.

Σε όλες τις βάσεις δεδομένων υπάρχει μια πιθανότητα παραβίασης του απορρήτου των ευαίσθητων δεδομένων από οποιονδήποτε επιτιθέμενο καταφέρει να έχει πρόσβαση σε αυτά. Η l-διαφορετικότητα προσπαθεί να ορίσει το πλήθος των ευαίσθητων τιμών σε κάθε κλάση ισοδυναμίας έτσι ώστε να μην μπορεί κάποιος κακόβουλος τρίτος να συσχετίσει μία εγγραφή με μία τιμή. Έτσι ακόμα και εάν ο επιτιθέμενος γνωρίζει την κλάση ισοδυναμίας στην οποία ανήκει το άτομο που αναζητά, δεν μπορεί να ανακαλύψει την ευαίσθητη τιμή της αφού η εγγραφή θα συσχετίζεται με τουλάχιστον l ευαίσθητες τιμές που εμφανίζονται στην συγκεκριμένη κλάση ισοδυναμίας.

Ο ορισμός της l-διαφορετικότητας δηλώνει ότι ο αρχικός πίνακας δεδομένων $RT(A_1,\ldots A_n,S)$ περιέχει ένα ευαίσθητο γνώρισμα S και μετά από γενίκευση προκύπτει ο πίνακας $RT^*(A_1,\ldots A_n,S)$. Οι εγγραφές χωρίζονται σε κλάσεις ισοδυναμίας (q* -blocks) ως προς τις τιμές των γνωρισμάτων του ψευδό-αναγνωριστικού τους. Μια κλάση ισοδυναμίας ορίζεται ως l-διαφορετική (l-diverse) αν

περιέχει τουλάχιστον l >= 2 «καλώς ορισμένες τιμές» για το ευαίσθητο γνώρισμα και ένας πίνακας είναι l-διαφορετικός εάν κάθε κλάση ισοδυναμίας του είναι l-διαφορετική.

Συμπερασματικά, η l-διαφορετικότητα προτιμάται σε σύγκριση με την k-ανωνυμία, όταν το μοντέλο δεδομένων εμφανίζει ευαίσθητα γνωρίσματα επειδή:

- Δεν είναι αναγκαίο ο κάτοχος των δεδομένων να έχει την ίδια γνώση με τον επιτιθέμενο για να αποτρέψει την επίθεση γιατί καμία εγγραφή δεν μπορεί να συσχετιστεί μοναδικά με μία συγκεκριμένη ευαίσθητη τιμή.
- Όσο μεγαλύτερο είναι το l, τόσο περισσότερη πρότερη γνώση πρέπει να έχει ο επιτιθέμενος σχετικά με το ευαίσθητο γνώρισμα για να εξάγει συμπεράσματα.
- Καλύπτει τις περιπτώσεις επίθεσης προς μεμονωμένα πρόσωπα, αφού μία εγγραφή μπορεί να συνδεθεί μόνο με τις ευαίσθητες τιμές της κλάσης ισοδυναμίας στην οποία ανήκει.

5.2 T-Εγγυητητα (T-Closeness)

Η t–εγγύτητα είναι μια έννοια ιδιωτικότητας που μπορεί να αναπαραστήσει το γενικό γνωστικό υπόβαθρο του επιτιθέμενου πάνω στην κατανομή των τιμών του ευαίσθητου γνωρίσματος. Ακόμα, η μέθοδος αυτή ορίστηκε για την αποτροπή των περιπτώσεων επίθεσης αλλοίωσης και ομοιότητας. Η πρώτη επίθεση αντιμετωπίζεται συχνά σε βάσεις που περιέχουν ιατρικά δεδομένα όπου το ευαίσθητο γνώρισμα αποτελεί μία ασθένεια. Αντίστοιχα, η δεύτερη επίθεση εμφανίζεται σε βάσεις δεδομένων που αφορούν για παράδειγμα τη μισθοδοσία υπαλλήλων όπου σε μία κλάση ισοδυναμίας, οι εγγραφές εμφανίζουν κοντινές τιμές μεταξύ τους και έτσι συμπεραίνουν ότι η κλάση αυτή αφορά τους υψηλόμισθους εργαζόμενους.

Η διαφορά της t-εγγύτητας με την l-διαφορετικότητα, αν και οι δύο λαμβάνουν υπόψη την πρότερη γνώση του επιτιθέμενου, είναι πως η l-διαφορετικότητα προσπαθεί 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. Ενώ η t-εγγύτητα προσπαθεί να περιορίσει την διαφοροποίηση στην γνώση του επιτιθέμενου μεταξύ της γνώσης που αποκτά από τον δημοσιευμένο πίνακα 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.

5.3 Anatomia (Anatomy)

Κατά την ανωνυμοποίηση των δεδομένων, εμφανίζεται το πρόβλημα της απώλειας της πληροφορίας, με αποτέλεσμα να μην μπορούν να αξιοποιηθούν τα δεδομένα. Το γεγονός αυτό

οφείλεται στις γενικεύσεις που προκαλούνται στα δεδομένα προκειμένου να δημιουργηθούν κλάσεις ισοδυναμίας και ταυτόχρονα στο ότι δεν προστατεύεται η συσχέτιση της κάθε εγγραφής με την ευαίσθητη τιμή της.

Η μέθοδος της ανατομίας έχει στόχο την μείωση της απώλειας πληροφορία οπού το πετυχαίνει καθώς στα δημοσιευμένα δεδομένα παραμένουν οι αρχικές τιμές τόσο των τιμών του ψευδό-αναγνωριστικού, όσο και των τιμών του ευαίσθητου γνωρίσματος. Αυτό που αποκρύπτεται, είναι η συσχέτιση κάθε εγγραφής με την ευαίσθητη τιμή της.

Στον παρακάτω παράδειγμα έχουμε έναν πίνακα με σύνολο γνωρισμάτων {Ηλικία, Ταχυδρομικός Κώδικας, Φύλο} και ευαίσθητο γνώρισμα το {Μισθός}. Ο πίνακας ανωνυμοποιείται με την τεχνική της γενίκευσης έτσι ώστε να δημιουργούνται κλάσεις ισοδυναμίας που να ικανοποιούν την 4-ανωνυμία και 3-διαφορετικότητα.

Α/Α	Ηλικία	Ταχυδρομικός Κώδικας	Φύλο	Μισθός
1	25	10443	Άνδρας	500
2	26	10555	Άνδρας	1100
3	27	10443	Άνδρας	1100
4	30	10551	Γυναίκα	820
5	42	11851	Γυναίκα	950
6	42	11853	Άνδρας	920
7	51	11257	Άνδρας	920
8	45	11142	Άνδρας	700

Πίνακας 30: Αρχικός πίνακας

Α/Α	Ηλικία	Ταχυδρομικός Κώδικας	Φύλο	Μισθός
1	<=30	[10443-10555]	*	500
2	<=30	[10443-10555]	*	1100
3	<=30	[10443-10555]	*	1100
4	<=30	[10443-10555]	*	820
5	>30	[11142-11853]	*	950
6	>30	[11142-11853]	*	920
7	>30	[11142-11853]	*	920
8	>30	[11142-11853]	*	700

Πίνακας 31: Ανωνυμοποιημένος πίνακας 4-ανωνυμίας και 3-διαφορετικότητας

Ύστερα, δημιουργείται ένας καινούριος πίνακας με τις αρχικές τιμές για τα γνωρίσματα του ψευδό-αναγνωριστικού προσθέτοντας μια καινούρια στήλη η οποία περιέχει τον αριθμό της κλάσης ισοδυναμίας που ανήκει η κάθε εγγραφή. Τέλος, δημιουργείται ένας συγκεντρωτικός πίνακας στον οποίο περιέχονται οι αρχικές τιμές του ευαίσθητου γνωρίσματος, ο αριθμός της κλάσης ισοδυναμίας και ο αριθμός εμφανίσεων της συγκεκριμένης τιμής μέσα στην κλάση ισοδυναμίας.

Α/Α	Ηλικία	Ταχυδρομικός Κώδικας	Φύλο	Μισθός
1	25	10443	Άνδρας	1
2	26	10555	Άνδρας	1
3	27	10443	Άνδρας	1
4	30	10551	Γυναίκα	1
5	42	11851	Γυναίκα	2
6	42	11853	Άνδρας	2
7	51	11257	Άνδρας	2
8	45	11142	Άνδρας	2

Πίνακας 32: Αρχικός πίνακας με αριθμό κλάσης ισοδυναμίας

Ομάδα	Μισθός	Αριθμός Εμφανίσεων
1	500	1
1	1100	2
1	820	1
2	950	1
2	920	2
2	700	1

Πίνακας 33: Πίνακας ευαίσθητων τιμών του αρχικού πίνακα

Με τη χρήση της ανατομίας και τη δημοσίευση των δύο τελευταίων πινάκων, ο επιτιθέμενος γνωρίζοντας κάποιες τιμές του ψευδό-αναγνωριστικού, μπορεί μεν να προσδιορίσει αν το άτομο που αναζητά ανήκει σε κάποια εγγραφή, αλλά δεν μπορεί να συσχετίσει με απόλυτη βεβαιότητα καμία εγγραφή με την ευαίσθητη τιμή της κλάσης ισοδυναμίας που ανήκει, αφού κάθε ομάδα ικανοποιεί την 3-διαφορετικότητα.

Γενικά, η ανατομία προτιμάται αντί της γενίκευσης και εφαρμόζεται στις περιπτώσεις όπου ο επιτιθέμενος γνωρίζει τις τιμές του ψευδό-αναγνωριστικού μιας εγγραφής και είναι βέβαιος ότι το άτομο που αναζητά βρίσκεται στις δημοσιευμένες εγγραφές.

5.4 M-ΑΜΕΤΑΒΛΗΤΟΤΗΤΑ (M-INVARIANCE)

Κατά την εφαρμογή των μεθόδων l-διαφορετικότητα, ανατομία και t-εγγυήτητα δεν υπάρχει εγγύηση ιδιωτικότητας για τα δυναμικά δεδομένα. Καμία από τις παραπάνω μεθοδολογίες δεν έχει τη δυνατότητα να δημοσιεύσει τα νέα δεδομένα μετά από πιθανές αλλαγές στη βάση δεδομένων. Αυτό αποτελεί σοβαρό πρόβλημα για μια βάση δεδομένων η οποία πρέπει να μένει πάντα ενημερωμένη. Έτσι, για την επίλυση αυτού του προβλήματος κατάλληλη είναι η χρήση της μεθοδολογίας m-Invariance. Δηλαδή, η m-Invariance αποτελεί επέκταση της μεθόδου l-diversity έτσι ώστε να μπορούν να ανωνυμοποιηθούν τόσο τα σταθερά όσο και τα δυναμικά δεδομένα.

Έστω ότι ένα νοσοκομείο δημοσιεύει τα δεδομένα των ασθενών του κάθε εξάμηνο. Ο πίνακας RT(1) είναι ο αρχικός πίνακας ενώ ο πίνακας RT(1)* είναι ο ανωνυμοποιημένος πίνακας. Μετά από έξι μήνες με βάση τα δεδομένα του αρχικού πίνακα RT(2) δημοσιεύεται ο πίνακας RT*(2).

Όνομα	Ηλικία	Ταχυδρομικός Κώδικας	Ασθένεια
Γιάννης	20	10443	Πόνος στη πλάτη
Μάκης	22	10551	Πόνος στο χέρι
Έλενα	30	10556	Γρίπη
Κατερίνα	31	10558	Γαστρεντερίτιδα
Τάκης	32	11141	Βρογχίτιδα
Στράτος	34	11143	Πόνος στη πλάτη
Μαρία	46	11255	Πόνος στο χέρι
Ιωάννα	47	11255	Βρογχίτιδα
Δήμητρα	49	11257	Γρίπη
Δημήτρης	51	11356	Βρογχίτιδα
Μαίρη	57	11359	Γαστρεντερίτιδα

Πίνακας 34: Πίνακας RT(1)

Κλάση	Ηλικία	Ταχυδρομικός Κώδικας	Ασθένεια
1	[20-22]	[10443-10551]	Πόνος στη πλάτη
1	[20-22]	[10443-10551]	Πόνος στο χέρι
2	[30-31]	[10556-10558]	Γρίπη
2	[30-31]	[10556-10558]	Γαστρεντερίτιδα
3	[32-34]	[11141-11143]	Βρογχίτιδα
3	[32-34]	[11141-11143]	Πόνος στη πλάτη
4	[46-49]	[11255-11257]	Πόνος στο χέρι
4	[46-49]	[11255-11257]	Βρογχίτιδα
4	[46-49]	[11255-11257]	Γρίπη
5	[51-57]	[11356-11359]	Βρογχίτιδα
5	[51-57]	[11356-11359]	Γαστρεντερίτιδα

Πίνακας 35: Πίνακας RT(1)

Οι ασθενείς Μάκης, Έλενα, Στράτος, Ιωάννα και Δημήτρης έχουν διαγραφεί από τη βάση δεδομένων και στη θέση τους έχουν προστεθεί οι ασθενείς Θανάσης, Μαριάννα, Χριστίνα, Νικολέτα και Χρήστος. Οι πίνακες RT(1)* και RT(2)* ικανοποιούν το 2-anonymity και το 2-diversity. Ωστόσο, ο επιτιθέμενος μπορεί να προσδιορίσει μοναδικά την ταυτότητα ενός ασθενή, με τη σύνδεση των δύο δημοσιευμένων πινάκων.

Όνομα	Ηλικία	Ταχυδρομικός Κώδικας	Ασθένεια
Γιάννης	20	10443	Πόνος στη πλάτη
Θανάσης	25	10445	Γρίπη
Μαριάννα	30	10551	Γρίπη
Κατερίνα	31	10558	Γαστρεντερίτιδα
Τάκης	32	11141	Βρογχίτιδα
Χριστίνα	35	11143	Πόνος στο χέρι
Μαρία	46	11255	Πόνος στο χέρι
Νικολέτα	47	11256	Πόνος στη πλάτη

Δήμητρα	49	11257	Γρίπη
Χρήστος	52	11357	Βρογχίτιδα
Μαίρη	57	11359	Γαστρεντερίτιδα

Πίνακας 36: Πίνακας RT(2)

Για παράδειγμα έστω ότι, ο επιτιθέμενος γνωρίζει την ηλικία και τον ταχυδρομικό κώδικα του Γιάννη. Επίσης, γνωρίζει ότι η θεραπεία του διήρκησε πάνω από έξι μήνες, αυτό σημαίνει ότι τα στοιχεία του καταγράφονται και στους δυο δημοσιευμένους πίνακες. Από τον πίνακα RT(1)* ο επιτιθέμενος μπορεί να συμπεράνει ότι ο Γιάννης πάσχει από πόνος στη πλάτη ή πόνο στο χέρι. Ταυτόχρονα από τον πίνακα RT(2)* μπορεί να συμπεράνει ότι ο Γιάννης πάσχει είτα από πόνο στη πλάτη ή από γρίπη. Έτσι, με τη σύνδεση των δυο δημοσιευμένων πινάκων ο επιτιθέμενος ανακαλύπτει ότι ο Γιάννης πάσχει από πόνος στη πλάτη.

Κλάση	Ηλικία	Ταχυδρομικός Κώδικας	Ασθένεια
1	[20-25]	[10443-10445]	Πόνος στη πλάτη
1	[20-25]	[10443-10445]	Γρίπη
2	[30-31]	[10551-10558]	Γρίπη
2	[30-31]	[10551-10558]	Γαστρεντερίτιδα
3	[32-35]	[11141-11143]	Βρογχίτιδα
3	[32-35]	[11141-11143]	Πόνος στο χέρι
4	[46-49]	[11255-11257]	Πόνος στο χέρι
4	[46-49]	[11255-11257]	Πόνος στη πλάτη
4	[46-49]	[11255-11257]	Γρίπη
5	[52-57]	[11357-11359]	Βρογχίτιδα
5	[52-57]	[11357-11359]	Γαστρεντερίτιδα

Πίνακας 37: Πίνακας RT(2)*

Εφαρμόζοντας την μέθοδο του m-invariance, ο πίνακας RT(2)* αντικαθίσταται με τον πίνακα RT(3) ο οποίος περιλαμβάνει τις γενικευμένες τιμές του πίνακα RT(2) μαζί με δύο εικονικές ή αλλιώς πλαστές εγγραφές Π1 και Π2. Οι 13 εγγραφές χωρίζονται σε 6 κλάσεις ισοδυναμίας.

Εφαρμόζοντας αυτή τη μεθοδολογία, ο επιτιθέμενος δεν μπορεί να διαχωρίσει τις πλαστές από τις πραγματικές εγγραφές στην ίδια κλάση ισοδυναμίας. Για παράδειγμα, οι κλάσεις ισοδυναμίας στους πίνακες RT(1)* και RT(3) έχουν το ίδιο σύνολο ευαίσθητων εγγραφών {πόνος στη πλάτη, πόνος στο χέρι}, οπότε ο επιτιθέμενος δεν μπορεί να προσδιορίσει με βεβαιότητα πάνω από 50% την ασθένεια του Γιάννη.

Τέλος, για την ακρίβεια το m-invariance προϋποθέτει την ικανοποίηση του m-diversity και μία εγγραφή να ανήκει πάντα σε μια κλάση ισοδυναμίας, η οποία έχει το ίδιο σύνολο ευαίσθητων ιδιοτήτων, για όλες τις δημοσιεύσεις.

Όνομα	Κλάση	Ηλικία	Ταχυδρομικός Κώδικας	Ασθένεια
Γιάννης	1	[20-24]	[10443-10444]	Πόνος στη πλάτη
Π1	1	[20-24]	[10443-10444]	Γρίπη
Θανάσης	2	[25-30]	[10445-10551]	Γρίπη
Μαριάννα	2	[25-30]	[10445-10551]	Γρίπη
Π2	2	[25-30]	[10445-10551]	Βρογχίτιδα
Κατερίνα	3	[31-32]	[10558-11141]	Γαστρεντερίτιδα
Τάκης	3	[32-32]	[10558-11141]	Βρογχίτιδα
Χριστίνα	4	[35-46]	[11143-11255]	Πόνος στο χέρι
Μαρία	4	[35-46]	[11143-11255]	Πόνος στο χέρι
Νικολέτα	5	[47-49]	[11256-11257]	Πόνος στη πλάτη
Δήμητρα	5	[47-49]	[11256-11257]	Γρίπη
Χρήστος	6	[52-57]	[11357-11359]	Βρογχίτιδα
Μαίρη	6	[52-57]	[11357-11359]	Γαστρεντερίτιδα

Πίνακας 38: Πίνακας RT(3)

5.5 Δ-Παρουσια (Δ-Presence)

Η μέθοδος δ-παρουσία εγγυάται ότι με την ανωνυμοποίηση του αρχικού πίνακα, ένας επιτιθέμενος δεν έχει τη δυνατότητα να προσδιορίσει αν κάποιο άτομο συμπεριλαμβάνεται στη συγκεκριμένη βάση με βεβαιότητα μεγαλύτερη από δ.

Για να ικανοποιούν την προστασία της ιδιωτικότητας οι μέθοδοι k-Anonymity και του l-diversity απαιτείται ένα σενάριο επίθεσης κατά το οποίο ένας κακόβουλος τρίτος γνωρίζει πληροφορίες για ένα άτομο και είναι σίγουρος ότι τα στοιχεία του ατόμου αυτού είναι δημοσιευμένα στο σύνολο εγγραφών. Ωστόσο, υπάρχουν και περιπτώσεις όπου ο επιτιθέμενος δεν πρέπει να είναι βέβαιος ότι το άτομο που αναζητά συμπεριλαμβάνεται σε αυτό τον πίνακα, γιατί τότε μπορεί με βεβαιότητα να αποκαλύψει την ευαίσθητη τιμή του.

Τελικά, μέσω της ανάλυσης του κινδύνου επιβεβαίωσης της συμμετοχής ή όχι ενός φυσικού προσώπου στα ανωνυμοποιημένα δεδομένα, εξετάζεται πότε τα δεδομένα είναι ικανοποιητικά ανώνυμα έτσι ώστε να μην αποκαλυφθεί η ευαίσθητη και προσωπική πληροφορία.

5.6 Κ^{m}-ΑΝΩΝΥΜΙΑ (Κ^{m}-ANONYMITY)

Παρ' όλες τις διάφορες τεχνικές, μεθοδολογίες και εγγυήσεις που έχουν προταθεί, πολλοί κίνδυνοι για την ιδιωτικότητα παραμένουν χωρίς αποτελεσματική αντιμετώπιση. Οι πληροφορίες που έχει στη διάθεση του ένας επιτιθέμενος μπορεί να έχουν πολλαπλές μορφές και ταυτόχρονα τα μοντέλα των δημοσιευμένων δεδομένων μπορεί να διαφέρουν κάθε φορά, με αποτέλεσμα η κάθε περίπτωση να απαιτεί διαφορετική διαχείριση προκειμένου να εξασφαλίζεται η ιδιωτικότητα των βάσεων δεδομένων.

Για την επίλυση τέτοιων περιπτώσεων χρησιμοποιείται το μοντέλο της k^{m}-ανωνυμίας, με βάση την μελέτη που έγινε από τον (Αγγέλη, 2014), το σενάριο αυτού του προβλήματος είναι το εξής: κάθε εγγραφή αποτελείται από σύνολα δεδομένων που παίρνουν τιμές από ένα κοινό πεδίο τιμών και υπάρχει ένας αντίπαλος ο οποίος κατέχει μερική γνώση για τα δεδομένα, δηλαδή γνωρίζει m τιμές μιας εγγραφής, και προσπαθεί να εντοπίσει τις υπόλοιπες τιμές της εγγραφής και να τις αντιστοιχίσει με ένα φυσικό πρόσωπο.

Στο μοντέλο της k^{m}-ανωνυμίας δεν υπάρχει σαφής διαχωρισμός μεταξύ των ευαίσθητων γνωρισμάτων και του ψευδό-αναγνωριστικού. Κάθε φορά ένα υποσύνολο των τιμών της εγγραφής σχηματίζει το σύνολο του ψευδό-αναγνωριστικού και οι υπόλοιπες τιμές σχηματίζουν το σύνολο των ευαίσθητων γνωρισμάτων. Επίσης, κάθε εγγραφή έχει διαφορετικό μέγεθος ενώ στις σχεσιακές βάσεις δεδομένων το μέγεθος της κάθε εγγραφής είναι σταθερό.

Η k^{m}-ανωνυμία αποτελεί το μεταγενέστερο μοντέλο της k-ανωνυμίας οπού κάθε συνδυασμός τιμών μεγέθους m, εμφανίζεται τουλάχιστον k φορές στο σύνολο δεδομένων. Έτσι, εάν για παράδειγμα υπάρχει ένας επιτιθέμενος που γνωρίζει ένα μέρος της πληροφορίας και θέλει να εντοπίσει και τα υπόλοιπα στοιχεία του ατόμου για να τον εμποδίσουμε χρησιμοποιούμε km-ανωνυμία. Σύμφωνα με το μοντέλο αυτό, εάν ο αντίπαλος γνωρίζει το πολύ m τιμές από μια εγγραφή, δεν θα μπορεί να εντοπίσει την εγγραφή αυτή γιατί θα υπάρχουν άλλες k-1 εγγραφές με τις ίδιες τιμές στο πίνακα.

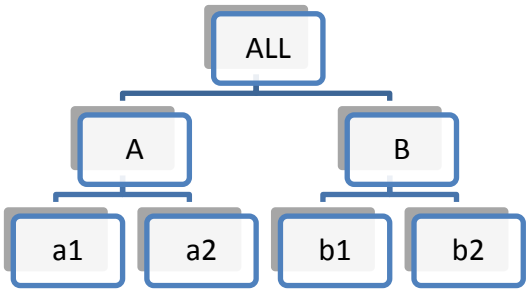
5.6.1 ΜΟΝΤΕΛΟ ΓΕΝΙΚΕΥΣΗΣ ΤΗΣ Κ^{m}-ΑΝΩΝΥΜΙΑΣ

Για την διαδικασία της ανωνυμοποίησης επιλέγεται η *τεχνική της ολικής γενίκευσης*, σύμφωνα με την οποία μια τιμή στη βάση δεδομένων αντικαθίσταται με μια πιο γενική τιμή η οποία περιέχει την αρχική, χωρίς να αλλάζει η σημασιολογία της. Στο παράδειγμα των καθημερινών αγορών μιας υπεραγοράς, η τιμή «γάλα» θα μπορούσε να γενικευτεί σε «γαλακτοκομικά προϊόντα» και η τιμή «ρύζι» σε «δημητριακά».

Το σύνολο των δυνατών γενικεύσεων μιας βάσης δεδομένων αποτελεί το δέντρο της ιεραρχίας γενίκευσης όπως φαίνεται παρακάτω. Όσο πιο ψηλά βρίσκεται η γενίκευση τόσο μεγαλύτερη είναι η απώλεια πληροφορίας που παρουσιάζουν τα δεδομένα. Σε μια συλλογή δεδομένων, όλες οι τιμές που υπάρχουν στη βάση πρέπει να βρίσκονται στο δέντρο ιεραρχίας όπως δείχνει από την εικόνα 2 πιο κάτω.

A	ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ
1	{a1,b1,b2}
2	{a2, b1}
3	{a2, b1, b2}
4	{a1, a2, b2}

Πίνακας 39: Σύνολο Δεδομένων D



Εικόνα 10: Ιεραρχία Γενίκευσης

Στο συγκεκριμένο παράδειγμα το σύνολο δεδομένων D δεν ικανοποιεί την k^m- ανωνυμία, αφού για k=2 και m=2 ο συνδυασμός τιμών {a1,b1} εμφανίζεται μόνο μια φορά. Η εφαρμογή της γενίκευσης {a1, a2}→ *A* στη βάση δεδομένων μπορεί να δώσει λύση στο πρόβλημα, αφού πλέον οποιοσδήποτε συνδυασμός m=2 τιμών στην βάση, εμφανίζεται σε τουλάχιστον k=2 εγγραφές.

A	ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ
1	{A,b1,b2}
2	{A, b1}
3	{A, b1, b2}
4	{A, A, b2}

Πίνακας 40: Σύνολο ανωνυμοποιημένων δεδομένων D'

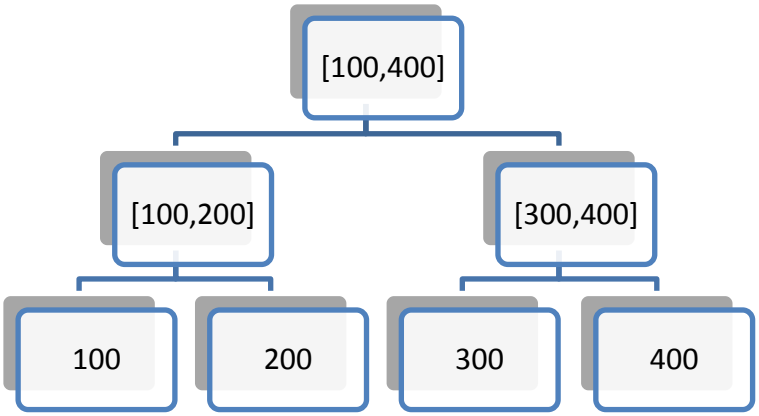
5.6.2 ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΟΥ ΜΕ ΙΕΡΑΡΧΙΑ ΓΕΝΙΚΕΥΣΗΣ

Με τη χρήση της k^m-ανωνυμοποίησης με ιεραρχία γενίκευσης ο αλγόριθμος θα δημιουργούσε για τον παρακάτω πίνακα ένα δέντρο ιεραρχίας με την πιο κάτω μορφή.

A	ΣΥΝΟΛΟ ΤΙΜΩΝ
1	{100,100,200,400,400}
2	{100,300,400}
3	{100,100,400,400}

Πίνακας 41: Κανονικοποιημένη Ποινή Βεβαιότητας

Δέντρο Ιεραρχίας:



Εικόνα 11: Ιεραρχία Γενίκευσης για το Σύνολο Δεδομένων από τον Πίνακα 12

Κατά τη διαδικασία της ανωνυμοποίησης για m=2 και k=2, ο αλγόριθμος θα έπρεπε να γενικεύσει τις τιμές {200, 300}, οι οποίες παραβιάζουν την ιδιωτικότητα στη βάση δεδομένων. Η τιμή {200} σύμφωνα με το δέντρο ιεραρχίας, γενικεύεται σε {100, 200}. Λόγω του ότι ο αλγόριθμος χρησιμοποιεί τεχνική ολικής ανακωδικοποίησης στα δεδομένα, όλες οι τιμές {100, 200} θα πρέπει να αντικατασταθούν με την γενικευμένη τιμή [100, 200]. Αυτό έχει σαν αποτέλεσμα να ανωνυμοποιηθεί η τιμή {100} η οποία δεν προκαλούσε στη βάση οποιαδήποτε παράβαση ιδιωτικότητας. Με την ίδια λογική η τιμή {300} γενικεύεται σε {300, 400} επηρεάζοντας και την τιμή {400}. Με το πέρας του αλγόριθμου k^m-ανωνυμοποίησης θα προκύψει το πιο κάτω ανωνυμοποιημένο σύνολο δεδομένων.

A	ΓΕΝΙΚΕΥΜΕΝΟ ΣΥΝΟΛΟ ΤΙΜΩΝ
1	{[100,200] [100,200] [100,200] [300,400] [300,400]}
2	{[100,200] [300,400] [300,400]}
3	{[100,200] [100,200] [300,400] [300,400]}

Πίνακας 42: Σύνολο Ανωνυμοποιημένων Δεδομένων

Στην περίπτωση αυτή ο επιτιθέμενος δε μπορεί να αναγνωρίσει με βεβαιότητα καμία εγγραφή βάσει της μερικής του γνώσης. Ο αλγόριθμος k^m-ανωνυμίας εξαιτίας της γενίκευσης πλήρους πεδίου με χρήση ιεραρχίας 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.

5.6.3 ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΟΥ ΧΩΡΙΣ ΙΕΡΑΡΧΙΑ ΓΕΝΙΚΕΥΣΗΣ

Στην περίπτωση της δημοσίευσης δεδομένων με συνεχή γνωρίσματα, είναι δυνατή η k^m-ανωνυμοποίηση με σημαντικά καλύτερη απόδοση σύμφωνα με μελέτες που έχουν γίνει. Η πιθανή προτεινόμενη λύση εφαρμόζει ολική γενίκευση με σκοπό την k^m-ανωνυμοποίηση των δεδομένων, όπως ακριβώς και στο πιο πάνω παράδειγμα. Η διαφορά με την προηγούμενη λύση, βρίσκεται στο ότι η κατάλληλη γενίκευση για κάθε γνώρισμα δεν ακολουθεί κάποια προκαθορισμένη ιεραρχία. Ο αλγόριθμος εκμεταλλευόμενος την φύση των δεδομένων προσπαθεί κάθε φορά να βρει τη λύση διευρύνοντας το διάστημα τιμών κάθε γνωρίσματος, με τέτοιο τρόπο ώστε να διατηρείται η Κανονικοποιημένη Ποινή Βεβαιότητας μικρότερη από μια μέγιστη επιτρεπτή ποινή. Η εφαρμογή του αλγορίθμου για το ίδιο σύνολο δεδομένων παρουσιάζεται στον παρακάτω πίνακα.

A	ΓΕΝΙΚΕΥΜΕΝΟ ΣΥΝΟΛΟ ΤΙΜΩΝ
1	{[100, 100, [200, 300], 400, 400}
2	{100, 400, [200, 300]}
3	{100, 100 ,400, 400}

Πίνακας 43: Γενικευμένο Σύνολο Ανωνυμοποιημένων Δεδομένων

Κατά τη διαδικασία της ανωνυμοποίησης για m=2 και k=2, ο αλγόριθμος θα πρέπει να γενικεύσει τις τιμές {200,300}, οι οποίες παραβιάζουν την ιδιωτικότητα στη βάση δεδομένων. Ο αλγόριθμος ελέγχοντας τις πιθανές γενικεύσεις βρίσκει ότι, η τιμή {200} μπορεί να γενικευτεί σε {200,300} διατηρώντας την Κανονικοποιημένη Ποινή Βεβαιότητας σε χαμηλά επίπεδα, μικρότερη κάθε φορά από μια παράμετρο μέγιστης επιτρεπτής ποινής.

Στην περίπτωση της λύσης αυτής, ο επιτιθέμενος και πάλι δε μπορεί να αναγνωρίσει με βεβαιότητα καμία εγγραφή βάσει της μερικής του γνώσης. Ο αλγόριθμος εξαιτίας της γενίκευσης

χωρίς χρήση ιεραρχίας, ικανοποιεί την k-ανωνυμία στο σύνολο των δεδομένων, χωρίς να υπεργενικεύει τα δεδομένα. Αυτό έχει σαν αποτέλεσμα, οι γενικευμένες τιμές να διατηρούνται κοντά στις αρχικές και κατ' επέκταση το ανωνυμοποιημένο σύνολο δεδομένων να εμφανίζει μικρή απώλεια πληροφορίας.

6 ΑΛΓΟΡΙΘΜΟΙ ΑΝΩΝΥΜΟΠΟΙΗΣΗΣ

6.1 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Στην παρούσα πτυχιακή εργασία και με βάση το άρθρο της (Sweeney, 2002) μελετώνται αλγόριθμοι οι οποίοι επιχειρούν την ικανοποίηση της k-ανωνυμίας δημοσιευμένων συλλογών δεδομένων που αποτελούνται από συνεχή γνωρίσματα. Εφαρμόζουν τοπική γενίκευση στις τιμές των γνωρισμάτων των εγγραφών, υπολογίζοντας την ζητούμενη διαμέριση κάθε γνωρίσματος ξεχωριστά.

Η k-ανωνυμία απαιτεί κάθε εγγραφή που εμφανίζεται στον πίνακα, να μην μπορεί να αναγνωρισθεί ανάμεσα από τουλάχιστον k εγγραφές. Αυτό επιτυγχάνεται αν k τουλάχιστον εγγραφές από το σύνολο, εμφανίζουν τις ίδιες τιμές ή αντίστοιχα τα ίδια διαστήματα τιμών σε όλα τους τα γνωρίσματα. Οι αλγόριθμοι ή αλλιώς συστήματα πραγματικού κόσμου που παρουσιάζονται εξασφαλίζουν την k-ανωνυμία ως προς την τιμή της 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 συνάρτησης στα δημοσιευμένα δεδομένα. Εφαρμόζουμε γενίκευση, όπου κάθε εμφάνιση κάποιας τιμής του πεδίου τιμών των γνωρισμάτων αντικαθιστάται με την κατάλληλη γενικευμένη τιμή ενώ δεν αντιστοιχίζονται όλες οι εμφανίσεις της τιμής στον πίνακα στην ίδια γενικευμένη τιμή.

6.1.1 ΕΛΑΧΙΣΤΗ ΠΑΡΑΜΟΡΦΩΣΗ ΕΝΟΣ ΠΙΝΑΚΑ

Όταν υπάρχουν διαφορετικές k-ελάχιστες γενικεύσεις, δημιουργούνται κάποια κριτήρια ώστε να επιλέξουν μια λύση μεταξύ τους. Αν γενικεύσουμε έναν αρχικό πίνακα T, καταλήγουμε σε έναν νέο πίνακα T' όπου έχει συνήθως λιγότερες πληροφορίες από τον αρχικό πίνακα, άρα ο πίνακας T' θεωρείται λιγότερο χρήσιμος. Για να συλλέξουμε την απώλεια πληροφοριών ορίζουμε μια μέτρηση πληροφοριών που αναφέρει το ποσό της στρέβλωσης σε έναν γενικευμένο πίνακα. Σε ένα κελί ενός γενικευμένου πίνακα, η αναλογία του πεδίου της τιμής που βρέθηκε στο κελί προς το ύψος της ιεραρχίας του χαρακτηριστικού αναφέρεται στο ποσό της γενίκευσης και ως εκ τούτου στα μέτρα παραμόρφωσης του κελιού.

Για παράδειγμα, χρησιμοποιώντας τις ιεραρχίες στις εικόνες 2 και το 3 αντίστοιχα, η ακρίβεια των γενικεύσεων του PT που έχει δειχθεί στους πίνακες του κεφαλαίου 3 (πίνακες 13, 14) είναι η εξής: Prec ($GT_{[1,0]}$) = 0,75, Prec ($GT_{[1,1]}$) = 0,58, Prec ($GT_{[0,2]}$) = 0,67, και Prec ($GT_{[0,1]}$) = 0,83. Κάθε ένα από αυτά πληρούν την k-ανωνυμίας για k = 2. Παρατηρούμε ότι οι πίνακες $GT_{[1,0]}$ και $GT_{[0,1]}$, γενικεύουν τις τιμές από ένα επίπεδο και πάνω, επειδή $|DGH_{Race}| = 2$ και $|DGH_{ZIP}| = 3$, Prec ($GT_{[0,1]}$)> Prec ($GT_{[1,0]}$).

Οι γενικεύσεις που βασίζονται σε χαρακτηριστικά με υψηλότερες ιεραρχίες γενίκευσης τυπικά διατηρούν την μετρική ακρίβεια καλύτερα από γενικεύσεις που βασίζονται σε χαρακτηριστικά με μικρότερες ιεραρχίες. Περαιτέρω, ιεραρχίες με διαφορετικά ύψη μπορούν να προσφέρουν διαφορετικά μέτρα μετρικής ακρίβειας του ίδιου πίνακα. Έτσι, η κατασκευή των ιεραρχιών γενίκευσης είναι μέρος των κριτηρίων προτίμησης. Η μετρική ακρίβεια Prec μετρά καλύτερα την ποιότητα των δεδομένων, όταν το σύνολο των ιεραρχιών που χρησιμοποιούνται περιέχουν μόνο τιμές σημασιολογικά χρήσιμες. Δεν υπάρχει καμία ανάγκη να αυξηθούν αυθαίρετα τα ύψη των ιεραρχιών πάρα μόνο να προτιμηθεί ένα χαρακτηριστικό από ένα άλλο. Αντ' αυτού, το βάρος μπορεί να τεθεί στα «χαρακτηριστικά» του Prec ώστε να κάνει την προτίμηση ρητή.

Ενώ η εντροπία είναι το κλασικό μέτρο που χρησιμοποιείται στη θεωρία πληροφοριών που χαρακτηρίζουν την καθαρότητα των δεδομένων, μια μέτρηση που βασίζεται στη σημασιολογία της γενίκευσης μπορεί να είναι πιο διακριτική από την άμεση σύγκριση των μηκών κωδικοποίησης των τιμών που είναι αποθηκευμένες στον πίνακα. Όπως προαναφέρθηκε, όλες οι k-ελάχιστες γενικεύσεις δεν είναι στρεβλωμένες και η προτίμηση μπορεί να βασίζεται στην k-ελάχιστη γενίκευση που έχει την μεγαλύτερη ακρίβεια.

Έστω ότι T_i $\{A_1, ..., A_n\}$ και T_m $\{A_1, ..., A_n\}$ είναι δύο πίνακες, έτσι ώστε T_i $[Q_{IT}] \leq T_m$ $[QI_T]$, όπου QI_T = {Ai, ..., Aj} είναι το ψευδο-αναγνωριστικό που σχετίζεται με τους πινάκες, $\{A_i,...A_j\} \subseteq \{A_1, ..., A_n\}$ και για κάθε x=i,…,j DGH_{AX} είναι πεδία ιεραρχιών γενίκευσης για το QI_T. Ο πίνακας T_m λέγεται ότι είναι μια ελάχιστη παραμόρφωση ενός πίνακα Τα όπου ικανοποιεί την k-ανωνυμίας με το ψευδο-αναγνωριστικό QI_T αν και μόνο αν:

- Tm ικανοποιεί την απαίτηση της κ-ανωνυμίας σε σχέση με το QI_T
- Για κάθε T_z τέτοιο ώστε $Prec(T_i) \geq Prec(T_z)$, $Prec(T_{z)} \geq Prec(T_m)$, T_z ικανοποιεί την k-ανωνυμίας σε σχέση με το QI_T => T_z[Ai, ..., Aj] = T_m [A_i, ..., A_j].

Εξετάστε τις τιμές που αναφέρονται στο παραπάνω παράδειγμα για τους πίνακες 13, 14 αντίστοιχα. Μόνο ο πίνακας $GT_{[0,1]}$ είναι μια k-ελάχιστη παραμόρφωση του αρχικού πίνακα PT. Η k-ελάχιστη παραμόρφωση είναι ειδική για έναν πίνακα, ένα ψευδο-αναγνωριστικό, ένα σύνολο ιεραρχιών γενίκευσης πεδίου, για τα χαρακτηριστικά του ψευδο-αναγνωριστικού και για μια Prec (ή μια σταθμισμένη Prec). Είναι τετριμμένο να δούμε ότι ένας πίνακας που ικανοποιεί την k-ανωνυμία έχει μια μοναδική k-ελάχιστη παραμόρφωση, από μόνος του. Επίσης, είναι εύκολο να δει κανείς ότι ένας γενικευμένος πίνακας RT που είναι μια k-ελάχιστη παραμόρφωση του πίνακα PT είναι επίσης και μια k-ελάχιστη γενίκευση του PT.

6.2 ΑΛΓΟΡΙΘΜΟΣ MINGEN

Ο αλγόριθμος που παρουσιάζεται σε αυτήν την ενότητα συνδυάζει αυτούς τους τυπικούς ορισμούς σε ένα θεωρητικό μοντέλο βάσει του οποίου τα συστήματα του πραγματικού κόσμου θα συγκριθούν.

Στο παρακάτω σχήμα παρουσιάζεται ο αλγόριθμος, όπου ονομάζεται Mingen, κατά τον οποίο: δίνεται ένας πίνακας PT $\{A_x,...,A_y\}$, ένα QI ψευδο-αναγνωριστικό = $\{A_1,...,A_n\}$, όπου $\{A_1,...,A_n\} \subseteq \{A_x,...,A_y\}$, ικανοποιεί την k-ανωνυμία, και το πεδίο γενίκευσης ιεραρχιών DGHAi, που παράγει ένα MGT πίνακα ο οποίος είναι μια k-ελάχιστη παραμόρφωση του $PT_{[QI]}$. Υποθέτει ότι k <| PT |, το οποίο αποτελεί απαραίτητη προϋπόθεση για την ύπαρξη μιας k-ελάχιστη γενίκευσης.

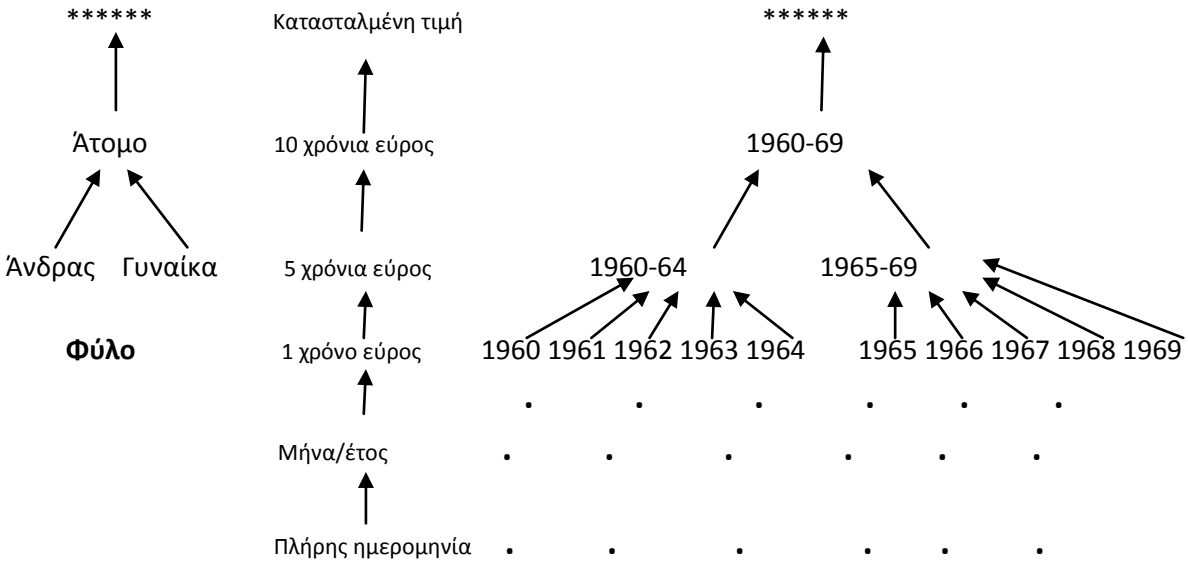
Είσοδος: Αρχικός πίνακα PT με ψευδο-αναγνωριστικό QI = $\{A_1, ..., A_n\}$, k περιορισμός ιεραρχίας γενίκευσης πεδίου DGHAi, όπου i = 1, ..., n, και οι προτιμώμενες προδιαγραφές.
Έξοδος: MGT, μια ελάχιστη παραμόρφωση του PT [QI] σε σχέση με k επιλεγμένη σύμφωνα με τις προδιαγραφές προτίμησης
Υποθέσεις: | PT | ≥ k
Μέθοδος: **1.** An PT ικανοποιεί τις απαιτήσεις της k-ανωνυμίας σε σχέση με το k τότε κάνουμε
1.1. MGT ← {PT} //PT είναι η λύση
2.1. Allgen ← {Ti: Ti είναι η γενίκευση του PT πάνω από QI}
2.2. Protected ← {Ti: Ti E allgens και ο πίνακας Ti ικανοποιεί την κ-ανωνυμία για k}
2.3. MGT ← {Ti: Ti E protected} όπου δεν υπάρχει Tz E protected
2.4. MGT ← preferred(MGT) // επιλεγώ τη σωστή λύση
3. επιστροφή MGT

Εικόνα 12: Ελάχιστη Γενίκευση του Αλγορίθμου Mingen

Τα βήματα του αλγορίθμου Mingen είναι απλά. Στο βήμα 1 προσδιορίζεται αν ο αρχικός πίνακα PT, ικανοποιεί την k-ανωνυμία, τότε είναι k-ελάχιστη παραμόρφωση. Σε όλες τις άλλες περιπτώσεις εκτελέστε το βήμα 2. Όπου στο βήμα 2.1 αποθηκεύεται το σύνολο από όλες τις πιθανές γενικεύσεις του PT πάνω στο QI σε allgens, δηλαδή όλες οι γενικεύσεις. Στη συνέχεια στο βήμα 2.2 αποθηκεύονται αυτές οι γενικεύσεις από allgens που πληρούν την προϋπόθεση της k-ανωνυμίας. Έπειτα στο βήμα 2.3, αποθηκεύονται οι k-ελάχιστες στρεβλώσεις από το να προστατευτούν σε MGT. Είναι εγγυημένο ότι | MGT | ≥ 1. Τέλος, στο βήμα 2.4, η λειτουργία που προτιμάται επιστρέφει μια ενιαία K-ελάχιστη παραμόρφωση από τον MGT με βάση τις προδιαγραφές που ορίζονται από το χρήστη.

Στο παράδειγμα που ακολουθεί, οι ιεραρχίες των εικόνων 2, 3 και 5 αντίστοιχα για το ψευδο-αναγνωριστικό QI = {Φυλή, ημερομηνία γέννησης, Φύλο, ΤΚ}, τον αρχικό πίνακα PT (πίνακας 25), και την k-ανωνυμίας για k = 2 να παρέχονται στον Mingen. Μετά το βήμα 2.3

του αλγορίθμου Mingen, έχουμε MGT = {GT}, όπου GT δείχνεται στην εικόνα 12. Το GT είναι μια k-ελάχιστη παραμόρφωση του PT επί QI και k-ελάχιστη γενίκευση του PT επί QI. Μπορεί να αποδειχθεί ότι η γενίκευση ενός πίνακα T σε ένα QI ψευδο-αναγνωριστικό, που ικανοποιεί μια συγκεκριμένη απαίτηση της k-ανωνυμίας, και έχει το ελάχιστο ποσό των στρεβλώσεων από όλες τις δυνατές γενικεύσεις του T πάνω από QI, είναι μια k-ελάχιστη παραμόρφωση του T έναντι του QI σε σχέση με το Prec. Από το θεώρημα 2, όπου δίνονται δυο πίνακες T_I και T_m έτσι ώστε $T_i \leq T_m$ και ο πίνακας T_m να ικανοποιεί την k-ανωνυμία. Ο Tm είναι μια k-ελάχιστη στρέβλωση του T_i. Ο T_m είναι μια k-ελάχιστη γενίκευση του T_i, άρα μια ενδεχόμενη λύση είναι k-ελάχιστη γενίκευση του πίνακα T επί το ψευδο-αναγνωριστικό QI.



Εικόνα 13: Ιεραρχία Γενίκευσης τιμών {Φύλο, Ημερομηνία Γέννησης} με Συμπίεση

Η προτιμώμενη λειτουργία επιστρέφει μόνο έναν πίνακα ως λύση. Η απαίτηση ενιαίας λύσης αποτελεί απαραίτητη προϋπόθεση, διότι η επιλεγείσα λύση γίνεται μέρος της συνένωσης των εξωτερικών πληροφοριών κατά την οποία η μετέπειτα σύνδεση πρέπει να προστατεύεται.

Όσον αφορά την πολυπλοκότητα, ο αλγόριθμος Mingen δεν κάνει καμία αξίωση για να είναι αποτελεσματικός. Σαφώς, η εξαντλητική αναζήτηση όλων των πιθανών γενικεύσεων είναι ανέφικτη ακόμα και σε μέτριου μεγέθους πίνακες. Έτσι, πώς γίνεται να βρούμε λύσεις σε συστήματα του πραγματικού κόσμου, σε πραγματικό χρόνο;

PT Prec=1.00				
Φυλή	**Ημερομηνία Γεννήσεως**	**Φύλο**	**Τ.Κ.**	**Ασθένεια**
Έγχρωμος	9/20/1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	2/14/1965	Άνδρας	02141	Πόνος στο στήθος
Έγχρωμος	10/23/1965	Γυναίκα	02138	Πόνος στο μάτι
Έγχρωμος	8/24/1965	Γυναίκα	02138	Βρογχιολίτιδα
Έγχρωμος	11/7/1964	Γυναίκα	02138	Παχυσαρκία
Έγχρωμος	12/1/1964	Γυναίκα	02138	Πόνος στο στήθος
Λευκός	10/23/1964	Άνδρας	02138	Δύσπνοια
Λευκός	3/15/1965	Γυναίκα	02139	Υπέρταση
Λευκός	8/13/1964	Άνδρας	02139	Παχυσαρκία
Λευκός	5/5/1964	Άνδρας	02139	Πυρετός
Λευκός	2/13/1967	Άνδρας	02138	Εμετός
Λευκός	3/21/1967	Άνδρας	02138	Πόνος στην πλάτη

Πίνακας 44: k- ελάχιστη παραμόρφωση για τον πίνακα PT όπου k=2

GT Prec=0.90				
Φυλή	**Ημερομηνία Γεννήσεως**	**Φύλο**	**Τ.Κ.**	**Ασθένεια**
Έγχρωμος	1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	1965	Άνδρας	02141	Πόνος στο στήθος
Άτομο	1965	Γυναίκα	0213*	Πόνος στο μάτι
Άτομο	1965	Γυναίκα	0213*	Βρογχιολίτιδα
Έγχρωμος	1964	Γυναίκα	02138	Παχυσαρκία
Έγχρωμος	1964	Γυναίκα	02138	Πόνος στο στήθος
Λευκός	1960-69	Άνδρας	0213*	Δύσπνοια
Άτομο	1965	Γυναίκα	0213*	Υπέρταση
Λευκός	1964	Άνδρας	0213*	Παχυσαρκία
Λευκός	1964	Άνδρας	0213*	Πυρετός
Λευκός	1960-69	Άνδρας	02138	Εμετός
Λευκός	1960-69	Άνδρας	02138	Πόνος στην πλάτη

Πίνακας 45: Πίνακας GT με μετρική ακρίβεια 0.90

6.3 ΣΥΣΤΗΜΑΤΑ ΠΡΑΓΜΑΤΙΚΟΥ ΚΟΣΜΟΥ

Δύο συστήματα πραγματικού κόσμου επιδιώκουν να παρέχουν προστασία της k-ανωνυμίας χρησιμοποιώντας γενίκευση και συμπίεση. Τα συστήματα αυτά είναι το Datafly Systems και στατιστικά του m-Argus συστήματος. Το κεφάλαιο αυτό δείχνει ότι μπορεί το Datafly να παραποιεί τα δεδομένα και το m-Argus να μην μπορεί παρέχει επαρκή προστασία.

6.3.1 DATAFLY SYSTEM

Η ρύθμιση στην οποία λειτουργεί ο αλγόριθμος core-Datafly**.** Ο κάτοχος των δεδομένων δηλώνει τα συγκεκριμένα χαρακτηριστικά και τις πλειάδες στον αρχικό πίνακα PT ως επιλέξιμες για δημοσίευση. Επίσης ομαδοποιεί ένα υποσύνολο των χαρακτηριστικών γνωρισμάτων του PT σε ένα ή περισσότερα ψευδο-αναγνωριστικά Q_{Ii}, εκχωρεί βαρύτητα κλίμακας από 0 έως 1 για κάθε χαρακτηριστικό και σε κάθε Q_{Ii} καθορίζει την πιθανότητα που θα χρησιμοποιηθεί το χαρακτηριστικό για τη σύνδεση. Ο κάτοχος των δεδομένων

καθορίζει ένα ελάχιστο επίπεδο ανωνυμίας που υπολογίζει μια τιμή για το k. Τέλος, η βαρύτητα από 0 έως 1 έχει εκχωρηθεί σε κάθε χαρακτηριστικό, σε κάθε Q_{Ii} για να δηλώσει την προτίμησή του όσον αφορά πιο χαρακτηριστικό θα παραποιήσει. Η τιμή 0 σημαίνει ότι ο παραλήπτης των δεδομένων θα προτιμούσε οι τιμές να μην αλλαχθούν και η τιμή 1 σημαίνει ότι μέγιστη παραμόρφωση θα μπορούσε να γίνει ανεκτή. Ένα ψευδο-αναγνωριστικό, όπου όλα τα χαρακτηριστικά του έχουν ίση προτίμηση και μια ίση πιθανότητα για τη σύνδεση τότε το βάρος μπορεί να θεωρηθεί ότι δεν είναι του παρόντος.

Στην εικόνα 12 παρουσιάζεται ο αλγόριθμος core-Datafly, όπου ο πίνακα PT (A_χ,...,A_y) με το ψευδο-αναγνωριστικό QI = {A_1, ..., A_n}, όπου {A_1, ..., A_n} $\subseteq$ {A_x, ..., A_y}, και ένα πεδίο ιεραρχιών γενίκευσης DGHAi, παράγει ένα MGT πίνακα που είναι μια γενίκευση του PT [QI] που ικανοποιεί την k-ανωνυμία.

Είσοδος: Αρχικός πίνακα PT με ψευδο-αναγνωριστικό QI = {A_1, ..., A_n}, k περιορισμός ιεραρχίας γενίκευσης πεδίου DGHAi, όπου i = 1, ..., n,
Έξοδος: MGT, μια γενίκευση του πίνακα PT [QI] σε σχέση με k
Υποθέσεις: | PT | $\geq$ k
Μέθοδος: **1.** Freq ← ακολουθία τιμών που περιλαμβάνει τις διακριτές ακολουθίες τιμών του PT [QI] μαζί με τον αριθμό των εμφανίσεων κάθε ακολουθίας.
2. καθώς υπάρχουν ακολουθίες στην freq που εμφανίζονται λιγότερο από k φορές, αυτό μετράει για περισσότερο από k πλειάδες.
2.1. το Aj χαρακτηριστικό στην freq όπου έχει τον μεγαλύτερο αριθμό διακριτών τιμών
2.2.freq ← γενίκευση των τιμών του A_j στην freq.
3. freq ← καταστέλλει ακολουθίες σε freq που εμφανίζονται λιγότερο από k φορές
4. freq ← επιβάλλει την k απαίτηση στις κατασταλμένες τιμές στην freq
5 .επιστροφή στον MGT ← δημιουργία πίνακα από freq

Εικόνα 14: Αλγόριθμος core-Datafly

Ο αλγόριθμος core-Datafly έχει λίγα βήματα. Στο βήμα 1, η κατασκευή freq είναι μια λίστα συχνοτήτων που περιέχει διακριτές ακολουθίες τιμών από τον πίνακα PT [QI], μαζί με τον αριθμό των εμφανίσεων της κάθε ακολουθίας. Κάθε αλληλουχία freq αντιπροσωπεύει μια ή περισσότερες πλειάδες σε έναν πίνακα. Εν συνεχεία του βήματος 2.1, χρησιμοποιείται μια ευρετική καθοδήγηση γενίκευσης. Τα χαρακτηριστικά που έχουν το μεγαλύτερο αριθμό των διακριτών τιμών σε freq είναι γενικευμένα. Η γενίκευση συνεχίζεται μέχρι να εξακολουθούν να υπάρχουν k ή λιγότερες πλειάδες που να έχουν διακριτές ακολουθίες στο freq. Η καταστολή, στο βήμα 3, οποιασδήποτε ακολουθίας συχνοτήτων εμφανίζεται κάτω από k φορές. Επίσης, καταστολή εκτελείται στο βήμα 4 έτσι ώστε ο αριθμός των κατασταλμένων πλειάδων να ικανοποιεί την k-ανωνυμία. Τέλος, το βήμα 5 παράγει έναν πίνακα MGT, με βάση την freq έτσι ώστε οι τιμές που αποθηκεύονται ως μία αλληλουχία συχνοτήτων να εμφανίζονται ως πλειάδες στον MGT αναπαραγόμενες σύμφωνα με την καθορισμένη συχνότητα.

Έστω ότι έχουμε τις ιεραρχίες στην εικόνα 2, 3 και 5 αντίστοιχα για το ψευδο-αναγνωριστικό QI = {Φυλή, ημερομηνία γέννησης, Φύλο, ΤΚ}, τον πίνακα PT (πίνακας 25), όπου ικανοποιεί την k-ανωνυμία με k = 2 να παρέχονται στον Datafly. Στον πίνακα 27A δείχνει τα περιεχόμενα του freq μετά το βήμα 1. Το χαρακτηριστικό «Ημερομηνία Γέννησης» έχει τους περισσότερους αριθμούς από τις διακριτές τιμές (12), έτσι ώστε οι τιμές της γενικεύονται. Ο πίνακας 27B δείχνει τα περιεχόμενα του freq μετά το βήμα 2.Οι πλειάδες t7 και t8 θα κατασταλούν. Ο πίνακας MGT στον πίνακα 28 είναι το τελικό αποτέλεσμα. Ο MGT ικανοποιεί την k-ανωνυμία για k = 2 με Prec (MGT) = 0.75. Ωστόσο, από το παράδειγμα και από τους πίνακες 25 και 26 αντίστοιχα, ο GT είναι μια k-ελάχιστη παραμόρφωση για PT με Prec (GT)= 0.90. Έτσι, ο Datafly αλλοίωσε τα αποτελέσματα περισσότερο από ό, τι χρειαζόταν.

Εθνικότητα	Ημερομηνία Γέννησης	Φύλο	Ταχυδρομικός Κώδικας	Εμφανίζεται	
Έγχρωμος	9/20/65	Άνδρας	02141	1	t1
Έγχρωμος	2/14/65	Άνδρας	02141	1	t2
Έγχρωμος	10/23/65	Γυναίκα	02138	1	t3
Έγχρωμος	8/24/65	Γυναίκα	02138	1	t4
Έγχρωμος	11/7/64	Γυναίκα	02138	1	t5
Έγχρωμος	12/1/64	Γυναίκα	02138	1	t6
Λευκός	10/23/64	Άνδρας	02138	1	t7
Λευκός	3/15/65	Γυναίκα	02139	1	t8
Λευκός	8/13/64	Άνδρας	02139	1	t9
Λευκός	5/5/64	Άνδρας	02139	1	t10
2	**12**	**2**	**3**		

A

Εθνικότητα	Ημερομηνία Γέννησης	Φύλο	Ταχυδρομικόςς Κώδικαςς	Εμφανίζεται	
Έγχρωμος	1965	Άνδρας	02141	2	t1, t2
Έγχρωμος	1965	Γυναίκα	02138	2	t3,t 4
Έγχρωμος	1964	Γυναίκα	02138	2	t5,t 6
Λευκός	1964	Άνδρας	02138	1	t7
Λευκός	1965	Γυναίκα	02139	1	t8
Λευκός	1964	Άνδρας	02139	2	t9, t10
2	**3**	**2**	**3**		

B

Πίνακας 46: Ενδιάμεσα στάδια του αλγορίθμου core-Datafly, A, B.

Ο πίνακας MGT προκύπτει από το σύστημα Datafly, με ψευδο-αναγνωριστικό QI={Εθνικότητα, Ημερομηνία Γέννησης, Φύλο, Ταχυδρομικό Κώδικα, Ασθένεια} και k=2.

Εθνικότητα	Ημερομηνία Γέννησης	Φύλο	Ταχυδρομικός Κώδικας	Ασθένεια
Έγχρωμος	1965	Άνδρας	02141	Δύσπνοια
Έγχρωμος	1965	Άνδρας	02141	Πόνος στο στήθος
Έγχρωμος	1965	Γυναίκα	02138	Πόνος στο μάτι
Έγχρωμος	1965	Γυναίκα	02138	Πυρετός
Έγχρωμος	1964	Γυναίκα	02138	Υπέρταση
Έγχρωμος	1964	Γυναίκα	02138	Εμετός
Λευκός	1964	Άνδρας	02138	Παχυσαρκία
Λευκός	1965	Γυναίκα	02139	Πόνος στη πλάτη
Λευκός	1964	Άνδρας	02139	Πυρετός
Λευκός	1964	Άνδρας	02139	Δύσπνοια

Πίνακας 47: Πίνακας MGT

Ο αλγόριθμος core- Datafly δεν παρέχει απαραιτήτως k-ελάχιστες γενικεύσεις ή k-ελάχιστες στρεβλώσεις, ακόμη και αν μπορεί να αποδειχθεί ότι οι λύσεις του ικανοποιούν πάντοτε την k-ανωνυμία. Ένα από τα προβλήματα είναι το σύστημα Datafly παίρνει άτεχνες αποφάσεις γενικεύοντας όλες τις τιμές που συνδέονται με ένα χαρακτηριστικό και καταστέλλει όλες τις τιμές μέσα σε μια πλειάδα. Ο Mingen λαμβάνει αποφάσεις σε επίπεδο κελιών και με αυτόν τον τρόπο, μπορεί να δώσει αποτελέσματα με μεγαλύτερη ακρίβεια. Ένα άλλο πρόβλημα είναι η ευρετική που επιλέγει το χαρακτηριστικό με τον μεγαλύτερο αριθμό διακριτών τιμών ως αυτό που θα γενικευτεί. Αυτό μπορεί να είναι υπολογιστικά αποτελεσματικό, αλλά μπορεί να δειχθεί ότι εκτελεί περιττές γενικεύσεις. Εν ο λίγοις, ο Datafly παράγει γενικεύσεις που ικανοποιούν την k-ανωνυμία, αλλά τέτοιες γενικεύσεις δεν μπορεί να είναι k-ελάχιστες στρεβλώσεις.

6.3.2 M-ARGUS SYSTEM

Στο σύστημα m-Argus, ο κάτοχος των δεδομένων παρέχει μια τιμή για το k και προσδιορίζει τα χαρακτηριστικά που είναι ευαίσθητα προσδίδοντας μια τιμή μεταξύ 0 και 3 σε κάθε χαρακτηριστικό. Το m-Argus System προσδιορίζει στη συνέχεια σπάνια και επομένως ανασφαλείς συνδυασμούς δοκιμάζοντας 2 και 3 συνδυασμούς ιδιοτήτων. Οι ανασφαλείς συνδυασμοί εξαλείφονται με τη γενίκευση των χαρακτηριστικών εντός του συνδυασμού και με την καταστολή των κελιών. Αντί να απομακρύνει ολόκληρες πλειάδες, το m-Argus καταστέλλει τις τιμές σε επίπεδο κελιών. Τα προκύπτοντα δεδομένα συνήθως περιέχουν όλες τις πλειάδες και τα χαρακτηριστικά των αρχικών δεδομένων, αν και οι τιμές μπορεί να λείπουν σε ορισμένες περιοχές των κελιών.

Στη συνέχεια θα μελετήσουμε τον αλγόριθμο m-Argus. Λαμβάνοντας υπόψη έναν πίνακα PT $\{A_\chi, ..., A_\gamma\}$, ένα ψευδο-αναγνωριστικό QI = $\{A_1, ..., A_n\}$, όπου $\{ A_1, ..., A_n \} \subseteq \{ A_\chi, ..., A_\gamma\}$, ανεξάρτητα υποσύνολα του QI γνωστά ως προσδιοριστικό, MORE (περισσότερο), και MOST (ακόμα περισσότερο) όπου QI= Identifying ∪ More ∪ Most, τον περιορισμό που

πληροί την k-ανωνυμίας και τη γενίκευση πεδίου ιεραρχιών DGHAi. Ο m-Argus παράγει έναν πίνακα MT που είναι μια γενίκευση του PT [QI].

Τα βασικά βήματα του αλγορίθμου m-Argus φαίνονται στην εικόνα 13. Στην υλοποίηση του αλγορίθμου βρέθηκαν κάποιες ελλείψεις της πραγματικής εφαρμογής του m-Argus. Σε γενικές γραμμές, ο κατασκευασμένος αλγόριθμος δημιουργεί λύσεις που είναι καλύτερα προστατευμένες από εκείνες που εκδίδονται από το πραγματικό πρόγραμμα.

Είσοδος: Αρχικός πίνακα PT, ψευδο-αναγνωριστικό QI = {A_1, ..., A_n} ανεξάρτητα υποσύνολα: Identifying, More και Most.
Έξοδος: MT, που περιέχει μια γενίκευση του PT[QI]
Υποθέσεις: | PT | ≥ k
Μέθοδος: **1.** Freq ← η ακολουθία τιμών περιλαμβάνει τις διακριτές ακολουθίες τιμών του πίνακα PT [QI] μαζί με τον αριθμό των εμφανίσεων κάθε ακολουθίας.
2. Γενίκευση κάθε $A_i \in QI$ στη freq, τιμές που αποδίδονται σε αυτό να ικανοποιούν το k.
3. Δοκιμή 2 και 3 συνδυασμών των Identifying, More και Most και αφήνουμε τις ακραίες τιμές να αποθηκεύσουν τους συνδυασμούς των κελιών που δεν έχουν k εμφανίσεις.
4. Ο κάτοχος των δεδομένων θα αποφασίσει η όχι αν θα γενικοποιήσει ένα $A_j \in QI$ βασισμένο σε ακραίες τιμές. Και αν ναι να προσδιορίζει το A_j για γενίκευση. Η freq περιέχει το γενικευμένο αποτέλεσμα.
5. Επανέλαβε τα βήματα 3 και 4 μέχρι ο κάτοχος των δεδομένων να μην μπορεί να εκλέξει ή να γενικοποιήσει.
6. Αυτόματα κατέστειλε μια τιμή που έχει ένα συνδυασμό ακραίων τιμών όπου προτεραιότητα δίνεται στην τιμή που εμφανίζει το μεγαλύτερο αριθμό ακραίων τιμών.

Εικόνα 15: Αλγόριθμος m-Argus

Το πρόγραμμα ξεκινά στο βήμα 1 με την κατασκευή freq, που είναι ένας κατάλογος συχνοτήτων που περιέχει διακριτές ακολουθίες τιμών από τον αρχικό πίνακα PT [QI], μαζί με τον αριθμό των εμφανίσεων της κάθε ακολουθίας. Στο βήμα 2, οι τιμές του κάθε χαρακτηριστικού γενικεύονται αυτόματα έως ότου κάθε τιμή που σχετίζεται με ένα χαρακτηριστικό στο ψευδο-αναγνωριστικό QI να εμφανιστεί τουλάχιστον k φορές. Αυτό αποτελεί απαραίτητη προϋπόθεση για την k-ανωνυμία. Στο βήμα 3, το πρόγραμμα ελέγχει αυτόματα συνδυασμούς των χαρακτηριστικών, για να εντοπιστούν οι συνδυασμοί των χαρακτηριστικών, των οποίων οι τιμές έχουν ανατεθεί σε συνδυασμό και δεν εμφανίζονται τουλάχιστον k φορές. Οι συνδυασμοί αυτοί αποθηκεύονται σε ακραίες τιμές. Στη συνέχεια, ο κάτοχος των δεδομένων, στα βήματα 4 και 5, αποφασίζει αν θα γενικεύσει ή όχι ένα χαρακτηριστικό στο QI που έχει ακραίες τιμές. Αν αποφασίσει να κάνει γενίκευση, επιλέγει το χαρακτηριστικό προς γενίκευση. Τέλος, στο βήμα 6, ο αλγόριθμος m-Argus καταστέλλει αυτόματα την τιμή κάθε συνδυασμού σε ακραίες τιμές. Προτεραιότητα δίνεται στην τιμή που εμφανίζεται πιο συχνά, προκειμένου να μειωθεί ο συνολικός αριθμός των καταστολών.

Ένα μειονέκτημα της πραγματικής εφαρμογής του m-Argus εμφανίζεται στο βήμα 3 (εικόνα 15). Το πραγματικό πρόγραμμα δεν δοκιμάζει όλους τους 2 και 3 συνδυασμούς. Αυτό μπορεί να είναι ένα σφάλμα προγραμματισμού.

Στη συνέχεια απεικονίζονται συνδυασμοί που το m-Argus ελέγχει έξι συνδυασμούς οι οποίοι δεν περιλαμβάνονται σε λίστα. Είναι εύκολο να έχουμε πίνακες στους οποίους οι τιμές εμφανίζονται σε συνδυασμούς που δεν εξετάστηκαν από τον m-Argus.

Ελεγχόμενοι Συνδυασμοί

Identifying x More x Most, Identifying x Most x Most, Most x Most x Most,
Identifying x More, Identifying x Most, More x Most, Most x Most

Ελεγχόμενοι Συνδυασμοί μόνο αν Identifying>1

More x More x Most, Most x Most x More, More x More

Εικόνα 16: Ελεγχόμενοι Συνδυασμοί από τον Αλγόριθμο m-Argus

Έστω ότι έχουμε τις ιεραρχίες στις εικόνες 2,3 και 5 αντίστοιχα, το ψευδο-αναγνωριστικό QI= {Εθνικότητα, Ημερομηνία Γέννησης, Φύλο, Ταχυδρομικός Κώδικας} όπου MOST= {Ημερομηνία Γέννησης}, MORE= {Φύλο, Ταχυδρομικός Κώδικας} και IDENTIFYING= {Εθνικότητα}, τον πίνακας PT (πίνακας 26), και τον περιορισμό της k-ανωνυμίας όπου k=2 να παρέχονται στον m-Argus. Στον παρακάτω πίνακα, το V δείχνει το αποτέλεσμα της δοκιμής MOST x MORE, και το freq ενημερώνεται για να δείξει {Ημερομηνία Γέννησης, Ταχυδρομικό Κώδικα} για t8 που δεν ικανοποιεί το k.

Ημερομηνία Γέννησης	Ταχυδρομικός Κώδικας	Εμφανίζεται	Sid	Ακραίες τιμές
1965	02141	2	t1,t2	{}
1965	02138	2	t3,t4	{}
1964	02138	3	t5,t6,t7	{}
1965	02139	1	t8	{}
1964	02139	2	t9,t10	{}
1967	02138	2	t11,t12	{}

V

Εθνικότητα	Ημερομηνία Γέννησης	Φύλο	Ταχυδρομικός Κώδικας	Εμφανίζεται	Sid	Ακραίες τιμές
Έγχρωμος	1965	Άνδρας	02141	2	t1,t2	{}
Έγχρωμος	1965	Γυναίκα	02138	2	t3,t4	{}
Έγχρωμος	1964	Γυναίκα	02138	2	t5,t6	{}
Λευκός	1964	Άνδρας	02138	1	t7	{}
Λευκός	1965	Γυναίκα	02139	1	t8	{{**Ημερομηνία Γέννησης, Ταχυδρομικός Κωδικός**}}
Λευκός	1964	Άνδρας	02139	2	t9,t10	{}
Λευκός	1967	Άνδρας	02138	2	t11,t12	{}

Freq

Πίνακας 48: Most x More: Συνδυασμός ελέγχου και Αποτελέσματος freq

Ο παρακάτω πίνακας δείχνει το freq πριν από το βήμα 6. Οι τιμές που πρέπει να κατασταλούν είναι υπογραμμισμένες. Ο πίνακας MT, στον επόμενο πίνακα, είναι το τελικό αποτέλεσμα από τον m-Argus που παρέχεται στην εικόνα 7. Για την επίτευξη της k-ανωνυμίας δεν επιβάλλονται οι κατασταλμένες τιμές, καθιστώντας το ευάλωτο σε σύνδεση, καθώς και σε μια «επίθεση συμπεράσματος» χρησιμοποιώντας περίληψη δεδομένων. Για παράδειγμα, γνωρίζοντας το συνολικό αριθμό των ανδρών και των γυναικών επιτρέπει στις κατασταλμένες τιμές στα δύο φύλα να συναχθούν.

Εθνικότητα	Ημερομηνία Γέννησης	Φύλο	Ταχυδρομικός Κώδικας	Εμφανίζεται	Sid	Ακραίες τιμές
Έγχρωμος	1965	Άνδρας	02141	2	t1,t2	{}
Έγχρωμος	1965	Γυναίκα	02138	2	t3,t4	{}
Έγχρωμος	1964	Γυναίκα	02138	2	t5,t6	{}
Λευκός	1964	Άνδρας	02138	1	t7	{{**Ημερομηνία Γέννησης**, Φύλο, Ταχυδρομικό Κώδικα}, {Εθνικότητα, **Ημερομηνία Γέννησης**, Ταχυδρομικό Κώδικα}}
Λευκός	1965	Γυναίκα	02139	1	t8	{{**Ημερομηνία Γέννησης**, Ταχυδρομικός Κωδικός}, {**Φύλο**, Ταχυδρομικό Κώδικα}, {**Ημερομηνία Γέννησης, Φύλο** Ταχυδρομικό Κώδικα}, {Εθνικότητα, **Ημερομηνία Γέννησης, Φύλο**}, {Εθνικότητα, **Ημερομηνία Γέννησης**, Ταχυδρομικό Κώδικα}, {Εθνικότητα, **Φύλο**}, {Εθνικότητα, **Ημερομηνία Γέννησης**}}

Λευκός	1964	Άνδρας	02139	2	t9,t10	{}
Λευκός	1967	Άνδρας	02138	2	t11,t12	{}

Πίνακας 49: Freq πριν την καταστολή

Το πραγματικό πρόγραμμα m-Argus παρέχει MTactual όπως φαίνεται στον πίνακα. Η πλειάδα αναγνωρισμένη ως t7 ["Λευκός", "1964", "Άνδρας", "02138"], η οποία είναι μοναδική σε όλο το MTactual [QI]. Ως εκ τούτου, το MTactual δεν ικανοποιεί την απαίτηση για k = 2.

Ένα μειονέκτημα του m-Argus απορρέει από τη μη εξέταση όλων των συνδυασμών των χαρακτηριστικών στο ψευδο-αναγνωριστικό. Μόνο 2- και 3- συνδυασμοί εξετάστηκαν. Μπορεί να υπάρχουν 4-συνδυασμοί ή μεγαλύτεροι που είναι μοναδικοί. Άμεσα επεκτείνοντας τον m-Argus για να υπολογίσουμε ότι σε όλους τους συνδυασμούς χάνει την υπολογιστική αποδοτικότητα του. Έτσι, οι γενικεύσεις από τον αλγόριθμο m-Argus δεν μπορεί να ικανοποιούν πάντα την k-ανωνυμία, έστω και αν όλες οι γενικεύσεις από το σύστημα Datafly ικανοποιούν την k -ανωνυμία. Και οι δύο αλγόριθμοι μπορούν να παρέχουν γενικεύσεις που δεν είναι k-ελάχιστες στρεβλώσεις, διότι και οι δυο επιβάλλουν γενίκευση στο επίπεδο του χαρακτηριστικού. Αυτό καθιστά αδρά μέτρα μετρικής ακριβείας (Prec). Ενδέχεται να υπάρχουν τιμές στον πίνακα που όταν είναι γενικευμένες σε επίπεδο κελιών , ικανοποιούν το k χωρίς να τροποποιούν όλες τις τιμές στο χαρακτηριστικό. Εν ολίγοις, χρειάζεται περισσότερη δουλειά για να διορθώσει κανείς αυτές τις ευρετικής βάσεως προσεγγίσεις.

id	Εθνικότητα	Ημερομηνία Γέννησης	Φύλο	Ταχυδρομικός Κώδικας
t1	Έγχρωμος	1965	Άνδρας	02141
t2	Έγχρωμος	1965	Άνδρας	02141
t3	Έγχρωμος	1965	Γυναίκα	02138
t4	Έγχρωμος	1965	Γυναίκα	02138
t5	Έγχρωμος	1964	Γυναίκα	02138
t6	Έγχρωμος	1964	Γυναίκα	02138
t7	Λευκός		Άνδρας	02138
t8	Λευκός			02139
t9	Λευκός	1964	Άνδρας	02139
t10	Λευκός	1964	Άνδρας	02139
t11	Λευκός	1967	Άνδρας	02138
t12	Λευκός	1967	Άνδρας	02138

MT

id	Εθνικότητα	Ημερομηνία Γέννησης	Φύλο	Ταχυδρομικός Κώδικας
t1	Έγχρωμος	1965	Άνδρας	02141
t2	Έγχρωμος	1965	Άνδρας	02141
t3	Έγχρωμος	1965	Γυναίκα	02138
t4	Έγχρωμος	1965	Γυναίκα	02138
t5	Έγχρωμος	1964	Γυναίκα	02138
t6	Έγχρωμος	1964	Γυναίκα	02138
t7	Λευκός	1964	Άνδρας	02138
t8	Λευκός		Γυναίκα	02139
t9	Λευκός	1964	Άνδρας	02139
t10	Λευκός	1964	Άνδρας	02139
t11	Λευκός	1967	Άνδρας	02138
t12	Λευκός	1967	Άνδρας	02138

MTactual

Πίνακας 50: Αποτέλεσμα από τον αλγόριθμο m-Argus και το πρόγραμμα

7 ΕΠΙΛΟΓΟΣ

7.1 ΣΥΝΟΨΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Η εργασία αυτή ασχολήθηκε με το πρόβλημα της διασφάλισης της ιδιωτικότητας των εγγραφών σε βάσεις δεδομένων. Θεωρήσαμε την περίπτωση επίθεσης κατά την οποία ο επιτιθέμενος έχει μερική γνώση τιμών μιας εγγραφής του συνόλου δεδομένων 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.

Μελετήθηκαν ευριστικοί αλγόριθμοι που εγγυώνται την ικανοποίηση της k-ανωνυμίας του συνόλου των δεδομένων. Οι αλγόριθμοι βασίζονται στην χρήση της ολικής γενίκευσης χωρίς τη χρήση κάποιας ιεραρχίας. Παράλληλα αναλύθηκε αλγόριθμος της k^m-ανωνυμίας με χρήση ιεραρχιών γενίκευσης. Η απόδοσή των αλγορίθμων εξετάστηκαν ως προς την μετρική απώλειας πληροφορίας (Κανονικοποιημένη Ποινή Βεβαιότητας) σε διαφορετικά δεδομένα εισόδου.

Όπως φαίνεται και από τα αποτελέσματα, η υπεροχή του αλγορίθμου που παρουσιάζεται για το πρόβλημα αυτό είναι ξεκάθαρη, καθώς επιτυγχάνει πολύ μικρότερη απώλεια πληροφορίας από τον αλγόριθμο της k^m-ανωνυμίας με χρήση ιεραρχίας. Αυτό οφείλεται στο γεγονός ότι ο προτεινόμενος αλγόριθμος εκμεταλλεύεται τις ιδιότητες των συνεχών γνωρισμάτων της βάσης.

Συμπεραίνεται πως για το πρόβλημα της προστασίας της ιδιωτικότητας σε σύνολα δεδομένων της παραπάνω μορφής από επιθέσεις με μερική γνώση σε κάποιες τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού μιας εγγραφής, ο αλγόριθμος που αναπτύχθηκε επιτυγχάνει μικρότερη απώλεια πληροφορίας συγκριτικά με τον αλγόριθμο της k^m-ανωνυμίας με χρήση ιεραρχιών γενίκευσης.

7.2 ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Μια χρήσιμη επέκταση του αλγόριθμου είναι η χρήση τοπικής γενίκευσης στο σύνολο δεδομένων. Αυτό θα έχει σαν αποτέλεσμα την μείωση της απώλειας πληροφορίας στα δημοσιευμένα σύνολα.

Εκτός αυτού, ο αλγόριθμος λόγω της πρακτικής χρησιμότητας του μπορεί να επεκταθεί και σε διαφορετικά μοντέλα επιθέσεων. Μια χρήσιμη επέκταση αφορά την μελέτη επιθέσεων του επιτιθέμενου με σύνθετη μερική και συναθροιστική γνώση. Σε αυτό το μοντέλο, ο επιτιθέμενος έχει σαν γνωστικό υπόβαθρο (i) ένα σύνολο m τιμών μιας εγγραφής του συνόλου δεδομένων και (ii) μια συναθροιστική γνώση πάνω στο σύνολο των γνωρισμάτων του ψευδο-αναγνωριστικού.

Χαρακτηριστικό παράδειγμα του συγκεκριμένου προβλήματος εμφανίζεται κατά τη δημοσίευση μιας βάσης ιατρικών δεδομένων που περιέχουν τις διάφορες ασθένειες κάθε ασθενούς ατόμου. Σε περίπτωση της δημοσίευσης των αρχικών τιμών υπάρχει ο κίνδυνος αναγνώρισης κάποιας εγγραφής από κάποιον επιτιθέμενο όπου γνωρίζει ένα μέρος των πληροφοριών ενός φυσικού προσώπου και ταυτόχρονα γνωρίζει και τον αριθμό μητρώου (Α.Μ.) του ασθενή.

Τέλος, ένα διαφορετικό μοντέλο επίθεσης θα μπορούσε να αφορά ένα σύνολο δεδομένων τέτοιο ώστε ένα ή περισσότερα γνωρίσματα του ψευδο-αναγνωριστικού να είναι ευαίσθητα. Σε αυτήν την περίπτωση μπορεί να επιχειρηθεί η ικανοποίηση της *l*- διαφορετικότητας για τα ευαίσθητα γνωρίσματα σε συνδυασμό με την km-ανωνυμία.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Abou-el-ela Abdou Hussien, Nermin Hamza και Hesham A. Hefny Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing [Επιθεώρηση]. - 2013.

Aggarwal Gagan [και συν.] Approximation Algorithms for k-Anonymity [Επιθεώρηση]. - 2005.

Dalenius Tore Finding a Needle In a Haystack or Identifying Anonymous Ceusus Records [Επιθεώρηση] // Journal of Official Statistics Vo. 2, No. 3. - 1986. - σσ. 329-336.

Fung Benjamin C.M. [et al.] Privacy-Preserving data publishing: A survey of recent developments [Journal] // ACM Computing Surveys Vo. 42, No. 4, Article 14. - 2010. - p. 53 pages.

Jing Zhang [και συν.] An Improved Algorithm for K-Anonymity. - Βερολίνο : [s.n.], 2012.

Ryan William και Manuel Blum K-Anonymity. - 2007.

Sweeney Latanya Achieving k-anonymity privacy protection using generalization and suppression [Επιθεώρηση] // International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. - 2002. - σσ. 571-588.

Sweeney Latanya k-anonymity: a model for protecting privacy [Επιθεώρηση] // International Journal on Uncertainty,Fuzziness and Knowledge-based Systems. - 2002. - σσ. 557-570.

Αγγέλη Σωτήρη k-Ανωνυμοποίηση Συλλογών Δεδομένων. - Αθήνα : [s.n.], 2014.

Γιαννακόπουλος Ιωάννης Κ. Ανωνυμοποίηση σχεσιακών δεδομένων σε κατανεμημένα περιβάλλοντα [Έκθεση]. - Αθήνα : Εθνικό Μετσόβιο Πολυτεχνείο , 2012.

Λεπενιώτη Αικατερίνη Προστασία ιδιωτικότητας από επιτιθέμενους με συναθροιστική γνώση. - Αθήνα : [s.n.], 2013.

Ριζομυλιώτης [Ηλεκτρονικό]. - Πανεπιστήμιο Αιγαίου, 24 Ιανουάριος 2012. - http://www.icsd.aegean.gr/website_files/metaptyxiako/391971949.pdf.

Σπίνου Μαρία Διασφάλιση Εμπιστευτηκότητας σε δυναμικά χωρικά δεδομένα υλοποιώντας το προτεινόμενο βασικό περιβάλλον για ικανοποίηση k-anonymity. - Σάμος : [s.n.], Σεπτέμβριος 2011.

ΓΛΩΣΣΑΡΙ

Attribute	Στήλη – γνώρισμα
Cloaking region (CR)	Περιοχή απόκρυψης
Complementary release attack	Επίθεση συμπληρωματικής έκδοσης
Domain Generalization Hierarchy (DGH_A)	Ιεραρχία γενίκευσης με βάση τον τομέα
Equivalence class	Κλάση ισοδυναμίας
Generalization	Γενίκευση
K-anonymity	Κ-ανωνυμία
Key Attribute	Ιδιότητες κλειδιά
k^m-anonymity	k^m-ανωνυμία
l-diversity	l-διαφορετικότητα
Minimal attack	Επίθεση ελαχιστοποίησης
Minimal generalization	Ελάχιστη γενίκευση
Presision Metric (Prec)	Μετρική ακρίβεια
Quasi-Identifier	Ψευδό-αναγνωριστικό
Reciprocal spatial k-anonymity	Αμοιβαία χωρική k-ανωνυμία
Sensitive attributes	Ευαίσθητα γνωρίσματα
Suppression	Συμπίεση
Temporal attack	Χρονική επίθεση
Tuple	Πλειάδα
Unsorted matching attack	Επίθεση μη ταξινομημένης αντιστοίχισης
Value Generalization Hierarchy (VGH_A)	Ιεραρχία γενίκευσης του πεδίου τιμών