ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

# Διασφάλιση απορρήτου σε δεδομένα γράφων και σχεσιακών βάσεων δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**ΨΑΡΑΥΤΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ**

**Επιβλέπων : Ασκούνης Δημήτριος**

Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

# Διασφάλιση απορρήτου σε δεδομένα γράφων και σχεσιακών βάσεων δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**ΨΑΡΑΥΤΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ**

**Επιβλέπων : Ασκούνης Δημήτριος**

Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 2ή Μαρτίου 2016.

| | | |
|---|---|---|
| ..... | ..... | ..... |
| Δημήτριος Ασκούνης | Ιωάννης Ψαρράς | Γρηγόριος Μέντζας |
| Αναπ. Καθηγητής Ε.Μ.Π | Καθηγητής Ε.Μ.Π | Καθηγητής Ε.Μ.Π. |

Αθήνα, Μάρτιος 2016

…….....

Ψαραύτης Κωνσταντίνος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

# Περίληψη

Ένας μεγάλος αριθμός από επιχειρήσεις, οργανισμούς και άλλες οντότητες συλλέγουν και επεξεργάζονται προσωπικά δεδομένα από ανθρώπους, τα οποία πολλές φορές δημοσιεύονται για ερευνητικούς ή άλλους προωθητικούς σκοπούς. Η διανομή και χρήση αυτών των δεδομένων μπορεί να μην είναι ασφαλής και να αποκαλυφθούν ευαίσθητες πληροφορίες ανθρώπων.

Η παρούσα διπλωματική πραγματεύεται την διασφάλιση της ανωνυμίας στα κοινωνικά δίκτυα με την απεικόνιση και αποθήκευση δεδομένων σε γράφους καθώς και σε εφαρμογές που κάνουν χρήση τις σχεσιακές βάσεις δεδομένων. Σε κάθε περίπτωση προτείνονται μοντέλα τα οποία εφαρμόζουν αλγορίθμους ανωνυμοποίησης ανάλογα την ιδιαιτερότητα του κάθε ζητήματος.

Η σωστή επιλογή του μοντέλου με την εφαρμογή του εξασφαλίζει την καλή προστασία των προσωπικών δεδομένων και τα χαμηλά ποσοστά απώλειας πληροφορίας. Στα πλαίσια αυτά και για τις ανάγκες ανωνυμοποίησης μιας σχεσιακής βάσης δεδομένων αναπτύχθηκε εργαλείο που είναι ικανό να προτείνει ένα κατάλληλο μοντέλο με την προβολή και την απάντηση από τον χρήστη προσεκτικά επιλεγμένων ερωτήσεων.

**Λέξεις κλειδιά**: Ανωνυμοποίηση, Προσωπικά δεδομένα, Σχεσιακές βάσεις δεδομένων, Κοινωνικά δίκτυα, Προστασία απορρήτου

# Abstract

A large number of companies, organizations and other entities collect and elaborate personal data from people, which are frequently published for research or other promotional purposes. The distribution and use of this data may not be safe, and sensitive personal data may be disclosed.

This thesis deals with the guarantee of anonymity in social networks with the display and storage of data in graphs, as well as in applications that use relational databases. In any case, models are proposed which apply anonymization algorithms depending on the particularity of each issue.

The correct choice of a model along with its application guarantees the effective protection of personal data and the low percentage of information loss. In this context and, as far as the need of anonymization of a relational database is concerned, a tool has been developed, which is able to propose a suitable model after the user has viewed and answered questions which have been carefully selected.

**Key words**: Anonymization, Personal data, Relational databases, Social networks, Privacy protection

*Ευχαριστώ θερμά την οικογένειά μου για την ηθική και υλική τους στήριξη προκειμένου να μπορέσω να ολοκληρώσω τις σπουδές μου και να εκπληρώσω απερίσπαστα τους στόχους μου.*

*Ευχαριστώ ιδιαίτερα τον καθηγητή μου κ. Δημήτριο Ασκούνη για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο.*

*Ευχαριστώ  τον διδάκτορα Δημήτριο Παπασπύρο για την καθοδήγηση και την ά-μεση βοήθεια του σε κάθε βήμα της παρούσας διπλωματικής εργασίας.*

# *Περιεχόμενα*

# *Κατάλογος Σχημάτων*

# *Κατάλογος Πινάκων*

# *Κεφάλαιο 1. Εισαγωγή*

Η δημοσίευση πληροφοριών είναι ζωτικής σημασίας για την αξιοποίηση τους από πολλές επιστήμες, εκπαιδευτικά ιδρύματα, εταιρίες, οργανισμούς, καθώς και ερευνητές. Παράλληλα με την εύκολη, γρήγορη και οικονομική πρόσβαση στο διαδίκτυο και τους αποθηκευτικούς χώρους, ο όγκος των πληροφοριών που είναι διαθέσιμος στο ευρύ κοινό σημειώνει ραγδαία αύξηση. Η διανομή και χρήση αυτών των δεδομένων μπορεί βλάψει την ιδιωτικότητα των ατόμων, αφήνοντας εκτεθειμένα ευαίσθητα προσωπικά τους στοιχεία. Για το λόγο αυτό έχει αναπτυχθεί τόσο η Ελληνική και η Ευρωπαϊκή νομοθεσία, όσο και τεχνικές ανωνυμοποίησης των δεδομένων προς δημοσίευση, που ως στόχο τους έχουν την προστασία της ταυτότητας του ατόμου και άλλων ευαίσθητων στοιχείων του. Σε μελέτη μάλιστα που πραγματοποιήθηκε από τους Alessandro [1], αποδείχθηκε πως η ελάχιστη τιμή στην οποία θα ήταν διατεθειμένος ένας πολίτης να πουλήσει προσωπικά του δεδομένα είναι υψηλότερη από την τιμή που δέχεται να πληρώσει για να προστατεύσει την ανωνυμία του. Ο καθένας μας έχει το δικαίωμα να προστατεύεται νομικά από επεμβάσεις στην ιδιωτική του ζωή, την οικογένεια, την κατοικία ή την αλληλογραφία του.

## *1.1 Οργάνωση κειμένου*

Η παρούσα εργασία αποτελείται από 8 κεφάλαια. Πέρα από το παρόν εισαγωγικό κεφάλαιο, το περιεχόμενο των υπολοίπων κεφαλαίων συνοψίζεται ως εξής:

Στο δεύτερο κεφάλαιο αναλύεται το θεσμικό πλαίσιο για την προστασία της ιδιωτικότητας στην Ευρώπη γενικά και στην Ελλάδα ειδικά. Περιγράφονται οι βασικές αρχές προστασίας των προσωπικών δεδομένων όπως έχουν διαμορφωθεί και οι συνθήκες κάτω από τις οποίες επιτρέπεται η επεξεργασία των δεδομένων και τα δικαιώματα των πολιτών.

Στο τρίτο κεφάλαιο περιγράφεται το βασικό θεωρητικό υπόβαθρο για την κατανόηση του γνωστικού πεδίου της ανωνυμοποίησης δεδομένων που περιλαμβάνει την θεωρία

των γράφων, τις σχεσιακές βάσεις δεδομένων, τον τρόπο επιλογής των ψευδο-αναγνωριστικών και τις τεχνικές εξόρυξης δεδομένων.

Στο τέταρτο κεφάλαιο γίνεται μία ολοκληρωμένη ανάλυση της προστασίας των προσωπικών δεδομένων στα κοινωνικά δίκτυα.

Στο πέμπτο κεφάλαιο πραγματοποιείται ανάλυση στις σχεσιακές βάσεις δεδομένων, στις τεχνικές γενίκευσης, στις μετρικές απώλειας πληροφορίας, σε 18 μοντέλα ανωνυμοποίησης και σε μια σειρά αλγορίθμων ανωνυμοποίησης.

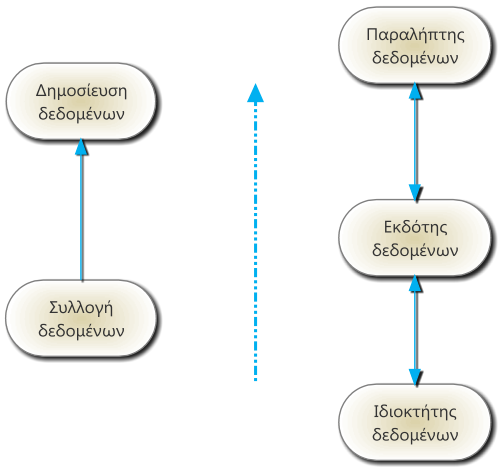
Στο έκτο κεφάλαιο παρουσιάζεται το εργαλείο εύρεσης κατάλληλου μοντέλου ανωνυμοποίησης και οι τεχνικές λεπτομέρειές του.

Στο έβδομο κεφάλαιο δίνεται ο επίλογος, με μία σύνοψη, ανασκόπηση των θεμάτων που μελετήθηκαν στην εργασία και προτάσεις για μελλοντικές επεκτάσεις.

Τέλος, στο Κεφάλαιο 8 παρουσιάζεται αναλυτικά η βιβλιογραφία που μελετήθηκε για την συγγραφή και ανάπτυξη του έργου.

## *1.2 Διαδικασία δημοσίευσης δεδομένων*

Η διαδικασία της προστασίας και έκδοσης προσωπικών δεδομένων συνοψίζεται σε δυο φάσεις [2]: την *συλλογή* και την *δημοσίευση* τους. Ο ανθρώπινος παράγοντας αποτελείται από τρεις ρόλους: τον *ιδιοκτήτη*, τον *εκδότη* και τον *παραλήπτη* των δεδομένων. Γραφικά, η σχέση των δύο φάσεων και των τριών ρόλων φαίνεται παρακάτω:




Σχήμα 1.1: Σχέση των δύο φάσεων και των τριών ρόλων

Στην πρώτη φάση, οι εκδότες συλλέγουν το σύνολο των πληροφοριών από τους ιδιοκτήτες τους και τις επεξεργάζονται. Στη συνέχεια, οι εκδότες δημοσιεύουν τις επεξεργασμένες βάσεις δεδομένων και τις κάνουν διαθέσιμες στους παραλήπτες, οι οποίοι, στο τέλος, έχουν το απαιτούμενο υλικό για ανάλυση και μελέτη. Για παράδειγμα, το νοσοκομείο συλλέγει τα δεδομένα από τους ασθενείς και δημοσιεύει τις εγγραφές τους σε κάποιο εξωτερικό ιατρικό κέντρο. Το νοσοκομείο είναι ο εκδότης, οι ασθενείς είναι οι ιδιοκτήτες και το ιατρικό κέντρο οι παραλήπτες των δεδομένων. Από την επεξεργασία των δεδομένων στο ιατρικό κέντρο μπορεί προκύψουν απλά συμπεράσματα, όπως το πλήθος των αντρών που διεγνώσθησαν με διαβήτη μέχρι και πιο σύνθετα, όπως οι λόγοι για τους οποίους οι συγκεκριμένοι ασθενείς δεν κολλάνε γρίπη.

Οι εκδότες πληροφοριών χωρίζονται σε δύο κατηγορίες βάσει των ακόλουθων μοντέλων [3]: στο *μη αξιόπιστο* μοντέλο, ο εκδότης δεν είναι έμπιστος και μπορεί να προσπαθήσει να αποκαλύψει τις ευαίσθητες πληροφορίες των ιδιοκτητών. Για το μοντέλο αυτό έχουν προταθεί λύσεις με μεθόδους κρυπτογραφίας, στατιστικής και ανώνυμης επικοινωνίας, ώστε να συλλέγονται οι πληροφορίες χωρίς να αφήνουν ίχνη για την μετέπειτα σύνδεσή τους με τους ιδιοκτήτες. Στο *αξιόπιστο* μοντέλο, ο εκδότης είναι έμπιστος και οι ιδιοκτήτες μπορούν άφοβα να τους αποκαλύψουν προσωπικές πληροφορίες. Ωστόσο, η ίδια εμπιστοσύνη δεν πρέπει να παρέχεται και στους παραλήπτες.

Οι παραλήπτες δεδομένων μπορεί να είναι ταυτόχρονα και οι επιτιθέμενοι. Μια φαρμακευτική εταιρεία, για παράδειγμα, μπορεί να είναι μια έμπιστη οντότητα, ωστόσο δεν μπορούμε να υποθέσουμε πως όλο το προσωπικό της είναι εξίσου έμπιστο. Με βάση αυτήν την υπόθεση έχει προταθεί η κρυπτογράφηση των δεδομένων και τα κλειδιά για την ανάγνωσή τους να τα έχουν μόνο οι έμπιστοι παραλήπτες. Το μεγάλο στοίχημα του τομέα της διατήρησης του απορρήτου στα δημοσιευμένα δεδομένα είναι να καταφέρει να προστατεύσει ταυτόχρονα την ιδιωτικότητα και την χρησιμότητα των πληροφοριών στις ανωνυμοποιημένες βάσεις δεδομένων.

# *Κεφάλαιο 2. Θεσμικό πλαίσιο*

Το κεφάλαιο αυτό παρουσιάζει συνοπτικά τη νομοθεσία και τις αρχές που διέπουν την προστασία των προσωπικών δεδομένων. Συγκεκριμένα περιγράφονται οι βασικές αρχές προστασίας των προσωπικών δεδομένων όπως έχουν διαμορφωθεί και γίνει αποδεκτές. Στη συνέχεια, παρατίθεται συνοπτική αλλά περιεκτική περιγραφή του Ευρωπαϊκού Δικαίου, το οποίο αποτελεί τη βάση για το Ελληνικό Εθνικό Δίκαιο. Τέλος γίνεται αναφορά στις συνθήκες κάτω από τις οποίες επιτρέπεται η επεξεργασία των δεδομένων και τα δικαιώματα των πολιτών.

## *2.1 Αρχές προστασίας απορρήτου*

Η κείμενη νομοθεσία και η διασφάλιση του απορρήτου διέπονται από τις παρακάτω βασικές αρχές όπως αυτές παρουσιάζονται στο [4]:

1. **Αρχή του περιορισμού της συλλογής**

   Η συλλογή προσωπικών δεδομένων πρέπει να υπόκειται σε περιορισμούς με σύννομα και δίκαια μέσα και, όταν είναι απαραίτητο, με τη συγκατάθεση του υποκειμένου των δεδομένων.

2. **Αρχή της ποιότητας των δεδομένων**

   Τα προσωπικά δεδομένα θα πρέπει να είναι σχετικά με το σκοπό για τον οποίο πρόκειται να χρησιμοποιηθούν και στο βαθμό που είναι απαραίτητο για το σκοπό αυτό θα πρέπει να είναι ακριβή, πλήρη και ενημερωμένα.

3. **Αρχή προσδιορισμού του σκοπού**

   Ο σκοπός για τον οποίο συλλέγονται προσωπικά δεδομένα θα πρέπει να προσδιορίζεται το αργότερο κατά τη χρονική στιγμή της συλλογής τους, ενώ η χρήση τους θα πρέπει να περιορίζεται στην εκπλήρωση του σκοπού αυτού ή κάποιου πλήρως συμβατού σκοπού.

4. **Αρχή περιορισμού της χρήσης:**

   Τα προσωπικά δεδομένα δε θα πρέπει να αποκαλύπτονται σε τρίτους ή να χρησιμοποιούνται για άλλο σκοπό εκτός από τον προσδιορισμένο σύμφωνα με την αρχή προσδιορισμού του σκοπού, εκτός εάν υπάρχει η σχετική συναίνεση του χρήστη ή η εξουσιοδότηση από το νόμο.

5. **Αρχή της προστασίας της ασφάλειας:**

   Τα προσωπικά δεδομένα θα πρέπει να προστατεύονται με χρήση των κατάλληλων μηχανισμών απέναντι σε κινδύνους όπως η μη εξουσιοδοτημένη πρόσβαση, καταστροφή, χρήση, τροποποίηση ή κοινοποίηση σε τρίτους.

6. **Αρχή της διαφάνειας:**

   Θα πρέπει να υπάρχει γενική διαφάνεια αναφορικά με τις πολιτικές και πρακτικές που σχετίζονται με τη συλλογή και επεξεργασία των προσωπικών δεδομένων καθώς και με την ταυτότητα του φορέα που διενεργεί τη συλλογή και επεξεργασία.

7. **Αρχή της συμμετοχής του ατόμου:**

   - το κάθε άτομο θα πρέπει να έχει το δικαίωμα, να αποκτά είτε απευθείας από τον υπεύθυνο της επεξεργασίας είτε μέσω κάποιου άλλου τρόπου επιβεβαίωση αναφορικά με το αν ο υπεύθυνος της επεξεργασίας διαθέτει δεδομένα που σχετίζονται με το εν λόγω άτομο.
   - να του ανακοινώνονται δεδομένα που σχετίζονται με αυτό, μέσα σε εύλογο χρονικό διάστημα, με εύλογο τρόπο, σε μορφή εύκολα κατανοητή και εφόσον η ανακοίνωση προϋποθέτει κόστος, αυτό να μην είναι υπερβολικό.
   - να του παρέχονται οι λόγοι για τους οποίους απορρίπτονται αιτήσεις του που αναφέρονται στις δύο παραπάνω παραγράφους και να διατηρεί στην περίπτωση αυτή τη δυνατότητα της αμφισβήτησης της απόρριψης και της περαιτέρω διεκδίκησης.
   - να αμφισβητεί προσωπικά δεδομένα που σχετίζονται με αυτό και σε περίπτωση επιτυχημένης αμφισβήτησης να μπορεί να προχωρεί σε εξάλειψη, διόρθωση ή ολοκλήρωση των δεδομένων αυτών.

8. **Αρχή της ευθύνης**:

   Κάθε υπεύθυνος της επεξεργασίας δεδομένων προσωπικού χαρακτήρα θα πρέπει να είναι υπόλογος αναφορικά με την εφαρμογή των μέτρων εκείνων που προάγουν τις παραπάνω αρχές, οι οποίες πρέπει να διέπουν την προστασία των προσωπικών δεδομένων.

## *2.2 Ευρώπη*

Ο ιδιωτικός βίος των ατόμων προστατεύεται εθιμικά εδώ και αιώνες, αλλά από κάποιο χρονικό σημείο άρχισε να περιβάλλεται και από τυπική ισχύ. Ειδικά κείμενα και ειδικοί νόμοι σε διάφορες ευρωπαϊκές χώρες προέβλεψαν τρόπους προστασίας αλλά και κυρώσεις για κάθε σχετική παραβίαση. Συγκεκριμένα για την προστασία των προσωπικών δεδομένων των Ευρωπαίων πολιτών οι κοινοτικοί νομοθέτες διαμόρφωσαν τις εξής οδηγίες στις χώρες μέλη της Ευρωπαϊκής ενώσεως [5]:

- Οδηγία 95/46/ΕΚ: Αφορά την προστασία των φυσικών προσώπων έναντι της επεξεργασίας δεδομένων προσωπικού χαρακτήρα και την ελεύθερη κυκλοφορία των δεδομένων αυτών

- Οδηγία 2002/58/ΕΚ: Αφορά την επεξεργασία των δεδομένων προσωπικού χαρακτήρα και την προστασία της ιδιωτικής ζωής στον τομέα των ηλεκτρονικών επικοινωνιών.

- Οδηγία 2006/24/ΕΚ: Αφορά τη διατήρηση δεδομένων που παράγονται ή υποβάλλονται σε επεξεργασία σε συνάρτηση με την παροχή διαθεσίμων στο κοινό υπηρεσιών ηλεκτρονικών επικοινωνιών ή δημοσίων δικτύων επικοινωνιών και την τροποποίηση της οδηγίας 2002/58/ΕΚ

- Οδηγία 2009/136/ΕΚ: Περιλαμβάνει την τροποποίηση της οδηγίας 2002/22/ΕΚ για την καθολική υπηρεσία και τα δικαιώματα των χρηστών όσον αφορά δίκτυα και υπηρεσίες ηλεκτρονικών επικοινωνιών. Επιπλέον περιλαμβάνει την τροποποίηση της οδηγίας 2002/58/ΕΚ σχετικά με την επεξεργασία των δεδομένων προσωπικού χαρακτήρα και την προστασία της ιδιωτικής ζωής στον τομέα των ηλεκτρονικών επικοινωνιών και του κανονισμού (ΕΚ) αριθ. 2006/2004 για τη συνεργασία μεταξύ

των εθνικών αρχών που είναι αρμόδιες για την επιβολή της νομοθεσίας για την προστασία των καταναλωτών.

## *2.3 Ελλάδα*

Η Ελλάδα υπήρξε από τις πρώτες χώρες που ενσωμάτωσαν τις κοινοτικές οδηγίες στο εσωτερικό δίκαιο με το νόμο 2472/97 [6] για την επεξεργασία δεδομένων προσωπικού χαρακτήρα, όπως ισχύει μετά τις τροποποιήσεις που κατά καιρούς εισήχθησαν.

Ο νόμος 2472/97 συνιστά ένα προστατευτικό πλαίσιο κανόνων που εδράζεται σε τέσσερις πυλώνες:

1. ένα σύστημα ουσιαστικών ρυθμίσεων που θέτει, αφενός τις προϋποθέσεις νομιμότητας της επεξεργασίας προσδιορίζοντας δεσμευτικά το σημείο ισορροπίας μεταξύ των αντιτιθεμένων δικαιωμάτων και συμφερόντων, και αφετέρου τις βασικές αρχές του νόμου με έμφαση στην αρχή του σκοπού και της αναλογικότητας (άρθρα 4-10),
2. την απονομή δικαιωμάτων στα πρόσωπα ώστε να προστατευτούν τα δικαιώματα και τα συμφέροντά τους (άρθρα 11-14),
3. την εισαγωγή και στην οργάνωση ανεξάρτητου θεσμικού ελέγχου της προστασίας προσωπικών δεδομένων ώστε να εξασφαλίζεται η εφαρμογή της νομοθεσίας (άρθρα 15-20) και
4. τους κανόνες που προβλέπουν διοικητικές, ποινικές και αστικές κυρώσεις σε περιπτώσεις παράβασης του νόμου (άρθρα 21-23).

## *2.4 Προσωπικά και ευαίσθητα δεδομένα*

Τα προσωπικά δεδομένα δε διαθέτουν όλα το ίδιο πληροφοριακό βάρος. Κάποια από αυτά παρουσιάζουν ιδιαίτερο βαθμό ευαισθησίας που τα διαφοροποιεί από τα κοινά, απλά δεδομένα, γι' αυτό και χρήζουν ιδιαίτερης προστασίας. Ο εθνικός νομοθέτης κατά τη διαμόρφωση του ν. 2472/1997 στηρίχθηκε για την κατηγοριοποίηση των δεδομένων στις διατάξεις της Οδηγίας 95/46/ΕΚ η οποία προβλέπει ότι απαγορεύεται η επεξεργασία ορισμένων κατηγοριών προσωπικών δεδομένων λόγω του ιδιαίτερου πληροφοριακού βάρους που αυτές φαίνεται να διαθέτουν, περιλαμβάνοντας όμως στο σύνολο αυτό

και άλλες κατηγορίες.

Οι κυριότερες κατηγορίες απλών προσωπικών δεδομένων θεωρούνται:

- Στοιχεία αναγνώρισης: Προσωπικά στοιχεία, επίσημα στοιχεία ληξιαρχείου, καταγωγή, στοιχεία ταυτότητας, (π.χ. υπηκοότητα), λοιπά στοιχεία αναγνώρισης.
- Προσωπικά χαρακτηριστικά: Φυσικά χαρακτηριστικά, ενδιαφέροντα, συνήθειες, μετακινήσεις, ταξίδια, στοιχεία προσωπικότητας, λοιπά στοιχεία προσωπικών χαρακτηριστικών.
- Οικογενειακές συνθήκες: Έγγαμος βίος, οικογενειακή κατάσταση, κοινωνικές επαφές, λοιπά στοιχεία οικογενειακών συνθηκών.
- Εκπαίδευση: Δεδομένα ακαδημαϊκής δραστηριότητας, τομείς ειδίκευσης και πιστοποιητικά, σπουδαστικό / μαθητικό αρχείο, εγγραφή σε επιτροπές, επαγγελματική ειδίκευση.
- Οικονομική κατάσταση: Έσοδα, περιουσιακά στοιχεία, επενδύσεις, απολογισμός εξόδων, δάνεια, υποθήκες, πιστώσεις, επιδόματα, εργασιακά προνόμια, επιχορηγήσεις, δεδομένα ασφάλισης, σύνταξη γήρατος, αγαθά και υπηρεσίες που προσφέρονται στο άτομο ή που προσφέρει το άτομο, τραπεζικοί λογαριασμοί, πιστωτικές κάρτες, κληρονομιά, αποζημίωση.
- Εργασία: Παρούσα εργασία, δεδομένα πρόσληψης, δεδομένα λύσης εργασιακής σχέσης, ιστορικό εργασίας, εργασιακή συμπεριφορά, περιγραφή εργασίας, αξιολόγηση εργασίας, εκπαιδευτικό αρχείο, δεδομένα ασφαλείας, αμοιβές και κρατήσεις, εργασιακές παροχές.

Οι κυριότερες κατηγορίες «ευαίσθητων» προσωπικών δεδομένων θεωρούνται:

- Φυλετική ή εθνική προέλευση: Προσωπικά δεδομένα που αποκαλύπτουν την ιθαγένεια / υπηκοότητα δεν εντάσσονται στα ευαίσθητα δεδομένα, καθώς η ιθαγένεια αποτελεί νομικό δεσμό του προσώπου με ένα κράτος, σε αντίθεση με την εθνικότητα που αποτελεί πραγματικό δεσμό μεταξύ ενός φυσικού προσώπου και του έθνους προέλευσής του. Σε κάθε περίπτωση επεξεργασίας κοινών δεδομένων φυλετικής ή εθνικής προέλευσης απαιτείται ο έλεγχος του σκοπού της επεξεργασίας ώστε αν αυτή κατατείνει σε περαιτέρω διάκριση εθνικού ή φυλετικού κριτηρίου να παρέχεται επαυξημένη προστασία.

- Πολιτικά φρονήματα: Στην κατηγορία των ευαίσθητων δεδομένων ανήκουν στοιχεία που αποκαλύπτουν άμεσα ή έμμεσα το πολιτικό φρόνημα και πεποιθήσεις, δηλαδή την πολιτική τοποθέτηση ενός ατόμου, όπως αυτή μπορεί να εκφράζεται με την κομματική ένταξη, ή την επιλογή ψήφου στις πολιτικές εκλογές.
- Θρησκευτικές ή φιλοσοφικές πεποιθήσεις: Η έννοια των θρησκευτικών πεποιθήσεων είναι ευρύτερη από την έννοια του θρησκεύματος καθώς μπορεί να περιλαμβάνει π.χ. την πίστη σε μη γνωστή θρησκεία ή και την αθεΐα.
- Συμμετοχή σε ένωση, σωματείο και συνδικαλιστική οργάνωση: Ως ένωση νοείται η ένωση προσώπων χωρίς νομική προσωπικότητα. Ως σωματείο νοείται η ένωση προσώπων που επιδιώκει σκοπό μη κερδοσκοπικό και έχει αποκτήσει νομική προσωπικότητα από και δια της εγγραφής της σε ειδικό δημόσιο βιβλίο που τηρείται στο Πρωτοδικείο της έδρας του. Ως συνδικαλιστική οργάνωση θεωρείται αυτή που ιδρύεται σύμφωνα με το ν. 1264/1982.
- Υγεία και κοινωνική πρόνοια: Ευαίσθητο είναι κάθε δεδομένο που αποκαλύπτει πληροφορίες σχετικές με τη βιολογική υπόσταση και την ψυχική υγεία του ανθρώπου. Με άλλα λόγια, αφορά στοιχεία για τη φυσική και πνευματική του κατάσταση, το ιατρικό του ιστορικό, τις ανικανότητες τις αναπηρίες του. Επίσης, αναφέρονται παροχές κοινωνικής πρόνοιας και συναφείς πληροφορίες όπως έγγραφα και στοιχεία κοινωνικών υπηρεσιών.
- Ερωτική ζωή: Προστατεύονται όλα τα δεδομένα που σχετίζονται με την κουλτούρα της ερωτικής συμπεριφοράς του προσώπου στο σύνολό της. Στοιχεία που περιορίζονται στο στενότερο πεδίο της γενετήσιας ελευθερίας της σεξουαλικής ζωής, όσο και δεδομένα που αφορούν το ευρύτερο πεδίο της ερωτικής δραστηριότητας, συνδέονται με τον πυρήνα του ιδιωτικού βίου, ο οποίος αποτελεί θεμελιώδες προστατευόμενο αγαθό του νόμου.
- Ποινικές διώξεις ή καταδίκες: Ευαίσθητα θεωρούνται τόσο τα δεδομένα που συλλέγονται αφού ασκηθεί ποινική δίωξη, όσο και εκείνα που περιλαμβάνονται σε μια ποινική προκαταρκτική δικογραφία. Η ποινική καταδίκη αποτελεί περιεχόμενο καταδικαστικής απόφασης και προστατεύεται ως προσωπικό δεδομένο. Αντιθέτως οι αθωωτικές αποφάσεις δεν αποτελούν ευαίσθητα δεδομένα χωρίς αυτό να σημαίνει ότι αποκλείεται να περιέχονται τέτοιου είδους ευαίσθητες πληροφορίες στο σκεπτικό τους.

## *2.5 Επεξεργασία Δεδομένων*

Ως επεξεργασία, ορίζεται από το νόμο η πραγματοποίηση κάθε εργασίας ή σειράς εργασιών με ή χωρίς τη βοήθεια αυτοματοποιημένων μεθόδων και η εφαρμογή τους σε δεδομένα προσωπικού χαρακτήρα. Οι κυριότερες ενέργειες [7] που συνδέονται με την επεξεργασία δεδομένων είναι η εξής:

- Συλλογή: Η φάση της συλλογής, είναι προγενέστερη σε σχέση με την επεξεργασία υπό τη στενή έννοια και συγκροτείται ως έννοια από το στάδιο της αναζήτησης των δεδομένων, της εύρεσης και λήψης τους. Από τη στιγμή που η φάση της συλλογής ξεκινά από την αναζήτηση των στοιχείων γίνεται φανερό ότι ο όρος νομιμότητα της συλλογής δεδομένων θα πρέπει να συντρέχει ήδη απ' αυτό το σημείο, διαφορετικά μιλούμε για παράνομη επεξεργασία δεδομένων. Αν η συλλογή γίνεται με αυτοματοποιημένο τρόπο θεωρείται ότι αυτομάτως στοιχειοθετείτε επεξεργασία.

- Καταχώρηση: Της συλλογής έπεται λογικά η καταχώριση των δεδομένων δηλαδή η ένταξη των προσωπικών δεδομένων στη σφαίρα γνώσης ή και επέμβασης του προσώπου που τελεί την επεξεργασία.

- Οργάνωση: Ειδική μορφή καταχώρησης προσωπικών δεδομένων αποτελεί η οργάνωσή τους δηλαδή η ένταξή τους με διαρθρωμένη μορφή σε έναν σχηματισμό που χαρακτηρίζεται από μια στοιχειώδη έστω ενότητα.

- Διατήρηση ή αποθήκευση: Η κατοχή προσωπικών δεδομένων σε ένα στοιχειώδες βάθος χρόνου συνιστά την έννοια της διατήρησης ή αποθήκευσης τους. Η διαφύλαξη προσωπικών δεδομένων σε σκληρό δίσκο, δισκέτα, μαγνητοταινία ηλεκτρονικού υπολογιστή ή σε κουτί ηλεκτρονικού ταχυδρομείου, όπως και η κατοχή φωτογραφιών, μαγνητοταινιών, κειμένων, φιλμ που περιέχουν καταγραφές προσωπικών δεδομένων, αποτελούν παραδείγματα διατήρησης ή αποθήκευσης προσωπικών δεδομένων.

- Τροποποίηση: Η τροποποίηση συνεπάγεται αλλοίωση του δεδομένου, όχι αναγκαστικά δραματικής έκτασης. Πολύ συχνά συντελείται επέμβαση στο φορέα εγγραφής του δεδομένου κατά τρόπο ώστε να επηρεάζεται είτε η αποτύπωσή του, είτε το διανόημα που αυτό εμπεριέχει με ή χωρίς επέμβαση στην αποτύπωσή του. Η τροποποίηση επέρχεται κυρίως με την αφαίρεση ή την προσθήκη στοιχείων στα υφιστάμενα δεδομένα.

- Εξαγωγή: Η έννοια της εξαγωγής αφενός σχετίζεται με την έννοια της ανάκτησης δεδομένων, με την εργασία της άντλησης, συλλογής και λήψης συνόλων δεδομένων από ευρύτερα σύνολα δεδομένων και τέλος εξαγωγή υφίσταται και στην περίπτωση πορισμάτων από τη συσχέτιση, διασύνδεση, διαβίβαση ή διάδοση προσωπικών δεδομένων.

- Χρήση: Η συγκεκριμένη κατηγορία, λοιπόν, εντασσόμενη στο κανονιστικό πλαίσιο της προστασίας από τη χρήση προσωπικών δεδομένων που εισάγεται από το Σύνταγμα, καλύπτει περιπτώσεις μη αυτοματοποιημένης επεξεργασίας για τις οποίες δεν υπάρχει αρχείο.

- Διαβίβαση: Η διαβίβαση προσωπικών δεδομένων συνίσταται στην μετάδοσή τους προς συγκεκριμένο αποδέκτη ή σε συγκεκριμένες ομάδες αποδεκτών. Η διαφορά με την διάδοση είναι ότι αυτή γίνεται προς κάθε κατεύθυνση ή προς απεριόριστο αριθμό αποδεκτών.

Σύμφωνα με την κείμενη νομοθεσία, κάθε ένας που επεξεργάζεται προσωπικά δεδομένα πρέπει να συμμορφώνεται με συγκεκριμένες αρχές που εξασφαλίζουν ότι τα προσωπικά δεδομένα:

- τυγχάνουν επεξεργασίας με τρόπο θεμιτό και νόμιμο,
- τηρούνται για σαφώς καθορισμένους σκοπούς,
- περιορίζονται στα απολύτως απαραίτητα για την επίτευξη των σκοπών αυτών,
- είναι ακριβή και επίκαιρα,
- τηρούνται για ορισμένο χρονικό διάστημα (ανάλογα με τους σκοπούς),
- προστατεύονται από επαρκή μέτρα ασφαλείας και
- δεν διαβιβάζονται σε χώρες που δεν εξασφαλίζουν ικανοποιητικό επίπεδο προστασίας.

Η επεξεργασία προσωπικών δεδομένων που δεν είναι ευαίσθητα επιτρέπεται μόνον όταν το υποκείμενο των δεδομένων έχει δώσει τη συγκατάθεσή του ή αν ισχύει κάποια από τις παρακάτω περιπτώσεις:

1. Τα δεδομένα είναι απαραίτητα για την εκτέλεση σύμβασης.

2. Η επεξεργασία προβλέπεται σε ειδικό νόμο.
3. Η επεξεργασία είναι απαραίτητη για τη διαφύλαξη της ζωής ατόμων που δεν είναι σε θέση να παρέχουν τη συγκατάθεσή τους.
4. Τα δεδομένα χρειάζονται για έργα δημοσίου συμφέροντος ή για άσκηση δημόσιας εξουσίας.
5. Η επεξεργασία απαιτείται για την ικανοποίηση έννομου συμφέροντος που επιδιώκει ο υπεύθυνος επεξεργασίας και το οποίο υπερτερεί των δικαιωμάτων και συμφερόντων των υποκειμένων των δεδομένων.

Όταν η επεξεργασία αφορά ευαίσθητα προσωπικά δεδομένα, επιτρέπεται μόνο μετά από ειδική άδεια της Αρχής Προστασίας Δεδομένων Προσωπικού Χαρακτήρα (Αρχή ή ΑΠΔΠΧ), εκτός ειδικών περιπτώσεων που προβλέπονται από το νόμο. Στην Αρχή θα αναφερθούμε στο επόμενο υποκεφάλαιο.

Για να είναι έγκυρη η συγκατάθεση των χρηστών για την τήρηση και επεξεργασία των προσωπικών τους δεδομένων πρέπει:

- Να είναι ελεύθερη, δηλαδή να παρέχεται χωρίς καμία πίεση ή εξάρτηση.
- Να είναι ειδική, δηλαδή να αφορά συγκεκριμένα προσωπικά δεδομένα, να παρέχεται για συγκεκριμένο σκοπό και να αφορά συγκεκριμένο υπεύθυνο επεξεργασίας.
- Να εκφράζεται σαφώς και όχι να προκύπτει με έμμεσους τρόπους.
- Να πραγματοποιείται κατόπιν ειδικής ενημέρωσης από τον υπεύθυνο επεξεργασίας.
- Να είναι έγγραφη ειδικά για την επεξεργασία ευαίσθητων δεδομένων.

Ο ν. 2472/1997 και η ευρωπαϊκή οδηγία 95/46/ΕΚ απονέμουν δικαιώματα που επιτρέπουν τους πολίτες να ελέγχουν ποιος, πότε και με τι τρόπο επεξεργάζεται δικά τους δεδομένα. Τα δικαιώματα αυτά είναι:

1. Το δικαίωμα ενημέρωσης: Όταν κάποιος επεξεργάζεται προσωπικά δεδομένα, οφείλει να ενημερώνει για την ταυτότητά του και την ταυτότητα του τυχόν εκπροσώπου του, το σκοπό της επεξεργασίας των δεδομένων που συλλέγει, τους αποδέκτες ή τις κατηγορίες αποδεκτών των δεδομένων και τον τρόπο με τον οποίο μπορεί ο ενδιαφερόμενος να ασκήσει το δικαίωμα πρόσβασης.

2. Το δικαίωμα πρόσβασης: Οι ενδιαφερόμενοι έχουν δικαίωμα να μαθαίνουν από τον υπεύθυνο επεξεργασίας αναλυτικές πληροφορίες για τα δεδομένα που τους αφορούν, όπως την προέλευση τους, τους σκοπούς της επεξεργασίας, τις μεθόδους επεξεργασίας και τους αποδέκτες τους.

3. Το δικαίωμα αντίρρησης: Οι χρήστες έχουν δικαίωμα να προβάλουν αντιρρήσεις με σχετικό αίτημα προς τον υπεύθυνο επεξεργασίας και, όπου είναι δυνατόν, να ζητήσουν τη διόρθωση ή διαγραφή των προσωπικών τους δεδομένων.

4. Το δικαίωμα δικαστικής προστασίας: Το δικαίωμα απονέμεται στο υποκείμενο των δεδομένων σε εντελώς συγκεκριμένες περιπτώσεις προσβολής των προσωπικών του δεδομένων.

## *2.6 Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα*

Η Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα [6], ιδρύθηκε με τον νόμο 2472/1997, ο οποίος ενσωμάτωσε την Ευρωπαϊκή Οδηγία 95/46, που θέτει κανόνες για την προστασία των προσωπικών δεδομένων σε όλες τις χώρες της Ευρωπαϊκής Ένωσης.

Αποστολή της Αρχής είναι η προστασία των δικαιωμάτων της προσωπικότητας και της ιδιωτικής ζωής του ατόμου στην Ελλάδα, σύμφωνα με τις διατάξεις των Ν. 2472/1997 και 3471/2006. Προστατεύει τον πολίτη από την παράνομη επεξεργασία των προσωπικών του δεδομένων, ταυτόχρονα και συνδράμει προς αυτόν σε κάθε περίπτωση που διαπιστώνεται παραβίαση των σχετικών δικαιωμάτων του σε κάθε επιχειρησιακό τομέα.

Συγκροτείται από έναν Πρόεδρο, έξι μέλη, και εξυπηρετείται από Γραμματεία που λειτουργεί σε επίπεδο Διεύθυνσης. Ο Πρόεδρος είναι απαραίτητα δικαστικός λειτουργός βαθμού Συμβούλου της Επικρατείας ή αντίστοιχου και άνω. Τόσο ο Πρόεδρος όσο και τα μέλη, καθώς και οι ισάριθμοι αναπληρωτές τους, διορίζονται με τετραετή θητεία που μπορεί να ανανεωθεί μία μόνο φορά. Η Γραμματεία της Αρχής αποτελείται από τρία Τμήματα: Ελεγκτών, Επικοινωνίας, Διοικητικών και Οικονομικών Υποθέσεων.

Οι κύριες διοικητικές και ελεγκτικές αρμοδιότητες της αρχής αφορούν την έκδοση Αδειών, την διενέργεια Διοικητικών Ελέγχων και την εξέταση προσφυγών, καταγγελιών και ερωτημάτων.

Τέλος πρέπει να αναφερθεί πως κάθε υπεύθυνος επεξεργασίας οφείλει να γνωστοποιεί στην Αρχή την τήρηση αρχείου με προσωπικά δεδομένα. Η Αρχή καταχωρεί τη γνωστοποίηση σε ειδικό μητρώο. Μερικές εκ των εξαιρέσεων που επιτρέπονται αναφέρονται παρακάτω:

- Υπάρχει γραπτή συγκατάθεση του υποκειμένου.
- Για τη διαφύλαξη ζωτικού συμφέροντος του υποκειμένου, εάν αυτό τελεί σε φυσική ή νομική αδυναμία να δώσει τη συγκατάθεση.
- Για την αναγνώριση, άσκηση ή υπεράσπιση δικαιώματος ενώπιον δικαστηρίου ή πειθαρχικού οργάνου.
- Για την ιατρική πρόληψη, διάγνωση, περίθαλψη ή τη διαχείριση υπηρεσιών υγείας και για την προστασία της Δημόσιας Υγείας.
- Για λόγους εθνικής ασφάλειας
- Για την εξυπηρέτηση των αναγκών εγκληματολογικής ή σωφρονιστικής πολιτικής και αφορά τη διακρίβωση εγκλημάτων, ποινικές καταδίκες ή μέτρα ασφαλείας
- Για την άσκηση δημοσίου φορολογικού ελέγχου ή δημοσίου ελέγχου κοινωνικών παροχών.
- Για ερευνητικούς και επιστημονικούς αποκλειστικά σκοπούς και υπό τον όρο ότι τηρείται η ανωνυμία και λαμβάνονται όλα τα απαιτούμενα μέτρα για την προστασία των δικαιωμάτων των προσώπων στα οποία αναφέρονται.
- Για δεδομένα δημοσίων προσώπων, εφόσον αυτά συνδέονται με την άσκηση δημοσίου λειτουργήματος ή τη διαχείριση συμφερόντων τρίτων, και πραγματοποιείται αποκλειστικά για την άσκηση του δημοσιογραφικού επαγγέλματος.

# *Κεφάλαιο 3. Θεωρητικό υπόβαθρο*

Στο κεφάλαιο αυτό θα γίνει μια συνοπτική αναφορά στην θεωρία των γράφων, τις σχεσιακές βάσεις δεδομένων, τον τρόπο επιλογής των ψευδο-αναγνωριστικών και τις τεχνικές εξόρυξης δεδομένων. Η ανάλυση και επεξήγηση όρων αναφορικά με τις παραπάνω κατηγορίες κρίνεται απαραίτητη για την ευκολότερη κατανόηση του περιεχομένου των επόμενων κεφαλαίων.

## *3.1 Γράφοι*

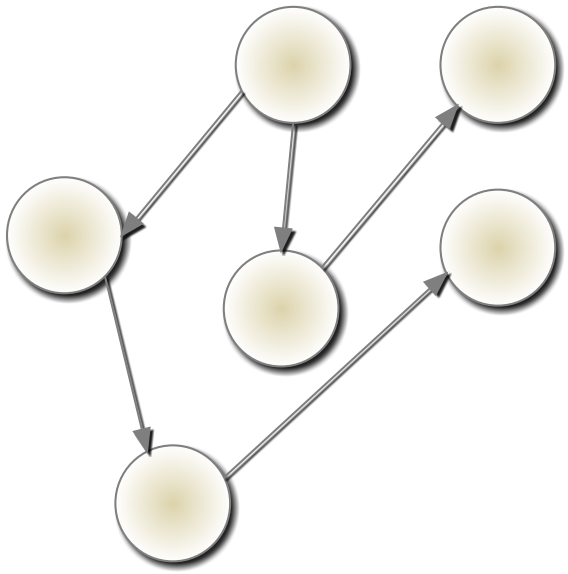
Ένας **γράφος** [8] G = (V, E) χρησιμοποιείται για την αναπαράσταση ενός συνόλου αντικειμένων που συνδέονται μεταξύ τους. Τα αντικείμενα καλούνται *κόμβοι ή κορυφές (vertices)* του γράφου και συμβολίζονται με το αγγλικό V. Οι συνδέσεις ανάμεσα σε ένα ζεύγος κόμβων ονομάζονται *ακμές (edges)* του γράφου και συμβολίζονται με E. Κάθε κόμβος ενός γράφου χαρακτηρίζεται από το *βαθμό* του και είναι ίσος με τον αριθμό των ακμών που συνδέονται με το συγκεκριμένο κόμβο. Δύο ακμές που συνδέονται στον ίδιο κόμβο καλούνται γειτονικές ακμές, ενώ δύο κόμβοι που συνδέονται μεταξύ τους μέσω μιας κοινής ακμής καλούνται γειτονικοί κόμβοι αντίστοιχα. Οι ακμές ενός γράφου μπορεί να είναι κυκλικές (loops), δηλαδή να καταλήγουν στον ίδιο κόμβο από τον οποίο ξεκινούν.

Οι γράφοι ταξινομούνται ανάλογα με τη δομή και τα χαρακτηριστικά τους. Μας ενδιαφέρουν οι παρακάτω:

1. Ένας *απλός γράφος* G = (V, E) ορίζεται ως ένα διάγραμμα αποτελούμενο από κόμβους και ακμές, ενώ μεταξύ δύο κόμβων του γράφου υφίσταται μία και μόνο μία ακμή και δεν περιλαμβάνει κυκλικές ακμές. Σε κάθε κόμβο αντιστοιχεί συνήθως ένας άνθρωπος ή μια ομάδα ανθρώπων. Η ακμή υποδηλώνει την σχέση μεταξύ τους.
2. Ένας γράφος ορίζεται ως *κατευθυνόμενος γράφος* G = (V, A) όταν οι ακμές που συνδέουν τους κόμβους του είναι προσανατολισμένες προς μια κατεύθυνση (φορά), οπότε και αυτές με τη σειρά τους χαρακτηρίζονται ως κατευθυνόμενες ακμές (directed edges). Μια κατευθυνόμενη ακμή a = (x, y) που ανήκει σε έναν

κατευθυνόμενο γράφο, έχει κατεύθυνση από τον κόμβο x προς τον κόμβο y. Ο κόμβος y καλείται άμεσος διάδοχος (direct successor) του x και ο κόμβος x άμεσος προκάτοχος (direct predecessor) του κόμβου y.

Ένα απλό παράδειγμα γραφικής απεικόνισης ενός γράφου:



Σχήμα 3.1: Παράδειγμα απεικόνισης γράφου

Ένα *δέντρο (tree)* είναι ένας μη κατευθυνόμενος απλός γράφος. Ονομάζεται *ριζωμένο* εάν μία κορυφή έχει οριστεί ως ρίζα. Η ρίζα μπορεί να έχει 0 ή περισσότερους κόμβους από κάτω της. Αυτοί οι κόμβοι ορίζουν τα *υποδέντρα (subtree)* της ρίζας. Οι ακμές του θα έχουν προσανατολισμό προς ή από τη ρίζα. Ο *πατέρας (parent)* μιας κορυφής είναι η αμέσως προηγούμενη κορυφή της διαδρομής από τη ρίζα σ' αυτήν. Κάθε κορυφή εκτός από τη ρίζα έχει ένα μοναδικό γονιό. Το *παιδί (child)* μιας κορυφής είναι μια κορυφή της οποίας αυτή είναι πατέρας. Μία κορυφή χωρίς παιδιά λέγεται *φύλλο (leaf)*. Μία *τερματική κορυφή* ενός δέντρου είναι μία κορυφή βαθμού 1. Σε ένα ριζωμένο δέντρο, όλα τα φύλλα είναι τερματικές κορυφές. Οι κόμβοι ενός δέντρου χωρίζονται σε *επίπεδα*, με την έννοια ότι κόμβοι που απέχουν το ίδιο από τη ρίζα βρίσκονται στο ίδιο επίπεδο.

## *3.2 Σχεσιακές Βάσεις Δεδομένων:*

- Ως **δεδομένα (data)** ορίζουμε τις προσωπικές πληροφορίες όπως είναι ο μισθός, ή η κομματική ταυτότητα, οι οποίες καταχωρούνται σε γραμμές και στήλες για την εύκολη και αποτελεσματική διαχείριση τους. Οι παρακάτω πίνακες αποτελούν πληροφορίες πολιτών και ασθενών ενός νοσοκομείου αντίστοιχα:

| Στοιχεία απογραφής πληθυσμού | | | |
|---|---|---|---|
| Όνομα | Ημερομηνία γέννησης | Φύλο | Ταχυδρομικός κώδικας |
| Κώστας | 1990/10/01 | Αρσενικό | 210044 |
| Μαντώ | 1980/05/22 | Θηλυκό | 210022 |
| Βαλάντης | 1982/06/23 | Αρσενικό | 210001 |
| Μαρίνα | 1992/03/12 | Θηλυκό | 210012 |

Πίνακας 3.1: Στοιχεία απογραφής πληθυσμού

| Δεδομένα ασθενών | | | | | |
|---|---|---|---|---|---|
| Αριθμός ταυτότητας | Επάγγελμα | Ημερομηνία γέννησης | Φύλο | Ταχυδρομικός κώδικας | Ασθένεια |
| 231001 | Μαθητής | 1990/10/01 | Αρσενικό | 210044 | Καρδιοπάθεια |
| 231002 | Λογιστης | 1980/05/22 | Θηλυκό | 210022 | Διαβήτης |
| 231003 | Μηχανικος | 1990/08/12 | Αρσενικό | 210021 | Ανεμοβλογιά |
| 231004 | Δικηγορος | 1980/02/25 | Θηλυκό | 210012 | Καρκίνος |

Πίνακας 3.2: Δεδομένα ασθενών

- **Πλειάδα (tuple)** είναι το σύνολο των πληροφοριών κάθε γραμμής που αντιστοιχούν τα στοιχεία ενός μοναδικού κάθε φορά ανθρώπου. Ένα τέτοιο σύνολο τιμών με βάση τον πίνακα 1 είναι: { Μαντώ, 1980/05/22, Θηλυκό, 210022 }. Η διαδικασία καταγραφής μιας πλειάδας ονομάζεται **εγγραφή (record).**
- Η κάθε στήλη αποδίδει μια σημασιολογική ιδιότητα σε κάθε κατηγορία από πληροφορίες και λέγεται **χαρακτηριστικό γνώρισμα (attribute)**.

- Το κάθε χαρακτηριστικό γνώρισμα παίρνει τις τιμές του από ένα διαθέσιμο πεδίο τιμών. Αυτό το πεδίο τιμών ονομάζεται **πεδίο (domain)**. Για παράδειγμα, σε μια βάση δεδομένων μια στήλη μπορεί να αφορά το φύλο ενός ανθρώπου. Τότε το πεδίο για την στήλη αυτή θα είναι: [“Αρσενικό”, “Θηλυκό”].

- **Κλάση ισοδυναμίας (equivalence class)** είναι το σύνολο των εγγραφών που έχουν ίδια ψευδο-αναγνωριστικά. Η κάθε κλάση ισοδυναμίας είναι μοναδική και τις επιμέρους κλάσεις τις συμβολίζουμε ως qid.

- Ως **πίνακα δεδομένων (table)** ορίζουμε το σύνολο όλων των δεδομένων. Τελικά ο κάθε πίνακας δεδομένων θα περιλαμβάνει μια πεπερασμένη σειρά από n-γραμμές με τιμές $<d_1,d_2,\ldots,d_n>$ όπου κάθε τιμή $d_j$ θα αφορά τον χώρο της j-στήλης. Προφανώς το j θα έχει πεδίο ορισμού το [ 1, n ]. Μαθηματικά ένας πίνακας δεδομένων θα ορίζεται ως $B(A_1,...,A_n)$ όπου τα $\{A_1,...,A_n\}$ είναι τα attributes του κάθε tuple. Το κάθε tuple μπορεί να μην είναι μοναδικό, θα αφορά όμως πάντα διαφορετικό άνθρωπο.

Ειδικότερα, τα χαρακτηριστικά γνωρίσματα ενός πίνακα τα κατηγοριοποιούμε [9] σε αναγνωριστικά ταυτότητας, ψευδο-αναγνωριστικά, ευαίσθητα γνωρίσματα και στα μη-ευαίσθητα γνωρίσματα.

1. Τα **αναγνωριστικά ταυτότητας (explicit-identifier)** είναι τιμές που μπορούν αμέσως και χωρίς καμιά άλλη γνώση να προσδιορίσουν την ταυτότητα ενός πολίτη. Προφανώς αυτές οι τιμές πρέπει να διαγράφονται κατευθείαν. Ένα τέτοιο παράδειγμα χαρακτηριστικού γνωρίσματος είναι από τον Πίνακα 3.1 το { Όνομα } και από τον πίνακα 3.2 ο { Αριθμός ταυτότητας } .

2. Τα **ψευδο-αναγνωριστικά (quasi-identifiers)** είναι ένα σύνολο από τα χαρακτηριστικά γνωρίσματα ενός πίνακα, όχι απαραίτητα μοναδικό, τα οποία μπορούν να χρησιμοποιήσουν οι επιτιθέμενοι για να τα διασταυρώσουν με άλλες δημοσιευμένες βάσεις δεδομένων. Με την μέθοδο αυτή μπορούν να άρουν την ανωνυμία και τελικά να αποκαλύψουν ευαίσθητες πληροφορίες. Έστω για παράδειγμα πως ο πίνακας 3.1 είναι γνωστός σε έναν επιτιθέμενο, τότε το σύνολο από χαρακτηριστικά γνωρίσματα: {Ημερομηνία γέννησης, Φύλο, Ταχυδρομικός κώδικας} είναι τα quasi-identifiers του και θα συμβολίζονται ως $QID_1$ .

3. Τα **ευαίσθητα γνωρίσματα (sensitive attributes)** περιλαμβάνουν ευαίσθητες

πληροφορίες όπως ασθένεια, μισθός, περιουσιακά στοιχεία. Η {ασθένεια} του Πίνακα 3.1 είναι ένα ευαίσθητο γνώρισμα.

4. Τα **μη ευαίσθητα γνωρίσματα (non sensitive attributes)** είναι τα χαρακτηριστικά γνωρίσματα που δεν μας ενδιαφέρει αν αποκαλυφθούν η όχι. Για τον λόγο αυτό πολλές φορές αυτά τα χαρακτηριστικά γνωρίσματα δεν τα εμφανίζουμε στην επεξεργασία δεδομένων. Έτσι γλιτώνουμε χρήση μνήμης και βελτιώνουμε πολύ την απόδοση του αλγορίθμου. Το { επάγγελμα } του πίνακα 3.1 είναι ένα παράδειγμα μη ευαίσθητου γνωρίσματος.

## *3.3 Επιλογή ψευδο-αναγνωριστικών*

Μια μεγάλη πρόκληση που καλείται να αντιμετωπίσει ο ιδιοκτήτης δεδομένων είναι η ταξινόμηση των χαρακτηριστικών γνωρισμάτων ενός πίνακα σε τρεις κατηγορίες [10]: στα ψευδο-αναγνωριστικά, στα ευαίσθητα γνωρίσματα και στα μη ευαίσθητα γνωρίσματα. Στα ψευδο-αναγνωριστικά πρέπει να περιλαμβάνεται ένα γνώρισμα Α εάν υπάρχει πιθανότητα ένας αντίπαλος να το έχει αποκτήσει από εξωτερικές πηγές. Έπειτα αφού έχουν καθοριστεί τα ψευδο-αναγνωριστικά, τα υπόλοιπα γνωρίσματα ομαδοποιούνται στα ευαίσθητα και μη ευαίσθητα με βάση το επίπεδο της ευαισθησία τους. Δεν υπάρχει μια ξεκάθαρη απάντηση στον τρόπο με τον οποίο ένας κάτοχος δεδομένων μπορεί να προσδιορίσει ακριβώς ποια γνωρίσματα προέρχονται από εξωτερικές πηγές, αλλά θα πρέπει να γνωρίζει ποιες είναι οι συνέπειες της εσφαλμένης ταξινόμησης. Στην περίπτωση που γίνει λάθος ταξινόμηση, ένα ευαίσθητο γνώρισμα S μπορεί να βρεθεί εκτεθειμένο σε επιθέσεις, αφού μπορεί ένας επιτιθέμενος με το γνώρισμα Α (που δεν εντάχθηκε όπως θα έπρεπε στα ψευδο-αναγνωριστικά) να κάνει σύνδεση πινάκων και τελικά να αποκαλύψει πληροφορίες. Από την άλλη πλευρά, αν ένα ευαίσθητο χαρακτηριστικό γνώρισμα S κατηγοριοποιηθεί εσφαλμένα στα ψευδο-αναγνωριστικά, θα υπήρχε άσκοπη απώλεια πληροφοριών. Η σωστή επιλογή των ψευδο-αναγνωριστικών πάντως παραμένει ακόμα ανοιχτό ζήτημα.

## *3.4 Εξόρυξη δεδομένων*

**Εξόρυξη Δεδομένων (data mining)** [11] είναι η ανάλυση συνήθως τεράστιων παρατηρούμενων συνόλων δεδομένων, έτσι ώστε να βρεθούν μη παρατηρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς, κατανοητούς και χρήσιμους στον κάτοχο των δεδομένων. Πρόκειται για μία σειρά από τεχνικές που βασίζονται σε ανάπτυξη αλγορίθμων και είναι χρήσιμες σε πολλούς και ετερόκλητους κλάδους. Ο όρος εξόρυξη δεδομένων χρησιμοποιείται συχνά αναφερόμενος σε δύο διαφορετικές διαδικασίες: την *ανακάλυψη γνώσης* και την *πρόβλεψη*. Η ανακάλυψη γνώσης παρέχει ρητή πληροφορία η οποία έχει αναγνώσιμη μορφή και μπορεί να γίνει κατανοητή από κάποιον χρήστη. Η πρόβλεψη παρέχει προβλέψεις για μελλοντικά συμβάντα. Σε κάποιες περιπτώσεις μπορεί να είναι διαφανής και αναγνώσιμη, όπως σε συστήματα βασισμένα σε κανόνες (rule based systems), ενώ σε άλλες περιπτώσεις μπορεί να είναι αδιαφανής, όπως σε νευρωνικά δίκτυα (neural networks). Ένας αριθμός μεθόδων εξόρυξης δεδομένων έχουν προταθεί για να ικανοποιήσουν τις απαιτήσεις διαφορετικών εφαρμογών. Όλες επιτυγχάνουν μια ομάδα από διεργασίες εξόρυξης δεδομένων για να προσδιορίσουν και να περιγράψουν ενδιαφέροντα πρότυπα γνώσης.

Στην παρούσα διπλωματική θα αναφερθούμε στις εξής κύριες τεχνικές Εξόρυξης Δεδομένων:

- **Ταξινόμηση (Classification):** Αποτελεί την πιο δημοφιλή και αποτελεσματική μέθοδο. Οι αλγόριθμοι ταξινόμησης εφαρμόζονται σε δεδομένα τα οποία έχουν πρώτα ταξινομηθεί σε κλάσεις, με σκοπό την εξαγωγή κανόνων που μπορούν αργότερα να χρησιμοποιηθούν για την ταξινόμηση νέων εισαγωγών στις κλάσεις. Ένα σύνολο εξαγόμενων κανόνων ονομάζεται ταξινομητής (classifier). Τα βήματα ενός αλγορίθμου ταξινόμησης είναι:

  1. Τροφοδοτούμε τον αλγόριθμο ταξινόμησης με ένα σύνολο από δεδομένα (dataset)
  2. Ο αλγόριθμος έπειτα "κατανοεί" τους κανόνες βάσει των οποίων ταξινομήθηκαν τα δεδομένα
  3. Τέλος βάσει των κανόνων αυτών, ο αλγόριθμος έχει την ικανότητα να ταξινομεί νέα δεδομένα

Ανάλογα με το είδος του ταξινομητή, οι αλγόριθμοι ταξινόμησης χωρίζονται σε αυτούς

που παράγουν λίστες αποφάσεων και σε αυτούς που παράγουν δέντρα αποφάσεων.

- **Συσχέτιση (association)**: Σκοπός της είναι η εύρεση των σημαντικότερων αλληλεξαρτήσεων μεταξύ των διαφόρων πεδίων ή χαρακτηριστικών γνωρισμάτων των πινάκων. Μέσω αυτών των συσχετίσεων διαμορφώνονται μοτίβα και εξάγονται συμπεράσματα.
- **Ομαδοποίηση (clustering):** Οι αλγόριθμοι ομαδοποίησης είναι ιδιαίτερα διαδεδομένοι και η λογική τους μοιάζει αρκετά με αυτή των αλγορίθμων ταξινόμησης. Εξετάζουν και προσδιορίζουν μία ή και παραπάνω κλάσεις και χαρακτηριστικά γνωρίσματα και ομαδοποιούν μεμονωμένες τιμές. Με αυτό το τρόπο, το σύνολο των εγγραφών χωρίζεται σε ομάδες έτσι ώστε οι εγγραφές της ίδιας ομάδας να έχουν περισσότερες ομοιότητες μεταξύ τους, με βάση κάποια προκαθορισμένα κριτήρια, από ότι με εγγραφές άλλων ομάδων. Η τεχνική της ομαδοποίησης μπορεί να είναι είτε στατιστική είτε αριθμητική, οπότε χρησιμοποιούνται διάφορα αριθμητικά κριτήρια ομοιότητας και ο προσδιορισμός των ομάδων τους βασίζεται στο νόημα και στις έννοιες που τα διάφορα αριθμητικά στοιχεία αντιπροσωπεύουν. Οι τιμές που προκύπτουν είναι κατηγορικές και όχι αριθμητικές.

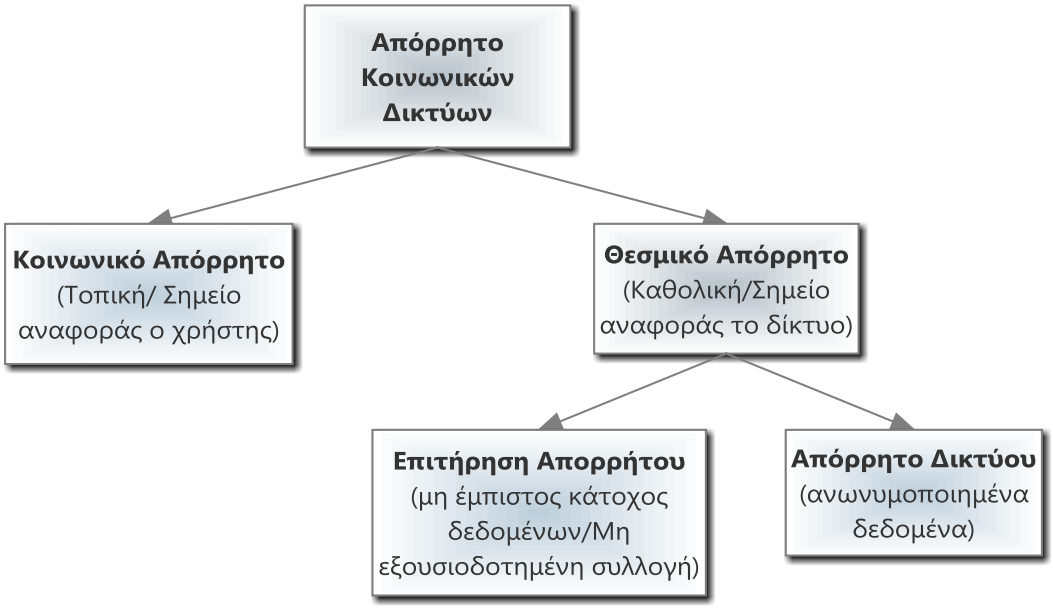
# *Κεφάλαιο 4. Απόρρητο στα κοινωνικά δίκτυα*

Η ανάπτυξη των σελίδων κοινωνικής δικτύωσης συνεχίζει να σημειώνει τεράστια πρόοδο χάρη στην δυνατότητα που προσφέρουν για ικανοποίηση κοινωνικών αναγκών όπως είναι η επικοινωνία μεταξύ των μελών τους και η ένταξη σε ομάδες κοινού ενδιαφέροντος. Οι σελίδες αυτές έχουν μετατραπεί ουσιαστικά σε μια επέκταση της ιδιωτικής μας ζωής και περιλαμβάνουν εκτός από τα αναγνωριστικά μας, λεπτομερή στοιχεία όπως είναι οι σκέψεις, οι πεποιθήσεις ή οι προτιμήσεις μας. Η ανησυχία σχετικά με την προστασία των προσωπικών δεδομένων που σχετίζονται με τις υπηρεσίες κοινωνικής δικτύωσης αποτελεί ένα πολυσυζητημένο θέμα. Οι σελίδες Facebook, LinkedIn και Twitter είναι ήδη γνωστές και διατηρούν ένα ευρύ κοινό όλων των ηλικιακών ομάδων. Πρόσφατα μάλιστα, την εμφάνισή τους έχουν κάνει οι σελίδες Pinterest, Instagram, Vine και Tumblr που απευθύνονται σε νεότερες ηλικίες. Με την αύξηση των χρηστών της κάθε σελίδας, νομοτελειακά αυξάνονται και τα δεδομένα που συλλέγονται αφού αποτελούν το βασικό συστατικό για την έρευνα και την προώθηση προϊόντων και υπηρεσιών. Ταυτόχρονα όμως τα δεδομένα αυτά περιλαμβάνουν ευαίσθητες πληροφορίες που δεν πρέπει να αποκαλυφθούν χωρίς εξουσιοδότηση. Οποιαδήποτε συλλογή και αποθήκευση πληροφοριών ανεξάρτητα από το περιεχόμενό τους και από την πρόθεση του κατόχου μπορεί να οδηγήσει σε παραβίαση της ιδιωτικότητας. Ένα πρόσφατο παράδειγμα αποτελεί η εφαρμογή AOL [12], η οποία, το 2006, προκειμένου να ικανοποιήσει τις ερευνητικές ανάγκες για πραγματικά διαδικτυακά δεδομένα δημοσίευσε το ιστορικό 20 εκατομμυρίων αναζητήσεων που αφορούσαν πάνω από 650000 χρήστες. Αν και τα δεδομένα δημοσιεύθηκαν ανωνυμοποιημένα, ωστόσο, υπήρξαν πολλές καταγγελίες από χρήστες πως αποκαλύφθηκαν τα αναγνωριστικά τους και συνδέθηκαν με τις αναζητήσεις που πραγματοποίησαν.

## *4.1 Κατηγορίες απορρήτου*

Για την καλύτερη δυνατή προστασία των προσωπικών δεδομένων στις σελίδες κοινωνικής δικτύωσης πρέπει να γίνει ένας διαχωρισμός με κριτήριο την οπτική γωνία του ζητήματος [12]. Οι δύο βασικές κατηγορίες είναι το *Κοινωνικό* και το *Θεσμικό Απόρρητο***.** Το Θεσμικό απόρρητο διακρίνεται στην *Επιτήρηση Απορρήτου* και στο *Απόρρητο*

*Δικτύου*. Στο παρακάτω σχήμα παρουσιάζονται γραφικά τα ζητήματα τα οποία θα αναλυθούν ξεχωριστά το καθένα:




Σχήμα 4.1: Κατηγορίες απορρήτου στα κοινωνικά δίκτυα

## *4.2 Κοινωνικό απόρρητο*

Η συνήθης τακτική των σελίδων κοινωνικής δικτύωσης είναι η συλλογή δεδομένων από τον χρήστη και η δημοσίευσή τους σε ένα κοινό βάσει κάποιων ρυθμίσεων απορρήτου που ο ίδιος ο χρήστης και ταυτόχρονα ιδιοκτήτης των δεδομένων έχει καθορίσει. Για παράδειγμα στο Facebook ο χρήστης έχει την δυνατότητα μέσα από μια σειρά ρυθμίσεων απορρήτου να επιλέγει εκείνος με ποιον θα μοιραστεί και τι ακριβώς. Αυτές οι επιλογές αποτελούν ένα βήμα παραπάνω από τα βασικά χαρακτηριστικά που προσφέρουν οι σελίδες κοινωνικής δικτύωσης. Επομένως το σημείο αναφοράς της ιδιωτικότητας εδώ είναι ο ίδιος ο χρήστης και ο τύπος απορρήτου καλείται *Κοινωνικό Απόρρητο* [12]. Τα κύρια ζητήματα που τον αφορούν είναι:

- η ευαισθητοποίηση χρήστη,
- η πολυπλοκότητα επιλογών απορρήτου,
- οι αλλαγές στις επιλογές απορρήτου και οι

- οι διενέξεις επιλογών απορρήτου

Η πρώτη και κύρια ανησυχία της κατηγορίας αφορά τους *μη ευαισθητοποιημένους χρήστες*, εκείνους δηλαδή που δεν αντιλαμβάνονται τον κίνδυνο που διατρέχουν, όταν αποφασίζουν να ανταλλάξουν πληροφορίες σε ένα κοινωνικό δίκτυο. Σε μελέτη μάλιστα που πραγματοποιήθηκε [13] το 2012, περίπου το 8% των Αμερικανών χρηστών του Facebook δεν είχαν καμία γνώση ύπαρξης των επιλογών απορρήτου. Ακόμα πιο ανησυχητικό είναι πως χρήστες που γνώριζαν για τις επιλογές αυτές δεν λάμβαναν τα ενδεδειγμένα μέτρα προστασίας. Επιπλέον το 28% από αυτούς μοιραζόταν προσωπικές πληροφορίες σε κοινό ευρύτερο από τους φίλους τους.

Το επόμενο ζήτημα που χρήζει αναφοράς είναι η μεγάλη πολλές φορές *πολυπλοκότητα των επιλογών του απορρήτου*, η οποία αποθαρρύνει τους χρήστες από το να ασχοληθούν και να επιλέξουν την πολιτική απορρήτου που οι ίδιοι επιθυμούν και ταιριάζει περισσότερο στις ανάγκες τους. Στο Facebook για παράδειγμα οι επιλογές απορρήτου απλώνονται σε έξι διαφορετικές ταμπέλες: Απόρρητο, Χρονολόγιο και Ετικέτες, Μπλοκάρισμα, Ακόλουθοι, Εφαρμογές και Διαφημίσεις. Πρέπει να προστεθεί ακόμη πως οι αναφερόμενες επιλογές δεν είναι εύκολα προσβάσιμες στην πολιτική ορών χρήσης, η οποία πολλές φορές μάλιστα παραλείπεται ή δεν αναλύεται επαρκώς.

Δεν είναι σπάνιο, επίσης, οι *επιλογές του απορρήτου να αλλάζουν* συχνά, γεγονός που συνήθως συμβάλει στην απώλεια ιδιωτικότητας. Πάλι θα αναφερθούμε στο Facebook, το οποίο το 2012 έκανε σημαντικές αλλαγές στις επιλογές και την πολιτική απορρήτου που οδήγησαν στην απλοποίησή τους. Ωστόσο, πολλοί σημειώνουν πως οι αλλαγές δεν ήταν προς το συμφέρον του χρήστη. Για παράδειγμα η επιλογή του Facebook να μην επιτρέπει στους χρήστες του να κρύβουν το Χρονολόγιο τους από άτομα που το αναζητούν. Επιπλέον κάποιες συντομεύσεις απορρήτου απενεργοποιήθηκαν και έγιναν διαθέσιμες μόνο από την κύρια σελίδα απορρήτου. Η νέα πολιτική απορρήτου [14] επίσης επιτρέπει στην σελίδα κοινωνικής δικτύωσης να διαθέτει δεδομένα σε εταιρείες τρίτων πολύ ευκολότερα από ότι η προηγούμενη.

Συγκεκριμένα η νέα πολιτική έχει ως εξής:

*"Μας δίνεται την άδεια να χρησιμοποιήσουμε το όνομά σας, τη φωτογραφία προφίλ, το περιεχόμενο και κάποιες πληροφορίες αναφορικά με διαφημίσεις, προωθήσεις, ή σχετικό περιεχόμενο (όπως την μάρκα της αρεσκείας σας), το οποίο σας προβάλλουμε."*

Ενώ η παλαιά σημείωνε:

*"Μπορείτε να χρησιμοποιήσετε τις επιλογές απορρήτου για να περιορίσετε την σύνδεση του ονόματός και της φωτογραφίας προφίλ σας με διαφημίσεις, προωθήσεις, ή σχετικό περιεχόμενο (όπως την μάρκα που σας αρέσει) που σας προβάλλουμε."*

Τέλος, σημειώνονται πολύ συχνά *διενέξεις στις επιλογές του απορρήτου*, όταν δύο επιλογές αφορούν την ίδια εφαρμογή. Αυτές οι διενέξεις υπάρχουν στα περισσότερα κοινωνικά δίκτυα [15] όπως το Facebook, το MySpace, το Orkut και άλλα. Αν, για παράδειγμα, ένας χρήστης στο Facebook επιλέξει η λίστα των φίλων του να είναι ιδιωτική, τι θα συμβεί όταν κάποιο μέλος της λίστας αυτής επιλέξει η δικιά του λίστα να είναι δημόσια; Ποια θα είναι η επιλογή που θα επικρατήσει; Οι λύσεις που προτείνονται για την διασφάλιση της ιδιωτικότητας σχετικά με την παρούσα κατηγορία αναλύονται στην συνέχεια. Αρχικά, οι νομοθέτες του κάθε κράτους μπορούν να απαιτήσουν οι ιστοσελίδες κοινωνικής δικτύωσης να διατηρούν αυστηρή πολιτική για την προστασία των προσωπικών δεδομένων και μια σειρά από επιλογές απορρήτου, οι οποίες να είναι κατάλληλες για κάθε τύπο δεδομένων. Μπορούν, επίσης, να απαιτήσουν να έχουν για τους χρήστες τους ένα καλό και ευκολονόητο εκπαιδευτικό πρόγραμμα που να τους ενημερώνει και να τους ευαισθητοποιεί στα ζητήματα απορρήτου. Οι ιστότοποι κοινωνικής δικτύωσης οφείλουν, επιπλέον, να είναι σε θέση να παρέχουν άμεσα και αποτελεσματικά λύσεις για τυχόν ζητήματα ιδιωτικότητας. Οι προεπιλεγμένες ρυθμίσεις απορρήτου να είναι όσο το δυνατόν πιο προστατευτικές. Ακόμα, οι επιλογές απορρήτου να είναι εύκολες στη χρήση και να αλλάζουν σπάνια ή και καθόλου. Επίσης, να επιτρέπεται η δημιουργία προφίλ με ψευδώνυμα, και να αποφεύγονται οι διενέξεις επιλογών απορρήτου. Τέλος, θα πρέπει οι χρήστες να είναι ιδιαίτερα προσεκτικοί με την αποκάλυψη προσωπικών τους πληροφοριών και να είναι περισσότερο ενήμεροι στα θέματα τεχνολογίας.

## *4.3 Θεσμικό απόρρητο*

Η δεύτερη κατηγορία, το Θεσμικό απόρρητο [12] έχει ως σημείο αναφοράς του το δίκτυο. Οποιαδήποτε σελίδα κοινωνικής δικτύωσης συλλέξει δεδομένα θα τα χρησιμοποιήσει για διάφορους σκοπούς όπως αυτή θα αναφέρει και στην πολιτική χρήσης της. Το Facebook για παράδειγμα έχει μια λεπτομερή πολιτική χρήσης στην οποία περιγράφει ακριβώς πως χρησιμοποιεί τις πληροφορίες που συλλέγει από τους χρήστες του.

Ενδεικτικά αναφέρει πως:

"Η εμπιστοσύνη σας για εμάς είναι σημαντική και για τον λόγο αυτό δεν δημοσιεύουμε πληροφορίες που σας αφορούν με τρίτους εκτός εάν έχουμε:

- Πάρει την άδειά σας
- Σας έχουμε ήδη ειδοποιήσει, όπως για παράδειγμα στους όρους χρήσης των υπηρεσιών μας ή
- Εάν έχουμε αφαιρέσει το όνομά σας και άλλα αναγνωριστικά προσωπικά σας στοιχεία"

Επομένως, όπως αναφέρεται και παραπάνω, τα δεδομένα κοινωνικών δικτύων πρώτα ανωνυμοποιούνται και έπειτα δημοσιεύονται ή διατίθενται σε εταιρείες. Ωστόσο όπως προκύπτει και από το παράδειγμα της AOL, η διαδικασία της ανωνυμοποίησης μπορεί να μην είναι πλήρως επιτυχής και ως εκ τούτου να ελλοχεύουν κίνδυνοι παραβίασης της ιδιωτικότητας πολλών εγγραφών. Παρακάτω θα αναφερθούμε στις δύο υποκατηγορίες του θεσμικού απορρήτου, την *Επιτήρηση Απορρήτου* και το *Απόρρητο Δικτύου.*

## *4.4 Επιτήρηση απορρήτου*

Η κατηγορία *επιτήρηση απορρήτου* [12] πραγματεύεται την περίπτωση ο ιδιοκτήτης των δεδομένων να μην είναι έμπιστος, καθώς και τη συλλογή χωρίς εξουσιοδότηση δεδομένων. Οι ιδιοκτήτες (στην παρούσα περίπτωση οι σελίδες κοινωνικής δικτύωσης) έχουν απεριόριστη πρόσβαση στα δεδομένα που έχουν συλλέξει από τους χρήστες τους και ως εκ τούτου η προστασία των δεδομένων είναι αδύνατη. Πολλές φορές επίσης συλλέγονται και αποθηκεύονται πληροφορίες από χρήστες χωρίς την πρότερη ενημέρωσή τους. Για παράδειγμα μια πρακτική μη εξουσιοδοτημένης συλλογής δεδομένων από την NSA αποκάλυψε πρόσφατα ο Edward Snowden. Η ιστορία του έχει γίνει πρωτοσέλιδο ανά τον κόσμο αφού αποκαλύφθηκαν πολλά σκάνδαλα. Τέλος άξιο σχολιασμού είναι το ότι δημοσιευμένα δεδομένα μπορεί να παραμένουν δημοσιευμένα ή αποθηκευμένα για πάντα. Είναι ιδιαίτερα δύσκολο να επιβάλλει ένας χρήστης το δικαίωμά του να διαγραφούν από οποιαδήποτε βάση δεδομένων. Η διαγραφή δεδομένων στα κοινωνικά δίκτυα ερμηνεύεται με τον όρο *λήθη* και κάθε χώρα έχει διαφορετική γνώμη και νομοθεσία γι' αυτήν.

Στην κατηγορία της επιτήρησης απορρήτου η πιο αποτελεσματική λύση είναι η αποφυγή δημοσίευσης οποιασδήποτε ευαίσθητης πληροφορίας σε σελίδα κοινωνικής δικτύωσης. Επειδή η λύση αυτή είναι πρακτικά αδύνατη, προτείνονται ορισμένες κρυπτογραφικές μέθοδοι. Κάποιες από αυτές είναι οι εξής:

**FlyByNight** [16]. Η εφαρμογή αυτή είναι σχεδιασμένη για το Facebook και κρυπτογραφεί τα δεδομένα του χρήστη πριν αποθηκευτούν στην σελίδα. Επειδή όμως το FlyByNight βασίζεται στον σέρβερ του Facebook, αποτυγχάνει να προστατέψει την συλλογή δεδομένων από τον πάροχο του δικτύου.

- **NOYB (none of your business) [17].** Η εφαρμογή NOYB είναι επίσης σχεδιασμένη για το Facebook και χρησιμοποιεί μεθόδους κρυπτογραφίας για την προστασία των αναγνωριστικών ταυτότητας των χρηστών. Συγκεκριμένα προστατεύει τον ίδιο τον χρήστη από το Facebook, όμως εφαρμόζεται μόνο για συγκεκριμένα γνωρίσματα του προφίλ και δεν επιτρέπει την κρυπτογράφηση ελεύθερου κειμένου.

- **FaceCloack [17].** Η εφαρμογή αποτελεί επέκταση του προγράμματος περιήγησης Firefox και χρησιμοποιεί ένα συμμετρικό κλειδί για να κρυπτογραφήσει προσωπικές πληροφορίες στο Facebook. Η μέθοδος αυτή κάνει χρήση δικών της σέρβερ οι οποίοι αποθηκεύουν μέρος του προφίλ του χρήστη σε κρυπτογραφημένη μορφή.

- **Scramble [18].** Η εφαρμογή αυτή είναι σχεδιασμένη για γενική χρήση. Τα δεδομένα πριν δημοσιευτούν κρυπτογραφούνται και οι μόνοι που μπορούν να δουν το περιεχόμενο είναι εκείνοι στους οποίος ο ίδιος ο χρήστης έχει διαθέσει το κλειδί για την αποκρυπτογράφηση.

- **Tor [19].** Είναι ένα σύστημα που δίνει στους χρήστες του τη δυνατότητα ανωνυμίας στο Διαδίκτυο. Κάνει δύσκολη την ανίχνευση διαδικτυακής δραστηριότητας του χρήστη. Το πετυχαίνει με την κρυπτογράφηση και δρομολόγηση τυχαίας επικοινωνίας μέσω ενός δικτύου από κόμβους που το λειτουργούν εθελοντές ανά την υφήλιο.

## *4.5 Απόρρητο δικτύου*

Η δεύτερη κατηγορία του θεσμικού απορρήτου αφορά δεδομένα κοινωνικών δικτύων που δημοσιεύονται από έμπιστους αυτή την φορά ιδιοκτήτες και καλείται *απόρρητο δικτύου* [12]. Δεν υπάρχουν πρόσφατες δημοσιεύσεις ανωνυμοποιημένων δεδομένων

κυρίως λόγω της αποτυχίας της AOL να διαφυλάξει με αποτελεσματικό τρόπο τα προσωπικά δεδομένα των χρηστών της. Στόχος των σελίδων κοινωνικής δικτύωσης και των διαφόρων εφαρμογών που αυτές υποστηρίζουν είναι τα δεδομένα που συλλέγουν να αναλυθούν, να επεξεργαστούν και τελικά να αξιοποιηθούν για προώθηση προϊόντων και υπηρεσιών. Η διάθεση των πληροφοριών αυτών, τις περισσότερες φορές πραγματοποιείται έπειτα από την ανωνυμοποίηση τους, η οποία αποτελεί και την βασική λύση της κατηγορίας. Θα αναφερθούμε σε αυτήν εκτενώς στο επόμενο υποκεφάλαιο.

## *4.6 Ανωνυμοποίηση γράφων*

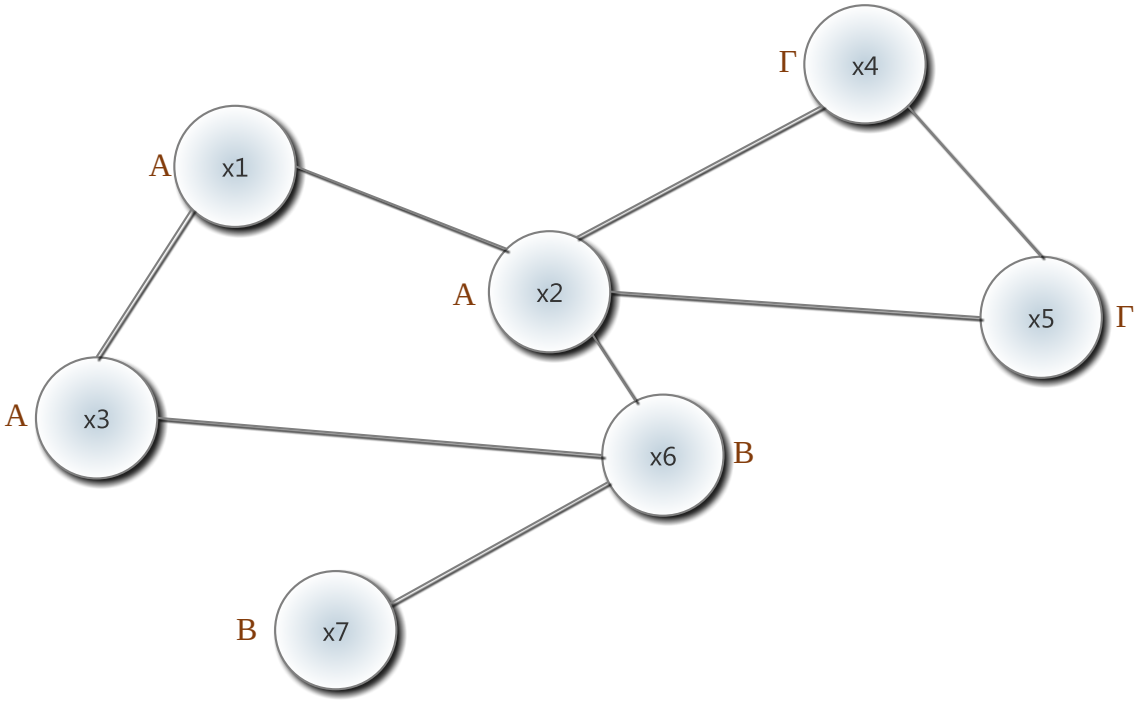
Όπως αναφέρθηκε παραπάνω η αποτελεσματική ανωνυμοποίηση αποτελεί το βέλτιστο τρόπο ικανοποίησης των εξωτερικών αναγκών για πληροφορίες και των χρηστών για την προστασία των ευαίσθητων προσωπικών πληροφοριών τους. Αρχικά, θα αναφερθούμε στα διάφορα μοντέλα με τα οποία απεικονίζονται και αποθηκεύονται οι πληροφορίες των κοινωνικών δικτύων στις βάσεις δεδομένων. Στην συνέχεια, θα γίνει αναφορά στους τύπους των επιθέσεων και τις γνώσεις υποβάθρου, και στο τέλος στα μοντέλα ανωνυμοποίησης των δεδομένων.

### *4.6.1 Μοντέλα δεδομένων*

Στους γράφους έχουν προταθεί τα εξής μοντέλα απεικόνισης δεδομένων [20]:

- Το *απλό μοντέλο*: Η βάση δεδομένων δημοσιεύεται με την μορφή γράφου induced, όπου V είναι το σύνολο των κόμβων και E το σύνολο των ακμών. Κάθε κόμβος εκπροσωπεί συνήθως έναν χρήστη ή μια ομάδα ανθρώπων. Η ακμή μαρτυρά την σχέση μεταξύ τους.

- Οι *κόμβοι με ετικέτες*: Εδώ ο γράφος G=(V, E), είναι εμπλουτισμένος με ετικέτες. Ο όρος L απεικονίζει το σύνολο των ετικετών και υπάρχει συνάρτηση αντιστοίχισης L: V→L η οποία αναθέτει σε κάθε κόμβο V από μια ετικέτα L. Για παράδειγμα ως ετικέτα μπορεί οριστεί το επάγγελμα του χρήστη. Το σχήμα 4.2 απεικονίζει ένα τέτοιο κοινωνικό δίκτυο με ετικέτες Α, Β και Γ.




Σχήμα 4.2: Απεικόνιση κοινωνικού δικτύου

- *Κόμβοι με συνημμένα γνωρίσματα ή δεδομένα*: Κάθε κόμβος ουσιαστικά αφορά ένα χρήστη και πάνω στον κόμβο προσκολλώνται όλα τα γνωρίσματα και οι πληροφορίες του χρήστη.
- *Ευαίσθητες ακμές*: Η βάση δεδομένων είναι ένας πολυ-γράφος $G=(V, E_1, E_2, \ldots, E_s)$, όπου V είναι το σύνολο των κόμβων και $E_i$ το σύνολο των ακμών. Το $E_s$ ορίζει τις ευαίσθητες πληροφορίες και σχέσεις.

### *4.6.2 Μοντέλα επιθέσεων*

Το πρόβλημα στην ανωνυμοποίηση των δεδομένων εξαρτάται από το γνωστικό υπόβαθρο και το μοντέλο επίθεσης που θα χρησιμοποιήσουν οι αντίπαλοι. Αρχικά να σημειώσουμε πως οι τύποι αποκάλυψης των προσωπικών πληροφοριών στα κοινωνικά δίκτυα είναι οι εξής [21]:

i. *Αποκάλυψη ταυτότητας*: Πρόκειται για τη σύνδεση ενός κόμβου από την ανωνυμοποιημένη βάση με τον άνθρωπο ή την οντότητα που ο κόμβος εκπροσωπεί.

ii. *Αποκάλυψη γνωρίσματος*: Αφορά την περίπτωση κατά την οποία ο αντίπαλος ανακαλύπτει καινούριες πληροφορίες του θύματος χωρίς όμως την ακριβή ταυτοποίηση του κόμβου στην βάση δεδομένων.

iii. *Αποκάλυψη ύπαρξης*: Αφορά την περίπτωση, ο επιτιθέμενος να αποκαλύψει την ύπαρξη μιας ευαίσθητης σχέσης μεταξύ δύο ανθρώπων που ανήκουν στο ίδιο κοινωνικό δίκτυο. Αυτός ο τύπος αποκάλυψης υποθέτει πως οι σχέσεις των κόμβων είναι ευαίσθητες και πως πρέπει να προστατευτούν.

Οι όροι *παθητικές* και *ενεργητικές* επιθέσεις χρησιμοποιούνται ώστε να ξεχωρίσουν τις περιπτώσεις που από την μία ο επιτιθέμενος απλώς παρατηρεί τα δεδομένα και δεν τα παραποιεί (παθητική επίθεση) ενώ από την άλλη ο επιτιθέμενος μπορεί να τα παραποιήσει για να πετύχει τον στόχο του (ενεργητική επίθεση).

## *Παθητικές επιθέσεις*

Στα κοινωνικά δίκτυα, μια εγγραφή μπορεί να αναγνωριστεί βάση των τιμών των χαρακτηριστικών γνωρισμάτων της και της σχέσης της με άλλες εγγραφές. Πριν τις παρουσιάσουμε πρέπει να εξετάσουμε δύο ορισμούς [22]:

Ορισμός 4.1 Επαγόμενος υπογράφος: Δεδομένου ενός γράφου G=(V, E), με ένα σύνολο κόμβων V και ένα σύνολο ακμών E, έστω S ένα σύνολο κόμβων, υποσύνολο του V, ο επαγόμενος υπογράφος του G από το S είναι ο G(S) = (S, E΄), όπου E΄= {(u, υ)|(u, υ) ∈ E, u ∈ V, υ ∈ V}

Ορισμός 4.2 1-γειτονιά**:** Δεδομένου ενός γράφου G=(V, E), εάν (u, υ) ∈ E τότε u και υ θα είναι γείτονες μεταξύ τους. Ας υποθέσουμε πως u ∈ V και W ένα σύνολο γειτόνων του u. Ο 1-γείτονας του u ∈ V είναι επαγόμενος υπογράφος του G στο {u} ∪ W.

## *$H_i$: Γνώσεις γειτονίας*

Ο καθηγητής Hay προσδιορίζει μια σειρά από συναρτήσεις, οι οποίες επιστρέφουν την τοπική δομή του γράφου που αφορούν το στοχευμένο κόμβο. Συγκεκριμένα:

- $H_o(x)$ επιστρέφει την ετικέτα του κόμβου x ( στην περίπτωση που δεν περιλαμβάνει ετικέτα, επιστρέφει το κενό),
- $H_1(x)$ επιστρέφει τον βαθμό του κόμβου x και
- $H_2(x)$ επιστρέφει συνολικά για όλους του γείτονες του x, τους βαθμούς τους.

Για παράδειγμα ο κόμβος x1 του σχήματος 4.2 επιστρέφει τα εξής αποτελέσματα:

$H_o(x) = A$

$H_1(x) = 2$

$H_2(x) = \{2, 4\}$

Οι συναρτήσεις μπορούν επίσης να οριστούν αναδρομικά, όπου $H_i(x)$ μπορεί να επιστρέφει το σύνολο των τιμών από την εξέταση του $H_{i-1}(x)$ πάνω στο σύνολο των κόμβων που συνορεύουν με τον x. Για παράδειγμα: $H_i(x) = \{ H_o(z_1), H_{i-1}(z_2),\ldots, H_{i-1}(z_m) \}$, όπου $z_1,\ldots,z_m$ είναι οι κόμβοι που συνορεύουν με τον x.

## *Γνώσεις υπογράφου*

Ο καθηγητής Hay [23] επίσης αποδεικνύει ότι πιο δυνατές και περισσότερο ρεαλιστικές συναρτήσεις από τις $H_i(x)$ είναι οι συναρτήσεις υπογράφου, οι οποίες επιστρέφουν τον υπογράφο γύρω από το στοχευμένο κόμβο. Η περιγραφική δύναμη της συνάρτησης μετριέται από τον αριθμό των ακμών του υπογράφου. Πρέπει να σημειωθεί πως αυτός ο τύπος της γνώσης που αποκομίζει ο αντίπαλος καλύπτει και την 1-γειτονία αφού η 1-γειτονία είναι μόνο ένας πιθανός υπογράφος περιφερειακά του κόμβου.

Στο σχήμα 4.2, εάν ο αντίπαλος γνωρίζει πως το θύμα βρίσκεται σε κέντρο-αστέρα με 4 γείτονες, τότε με τη μέθοδο της εις άτοπο απαγωγής μπορεί να τον εντοπίσει στον κόμβο x2, αφού είναι ο μοναδικός που έχει 4 γείτονες. Ωστόσο, εάν ο αντίπαλος γνωρίζει μονάχα πως το θύμα έχει την ετικέτα A και συνορεύει και με κόμβο που έχει επίσης ετικέτα A, τότε δεν μπορεί να είναι σίγουρος στην επίθεσή του, αφού 3 κόμβοι (x1, x2, x3) πληρούν τις προϋποθέσεις.

## *Ενεργές επιθέσεις*

Στα κοινωνικά δίκτυα, στις ενεργές επιθέσεις, οι επιτιθέμενοι έχουν την δύναμη να επεξεργαστούν την βάση δεδομένων του δικτύου πριν την ανωνυμοποιημένη έκδοσή του. Με το ζήτημα αυτό ασχολήθηκε η ομάδα του Backstrom [24]. Οι επιτιθέμενοι εδώ, αρχικά επιλέγουν αυθαίρετα ένα σύνολο από τους υποψήφιους στόχους ενός δικτύου. Στην συνέχεια κατασκευάζουν έναν μικρό αριθμό από ψεύτικους χρήστες, τους συνδέουν με τους στόχους και φροντίζουν να κατασκευάσουν ένα μοναδικό μοτίβο, που να μπορούν να το εντοπίσουν στην ανωνυμοποιημένη βάση. Η επίθεση για να πετύχει χρειάζεται να περιλαμβάνει την δημιουργία O(logN) ψεύτικων χρηστών, όπου N είναι

ο συνολικός αριθμός χρηστών.

Εναλλακτικά, ο επιτιθέμενος μπορεί να βρει ή να κατασκευάσει k κόμβους με το όνομα $\{ x_1,\ldots,x_k \}$, τις αντίστοιχες ακμές $\{ x_i, x_j \}$ και τελικά να αποκτήσει έναν τυχαίο γράφο H. Ο γράφος αυτός θα ανήκει στο δίκτυο G και θα πρέπει να πληρεί τις εξής προϋποθέσεις [25]:

- δεν πρέπει να υπάρχει κανένας άλλος υπογράφος S στο G που να είναι ισομορφικός στον H, δηλαδή, ο H να είναι σαν το S αλλά με διαφορετικές ετικέτες. Με αυτόν το τρόπο ο H θα μπορεί να αναγνωριστεί μοναδικά
- επιπλέον δεν θα πρέπει να υπάρχει κανένας αυτομορφισμός στον H.

Ο επιτιθέμενος, τελικά, θα μπορεί να στοχεύει έναν κόμβο w στο H με ένα υποσύνολο κόμβων N, ώστε όταν εντοπιστεί ο H να εντοπίζεται αυτόματα και ο κόμβος w.

Οι ενεργές επιθέσεις, ευτυχώς, δεν είναι πρακτικές για μεγάλου εύρους επιθέσεις ιδιωτικότητας. Τους τρεις περιορισμούς αναλύουν οι Narayanan και Shmatikov.

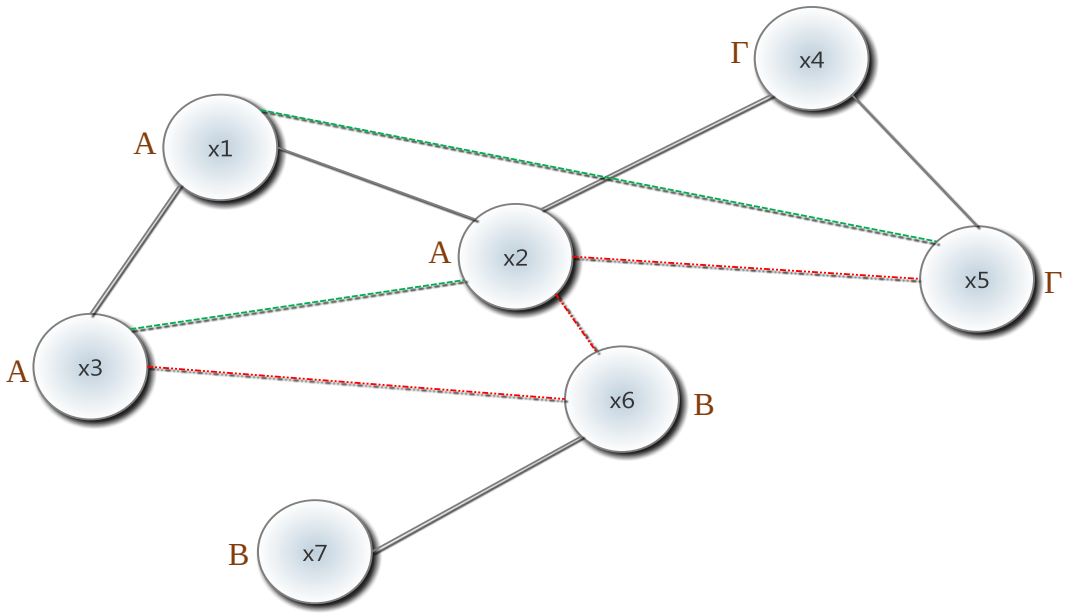
1. Οι ενεργές επιθέσεις περιορίζονται μόνο στα κοινωνικά δίκτυα και ο επιτιθέμενος πρέπει να δημιουργήσει πολλούς ψεύτικους χρήστες πριν την δημοσίευση της βάσης. Αυτό δεν είναι εύκολο καθώς οι περισσότεροι πάροχοι κοινωνικής δικτύωσης, όπως το Facebook, φροντίζουν το κάθε προφίλ να μην ενεργοποιείται εάν δεν περιλαμβάνει μοναδικό e-mail γεγονός που καθιστά την διαδικασία χρονοβόρα και αποτρεπτική.
2. Ο επιτιθέμενος μπορεί να δημιουργεί ψεύτικους χρήστες και να τους συνδέει με τους στόχους του αλλά δεν μπορεί να φαίνεται ενεργός στο μεγαλύτερο μέρος από αυτούς.
3. Ένας μεγάλος αριθμός από παρόχους κοινωνικής δικτύωσης δεν επιτρέπει την σύνδεση δύο χρηστών εάν δεν υπάρχει κοινή αποδοχή για την ενέργεια αυτή. Χάρη στον περιορισμό αυτόν, εάν το θύμα δε δεχτεί τη σύνδεση προστατεύεται από αυτές τις κακόβουλες επιθέσεις.

### *4.6.3 Μοντέλα ανωνυμοποίησης*

Παρακάτω αναλύονται τα μοντέλα ανωνυμοποίησης που προτείνονται στην βιβλιογραφία.

## *Προσθήκη, αφαίρεση ακμών και γενίκευση ετικέτας*

Οι καθηγητές Zhou και Pei [20] θεωρούν την 1-γειτονία ενός κόμβου ως τις γνώσεις του αντιπάλου. Ο στόχος του μοντέλου είναι η προστασία της αποκάλυψης της εγγραφής. Για την τροποποίηση του γράφου χρησιμοποιείται αλγόριθμος που προσθέτει, αφαιρεί ακμές και γενικεύει ετικέτες. Ένα παράδειγμα αυτού του μοντέλου παρουσιάζεται στο σχήμα 4.3, το οποίο είναι αποτέλεσμα τροποποίησης του σχήματος 4.2:




Σχήμα 4.3 Ανωνυμοποίηση με τροποποίηση ακμών

Ο στόχος εδώ είναι να αποφευχθούν οι επιθέσεις ανθρώπων που έχουν γνώσεις 1-γειτονίας για τον κόμβο που στοχεύετε και ο γράφος να γίνει 2-ανώνυμος. Θέλουμε, δηλαδή, για κάθε σύνολο 1-γειτονίας N, να υπάρχουν τουλάχιστον 2 ίδιοι κόμβοι που το περιλαμβάνουν. Στο παράδειγμά μας προστίθενται οι ακμές ($x_2$, $x_3$) και ($x_1$, $x_5$) και διαγράφονται οι ($x_2$, $x_5$), ($x_2$, $x_6$), ($x_3$, $x_6$). Τώρα, εάν ο στόχος των επιτιθέμενων είναι ο κόμβος $x_1$, παρατηρούμε πως ο $x_1$ έχει την ετικέτα Α, την οποία έχουν και οι δύο γείτονες του. Οι κόμβοι $x_6$ , $x_7$ έχουν και αυτοί την ίδια 1-γειτονία. Ο γράφος είναι 2-ανώνυμος και προστατεύει από επιθέσεις 1-γειτονίας.

Τα δύο βασικά βήματα για την ανωνυμοποίηση του γράφου:

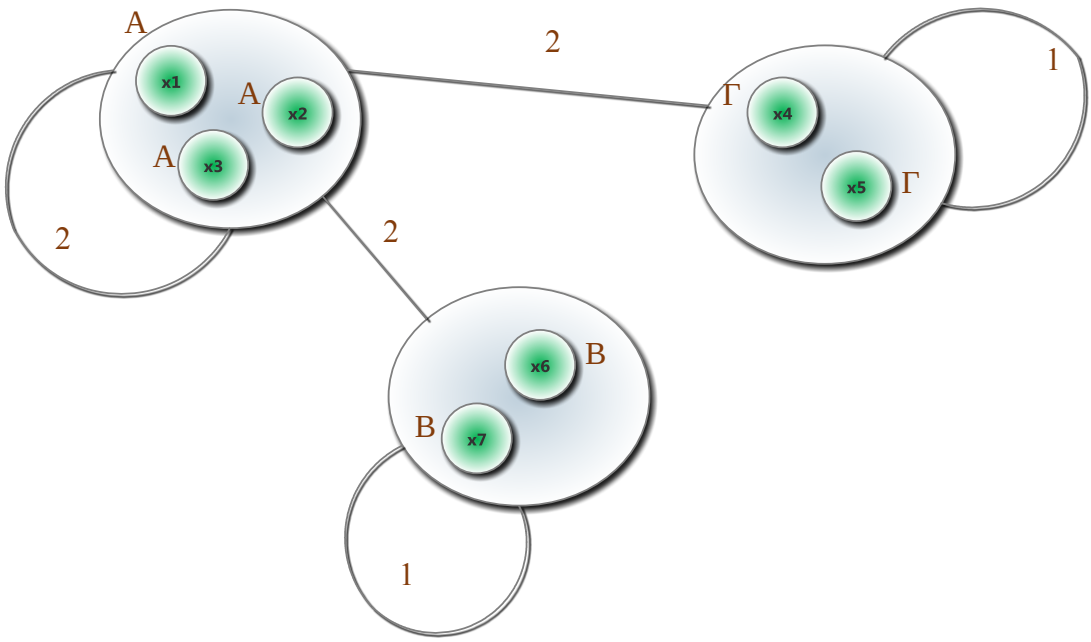
1. Βρίσκει την 1-γειτονία του κάθε κόμβου στο δίκτυο και
2. συγκεντρώνει τους κόμβους βάσει ομοιότητας και τροποποιεί τις γειτονίες αυτές με στόχο να περιλαμβάνουν τουλάχιστον k κόμβους με την ίδια 1-γειτονία.

## *Συσταδοποίηση κόμβων για k-ανωνυμία*

Ο καθηγητής Hay [23] προτείνει την ομαδοποίηση κόμβων σε μέγεθος τουλάχιστον k. Οι συστάδες διαμορφώνουν ισοδύναμες κλάσεις κόμβων, ώστε από το σύνολο να μην ξεχωρίζει κανένας κόμβος. Το μοντέλο προφυλάσσει το απόρρητο των εγγραφών από διάφορες επιθέσεις βασισμένες σε διαφορετικά είδη γνώσεων υποβάθρου του αντιπάλου, συμπεριλαμβανομένων των $H_i$, υπογράφου γειτονίας και 1-γειτονίας.

Κάθε συστάδα κόμβων καλείται υπερκόμβος και η ακμή που ενώνει δυο υπερκόμβους λέγεται υπερακμή. Οι υπερακμές μπορούν να επιστρέφουν ξανά στον εαυτό τους και σε κάθε μία από αυτές προστίθεται και ένας ακέραιος θετικός αριθμός που μαρτυρά το πλήθος των ακμών εντός του υπερκόμβου.

Ένα τέτοιο παράδειγμα είναι το σχήμα 4.3, το οποίο είναι μια επεξεργασμένη εκδοχή του αρχικού σχήματος 4.2 και απεικονίζει 3 συστάδες:




Σχήμα 4.3: Ανωνυμοποίηση με συσταδοποίηση κόμβων

Στον παραπάνω γράφο G=(V, E), του σχήματος 4.3 εντοπίζουμε τρεις υπερκόμβους, τους S1, S2, S3 που ορίζονται ως V = { S1, S2, S3 }. Η υπερακμή μεταξύ των S1 και S2 περιλαμβάνει τον αριθμό 2, αφού και από τον αρχικό γράφο παρατηρούμε πως υπάρχουν 2 ακμές μεταξύ των κόμβων. Ο αριθμός ορίζεται ως d( S1, S2) = 2. Ομοίως για τους υπερκόμβους S1 και S3 έχουμε d(S1, S3) = 2. Η υπερακμή του S1 που επιστρέφει ξανά στον S1 παίρνει τιμή 2, ενώ αντίστοιχα οι υπερακμές των S2, S3 την τιμή 1. Ο γράφος του σχήματος 4.3 είναι 2-ανώνυμος, αφού κάθε υπερκόμβος περιλαμβάνει δύο κόμβους του αρχικού γράφου και είναι έτσι ομαδοποιημένοι ώστε να μην ξεχωρίζει κανένας από αυτούς. Προκειμένου να διατηρηθεί στο μέγιστο η χρησιμότητα των πληροφοριών παρουσιάστηκε η παρακάτω ιδέα: Εάν δημοσιευτεί ο γράφος G=(V, E) του σχήματος 4.3 αντί του σχήματος 4.2 τότε θα υπάρχουν πολλαπλές πιθανότητες για το πως θα είναι ο αρχικός γράφος. Κάθε μία από αυτές τις πιθανότητες είναι και ένας πιθανός γράφος που παίρνει το όνομα κόσμος. Έστω W(G) να είναι το σύνολο όλων των πιθανών κόσμων. Εάν υποθέσουμε πως όλοι οι κόσμοι έχουν την ίδια πιθανότητα να εμφανιστούν, τότε η πιθανότητα ο αρχικός γράφος να είναι ένας από τους πιθανούς είναι 1/|W(G)|. Όσο μεγαλύτερη είναι η πιθανότητα, τόση μεγαλύτερη ποιότητα πληροφοριών επιτυγχάνουμε. Η διατύπωση της εξίσωσης W(G) είναι:

$$|\mathrm{W(G)}| = \prod_{X \in V} \binom{\frac{1}{2}|X|(|X|-1)}{d(X,X)} \prod_{\mathrm{X,Y \in V}} \binom{|X|,|Y|}{d(X,Y)}$$

Έτσι ο γράφος του σχήματος 4.3 δίνει το αποτέλεσμα:

$$\mathrm{W(G)} = \binom{\frac{1}{2}3*2}{2}\binom{\frac{1}{2}2*1}{1}\binom{\frac{1}{2}2*1}{1}\binom{3*2}{2}\binom{3*2}{2} = 675$$

Ο αλγόριθμος υλοποίησης των συστάδων κάνει χρήση ευριστικής μεθόδου για να επιστρέψει μία καλή λύση. Αρχικά, παίρνει τον γράφο και εντάσσει όλους τους κόμβους σε έναν και μοναδικό υπερκόμβο. Στην συνέχεια, ο αλγόριθμος αναζητά επαναλαμβανόμενα εναλλακτικούς τρόπους επίτευξης του ανωνυμοποιημένου γράφου με τις εξής τεχνικές:

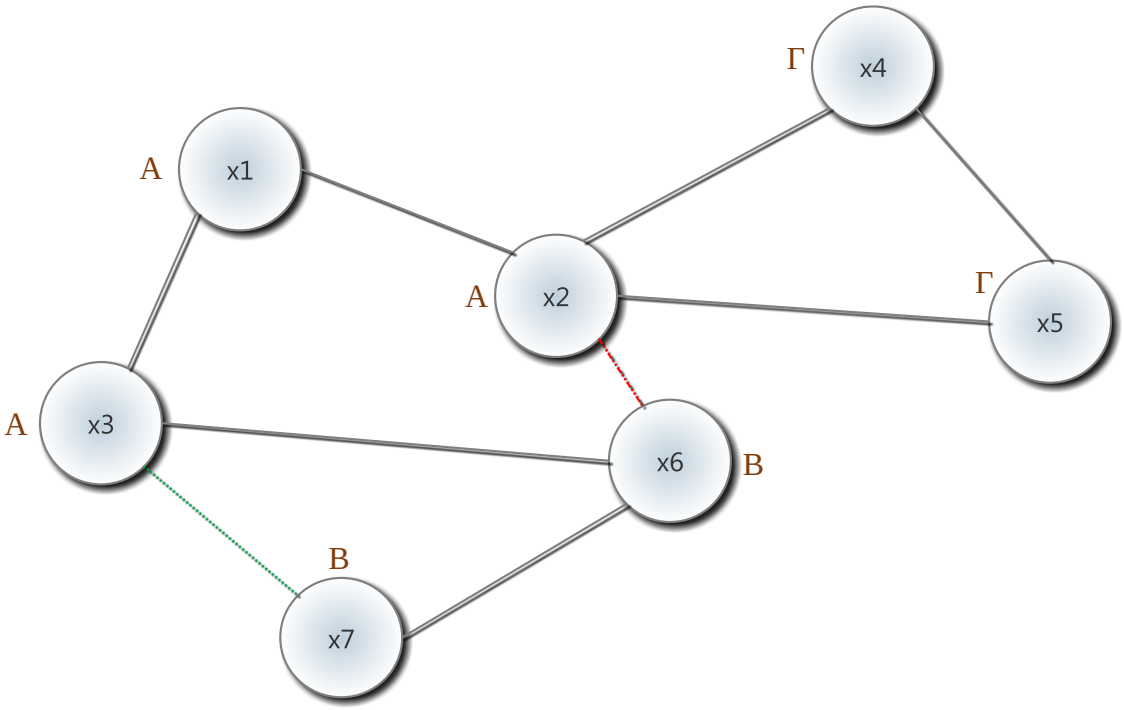
- την διχοτόμηση ενός υπερκόμβου,
- την συγχώνευση δυο υπερκόμβων και
- τη μετακίνηση ενός κόμβου από τον έναν υπερκόμβο στον άλλο.

Για κάθε μια εφαρμογή από τις παραπάνω τεχνικές, υπολογίζεται η πιθανότητα 1/ |W(G)| και εάν αυξάνεται, τότε η τελευταία εναλλακτική γίνεται η αποδεκτή. Ο αλγόριθμος ολοκληρώνει την διαδικασία όταν γίνονται αποδεκτές λιγότερες από το 10% των εναλλακτικών.

## *Παραγωγή υπεργράφου*

Στο μοντέλο ανωνυμοποίησης k-ανώνυμος γράφος που προτείνουν οι Liu και Terzi [26], υποθέτουν πως οι γνώσεις του αντιπάλου είναι το αποτέλεσμα της συνάρτησης $H_1$ δηλαδή ο βαθμός του κόμβου που στοχεύεται. Για την παραγωγή του ανωνυμοποιημένου γράφου ο αλγόριθμος που προτείνουν προσθέτει και αφαιρεί ακμές.

Αρχικά, δεδομένου ενός γράφου G=(V, E), γίνεται καταγραφή και ταξινόμηση των βαθμών του κάθε κόμβου σε μία λίστα d. Για να γίνει ο γράφος k-ανώνυμος θα πρέπει κάθε βαθμός κόμβου να παρουσιάζεται τουλάχιστον k φορές στην λίστα d. Για παράδειγμα, μια τυχαία λίστα d = { 5, 5, 3, 3, 2, 2, 2 } εκπροσωπεί έναν γράφο που είναι 2-ανώνυμος, ενώ η λίστα d = { 4, 3, 2, 2, 2, 1 } που εκπροσωπεί τον γράφο της εικόνας 1 δεν είναι k-ανώνυμος επειδή οι τιμές 4, 3, 1 παρουσιάζονται μονάχα μια φορά. Στο σχήμα 4.2 εάν διαγράψουμε την ακμή (x2, x6) και προσθέσουμε την ακμή (x3, x7) προκύπτει ο γράφος:




Σχήμα 4.4: Ανωνυμοποίηση σχήματος 4.2 με παραγωγή υπεργράφου

Τώρα η λίστα d του επεξεργασμένου γράφου έγινε d = { 3, 3, 2, 2, 2, 2, 2 } και ο γράφος 2-ανώνυμος. Επομένως εάν ο επιτιθέμενος γνωρίζει μονάχα τον βαθμό του κόμβου που στοχεύει, τότε θα υπάρχουν τουλάχιστον δύο κόμβοι με τον ίδιο βαθμό και δεν θα ξεχωρίζουν.

Για την μέτρηση της ποιότητας των δημοσιευμένων δεδομένων εφαρμόζεται η παρακάτω εξίσωση:

$$L_1(d' - d) = \sum_i |d'(i) - d(i)|$$

όπου d΄ είναι μία λίστα βαθμών που προέκυψε από την αρχική d και η d(i) αναφέρεται στην i-στη τιμή της λίστας d.

Ακολουθούν τα βασικά βήματα του αλγορίθμου στην ανωνυμοποίηση:

- Αφού του δοθεί η λίστα d, κατασκευάζει μια νέα λίστα βαθμών d΄, η οποία, πρώτον, είναι k-ανώνυμη και δεύτερον, παίρνει την ελάχιστη τιμή στην εξίσωση $D_A(d', d) = L_1(d' - d)$.
- Με την λίστα d΄ κατασκευάζει έναν γράφο G΄(V, E΄), τέτοιον ώστε $d'_{G'} = d'$ και E΄∩ E = E.

Για το πρώτο βήμα προτείνεται ένας δυναμικός αλγόριθμος που δίνει λύση στο πρόβλημα σε πολυωνυμικό χρόνο. Για το δεύτερο βήμα χρησιμοποιείτε ο αλγόριθμος ConstructionGraph από την [27] που κατασκευάζει τον γράφο. Ως είσοδο ο ConstructionGraph παίρνει την λίστα d΄από το βήμα 1 και η έξοδος G1 που θα δώσει γίνεται αποδεκτή μόνο εάν είναι υπεργράφος της G. Υπάρχει περίπτωση η G να περιλαμβάνει ακμές που δεν υπάρχουν στην G1 και τότε η λίστα βαθμών d΄, διατηρώντας πάντα την ιδιότητα της k-ανωνυμίας, επεξεργάζεται αυξάνοντας κάποιες τιμές της, και επιστρέφει εκ νέου στον αλγόριθμο του βήματος 2. Εάν πάλι δεν δοθεί αποδεκτό αποτέλεσμα, προτείνεται διαγραφή των ακμών που παραβιάζουν τις προϋποθέσεις για υπεργράφο.

### *4.6.4 Συμπεράσματα*

Η προστασία του απορρήτου στα κοινωνικά δίκτυα και η ανωνυμοποίηση των γράφων για την δημοσίευση τους είναι ακόμα σε πρώιμο στάδιο και υπάρχουν ακόμη πολλά περιθώρια βελτίωσης και αλλαγών.

Αρχικά με τις μέχρι τώρα μεθόδους ανωνυμοποίησης δεν μπορεί να βρεθεί μια βέλτιστη λύση, υπάρχει όμως η δυνατότητα να βρεθούν βελτιωμένα μοντέλα που να προκαλούν μικρότερη παραμόρφωση δεδομένων και απώλεια πληροφοριών. Έπειτα, δεδομένου ότι τα κοινωνικά δίκτυα είναι δυναμικά και έχουμε συνεχώς την ανανέωση της βάσης τους θα ήταν καλή ιδέα να προταθεί μοντέλο ανωνυμοποίησης για βάσεις δεδομένων πολλαπλών δημοσιεύσεων, όπως είναι το μοντέλο MultiRelational k-anonymity για τις σχεσιακές βάσεις δεδομένων. Επιπλέον, δεν έχει προταθεί ακόμα μοντέλο που να προφυλάσσει το απόρρητο όταν στο δίκτυο ο κάθε κόμβος περιλαμβάνει πολλαπλά ευαίσθητα γνωρίσματα. Τέλος η k-ανωνυμία δεν αρκεί πολλές φορές. Αν, για παράδειγμα, όλοι οι κόμβοι μίας κλάσης ισοδυναμίας περιλαμβάνουν το ίδιο ευαίσθητο γνώριμα, τότε ο επιτιθέμενος αρκεί να εντοπίσει την κλάση στην οποία βρίσκεται ο κόμβος που αναζητά.

# *Κεφάλαιο 5. Απόρρητο στις σχεσιακές βάσεις δεδομένων*

Οι Cox και Dalenius [28] ορίζουν την *ανωνυμοποίηση δεδομένων* ως την προστασία της ταυτότητας και των ευαίσθητων πληροφοριών των ιδιοκτητών τους.

Ακόμα και με την διαγραφή αναγνωριστικών ταυτότητας μπορεί κανείς να προσδιορίσει μονοσήμαντα ένα άτομο. Η καθηγήτρια του πανεπιστημίου του Harvard, Latanya Sweeny μας παραθέτει σε άρθρο της [9] ένα χαρακτηριστικό παράδειγμα το οποίο καταδεικνύει πόσο εύκολο είναι για έναν επιτιθέμενο, ο οποίος κατέχει ένα μέρος από γνώσεις για μια οντότητα του πραγματικού κόσμου να αποτελέσει απειλή για την εγγραφή του στα ανωνυμοποιημένα δεδομένα.

Συγκεκριμένα ένας απλός συνδυασμός δύο σχεσιακών βάσεων δεδομένων που η πρώτη αφορούσε ανωνυμοποιημένα ιατρικά δεδομένα από οργανισμό ασφάλισης υγείας και η δεύτερη εκλογικά στοιχεία των πολιτών από τους δημόσιους καταλόγους ψηφοφορίας, αρκούσε για να αποκαλυφθεί ο ιατρικός φάκελος του κυβερνήτη της Μασαχουσέτης. Ο κυβερνήτης ζούσε στο Cambridge της Μασαχουσέτης και τα στοιχεία του, ήταν καταχωρημένα στον οργανισμό ασφάλισης υγείας. Σύμφωνα με τους καταλόγους ψηφοφορίας 6 άτομα ήταν γεννημένα την ίδια ημερομηνία με τον κυβερνήτη, μόνο τρία από αυτά ήταν άντρες και μόνο ο κυβερνήτης είχε το συγκεκριμένο ταχυδρομικό κώδικα.

Η διαδικασία που περιγράφεται ονομάζεται *"αποκάλυψη"* ή αλλιώς *"σύνδεση"*. Κάθε ένα από τα ψευδο-αναγνωριστικά δεν μπορεί από μόνο του να αποκαλύψει μία ταυτότητα. Με τον συνδυασμό τους όμως υπάρχει μεγάλη πιθανότητα να αποκαλυφθεί. Συγκεκριμένα στο παράδειγμά μας έγινε συνδυασμός: QID = { Ημερομηνία γέννησης, Φύλο, Ταχυδρομικός κώδικας } όπως φαίνεται και στο σχήμα 5.1:


Σχήμα 5.1: Μέθοδος σύνδεσης δύο δημοσιευμένων βάσεων

Η καθηγήτρια πραγματοποίησε επίσης μια μελέτη πάνω σε δημοσιευμένα δημογραφικά στοιχειά και πώς αυτά μπορούν μοναδικά να προσδιορίσουν την ταυτότητα ενός πολίτη με βάση τα ψευδο-αναγνωριστικά. Τα συμπεράσματα της είναι εντυπωσιακά:

- Το 87.1% δηλαδή τα 216 από τα 248 εκατομμύρια του πληθυσμού των Ηνωμένων Πολιτειών της Αμερικής, μπορούν να προσδιοριστούν μονοσήμαντα με τη χρήση μόνο τριών χαρακτηριστικών γνωρισμάτων: του 5-ψήφιου ταχυδρομικού κώδικα, του φύλου και της ημερομηνίας γεννήσεως.
- Παραπάνω από τον μισό πληθυσμό των ΗΠΑ, συγκεκριμένα το 58.4% μπορεί να εξακριβωθεί ακόμα και με την γνώση γενικότερων γνωρισμάτων όπως είναι η πόλη, το φύλο και η ημερομηνία γεννήσεως του.
- Το 18.1% του πληθυσμού των ΗΠΑ μπορούν να προσδιοριστούν μοναδικά και με την χρήση ακόμα γενικότερων γνωρισμάτων όπως την χώρα , το φύλο και την

ημερομηνία γεννήσεως.

Από τα παραπάνω βλέπουμε πως ένας ιδιοκτήτης εγγραφής επαναπροσδιορίζεται με την αποκάλυψη των ψευδο-αναγνωριστικών του. Για να εφαρμόσει ένας επιτιθέμενος την μέθοδο της *αποκάλυψης* χρειάζεται δύο προηγούμενες γνώσεις, πρώτον να ξέρει ότι η εγγραφή του θύματος υπάρχει στην δημοσιευμένη βάση δεδομένων και δεύτερον να γνωρίζει τα ψευδο-αναγνωριστικά του. Αν για παράδειγμα ο επιτιθέμενος είναι εργαζόμενος σε μια εταιρεία και μάθει πως το αφεντικό του νοσηλεύτηκε θα συμπεράνει ασφαλώς πως η ιατρική του εγγραφή θα υπάρχει στην δημοσιευμένη βάση με τα δεδομένα των ασθενών του νοσοκομείου. Επίσης δεν θα του είναι δύσκολο να αποκτήσει τον ταχυδρομικό κώδικα, την ημερομηνία γεννήσεως και φυσικά το φύλο του αφεντικού του. Αυτά τα χαρακτηριστικά γνωρίσματα μπορούν να λειτουργήσουν ως ψευδο-αναγνωριστικά σε επιθέσεις "αποκάλυψης".

Για να αποτραπούν τέτοιου είδους επιθέσεις ο εκδότης δεδομένων παρέχει έναν ανωνυμοποιημένο πίνακα:

*T( QID', ευαίσθητα γνωρίσματα, μη-ευαίσθητα γνωρίσματα )*

όπου QID' είναι η ανωνυμοποιημένη έκδοση του συνόλου των ψευδο-αναγνωριστικών που προέκυψε με την επεξεργασία των χαρακτηριστικών γνωρισμάτων του αρχικού πίνακα και με μεθόδους που θα μελετήσουμε παρακάτω. Οι τεχνικές ανωνυμοποίησης κρύβουν ορισμένες λεπτομερείς πληροφορίες και διαμορφώνουν κλάσεις ισοδυναμίας με βάση το QID' . Κατά συνέπεια αν η εγγραφή ενός ανθρώπου συνδεθεί μέσω του QID' , επειδή ο ίδιος καθώς και άλλες εγγραφές θα έχουν τις ίδιες τιμές σε αυτό θα προστατεύονται τα ευαίσθητα χαρακτηριστικά τους. Εναλλακτικά οι τεχνικές ανωνυμοποίησης μπορούν να παράξουν έναν σύνθετο πίνακα T βασισμένο σε *στατιστικές* ιδιότητες του αρχικού πίνακα T ή να του προσθέσουν *θόρυβο*. Το μεγάλο πρόβλημα στην διαδικασία της ανωνυμοποίησης ενός πίνακα T είναι πως όσο μεγαλύτερο επίπεδο ιδιωτικότητας επιτυγχάνεται τόσο δυσκολότερο είναι να διατηρηθεί η χρησιμότητα των δεδομένων του. Χρησιμοποιείται σαν παράμετρος σύγκρισης των διάφορων μεθόδων ανωνυμοποίησης η απώλεια πληροφορίας που υπάρχει στα δημοσιευμένα δεδομένα έναντι των αρχικών ώστε να μπορούμε να επιλέξουμε την μέθοδο εκείνη που ταιριάζει καλύτερα.

## *5.1 Τεχνικές ανωνυμοποίησης*

Ο αρχικά ανεπεξέργαστος πίνακας δεδομένων δεν προστατεύει το απόρρητο των εγγραφών και πρέπει να υποστεί επεξεργασία πριν την δημοσίευσή του. Οι τροποποιήσεις του πίνακα επιτυγχάνονται με μια σειρά από τεχνικές ανωνυμοποίησης. Οι συνηθέστερες από αυτές είναι η **γενίκευση**, η **κατάπνιξη**, η **ανατομία**, η **μετάθεση** και η **διαταραχή**. Η γενίκευση και η κατάπνιξη αντικαθιστά πληροφορίες, συνήθως από τα ψευδο-αναγνωριστικά, με άλλες λιγότερο περιγραφικές. Η ανατομία και η μετάθεση καταργεί τις σχέσεις μεταξύ ψευδο-αναγνωριστικών και ευαίσθητων στοιχείων με ομαδοποιήσεις και μετακυλήσεις. Η διαταραχή παραποιεί τα δεδομένα, προσθέτοντας θόρυβο, συγκεντρώνοντας ή ανταλλάσσοντας τις τιμές με αποτέλεσμα να παράγει συνθετικά δεδομένα βασιζόμενη στις στατιστικές ιδιότητες του πίνακα.

### *5.1.1 Γενίκευση και κατάπνιξη*

Σε μια βάση δεδομένων το κάθε χαρακτηριστικό γνώρισμα αποδίδει διαφορετική κάθε φορά σημασιολογική ιδιότητα και τα πεδία της περιγράφουν το πεδίο ορισμού. Με την μέθοδο της **γενίκευσης** αντικαθιστούμε την αρχική τιμή ενός πεδίου με μια άλλη γενικότερη διατηρώντας το είδος της πληροφορίας. Στην συνέχεια μπορούμε να γενικεύσουμε ξανά το γενικευμένο πεδίο και ούτω καθεξής. Το σύνολο όλων των επιπέδων γενίκευσης ονομάζεται **ιεραρχία γενίκευσης**. Κάθε φορά που ακολουθούμε την διαδικασία της γενίκευσης οι πολλαπλές τιμές που οδηγούν σε μια γενικευμένη τιμή μειώνει τον αριθμό των ξεχωριστών πλειάδων και κατ' επέκταση διαμορφώνει ολοένα και μεγαλύτερες σε μέγεθος κλάσεις ισοδυναμίας. Θα πρέπει να κάνουμε τις εξής διακρίσεις:

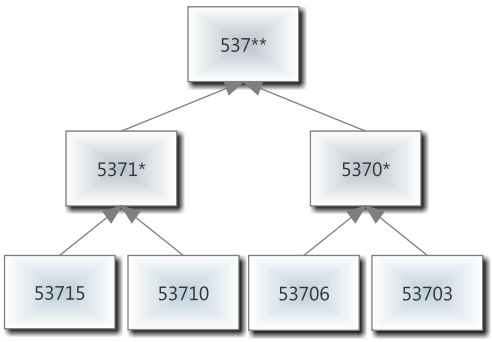
Με κριτήριο τον τρόπο απεικόνισης της ιεραρχίας γενίκευσης διακρίνουμε [29]:

- την **ιεραρχία γενικευμένου πεδίου** που συμβολίζεται ως DGH (domain generalization hierarchy) και στην οποία το κάθε επίπεδο γενίκευσης των πεδίων εμφανίζεται ως ένα σύνολο τιμών που δύναται να μεταβεί στο επόμενο και
- την **ιεραρχία γενικευμένης τιμής** που συμβολίζεται ως VGH (value generalization hierarchy) και στην οποία κάθε τιμή των πεδίων έχει προκαθορισμένη και μοναδική γενικευμένη τιμή στο επόμενο επίπεδο. Απεικονίζεται συνήθως ως ένα **ταξινομημένο δέντρο** με την γενίκευση να αυξάνεται από τα φύλλα έως την ρίζα.

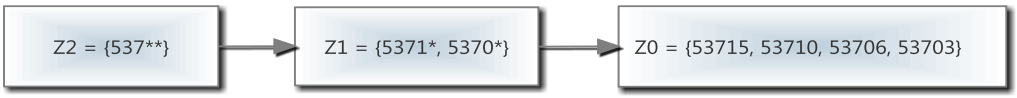
Με κριτήριο το είδος του πεδίου τιμών διακρίνουμε τις:

- αριθμητικές τιμές όπου κατά την γενίκευση τους θα αντιστοιχίζονται σε ένα διάστημα τιμών ή
- κατηγορικές τιμές, όπου βάση της σημασιολογίας του χαρακτηριστικού γνωρίσματος γενικεύονται σε κάτι ευρύτερο.

Συνοπτικά οι παραπάνω έννοιες απεικονίζονται στα σχήματα παρακάτω:

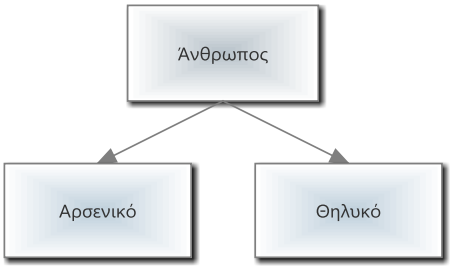



Σχήμα 5.2: DGH Ταχυδρομικού κώδικα




Σχήμα 5.3: VGH Ταχυδρομικού κώδικα


Σχήμα 5.4: Παράδειγμα DGH



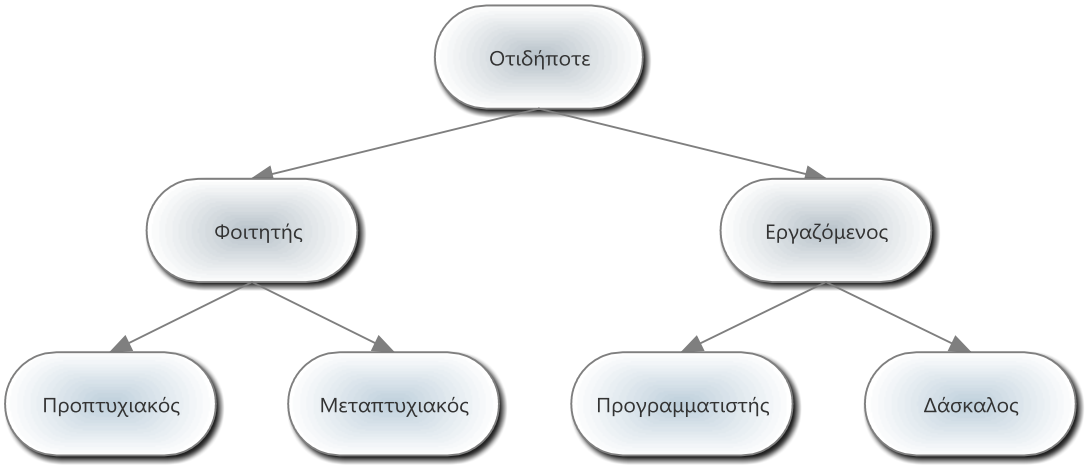

Σχήμα 5.5: Παράδειγμα VGH

Τα σχήματα 5.2, 5.3 αφορούν αριθμητικά παραδείγματα ενός χαρακτηριστικού γνωρίσματος={ΤΑΧΥΔΡΟΜΙΚΟΣ ΚΩΔΙΚΑΣ} με το πρώτο να απεικονίζεται σε ιεραρχία γενικευμένου πεδίου ενώ το δεύτερο σε ιεραρχία γενικευμένης τιμής.

Τα σχήματα 5.4, 5.5 αφορούν κατηγορικά παραδείγματα ενός χαρακτηριστικού γνωρίσματος ={ ΦΥΛΟ } με το πρώτο να απεικονίζεται σε ιεραρχία γενικευμένου πεδίου ενώ το δεύτερο σε ιεραρχία γενικευμένης τιμής.

Τα σύνολα Z0 και S0 της ιεραρχίας γενικευμένου πεδίου αποτελούν τα αρχικά γνωρίσματα μιας σχεσιακής βάσης δεδομένων.

Στην συνέχεια της εργασίας θα παρουσιάζουμε τα επίπεδα γενίκευσης σε ταξινομημένα δέντρα καθώς αυτά έχουν επικρατήσει λόγω της απλότητας και της ευχρηστίας τους.

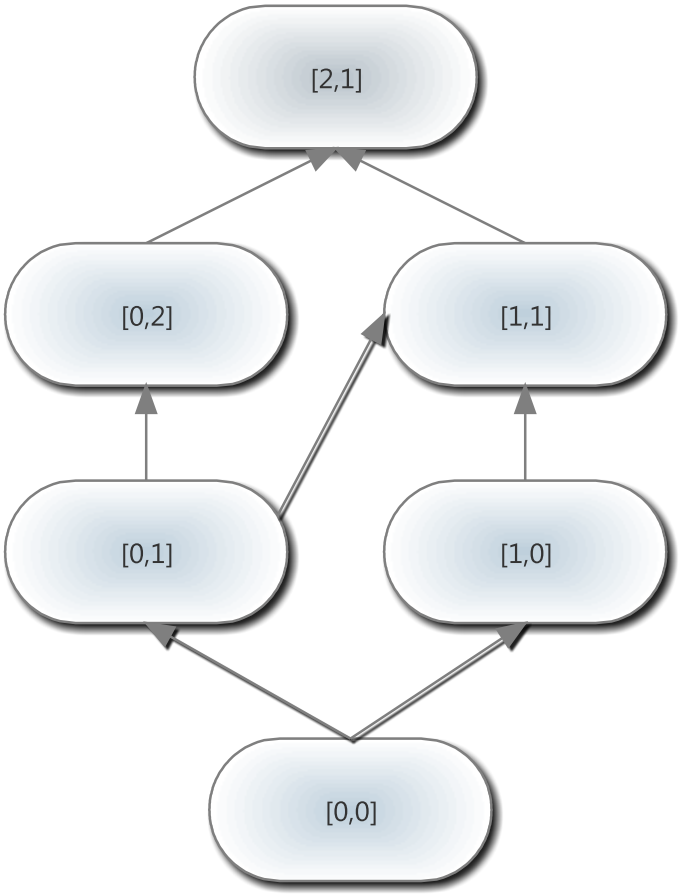



Σχήμα 5.6: Ταξινομημένη δομή δέντρου { Εργασία }

Με τον όρο **διάνυσμα γενίκευσης** [30] θα εννοούμε το επίπεδο γενίκευσης που έχει επιτευχθεί σε κάθε χαρακτηριστικό γνώρισμα. Για παράδειγμα το διάνυσμα [1,0,2] μας λέει πως το πρώτο γνώρισμα έχει γενικευτεί στο πρώτο επίπεδο, το δεύτερο καθόλου ενώ το τρίτο στο δεύτερο επίπεδο.

Με τον όρο **πλέγμα** θα εννοούμε την συλλογή από διανύσματα γενίκευσης και τις διασυνδέσεις τους. Στο πλέγμα παρουσιάζονται σχηματικά όλοι οι δυνατοί συνδυασμοί μεταξύ των επιπέδων των ιεραρχιών γενίκευσης των γνωρισμάτων του ψευδο-αναγνωριστικού, όπου εκφράζονται ουσιαστικά όλες οι δυνατές γενικεύσεις των πλειάδων. Θα απεικονίζονται από κάτω με το αρχικό επίπεδο και προοδευτικά προς τα πάνω μέχρι την μέγιστη δυνατή γενίκευση τους. Ένα τέτοιο πλέγμα που αφορά τα 2

γνωρίσματα του σχήματος 5.5 και 5.6 είναι:




Σχήμα 5.7: Πλέγμα γενίκευσης

Οι κύριες μορφές γενίκευσης είναι πέντε [30, 31]. Διακρίνονται σε:

1. **Ολική γενίκευση πεδίων.** Οι τιμές των ψευδο-αναγνωριστικών γενικεύονται όλες στο ίδιο επίπεδο του δοθέντος ταξινομημένου δέντρου. Προτού γίνει οποιαδήποτε γενίκευση όλες οι τιμές παραμένουν στην βάση του δέντρου. Έπειτα κατά την επεξεργασία της βάσης δεδομένων (από το σχήμα 5.6) αν ο κόμβος *Προπτυχιακός* γενικευτεί στον πατέρα του (κόμβος *Φοιτητής*), πρέπει την ίδια στιγμή ο κόμβος *Μεταπτυχιακός* να γενικευτεί και αυτός στον πατέρα του και με την σειρά τους οι κόμβοι *Προγραμματιστής, Δάσκαλος* στον κόμβο *Εργαζόμενος*. Η μέθοδος αυτή οδηγεί σε ευρείας κλίμακα παραμόρφωση της βάσης δεδομένων.
2. **Γενίκευση υπόδεντρου.** Τα όρια της γενίκευσης αυτής είναι μικρότερα από την ολική και προκαλούν μικρότερη παραμόρφωση των πληροφοριών. Όταν ένας κόμβος γενικεύεται στον πατέρα του τότε είναι υποχρεωτικό να γενικευτούν και τα

υπόλοιπα αδέρφια αυτού του κόμβου. Αν δηλαδή γενικευτεί ο κόμβος *Προπτυχιακός* στον πατέρα του *Φοιτητής* τότε πρέπει να γενικευτεί και ο κόμβος *Μεταπτυχιακός*. Οι κόμβοι όμως *Προγραμματιστής*, *Δάσκαλος* δεν χρειάζεται να υποστούν επεξεργασία και μπορούν να παραμείνουν ως έχουν.

3. **Γενίκευση αδερφού.** Είναι παρόμοια με την γενίκευση υπόδεντρου, με την διαφορά πως ορισμένα αδέρφια μπορούν να παραμείνουν στο ίδιο επίπεδο. Η τιμή του πατέρα τότε θα εκπροσωπεί όλες τις τιμές των παιδιών που λείπουν. Για παράδειγμα αν ο κόμβος *Προγραμματιστής* γενικευτεί στον πατέρα του *Εργαζόμενος* τότε όλες οι εργασίες θα καλύπτονται από την τιμή *Εργαζόμενος* εκτός από την τιμή *Δάσκαλος*. Η τεχνική αυτή προκαλεί ακόμα μικρότερη παραμόρφωση του πίνακα από ότι η γενίκευση του υπόδεντρου.

4. **Γενίκευση εγγραφής.** Η μέθοδος αυτή διαφέρει ελαφρώς από τις παραπάνω. Ενώ στην ολική γενίκευση παραμορφώνονται όλες οι εγγραφές, εδώ περιοριζόμαστε στην γενίκευση μιας μοναδικής πλειάδας της βάσης δεδομένων. Για παράδειγμα μπορούμε σε μία εγγραφή να γενικεύσουμε μόνο τον κόμβο *Προπτυχιακός* στον πατέρα του και τίποτα άλλο.

5. **Γενίκευση n-διαστάσεων**. Εδώ δίνεται έμφαση σε διαφορετικές μεθόδους και συνδυασμούς γενίκευσης των ψευδο-αναγνωριστικών. Θεωρούμε ως $D_i$ το πεδίο ενός χαρακτηριστικού γνωρίσματος $A_i$. Μια 1-διαστάσεων γενίκευση όπως είναι η ολική γενίκευση πεδίων και η γενίκευση υπόδεντρου ορίζεται από μια συνάρτηση $f_i : D_{Ai} \rightarrow D'$ για κάθε γνώρισμα $A_i$ των ψευδο-αναγνωριστικών. Η n-διαστάσεων γενίκευση αντίθετα, ορίζεται από την συνάρτηση $f: D_{A1} \times \ldots \times D_{An} \rightarrow D'$ και χρησιμοποιείται για να γενικεύσει την κάθε κλάση $qid = < \upsilon_1,...,\upsilon_n >$ σε $qid' = < u_1,...,u_n >$ όπου κάθε $\upsilon_i$ θα είναι ίσο με $u_i$ ή θα είναι παιδί του $u_i$ στο ταξινομημένο δέντρο που ανήκει. Για παράδειγμα από τα σχήματα 5.5, 5.6 ο συνδυασμός [Προπτυχιακός, Θηλυκό] μπορεί να γενικευτεί σε [Φοιτητής, Άνθρωπος], ή ο [Μεταπτυχιακός, Αρσενικό] σε [Οτιδήποτε, Αρσενικό]. Με την σωστή επιλογή των τιμών που θα γενικεύσουμε πετυχαίνουμε μικρότερη παραμόρφωση δεδομένων και μεγαλύτερη πληρότητα δεδομένων.

Η **κατάπνιξη (suppression)** [32, 33] είναι η απόλυτη μορφή γενίκευσης, η οποία χρησιμοποιεί ειδικούς σύμβολο-χαρακτήρες όπως *, &, # για να αντικαταστήσει αρχικές τιμές. Όπως στην μέθοδο της γενίκευσης υπάρχουν πέντε είδη κατάπνιξης:

1. **Κατάπνιξη χαρακτηριστικού γνωρίσματος** στην οποία διαγράφονται ολόκληρες στήλες.
2. **Κατάπνιξη Εγγραφής** στην οποία διαγράφονται όλα τα δεδομένα μιας η περισσοτέρων πλειάδων.
3. **Κατάπνιξη τιμών** όπου διαγράφονται μεμονωμένα πληροφορίες εγγραφών.
4. **Κατάπνιξη κελιών** δηλαδή διαγραφή των όποιων εγγραφών περιλαμβάνουν μια συγκεκριμένη πληροφορία.
5. **Κατάπνιξη n-διαστάσεων** στην οποία συνδυάζονται οι παραπάνω μέθοδοι κατάπνιξης για την επίτευξη του βέλτιστου αποτελέσματος.

Οι μέθοδοι αυτές χρησιμοποιούνται τις περισσότερες φόρες παράλληλα με την γενίκευση. Συγκεκριμένα οι αλγόριθμοι ελέγχουν αν μετά την εφαρμογή μιας μεθόδου γενίκευσης εμφανίζονται εγγραφές λιγότερο από k φορές. Τότε μας συμφέρει συνήθως να τις απομακρύνουμε παρά να προχωρήσουμε την γενίκευση των πεδίων σε υψηλότερο επίπεδο. Τελικά προκύπτει πως συνδυάζοντας κάποια από τις μεθόδους γενίκευσης και κάποια από τις κατάπνιξης ελαχιστοποιούμε τα επίπεδα γενίκευσης των χαρακτηριστικών γνωρισμάτων και πετυχαίνουμε λιγότερη απώλεια πληροφορίας.

### *5.1.2 Ανατομία και Μετάθεση*

Σε αντίθεση με την γενίκευση και κατάπνιξη η ανατομία [34] δεν τροποποιεί τα ψευδο-αναγνωριστικά ή τις ευαίσθητες πληροφορίες αλλά αποκρύπτει τις συσχετίσεις μεταξύ της κάθε εγγραφής με την ευαίσθητη τιμή της. Συγκεκριμένα η μέθοδος της ανατομίας αρχικά δημοσιεύει δύο πίνακες: τους QIT (5.3) και ST (5.4), οι οποίοι περιλαμβάνουν τα ψευδο-αναγνωριστικά και τα ευαίσθητα στοιχεία αντίστοιχα, ενώ και οι δύο πίνακες έχουν ένα κοινό γνώρισμα, τον Αριθμό κλάσης ισοδυναμίας. Όλες οι εγγραφές της κάθε κλάσης ισοδυναμίας και οι αντίστοιχες ευαίσθητες τιμές τους θα έχουν και στους δύο πίνακες την ίδια τιμή στον Αριθμό κλάσης ισοδυναμίας. Αν μια κλάση έχει $\ell$ διακριτές ευαίσθητες τιμές, τότε για κάθε μία από αυτές η πιθανότητα να συνδεθεί με μια εγγραφή είναι το πολύ ( $1/\ell$ ). Ας υποθέσουμε πως ο κάτοχος θέλει να εκδώσει τον πίνακα 5.1, όπου η ασθένεια είναι το ευαίσθητο χαρακτηριστικό γνώρισμα ενώ τα QID = { Ηλικία, Φύλο }.

| Ανατομία: Αρχικός πίνακας ασθενών | | |
|---|---|---|
| Ηλικία | Φύλο | Ασθένεια |
| 30 | Αρσενικό | Ηπατίτιδα |
| 30 | Αρσενικό | Ηπατίτιδα |
| 30 | Αρσενικό | Aids |
| 32 | Αρσενικό | Ηπατίτιδα |
| 32 | Αρσενικό | Aids |
| 32 | Αρσενικό | Aids |
| 36 | Θηλυκό | Γρίπη |
| 38 | Θηλυκό | Γρίπη |
| 38 | Θηλυκό | Καρδιά |
| 38 | Θηλυκό | Καρδιά |

Πίνακας 5.1 Αρχικός πίνακας ασθενών

| Ανατομία: Ενδιάμεσος πίνακας με τις κλάσεις ισοδυναμίας | | |
|---|---|---|
| Ηλικία | Φύλο | Ασθένεια |
| [30-35) | Αρσενικό | Ηπατίτιδα |
| [30-35) | Αρσενικό | Ηπατίτιδα |
| [30-35) | Αρσενικό | Aids |
| [30-35) | Αρσενικό | Ηπατίτιδα |
| [30-35) | Αρσενικό | Aids |
| [30-35) | Αρσενικό | Aids |
| [35-40) | Θηλυκό | Γρίπη |
| [35-40) | Θηλυκό | Γρίπη |
| [35-40) | Θηλυκό | Καρδιά |
| [35-40) | Θηλυκό | Καρδιά |

Πίνακας 5.2 Ενδιάμεσος πίνακας

Πρώτα ομαδοποιεί τις εγγραφές σε κλάσεις ισοδυναμίας με τρόπο ώστε το πολύ το ( $1/\ell$ ) των εγγραφών να έχει την ίδια ασθένεια. Ο ενδιάμεσος πίνακας 5.2 περιλαμβάνει δύο κλάσεις ισοδυναμίας: < [30-35), Αρσενικό > και < [35-40), Θηλυκό >.

| Ανατομία: Πίνακας κλάσεων ισοδυναμίας | | |
|---|---|---|
| Ηλικία | Φύλο | Αριθμός κλάσης ισοδυναμίας |
| 30 | Αρσενικό | 1 |
| 30 | Αρσενικό | 1 |
| 30 | Αρσενικό | 1 |
| 32 | Αρσενικό | 1 |
| 32 | Αρσενικό | 1 |
| 32 | Αρσενικό | 1 |
| 36 | Θηλυκό | 2 |
| 38 | Θηλυκό | 2 |
| 38 | Θηλυκό | 2 |
| 38 | Θηλυκό | 2 |

Πίνακας 5.3: Πίνακας κλάσεων ισοδυναμίας QIT

Στην συνέχεια δημιουργείται ο πίνακας QIT (5.3) που περιλαμβάνει τις αρχικές τιμές για τα γνωρίσματα του ψευδο-αναγνωριστικού προσθέτοντας μια καινούρια στήλη η οποία περιέχει τον αριθμό της κλάσης ισοδυναμίας που ανήκει η κάθε εγγραφή.

| Ανατομία: Πίνακας ευαίσθητων τιμών | | |
|---|---|---|
| Αριθμός κλάσης ισοδυναμίας | Ασθένεια | Αριθμός εμφανίσεων |
| 1 | Ηπατίτιδα | 3 |
| 1 | Aids | 3 |
| 2 | Γρίπη | 2 |
| 2 | Καρδιά | 2 |

Πίνακας 5.4: Πίνακας ευαίσθητων τιμών ST

Στο τελευταίο βήμα σχηματίζεται συγκεντρωτικός πίνακας ST (5.4), ο οποίος περιλαμβάνει τις αρχικές τιμές του ευαίσθητου γνωρίσματος, μαζί με τον αριθμό της κλάσης ισοδυναμίας, και τον αριθμό εμφανίσεων της συγκεκριμένης τιμής μέσα στην κλάση ισοδυναμίας.

Με τη χρήση της ανατομίας και τη δημοσίευση των δύο τελευταίων πινάκων, ο επιτιθέμενος γνωρίζοντας κάποιες τιμές του ψευδο-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 στην οποία ανήκει, αφού οι πίνακες 5.3 και 5.4 ικανοποιούν τις προϋποθέσεις για ιδιωτικότητα με $\ell \leq 2$. Η πιθανότητα αποκάλυψης κάθε ευαίσθητης τιμής του πίνακα ST με τις κλάσεις ισοδυναμίας του πίνακα 5.3 είναι το πολύ $1/\ \ell = 1/2 = 50\%$. Το μεγάλο πλεονέκτημα της μεθόδου είναι πως τα δεδομένα στους πίνακες 5.3 και 5.4 δεν τροποποιούνται και επομένως η ποιότητα των πινάκων παραμένει υψηλή. Για να το δούμε πρακτικά, ας υποθέσουμε ότι ο παραλήπτης θέλει να μετρήσει τον αριθμό των ασθενών που είναι 38 και έχουν πρόβλημα στην καρδιά. Η σωστή απάντηση από τον αρχικό πίνακα 3 είναι 2 ασθενείς. Από τους πίνακες ανατομίας QIT και ST βρίσκουμε πως είναι 3 x (2/4) = 1.5 επειδή 2 από τις 4 εγγραφές στο γνώρισμα { Αριθμός Εμφανίσεων } έχουν πρόβλημα καρδιάς. Από τον γενικευμένο πίνακα 5.2 προκύπτει πως είναι 2 x (1/5) = 0.4, όπου (1/5) επειδή 2 ασθενείς έχουν ίση πιθανότητα να είναι ηλικίας { 35, 36, 37, 38, 39 } και να έχουν πρόβλημα καρδιάς. Προφανώς με την τεχνική της ανατομίας επιτεύχθει μεγαλύτερη ακρίβεια.

Ωστόσο πρέπει να σημειωθεί πως επειδή στην τεχνική της ανατομίας δημοσιεύονται δύο πίνακες απαιτούνται εξιδεικευμένοι αλγόριθμοι ταξινόμησης, ομαδοποίησης και συσχέτισης για την εξόρυξη δεδομένων. Ακόμη η ανατομία σε αντίθεση με την τεχνική της γενίκευσης δεν ταιριάζει σε βάσεις δεδομένων που πρέπει συχνά να ανανεώνουν τις εγγραφές τους.

Η τεχνική της **μετάθεσης** [35] είναι παρόμοια με την ανατομία και επίσης διατηρεί αναλλοίωτη την ακρίβεια των τιμών των δεδομένων. Η κεντρική ιδέα είναι η διάσπαση των δεσμών μεταξύ των ψευδο-αναγνωριστικών και των αριθμητικά ευαίσθητων στοιχείων με την ομαδοποίηση των εγγραφών σε κλάσεις ισοδυναμίας και το ανακάτωμα των ευαίσθητων στοιχείων της κάθε κλάσης.

### *5.1.3 Διαταραχή*

Η τεχνική της **διαταραχής (perturbation)** [36] χρησιμοποιείται ευρύτατα σε στατιστικούς πίνακες εξαιτίας της ευκολίας, της αποτελεσματικότητας και της ικανότητάς της να διατηρεί στατιστικές πληροφορίες. Η γενική ιδέα είναι η αντικατάσταση των αρχικών τιμών ενός πίνακα με άλλες πιο σύνθετες με τρόπο ώστε οι στατιστικές πληροφορίες του τελικού πίνακα να διαφέρουν ελάχιστα από τις αρχικές. Ανάλογα με τον βαθμό επεξεργασίας του "διαταραγμένου" πίνακα οι ευαίσθητες πληροφορίες των εγγραφών μπορεί να μην ανταποκρίνονται στην πραγματικότητα και κατ' επέκταση να αποτρέπονται οι επιθέσεις με την μέθοδο της αποκάλυψης.

Συγκριτικά με τις τεχνικές ανωνυμοποίησης που αναλύθηκαν παραπάνω ένας περιορισμός της διαταραχής είναι πως οι πληροφορίες των εγγραφών του τελικού πίνακα δεν ανταποκρίνονται ορθά στις οντότητες που αυτά εκπροσωπούν. Οι επεξεργασμένοι πίνακες ωστόσο παραμένουν ιδιαίτερα χρήσιμοι για τον παραλήπτη αν ο κύριος τομέας μελέτης είναι οι στατιστικές ιδιότητες του πίνακα. Μάλιστα ο παραλήπτης μπορεί να ανασυνθέσει με ακρίβεια τον πίνακα στην περίπτωση που γνωρίζει τις παραμέτρους που χρησιμοποιήθηκαν για την επεξεργασία του. Επειδή υπάρχει ο φόβος της ανασύνθεσης του επεξεργασμένου πίνακα πολλοί υποστηρίζουν πως δεν πρέπει να δημοσιεύεται και να γίνονται διαθέσιμες μόνο οι στατιστικές ιδιότητες και τα συμπεράσματα εξόρυξης των δεδομένων του.

Παρακάτω θα αναλυθούν οι τρεις πιο γνωστές μέθοδοι της διαταραχής: η *προσθήκη θορύβου*, η *εναλλαγή δεδομένων* και η *συνθετική παραγωγή δεδομένων*.

1. **Προσθήκη Θορύβου**: Είναι μια ευρέως χρησιμοποιούμενη μέθοδος των στατιστικών πινάκων για την προστασία της ιδιωτικότητας. Χρησιμοποιείται συχνά για την απόκρυψη ευαίσθητων αριθμητικών τιμών όπως είναι ο μισθός. Βάσει της γενικής ιδέας αντικαθίσταται η κάθε ευαίσθητη τιμή s με s + r, όπου r είναι μία τυχαία τιμή και παράμετρος της μεθόδου. Η ιδιωτικότητα ποσοτικά υπολογίζεται ως η απόσταση των αρχικών από τις επεξεργασμένες τιμές. Μετρήσεις και πειράματα της μεθόδου έδειξαν πως βασικές στατιστικές ιδιότητες όπως μέση τιμή ή διακύμανση και συμπεράσματα εξόρυξης πληροφοριών δεν επηρεάζονται καθόλου.

2. **Εναλλαγή δεδομένων** [37]: Η κεντρική ιδέα της μεθόδου για την ανωνυμοποίηση του πίνακα είναι η ανταλλαγή ευαίσθητων τιμών μεταξύ των εγγραφών. Οι

ανταλλαγές γίνονται με τρόπο ώστε να διατηρούνται τα βασικά χαρακτηριστικά που διέπουν τον πίνακα για στατιστική ανάλυση. Χρησιμοποιείται για την προστασία αριθμητικών και κατηγορικών χαρακτηριστικών γνωρισμάτων. Μια εναλλακτική μορφή εναλλαγής δεδομένων είναι η μέθοδος εναλλαγή κλάσης. Πρώτα κατατάσσει τις τιμές ενός γνωρίσματος Α σε αύξουσα σειρά. Ύστερα για κάθε τιμή υ ∈ Α, ανταλλάσσει την υ με άλλη τιμή u ∈ Α εντός ενός προκαθορισμένου από τον κάτοχο εύρους p% της τιμής υ. Η μέθοδος εναλλαγή κλάσης διατηρεί ακόμα καλύτερα τις στατιστικές ιδιότητες του πίνακα από την εναλλαγή δεδομένων.

3. **Συνθετική παραγωγή δεδομένων**: Η μέθοδος χρησιμοποιείται για την διατήρηση της ιδιωτικότητας των εγγραφών και ταυτόχρονα των χρήσιμων στατιστικών ιδιοτήτων του πίνακα. Στην πράξη χτίζει αρχικά ένα στατιστικό πίνακα από τα δεδομένα και κρατάει μερικά κομμάτια του. Αυτά τα κομμάτια ενωμένα αποτελούν τον συνθετικό πίνακα.

## *5.2 Μετρική απώλειας πληροφορίας*

Ο τομέας της προστασίας της ιδιωτικότητας έχει δύο κύριες πλευρές: την προστασία του απορρήτου και την διατήρηση της ποιότητας των πληροφοριών που δημοσιεύονται. Τα σύνολα δεδομένων δημοσιεύονται κυρίως για την εκμετάλλευση της χρήσιμης πληροφορίας που περιέχουν σχετικά με το σύνολο του πληθυσμού που αντιπροσωπεύουν. Επομένως είναι σημαντικό οι πληροφορίες που περιέχονται σε αυτά να είναι ποιοτικές καθώς με την μελέτη τους μπορούν να αποδοθούν σημαντικά αποτελέσματα και συμπεράσματα σε έρευνες ή στατιστικές αναλύσεις.

Οι τεχνικές ανωνυμοποίησης που είδαμε σε προηγούμενο κεφάλαιο τροποποιούν τα αρχικά δεδομένα με στόχο την διαφύλαξη της ιδιωτικότητας των εγγραφών. Ένα μέρος της προσωπικής πληροφορίας αποκρύπτεται στα δημοσιευμένα. Συνεπώς σε κάθε περίπτωση τροποποίησης των δεδομένων με σκοπό την προστασία της ιδιωτικότητας σημαντικό ρόλο κατέχει και το ποσοστό χρήσιμης πληροφορίας που χάνεται.

Οι αλγόριθμοι που υλοποιούν της αρχές ανωνυμίας εξετάζονται ως προς την αποδοτικότητά τους αναφορικά με αυτήν. Έχουν αναπτυχθεί κατάλληλες μετρικές και εργαλεία έτσι ώστε να αξιολογούνται οι αλγόριθμοι και οι εγγυήσεις ιδιωτικότητας όχι

μόνο βάσει της προστασίας που προσφέρουν αλλά και βάσει της χρήσιμης πληροφορίας που διατηρούν στα δημοσιευμένα δεδομένα. Σύμφωνα με αυτό το κριτήριο και με χρήση των κατάλληλων μετρικών διακρίνεται η βέλτιστη τροποποίηση των δεδομένων μεταξύ των προτεινόμενων τεχνικών.

Σε προγραμματιστικό επίπεδο, οι βέλτιστες μετρικές κόστους απώλειας πληροφορίας είναι οι **γενικού σκοπού**. Προτιμούνται επειδή καλύπτουν ένα μεγάλο εύρος και προοδευτικά τείνουν να τυποποιηθούν και να εφαρμόζονται σε κάθε περίπτωση. Οι μετρικές αυτές μετρούν την ομοιότητα μεταξύ των αρχικών και ανωνυμοποιημένων δεδομένων βάσει της *αρχής ελάχιστης παραμόρφωσης* [38].

Διάκριση ανάμεσά τους μπορεί να γίνει σε **μετρικές δεδομένων**, οι οποίες μετρούν την ποιότητα των δεδομένων όλου του ανωνυμοποιημένου πίνακα συγκριτικά με τον αρχικό και σε **μετρικές εύρους**, οι οποίες εξετάζουν αν ο αλγόριθμος παράγει πίνακες με μέγιστη πληροφορία ή με ελάχιστη παραμόρφωση.

### *5.2.1 Ελάχιστη Παραμόρφωση*

Στην μετρική ελάχιστης παραμόρφωσης (MD) [38] επιβάλλεται ποινή για τιμή του πίνακα που γενικεύεται ή διαγράφεται. Έτσι η γενίκευση δέκα πεδίων με την τιμή Μηχανικός σε Επαγγελματία, μεταφράζεται σε 10 μονάδες παραμόρφωσης. Περαιτέρω γενίκευση των τιμών σε Οποιαδήποτε_Εργασία κοστίζει επιπλέον 10 μονάδες παραμόρφωσης. Αυτή η μετρική αφορά μονά χαρακτηριστικά γνωρίσματα.

### *5.2.2 Μετρική διάκρισης*

Η Μετρική διάκρισης (DM) [33] εξετάζει κατά πόσο οι εγγραφές γίνονται δυσδιάκριτες και ορίζει την απώλεια πληροφορίας με την επιβολή ποινής για κάθε μία εγγραφή ίση με το μήκος της κλάσης ισοδυναμίας της στο τετράγωνο. Το σύνολο της ποινής για τον κάθε ανωνυμοποιημένο πίνακα T* είναι:

$$\text{DM (T*)} = \sum_{\forall E\ s.t. |EQ| \geq k} |EQ^2| + \sum_{\forall E\ s.t. |EQ| < k} |T| * |EQ|$$

όπου T είναι ο αρχικός πίνακας, |T| το πλήθος των εγγραφών του και |EQ| το μήκος της κλάσης ισοδυναμίας έπειτα από την ανωνυμοποίηση. Η βασική ιδέα πίσω από την μετρική είναι ότι μεγαλύτερες κλάσεις ισοδυναμίας εκπροσωπούν μεγαλύτερη απώλεια

πληροφοριών.

Η DM από την MD έχουν διαφορές. Η MD επιβάλλει ποινή για κάθε γενίκευση τιμής χωρίς να λαμβάνει υπόψιν της το είδος της τιμής αυτής. Για παράδειγμα η ποινή στην MD για την γενίκευση 99 τιμών { Μηχανικός } και μία { Δικηγόρος } στον πατέρα τους { Επαγγελματίας } είναι ίδια με την γενίκευση 50 τιμών { Χορευτής } και 50 { Συγγραφείς } σε { Καλλιτέχνης }. Η DM σε αντίθεση μπορεί να διακρίνει τις δυο περιπτώσεις:

- DM (Επαγγελματίας) = $99^2 + 1^2 = 9802$
- DM (Επαγγελματίας) = $50^2 + 50^2 = 5000$

Η DM μπορεί να προσδιορίσει στον αλγόριθμο πως η δεύτερη περίπτωση κοστίζει λιγότερη από την πρώτη.

### *5.2.3 Μετρική μέσου μεγέθους κλάσης ισοδυναμίας*

Η μετρική μέσου μεγέθους κλάσης ισοδυναμίας (CAVG) [39] εξετάζει την σωστή διαμόρφωση των κλάσεων ισοδυναμίας. Δίνεται από την εξίσωση:

$$\text{CAVG (T} *) = \frac{(|\text{T}|)}{|\text{EQs}| * \text{k}}$$

όπου |EQs| είναι το πλήθος των κλάσεων ισοδυναμίας και k η παράμετρος της ανωνυμίας. Η βέλτιστη ποινή προκύπτει αν δωθεί αποτέλεσμα 1, οπότε συμπεραίνουμε πως έχει γίνει ιδανική ανωνυμοποίηση.

### *5.2.4 Γενικευμένη απώλεια πληροφοριών*

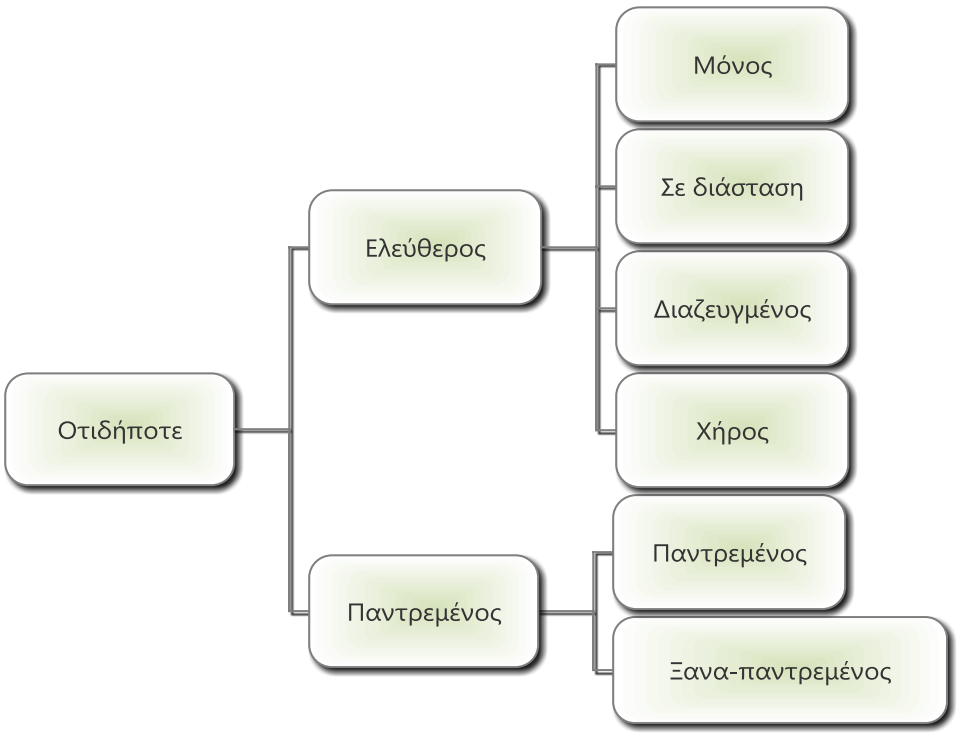
Αυτή η μετρική (GenILoss) [40] επιβάλλει ποινή όταν γενικεύεται ένα συγκεκριμένο χαρακτηριστικό γνώρισμα με την ποσοτικοποίηση του κλάσματος των γενικευμένων τιμών. Η ολική απώλεια πληροφοριών ενός ανωνυμοποιημένου πίνακα T* μετριέται με βάση τον τύπο:

$$\text{GenILoss (T} *) = \frac{1}{|T| * n} * \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{Uij - Lij}{Ui - Li}$$

όπου $L_i$ και $U_i$ είναι το πάνω και κάτω όριο ενός χαρακτηριστικού γνωρίσματος i και τα $L_{ij}$ , $U_{ij}$ το πάνω και κάτω όριο του επιπέδου της γενίκευσης ενός j κελιού.

Η μετρική GenILoss βασίζεται στην ιδέα πως κάθε τιμή ενός κελιού που εκπροσωπεί ένα μεγαλύτερο εύρος τιμών είναι λιγότερο ακριβή. Εάν η μετρική μας δώσει 0 τότε δεν έχει γίνει καμία τροποποίηση ενώ αν δώσει 1 τότε ο πίνακας έχει υποστεί την μέγιστη γενίκευση δεδομένων. Η μετρική μπορεί να εφαρμοστεί σε αριθμητικά αλλά και σε κατηγορικά δεδομένα. Στα αριθμητικά δεδομένα ο τρόπος είναι προφανής. Στα κατηγορικά δεδομένα ο αλγόριθμος αντιστοιχεί κάθε τιμή με έναν αριθμό.

Για παράδειγμα στο σχήμα 5.8 που απεικονίζεται σε ιεραρχία γενίκευσης τιμής η οικογενειακή κατάσταση, η τιμή Μόνος αντιστοιχίζεται στην τιμή 1, η τιμή Σε διάσταση στην τιμή 2 και συνεχίζει μέχρι την τιμή Ξανα-παντρεμένος που παίρνει την τιμή 6. Επομένως η κατάσταση Ελεύθερος εκπροσωπείται από τις τιμές [1-4].




Σχήμα 5.8: Αναλυτική ιεραρχία γενίκευσης στην { Οικογενειακή κατάσταση }

## *5.3 Μοντέλα επιθέσεων και ιδιωτικότητας*

Ο Dalenius [41] παραθέτει έναν πολύ αυστηρό ορισμό της ιδιωτικότητας:

*Η πρόσβαση στα δημοσιευμένα δεδομένα δεν πρέπει να επιτρέπει σε κανέναν επιτιθέμενο να μάθει οτιδήποτε καινούριο για τον οποιονδήποτε άνθρωπο, ακόμα και αν ο επιτιθέμενος έχει αποκτήσει πληροφορίες από άλλες πηγές.*

Πρακτικά η απόλυτη προστασία της ιδιωτικότητας είναι αδύνατη εξαιτίας του μεγάλου όγκου πληροφοριών που συνήθως κατέχουν οι επιτιθέμενοι. Θεωρούμε πως τα ανωνυμοποιημένα δεδομένα απειλούνται όταν ο επιτιθέμενος μπορεί να αποκαλύψει μία εγγραφή δεδομένων ή ένα ευαίσθητο χαρακτηριστικό γνώρισμα μιας δημοσιευμένης βάσης δεδομένων ή την παρουσία στην δημοσιευμένη βάση δεδομένων.

Με βάση λοιπόν τις επιθέσεις που μπορούν να δεχτούν οι βάσεις δεδομένων τα μοντέλα ιδιωτικότητας χωρίζονται σε τρεις κατηγορίες: *αποκάλυψη εγγραφής, αποκάλυψη χαρακτηριστικού γνωρίσματος* και *αποκάλυψη παρουσίας στον πίνακα*:

i. Αποκάλυψη εγγραφής: Το πρώτο είδος απειλής κατά τη δημοσίευση δεδομένων είναι η αποκάλυψη εγγραφής. Σε αυτή την περίπτωση, ο αντίπαλος, ανεξαρτήτως γνωστικού υποβάθρου, δεν θα πρέπει να είναι σε θέση να συνδέσει με υψηλό βαθμό βεβαιότητας μια συγκεκριμένη εγγραφή από τα δεδομένα σε μια οντότητα του πραγματικού κόσμου. Εδώ υποθέτουμε πως ο επιτιθέμενος γνωρίζει ήδη πως η εγγραφή της οντότητας που τον ενδιαφέρει περιλαμβάνεται στα δεδομένα.

ii. Αποκάλυψη χαρακτηριστικού γνωρίσματος: Το επόμενο είδος απειλής κατά τη δημοσίευση δεδομένων είναι η αποκάλυψη χαρακτηριστικού γνωρίσματος. Σε αυτή την περίπτωση, ο επιτιθέμενος, ανεξαρτήτως γνωστικού υπόβαθρου δεν θα πρέπει να είναι σε θέση να χρησιμοποιήσει τις εγγραφές στα δημοσιευμένα δεδομένα προκειμένου να ανακαλύψει με υψηλή πιθανότητα ευαίσθητα γνωρίσματα για μια οντότητα του πραγματικού κόσμου. Ομοίως με την περίπτωση της αποκάλυψης εγγραφής, η ύπαρξη της εγγραφής μιας οντότητας στα δημοσιευμένα δεδομένα θεωρείται πως είναι γνωστή στον εισβολέα από την αρχή.

iii. Αποκάλυψη παρουσίας στον πίνακα: Το τρίτο και τελευταίο είδος απειλής κατά τη δημοσίευση δεδομένων είναι η αποκάλυψη παρουσίας στον πίνακα. Σε αυτή τη περίπτωση, ο αντίπαλος, ανεξαρτήτως γνωστικού υπόβαθρου, δεν θα πρέπει να είναι σε θέση να συμπεράνει με υψηλό βαθμό βεβαιότητας ότι η εγγραφή που

αντιστοιχεί σε μια συγκεκριμένη οντότητα συμπεριλαμβάνεται στα δημοσιευμένα δεδομένα. Αυτή η απειλή συνήθως συμβαίνει όταν τα αρχεία που πρόκειται να δημοσιευτούν επιλέγονται κατόπιν κριτηρίων που σχηματίζουν ένα ή περισσότερα ευαίσθητα στοιχεία τα οποία με κάποιο τρόπο είναι γνωστά στον επιτιθέμενο.

Στην συνέχεια με τον όρο *θύμα* εννοούμε την οντότητα του πραγματικού κόσμου του οποίου οι ευαίσθητες πληροφορίες δέχονται την επίθεση. Θεωρούμε πως ένας πίνακας διατηρεί το απόρρητο των ανθρώπων που περιλαμβάνει, αν μπορεί να αποτρέψει αποτελεσματικά έναν επιτιθέμενο από το να εφαρμόσει μία από τις παραπάνω αποκαλύψεις.

Αξίζει να σημειωθεί επίσης πως τα περισσότερα μοντέλα ιδιωτικότητας επιδιώκουν να ικανοποιήσουν την *αρχή της ελάχιστης πληροφορίας*. Η δημοσιευμένη βάση δεδομένων θα πρέπει να παρέχει στον αντίπαλο ελάχιστες επιπλέον πληροφορίες συγκριτικά με αυτές που ήδη κατέχει. Αν οι προηγούμενες γνώσεις του αντιπάλου έχουν μεγάλη διαφορά από τις επόμενες τότε η επίθεση καλείται *πιθανολογική*. Σε αυτήν την κατηγορία τα περισσότερα μοντέλα ιδιωτικότητας κατατάσσουν τα ευαίσθητα γνωρίσματα μέσα στα ψευδο-αναγνωριστικά αντί να τα έχουν χωριστά.

Παρακάτω παρατίθεται ένας πίνακας που απεικονίζει περιληπτικά τα μοντέλα ιδιωτικότητας που ταιριάζουν στο κάθε είδος επίθεσης:

| ***Μοντέλο ιδιωτικότητας*** | ***Μοντέλο επίθεσης*** | | |
|---|---|---|---|
| | Αποκάλυψη Εγγραφής | Αποκάλυψη Χαρακτηριστικού Γνωρίσματος | Αποκάλυψη παρουσίας Πίνακα |
| ***k-ανωνυμία*** | ✓ | | |
| ***Πολυσχεσιακή k-ανωνυμία*** | ✓ | | |
| ***(c,t)-απομόνωση*** | ✓ | | |
| ***$k^m$-ανωνυμία*** | ✓ | | |
| ***l-ποικιλομορφία*** | ✓ | ✓ | |
| ***Οριοθέτηση δύναμης*** | | ✓ | |
| ***(X,Y)-ιδιωτικότητα*** | ✓ | ✓ | |
| ***(a, k)-ανωνυμία*** | ✓ | ✓ | |
| ***LKC-ιδιωτικότητα*** | ✓ | ✓ | |
| ***(k, e)-ανωνυμία*** | | ✓ | |
| ***(ε,m)-ανωνυμία*** | | ✓ | |
| ***t-εγγύτητα*** | | ✓ | |
| ***Εξατομικευμένη ιδιωτικότητα*** | | ✓ | |
| ***FF-ανωνυμία*** | | ✓ | |
| ***m-αμεταβλητότητα*** | ✓ | ✓ | |
| ***δ-παρουσία*** | | | ✓ |
| ***ε-διαφορική ιδιωτικότητα*** | | | ✓ |
| ***(d-g)-ιδιωτικότητα*** | | | ✓ |

Πίνακας 5.5: Μοντέλα ιδιωτικότητας

### *5.3.1 Αποκάλυψη εγγραφής*

Στην **αποκάλυψη εγγραφής** όπως είδαμε ο επιτιθέμενος προσπαθεί να απομονώσει και να ταυτοποιήσει μονοσήμαντα ευαίσθητες πληροφορίες ανθρώπων.

Ας υποθέσουμε πως ένα νοσοκομείο θέλει να δημοσιεύσει τις εγγραφές των ασθενών (Πίνακας 5.6) σε ένα ερευνητικό κέντρο :

| Αρχική ιατρική βάση | | | |
|---|---|---|---|
| Εργασία | Φύλο | Ηλικία | Ασθένεια |
| Μηχανικός | Αρσενικό | 35 | Αφυδάτωση |
| Μηχανικός | Αρσενικό | 38 | Αφυδάτωση |
| Δικηγόρος | Αρσενικό | 38 | Πυρετός |
| Συγγραφέας | Θηλυκό | 30 | Γρίπη |
| Συγγραφέας | Θηλυκό | 30 | Πυρετός |
| Χορευτής | Θηλυκό | 30 | Πυρετός |
| Χορευτής | Θηλυκό | 30 | Πυρετός |

Πίνακας 5.6: Ιατρική βάση δεδομένων

| Εξωτερική βάση δεδομένων | | | |
|---|---|---|---|
| Όνομα | Εργασία | Φύλο | Ηλικία |
| Αλίκη | Συγγραφέας | Θηλυκό | 30 |
| Βασίλης | Μηχανικός | Αρσενικό | 35 |
| Γεωργία | Συγγραφέας | Θηλυκό | 30 |
| Δημήτρης | Δικηγόρος | Αρσενικό | 38 |
| Ελευθερία | Χορευτής | Θηλυκό | 30 |
| Ζήσης | Μηχανικός | Αρσενικό | 38 |
| Ήβη | Χορευτής | Θηλυκό | 30 |
| Θαλής | Δικηγόρος | Αρσενικό | 39 |
| Ιωάννα | Χορευτής | Θηλυκό | 32 |

Πίνακας 5.7: Εξωτερική βάση δεδομένων

Επίσης ας υποθέσουμε πως το ερευνητικό κέντρο έχει πρόσβαση στον εξωτερικό πίνακα 5.7 και ξέρει επίσης πως κάθε εγγραφή του πίνακα 5.6 υπάρχει και στην εξωτερική βάση δεδομένων.

Με την ένωση των δύο πινάκων στα κοινά χαρακτηριστικά γνωρίσματα: { Εργασία, Φύλο, Ηλικία } μπορεί να αποκαλυφθεί η ταυτότητα του ανθρώπου και η ασθένεια του.

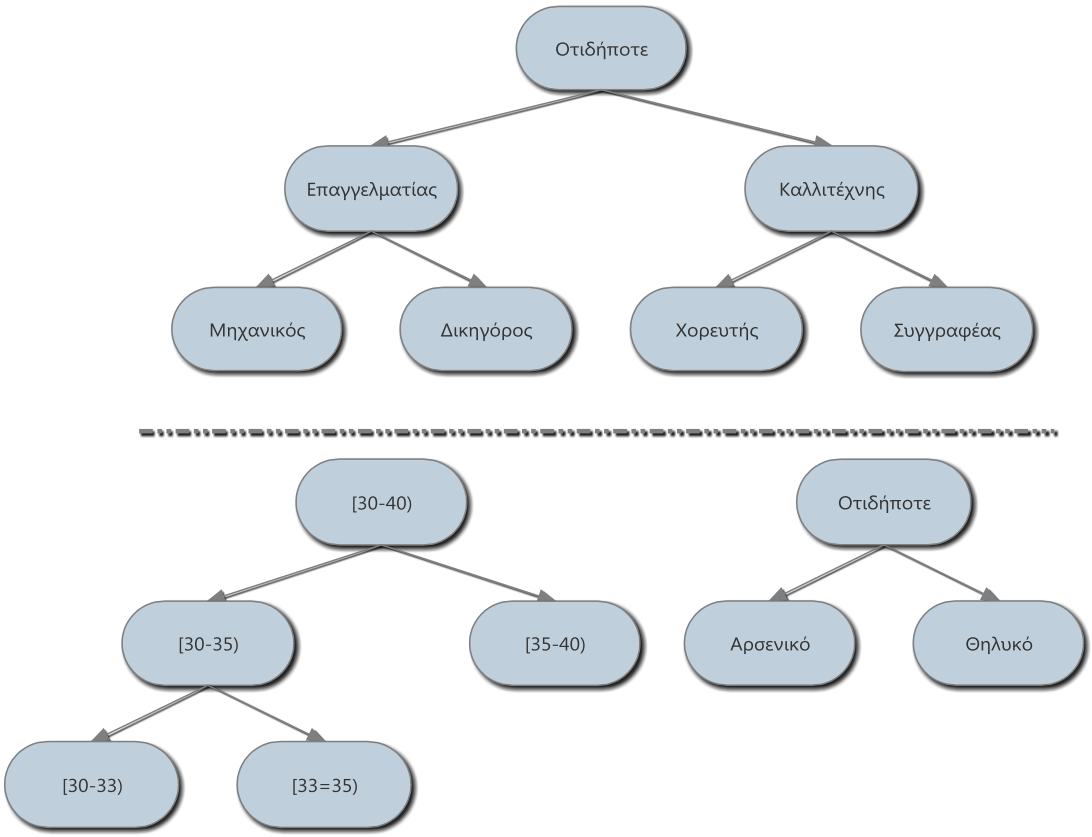
## *k-ανωνυμία*

Οι καθηγητές Latanya Sweeny και Samarati είναι οι πρώτοι που παρουσίασαν το μοντέλο **k-ανωνυμία (k-anonymity)** [9, 29, 40] για να αποκρούσουν την αποκάλυψη εγγραφής. Η ιδέα είναι πως για κάθε εγγραφή σε έναν δημοσιευμένο πίνακα δεδομένων θα υπάρχουν τουλάχιστον άλλες (k -1) εγγραφές ίδιες με την αρχική. Διαμορφώνονται λοιπόν κλάσεις ισοδυναμίας, όπου η κάθε μία θα είναι μεγέθους τουλάχιστον k. Έτσι κάθε προσπάθεια σύνδεσης με έναν εξωτερικό πίνακα θα επιστρέψει k ίδιες εγγραφές και η πιθανότητα να αποκαλυφθούν τα ευαίσθητα γνωρίσματα ιδιωτών μέσω των ψευδο-αναγνωριστικών δεν θα ξεπερνάει το (1/k). Κανείς δεν θα μπορεί δηλαδή να ξεχωρίσει με απόλυτη βεβαιότητα ένα άτομο ανάμεσα σε μια ομάδα k ανθρώπων, αφού θα έχουν τα ίδια στοιχεία. Ο πίνακας που ικανοποιεί αυτές της προϋποθέσεις αποκαλείται k-ανώνυμος (k-anonymous).

Να σημειωθεί πως το μοντέλο k-anonymity δεν μπορεί να αντικατασταθεί από κάποιο της αποκάλυψης χαρακτηριστικού γνωρίσματος γιατί η k-ανωνυμία δεν θεωρεί δεδομένη την ύπαρξη ευαίσθητων γνωρισμάτων.

Ορισμός k-anonymity: Έστω πίνακας B με χαρακτηριστικά γνωρίσματα $[A_1,\ldots A_n]$ και $QID_B$ τα ψευδο-αναγνωριστικά του με χαρακτηριστικά γνωρίσματα $[A_i,...A_j]$ που αποτελούν υποσύνολο του $[A_1,\ldots A_n]$. Αν κάθε εγγραφή του B[ $QID_B$] εμφανίζεται τουλάχιστον k φορές τότε λέμε πως ο πίνακας B ικανοποιεί το μοντέλο k-anonymity.

Παρακάτω απεικονίζονται τρία ταξινομημένα δέντρα στην θεματολογία: Εργασία, Φύλο, Ηλικία αντίστοιχα βασισμένα στα χαρακτηριστικά του πίνακα 5.6:




Σχήμα 5.9: Ταξινομημένα δέντρα στα γνωρίσματα { Εργασία, Ηλικία, Φύλο }

Επεξεργαστήκαμε τον πίνακα 5.6 σε 3-anonymous με την μέθοδο της γενίκευσης στα QID = { Εργασία, Ηλικία Φύλο, } με βάση τα ταξινομημένα δέντρα του σχήματος 5.9 και προκύπτει ο πίνακας 5.8:

| 3-anonymous ιατρική βάση | | | |
|---|---|---|---|
| Εργασία | Φύλο | Ηλικία | Ασθένεια |
| Επαγγελματίας | Αρσενικό | [35-40) | Αφυδάτωση |
| Επαγγελματίας | Αρσενικό | [35-40) | Αφυδάτωση |
| Επαγγελματίας | Αρσενικό | [35-40) | Πυρετός |
| Καλλιτέχνης | Θηλυκό | [30-35) | Γρίπη |
| Καλλιτέχνης | Θηλυκό | [30-35) | Πυρετός |
| Καλλιτέχνης | Θηλυκό | [30-35) | Πυρετός |
| Καλλιτέχνης | Θηλυκό | [30-35) | Πυρετός |

Πίνακας 5.8: 3-anonymous ιατρική βάση δεδομένων

| 4-anonymous εξωτερική βάση δεδομένων | | | |
|---|---|---|---|
| Όνομα | Εργασία | Φύλο | Ηλικία |
| Αλίκη | Καλλιτέχνης | Θηλυκό | [30-35) |
| Βασίλης | Επαγγελματίας | Αρσενικό | [35-40) |
| Γεωργία | Καλλιτέχνης | Θηλυκό | [30-35) |
| Δημήτρης | Επαγγελματίας | Αρσενικό | [35-40) |
| Ελευθερία | Καλλιτέχνης | Θηλυκό | [30-35) |
| Ζήσης | Επαγγελματίας | Αρσενικό | [35-40) |
| Ήβη | Καλλιτέχνης | Θηλυκό | [30-35) |
| Θαλής | Επαγγελματίας | Αρσενικό | [35-40) |
| Ιωάννα | Καλλιτέχνης | Θηλυκό | [30-35) |

Πίνακας 5.9: 4-anonymous εξωτερική βάση δεδομένων

Διαμορφώθηκαν δύο κλάσεις ισοδυναμίας με τιμές : <Επαγγελματίας, Αρσενικό, [35-40)> και <Καλλιτέχνης, Θηλυκό, [30-35)>. Αν κάνουμε τον πίνακα 5.7 σε 4-anonymous πάλι με την μέθοδο της γενίκευσης στα ίδια QID με του πίνακα 5.8, τότε προκύπτει ο πίνακας 5.9.

Παρατηρούμε πως αν συνδέσουμε τις εγγραφές στα ψευδο-αναγνωριστικά του πίνακα 5.7 με αυτές του 5.9, κάθε μία εγγραφή είτε δεν θα συνδέεται με καμία είτε με τουλάχιστον τρεις ακόμη εγγραφές.

Το μοντέλο της k-ανωνυμίας υποθέτει πως τα ψευδο-αναγνωριστικά είναι γνωστά σε εκείνον που θα επεξεργαστεί τα αρχικά δεδομένα. Υπάρχει η δυνατότητα να κατηγοριοποιήσουμε όλα τα χαρακτηριστικά γνωρίσματα ενός πίνακα, χρήση των οποίων μπορεί να κάνει ένας επιτιθέμενος, σε ψευδο-αναγνωριστικά. Όσα περισσότερα χαρακτηριστικά γνωρίσματα συμπεριλαμβάνονται σε αυτά τόσο μεγαλύτερη προστασία της ιδιωτικότητας επιτυγχάνεται. Από την άλλη όμως, αυτό θα σημαίνει πως θα έχουμε αυξημένη παραμόρφωση δεδομένων αφού τότε για να ενταχθούν οι εγγραφές σε κλάσεις ισοδυναμίας θα πρέπει να συμφωνούν σε περισσότερα χαρακτηριστικά γνωρίσματα. Με το ζήτημα αυτό ασχολήθηκε ο

καθηγητής Fung [22], ο οποίος επέκτεινε τον μοντέλο της k-ανωνυμίας και πρότεινε τον προσδιορισμό πολλαπλών ψευδο-αναγνωριστικών, με την προϋπόθεση όμως πως ο κάτοχος των δεδομένων θα γνωρίζει τα πιθανά QID. Ας δούμε ένα παράδειγμα του ζητήματος:

Ένας κάτοχος δεδομένων θέλει να εκδώσει έναν πίνακα T ( A, B, Γ, Δ, S ), όπου η στήλη S αποτελεί ένα ευαίσθητο χαρακτηριστικό γνώρισμα και ξέρει πως ο τελικός αποδέκτης αυτού του πίνακα έχει πρόσβαση στους ήδη δημοσιευμένους πίνακες T1( A, B, X ) και T2( Γ, Δ, Y ), όπου X, Y είναι ξένα για τον πίνακα T γνωρίσματα. Για να αποτραπεί η αποκάλυψη των εγγραφών του πίνακα T στις πληροφορίες X ή Y ο κάτοχος των δεδομένων πρέπει να εφαρμόσει την μέθοδο της k-ανωνυμίας στον πίνακα T για $QID_1$ = {A, B} και $QID_2$ = {Γ, Δ}. Έπειτα κάθε εγγραφή του πίνακα T θα είναι σε δύο ομάδες και σε κλάσεις ισοδυναμίας μεγέθους τουλάχιστον k, η πρώτη με βάση το $QID_1$ και η δεύτερη με βάση το $QID_2$ . Οι δύο ομάδες δεν είναι απαραίτητα ίδιες. Επίσης ο πίνακας T μπορεί να ικανοποιεί την k-ανωνυμία για $QID_1$ και για $QID_2$ αλλά όχι για QID = { A, B, Γ, Δ } .

Ένα δεύτερο σενάριο είναι να έχουμε δύο επιτιθέμενους, ο ένας να έχει πρόσβαση στον T1 και ο δεύτερος στον T2. Επειδή λοιπόν κανείς εκ των δύο δεν έχει πρόσβαση και στα τέσσερα χαρακτηριστικά γνωρίσματα A, B, Γ, και Δ είναι ανώφελο να γίνει ο πίνακας T k-ανώνυμος για QID = { A, B, Γ, Δ } αφού αυτό θα οδηγούσε σε υπερπροστασία και κατ' επέκταση σε μεγαλύτερη παραμόρφωση δεδομένων από ότι χρειάζεται.

Ο προσδιορισμός πολλαπλών ψευδο-αναγνωριστικών είναι πρακτικός μόνο αν ο ιδιοκτήτης των δεδομένων γνωρίζει τον τρόπο με τον οποίο ο επιτιθέμενος θα εφαρμόσει την μέθοδο της αποκάλυψης. Τέτοια γνώση συνήθως δεν έχει ένας ιδιοκτήτης. Η λάθος απόφαση μπορεί να προκαλέσει υψηλή απώλεια πληροφοριών και κίνδυνο άρσης της ανωνυμίας. Στην παρουσία πολλαπλών ψευδο-αναγνωριστικών, πολλά από αυτά μπορεί να είναι περιττά τα οποία με την παρακάτω ιδιότητα μπορούμε να τα εντοπίσουμε με ευκολία:

Παρατήρηση: (Ιδιότητα του υποσυνόλου): Έστω QID' ⊆ QID. Εάν ένας πίνακας T είναι k-ανώνυμος για ένα σύνολο QID τότε θα παραμένει k-ανώνυμος και σε ένα υποσύνολο QID' του QID. Το QID' καλύπτεται δηλαδή από το QID και μπορεί να αφαιρεθεί από την διαδικασία της ανωνυμοποίησης.

## *(X-Y) ανωνυμία*

Το μοντέλο k-anonymity υποθέτει πως η κάθε εγγραφή εκπροσωπεί έναν και μόνο άνθρωπο. Αν παραπάνω από μία εγγραφές ενός πίνακα εκπροσωπούν έναν μοναδικό άνθρωπό, τότε μία ομάδα από k εγγραφές εκπροσωπούν λιγότερους από k ανθρώπους και υπάρχει πιθανότητα τα στοιχεία του να μείνουν απροστάτευτα [42]. Το παρακάτω παράδειγμα επεξηγεί το ζήτημα.

Έστω ότι ένας πίνακας Α{ ΕΓ, Εργασία, Φύλο, Ηλικία, Ασθένεια} εκπροσωπεί ένα σύνολο από εγγραφές. Η κάθε εγγραφή προσδιορίζεται από το γνώρισμα ΕΓ που περιλαμβάνει ευαίσθητα στοιχεία και συνοδεύεται από τα υπόλοιπα γνωρίσματα { Εργασία, Φύλο, Ηλικία, Ασθένεια }. Ένας ασθενής μπορεί να έχει πολλαπλές εγγραφές, μία για κάθε ασθένεια. Σε αυτήν την περίπτωση, η επιλογή QID = { Εργασία, Φύλο, Ηλικία } είναι λάθος γιατί αποτυγχάνει να εξασφαλίσει ότι κάθε κλάση ισοδυναμίας θα περιλαμβάνει τουλάχιστον k διακριτούς ασθενείς. Αν για παράδειγμα κάθε ασθενής είχε τουλάχιστον τρεις ασθένειες και κατ' επέκταση τουλάχιστον τρεις εγγραφές, τότε μια κλάση ισοδυναμίας από k εγγραφές δεν θα περιλαμβάνει πάνω από k/3 ασθενείς.

Για την αντιμετώπιση του προβλήματος αυτού οι καθηγητές Wang και Fung προτείνουν το μοντέλο **(X, Y)-ανωνυμία**, όπου X και Y είναι σύνολα από χαρακτηριστικά γνωρίσματα. Σε έναν πίνακα T θεωρούμε την Π(T) ως την προβολή του, την ε(T) ως την επιλογή τιμών του, το att(T) ως το σύνολο των χαρακτηριστικών γνωρισμάτων του και ως |T| το πλήθος των διακριτών εγγραφών του.

Ορισμός (X-Y)-anonymity: Έστω x μια τιμή της X. Η ανωνυμία της x με βάση την Y, συμβολίζεται ως $\alpha_Y$ (x) και είναι το πλήθος διακριτών τιμών της Y που εμφανίζονται από κοινού με την x, με άλλα λόγια το $|\Pi_y \varepsilon_x(T)|$. Αν το Y είναι ευαίσθητο γνώρισμα της T τότε το $\alpha_Y$ (x) γράφεται και ως α (x) και ισούται με τον αριθμό των εγγραφών που περιλαμβάνουν το x. Έστω $A_Y(X)$ = ελάχιστο{ $\alpha_Y$ (x) | x ∈ X }. Ένας πίνακας T ικανοποιεί την (X-Y)-ανωνυμία για έναν συγκεκριμένο ακέραιο k αν $A_Y$ (x) ≥ k.

Με το μοντέλο (X-Y)-ανωνυμία κάθε τιμή του X συνδέεται με τουλάχιστον k διακριτές τιμές του Y. Η k-ανωνυμία είναι η ειδική περίπτωση όπου το X είναι το QID και το Y είναι το ευαίσθητο γνώρισμα του πίνακα T το οποίο εντοπίζει μοναδικά του ιδιοκτήτες των εγγραφών. Επίσης αν κάθε τιμή της X περιγράφει ένα σύνολο από ψευδο-

αναγνωριστικά (για παράδειγμα X = { Εργασία, Φύλο, Ηλικία }) και το Y περιγράφει τα ευαίσθητα γνωρίσματα (για παράδειγμα Y= { Ασθένεια }) τότε κάθε κλάση ισοδυναμίας συνδέεται με διάφορα ευαίσθητα γνωρίσματα κάνοντας το δύσκολο να συναχθεί σε μία συγκεκριμένη ευαίσθητη τιμή. Τελικά με την (X-Y)-ανωνυμία από το παράδειγμα παραπάνω αν εφαρμόσουμε το μοντέλο k-ανωνυμία με X = { Εργασία, Φύλο, Ηλικία } και Y = { ΕΓ } τότε κλάση ισοδυναμίας της X θα συνδέεται με τουλάχιστον k διακριτούς ασθενείς.

## *Πολυσχεσιακή k-ανωνυμία*

Οι περισσότερες προεκτάσεις του μοντέλου της k-ανωνυμίας στοχεύουν στην ανωνυμοποίηση ενός μοναδικού πίνακα. Ωστόσο κάθε βάση δεδομένων συνήθως περιλαμβάνει πολλαπλούς σχεσιακούς πίνακες. Το 2007 ο καθηγητής του Manchester Nergiz πρότεινε ένα νέο μοντέλο ιδιωτικότητας, την **πολυσχεσιακή k-ανωνυμία (Multi-Relational k-anonymity)** [43] για να εξασφαλίσει την k-ανωνυμία σε πολλαπλούς πίνακες. Το μοντέλο αυτό θεωρεί πως μια σχεσιακή βάση δεδομένων περιλαμβάνει έναν πίνακα Π με προσωπικές εγγραφές και ορισμένα ευαίσθητα χαρακτηριστικά γνωρίσματα και μια σειρά από πίνακες Ti για $i = [1 \leq i \leq n]$ που περιλαμβάνουν ψευδο-αναγνωριστικά και ευαίσθητα γνωρίσματα. Η γενική ιδέα για την επίτευξη της ιδιωτικότητας είναι να εξασφαλιστεί πως για κάθε εγγραφή που οι τιμές της περιλαμβάνονται στους πίνακες Π, T1,....Tv υπάρχουν τουλάχιστον k-1 εγγραφές με τα ίδια ψευδο-αναγνωριστικά.

## *(c, t)-απομόνωση*

Το μοντέλο **(c, t)-απομόνωση ((c, t)-isolation)** [44] αναπτύχθηκε ώστε η πρόσβαση ενός αντιπάλου σε δημοσιευμένους πίνακες να μην του επιτρέπει στο σύνολο των εγγραφών να απομονώσει μοναδικές εγγραφές ανθρώπων. Το μοντέλο ιδιωτικότητας δημιουργεί στατιστικές βάσεις δεδομένων. Ορίζουμε ως p τα δεδομένα του θύματος, ως v το θύμα, ως q τις πληροφορίες που έχει ήδη ο αντίπαλος για το θύμα και ως $\delta_p$ την απόσταση μεταξύ p και q. Λέμε πως το σημείο q (c, t)-απομονώνει το σημείο p αν το $B(q, c\delta_p)$ περιέχει λιγότερα από t σημεία του πίνακα, όπου $B(q, c\delta_p)$ είναι κύκλος ακτίνας cδp και κέντρου q. Το μοντέλο αξιοποιεί τις αποστάσεις μεταξύ των τιμών της

κάθε εγγραφής. Για τον λόγο αυτό ταιριάζουν καλύτερα τα αριθμητικά γνωρίσματα σε στατιστικές βάσεις δεδομένων.

## *$k^m$-ανωνυμία*

Η **$k^m$-ανωνυμία ($k^m$-anonymity)** [45] είναι μια πιο χαλαρή εγγύηση ιδιωτικότητας που προτάθηκε για την ανωνυμοποίηση συνόλων από τιμές δεδομένων τα οποία δεν υπακούνε σε κάποιο σχεσιακό σχήμα αλλά είναι σύνολα από μία δεξαμενή τιμών, όπως για παράδειγμα τα καλάθια των αγορών του σουπερμάρκετ.

Σε αυτό το πρόβλημα κάθε εγγραφή αποτελείται από σύνολα δεδομένων που παίρνουν τιμές από ένα κοινό πεδίο τιμών. Ο επιτιθέμενος κατέχει μερική γνώση πάνω στα δεδομένα, γνωρίζοντας m τιμές μιας εγγραφής, και προσπαθεί να εντοπίσει τις υπόλοιπες τιμές της εγγραφής και να τις αντιστοιχίσει με ένα φυσικό πρόσωπο. Σε αντίθεση με τις προηγούμενες εγγυήσεις ιδιωτικότητας, δεν υπάρχει σαφής διαχωρισμός μεταξύ ευαίσθητων γνωρισμάτων και ψευδο-αναγνωριστικού. Σε κάθε περίπτωση ένα υποσύνολο των τιμών της εγγραφής σχηματίζει το σύνολο του ψευδο-αναγνωριστικού και οι υπόλοιπες τιμές σχηματίζουν το σύνολο των ευαίσθητων γνωρισμάτων. Η κάθε εγγραφή έχει διαφορετικό μέγεθος, σε αντίθεση με τις σχεσιακές βάσεις δεδομένων, όπου το μέγεθος της κάθε εγγραφής είναι σταθερό.

Η εγγύηση που προτείνεται απαιτεί να μην μπορεί ένας επιτιθέμενος που γνωρίζει μέχρι m τιμές μιας εγγραφής να την διαχωρίσει από τουλάχιστον $k - 1$ εγγραφές με τις ίδιες τιμές. Η ιδέα αυτή βασίζεται στο ότι δεν είναι πάντα εύκολο να διαχωρίσει ποιές τιμές μπορεί να είναι ευαίσθητες και ποιές όχι. Έτσι όλα τα γνωρίσματα μπορούν να θεωρηθούν ταυτόχρονα ευαίσθητα και ψευδο-αναγνωριστικά, ανάλογα με την γνώση του επιτιθέμενου. Επίσης, το όριο στην γνώση του επιτιθέμενου σημαίνει ότι αν στην πραγματικότητα γνωρίζει όλες ή σχεδόν όλες τις τιμές μιας εγγραφής, τότε δεν απομένει πληροφορία να προστατευθεί από αυτόν.

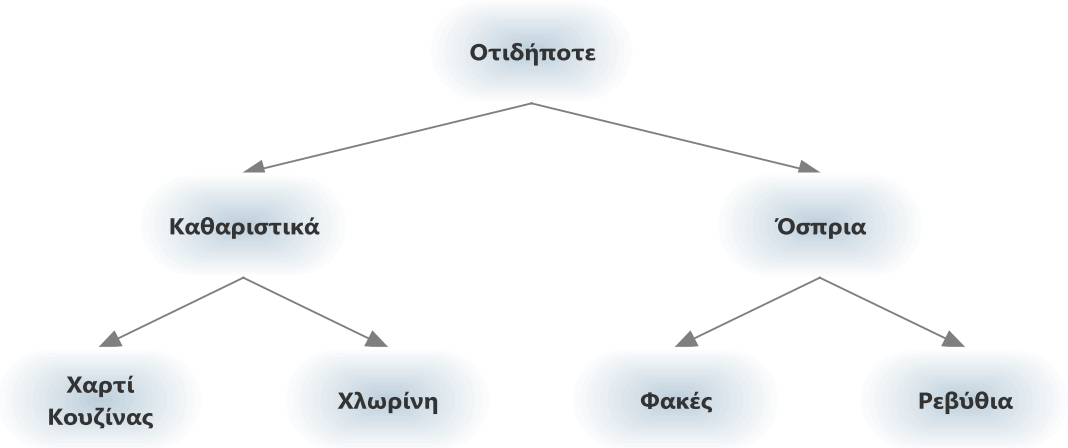
Ας υποθέσουμε, ότι το σούπερ-μάρκετ της γειτονιάς αποφασίζει να δώσει στην δημοσιότητα πληροφορίες σχετικές με το τι προϊόντα αγόρασαν οι καταναλωτές τον τελευταίο μήνα. Ένα παράδειγμα τέτοιας δημοσίευσης παρουσιάζεται στον πίνακα δεδομένων 5.10:

| Πελάτης | Αγορές |
|---|---|
| **Λουκάς** | { Χαρτί κουζίνας, Φακές, Ρεβύθια } |
| **Φανή** | { Χλωρίνη, Φακές } |
| **Δημήτρης** | { Χλωρίνη, Φακές, Ρεβύθια } |
| **Άννα** | { Χαρτί κουζίνας, Χλωρίνη, Ρεβύθια } |

Πίνακας 5.10: Αγορές πελατών Σούπερ-μάρκετ

Ας υποθέσουμε τώρα πως ένας φίλος του Λουκά, πήγε την ίδια μέρα με αυτόν, στο ίδιο Σούπερ-μάρκετ για να κάνει τα ψώνια του. Στο ταμείο συναντάει τον Λουκά και βλέπει ότι στο καλάθι του περιέχονται Χαρτί κουζίνας και Φακές. Όταν δημοσιεύονται οι πληροφορίες, ο φίλος του θα μπορεί χωρίς μεγάλη δυσκολία να αναγνωρίσει τον Λουκά στην δημοσιευμένη λίστα καθώς και να αποκαλύψει τα υπόλοιπα προϊόντα που αυτός αγόρασε. Το σύνολο δεδομένων επομένως δεν ικανοποιεί την $k^m$-ανωνυμία αφού για k=2 και m=2 ο συνδυασμός τιμών { Χαρτί κουζίνας, Φακές } εμφανίζεται μόνο μια φορά.

Για την διαδικασία της ανωνυμοποίησης επιλέγεται η τεχνική της ολικής γενίκευσης. Στο παράδειγμα των καθημερινών αγορών του πίνακα 5.10 η ιεραρχία είναι:




Σχήμα 5.10: Ιεραρχία γενίκευσης λίστας αγορών

Η εφαρμογή της γενίκευσης { Χαρτί κουζίνας, Χλωρίνη }→ { Καθαριστικά } στη βάση δεδομένων μπορεί να δώσει λύση στο πρόβλημα, αφού πλέον οποιοσδήποτε συνδυασμός m=2 τιμών στην βάση, εμφανίζεται σε τουλάχιστον k=2 εγγραφές.

| **Πελάτης** | **Αγορές** |
|---|---|
| **Λουκάς** | { Καθαριστικά, Φακές, Ρεβύθια } |
| **Φανή** | { Καθαριστικά, Φακές } |
| **Δημήτρης** | { Καθαριστικά, Φακές, Ρεβύθια } |
| **Άννα** | { Καθαριστικά, Ρεβύθια } |

Πίνακας 5.11: Ανωνυμοποιημένες αγορές πελατών Σούπερ-μάρκετ

### *5.3.2 Αποκάλυψη χαρακτηριστικού γνωρίσματος*

Τα μοντέλα k-ανωνυμία, (X,Y)-ανωνυμία αποτρέπουν την αποκάλυψη εγγραφής κρύβοντας την κάθε εγγραφή σε κλάσεις ισοδυναμίας με τα ίδια ψευδο-αναγνωριστικά. Παρόλα αυτά, αν οι περισσότερες εγγραφές σε μια κλάση ισοδυναμίας έχουν ίδιες ή και παρόμοιες τιμές στα ευαίσθητα χαρακτηριστικά τότε ο επιτιθέμενος μπορεί να συνδέσει έναν άνθρωπο με τις ευαίσθητες τιμές του χωρίς να απαιτείται να προσδιορίσει την ακριβή εγγραφή του. Αυτή ακριβώς η περίπτωση απεικονίζεται στον πίνακα 5.8 που είναι 3-ανώνυμος. Με την σύνδεση του πίνακα 5.8 και qid = < Καλλιτέχνης, Θηλυκό, [30-35) > ένας αντίπαλος μαθαίνει πως ο άνθρωπος με τις τιμές qid έχει 75% πιθανότητα πυρετό αφού τρεις στις τέσσερις εγγραφές αυτής της κλάσης προσδιορίζονται από την συγκεκριμένη ασθένεια. Επίσης, το μοντέλο (X,Y)-ανωνυμία υποθέτει ότι κάθε κλάση της X συνδέεται με τουλάχιστον k διακριτές από τις Y τιμές και δημιουργούνται συσχετισμοί εάν κάποιες από αυτές τις τιμές εμφανίζονται συχνότερα από κάποιες άλλες. Με τα παραπάνω ζητήματα ανοίγει το κεφάλαιο της επίθεσης με την μέθοδο της **αποκάλυψης χαρακτηριστικού γνωρίσματος**.

Στην επίθεση αυτή ο αντίπαλος ενδεχομένως να μην μπορεί να προσδιορίσει ακριβώς ποια είναι η εγγραφή του θύματος αλλά, βασιζόμενος στην κλάση ισοδυναμίας που

ανήκει θα μπορεί να τον αντιστοιχίσει με κάποια ευαίσθητη πληροφορία του δημοσιευμένου πίνακα. Τα μοντέλα που είδαμε παραπάνω δεν επαρκούν επομένως στις περιπτώσεις που οι ευαίσθητες πληροφορίες κυριαρχούν στις κλάσεις ισοδυναμίας.

## *l-ποικιλομορφία*

Για την αντιμετώπιση της αποκάλυψης χαρακτηριστικού γνωρίσματος στο πανεπιστήμιο Cornell [46] προτάθηκε από τους Machanavajjhala, Gehrke, Kifer και Venkitasubramaniam το μοντέλο της **l-ποικιλομορφίας (l-diversity)**. Η l-ποικιλομορφία επεξεργάζεται την βάση δεδομένων με τρόπο ώστε κάθε κλάση ισοδυναμίας να περιλαμβάνει τουλάχιστον $\ell$ ευαίσθητα γνωρίσματα που είναι μοναδικά. Επίσης ικανοποιεί τις προϋποθέσεις της k-ανωνυμίας όπου εδώ θέτουμε k = $\ell$, επειδή κάθε κλάση ισοδυναμίας περιλαμβάνει τουλάχιστον $\ell$ εγγραφές. Το μοντέλο αυτό δεν μπορεί να αντιμετωπίσει την κατηγορία των πιθανολογικών επιθέσεων επειδή κάποιες ευαίσθητες τιμές είναι από την φύση τους συχνότερες από άλλες και επιτρέπουν επομένως σε έναν επιτιθέμενο να κάνει υποθέσεις μεταξύ των τιμών αυτών και εγγραφών. Ο πυρετός για παράδειγμα είναι συχνότερος από τον καρκίνο.

Ορισμός 1: Ένας πίνακας ικανοποιεί την ***εντροπία l-ποικιλομορφίας*** αν σε κάθε κλάση ισοδυναμίας ισχύει:

$$-\sum_{s\in S} P(qid,s)\log(P(qid,s)) \geq \log(l)$$

όπου S είναι ευαίσθητο γνώρισμα, P(qid,s) είναι το ποσοστό του πλήθους των εγγραφών με την τιμή s ως προς το σύνολο των εγγραφών στην κλάση ισοδυναμία. Το αριστερό μέλος της συνάρτησης καλείται εντροπία του ευαίσθητου γνωρίσματος και έχει την ιδιότητα, όσο περισσότερο ισοκατανεμημένες να είναι οι ευαίσθητες τιμές στις κλάσεις ισοδυναμίας τόσο μεγαλύτερη τιμή να λαμβάνει. Ως εκ τούτου, μια μεγάλη τιμή κατωφλίου $\ell$ σε μία κλάση, συνεπάγεται μεγαλύτερη προστασία. Να σημειωθεί τέλος πως η βάση του log δεν επηρεάζει τα αποτελέσματα μας.

Παράδειγμα: Έστω ο πίνακας 5.8. Στην πρώτη κλάση ισοδυναμίας < Επαγγελματίας, Αρσενικό, [35-40) >, προκύπτει με βάση τον ορισμό 1:

$$\frac{-2}{3}\log(\frac{2}{3}) - \frac{1}{3}\log(\frac{1}{3}) = \log(1.9)$$

ενώ για την δεύτερη κλάση < Καλλιτέχνης, Θηλυκό, [30-35) > :

$$\frac{-3}{4}\log(\frac{3}{4}) - \frac{1}{4}\log(\frac{1}{4}) = \log(1.8)$$

Οπότε ο πίνακας ικανοποιεί την εντροπία l-diversity για $\ell \leq 1.8$ .

Για να μπορεί να εφαρμοστεί η εντροπία l-diversity θα πρέπει ο πίνακας να έχει μέγεθος τουλάχιστον log(l) αφού η εντροπία μιας κλάσης ισοδυναμίας είναι πάντα μεγαλύτερη ή ίση με την ελάχιστη εντροπία του υποσυνόλου των κλάσεων { $qid_1$,…,$qid_n$ } όπου κάθε κλάση ορίζεται ως qid = $qid_1$….$qid_n$, οπότε ισχύει: εντροπία(qid) ≥ min(εντροπία(qid1),…,εντροπία($qid_n$)).

Αυτή η υπόθεση είναι ιδιαίτερα δύσκολη να επιτευχθεί, κυρίως γιατί κάποιες ευαίσθητες τιμές εμφανίζονται πολύ συχνά στο ευαίσθητο χαρακτηριστικό γνώρισμα.

Η εντροπία l-diversity δεν έχει μετρικό σύστημα επικινδυνότητας της πιθανολογικής επίθεσης. Μας ενημερώνει για παράδειγμα ότι ο πίνακας 5.8 είναι 1.8-diverse αλλά όχι ότι ένας επιτιθέμενος έχει 75% πιθανότητα να συνδέσει την ευαίσθητη τιμή <Πυρετός>, με κάποια εγγραφή της κλάσης. Τέλος, πρέπει να σημειωθεί πως είναι δύσκολο να επιλεγούν τα κατάλληλα επίπεδα προστασίας ανάλογα με την συχνότητα και την ευαισθησία των τιμών.

Ορισμός 2: ***Η αναδρομική (c, l)-diverse***. Έστω c > 0 μία σταθερά. Το S ένα ευαίσθητο χαρακτηριστικό γνώρισμα. Οι $s_1$,….$s_m$ οι τιμές της S που απεικονίζονται στις κλάσεις. Τα $f_1$,…..,$f_m$ οι συχνότητες εμφάνισης των $s_1$,….$s_m$ αντίστοιχα. Τα $f_{(1)}$,….$f_{(m)}$ είναι οι αντίστοιχες συχνότητες σε φθίνουσα σειρά. Ένας πίνακας είναι αναδρομικός (c, l)-diverse όταν σε κάθε κλάση ισοδυναμίας ισχύει : $f(1) \leq c\sum_{i=l}^{m} f(i)$ για σταθερά c.

Το μοντέλο αναδρομικός (c, l)-diverse φροντίζει οι συχνότερες τιμές να εμφανίζονται σπανιότερα και οι λιγότερο συχνές τιμές να εμφανίζονται πιο συχνά. Μια κλάση ισοδυναμίας είναι αναδρομική (c, l)-diverse όταν η συχνότητα της ευαίσθητης τιμής που εμφανίζεται τις περισσότερες φορές είναι μικρότερη από το άθροισμα των υπολοίπων πολλαπλασιαζόμενο από μια σταθερά c που επιλέγει ο εκδότης, δηλαδή $f(1) \leq c\sum_{i=l}^{m} f(i)$ . Με αυτόν τον τρόπο ακόμα και αν ένας αντίπαλος με τις προηγούμενες του γνώσεις αποκλείσει κάποιες ευαίσθητες τιμές, πάλι τα δεδομένα που θα του παρέχονται δεν θα επαρκούν για να άρει την ανωνυμία. Ένας πίνακας είναι αναδρομικός (c, l)-diverse όταν ισχύει η παραπάνω ιδιότητα σε όλες τις κλάσεις

ισοδυναμίας. Να σημειωθεί τέλος πως η σταθερά c που είναι μια ανεξάρτητη παράμετρος επιτρέπει μεγαλύτερη ελευθερία από το μοντέλο της εντροπίας l-ποικιλομορφίας.

Ως επίλογο μπορούμε να πούμε πως η l-ποικιλομορφία θεωρεί έμμεσα ότι κάθε ευαίσθητο χαρακτηριστικό γνώρισμα ταξινομεί ομοιόμορφα τις τιμές του. Στην περίπτωση όμως που οι συχνότητες εμφάνισης των ευαίσθητων τιμών έχουν μεγάλες αποκλίσεις, η επίτευξη της θα προκαλέσει μεγάλη παραμόρφωση των δεδομένων. Για να δούμε πρακτικά τι σημαίνει αυτό, έστω ότι έχουμε έναν πίνακα δεδομένων που περιλαμβάνει για 1000 ασθενείς, ψευδο-αναγνωριστικά και ένα ευαίσθητο γνώρισμα= { Ασθένεια } με δύο πιθανές τιμές, Aids ή Γρίπη. Αν υπάρχουν μόνο πέντε ασθενείς με Aids τότε για να πετύχουμε 2-diversity θα πρέπει να υπάρχει σε κάθε κλάση ισοδυναμίας τουλάχιστον μια εγγραφή με αυτήν την τιμή. Επομένως μπορούν να διαμορφωθούν μονάχα πέντε κλάσεις ισοδυναμίας που όπως ήταν αναμενόμενο θα οδηγούσε σε υψηλή απώλεια πληροφοριών. Συνήθως στις επιθέσεις σύνδεσης περιλαμβάνονται δύο πηγές, ένας πίνακας T1 με αναγνωριστικά ταυτότητας και ψευδο-αναγνωριστικά καθώς και ένας πίνακας T2 με ευαίσθητα χαρακτηριστικά γνωρίσματα και ψευδο-αναγνωριστικά. Το μοντέλο k-anonymity ταιριάζει για την ανωνυμοποίηση του T1 και το l-diversity για τον T2. Υπό αυτή την έννοια, τα δύο μοντέλα δεν ανταγωνίζονται το ένα το άλλο, αλλά από κοινού χρησιμοποιούνται σε διαφορετικές περιπτώσεις.

## *Οριοθέτηση δύναμης*

Οι καθηγητές Wang, Fung και Yu προτείνουν μια εναλλακτική πρόταση του μοντέλου k-ανωνυμία, την **οριοθέτηση δύναμης (Confidence bounding)** [32]. Ο στόχος είναι να περιορίσουν τις δυνατότητες του αντιπάλου με τον προσδιορισμό ενός ή και παραπάνω *προτύπων ιδιωτικότητας* στην μορφή, $< QID \rightarrow s, h >$. Το s είναι η ευαίσθητη τιμή, τα QID είναι τα ψευδο-αναγνωριστικά και h είναι το κατώφλι. Έστω $Conf(QID \rightarrow s)$ να είναι το $\max\{conf(qid \rightarrow s)\}$ όλων των κλάσεων ισοδυναμίας και $conf(qid \rightarrow s)$ να είναι το ποσοστό των εγγραφών που περιέχουν την τιμή s στην κλάση ισοδυναμίας. Ένας πίνακας ικανοποιεί το $< QID \rightarrow s, h >$ αν ισχύει $Conf(QID \rightarrow s) \leq h$ σε όλες τις κλάσεις ισοδυναμίας. Με άλλα λόγια, το $< QID \rightarrow s, h >$ οριοθετεί την δυνατότητα του αντιπάλου να συνδέσει μία ευαίσθητη τιμή με έναν άνθρωπο το μέγιστο κατά h

τοις εκατό πιθανότητα.

Για παράδειγμα, με QID = { Εργασία, Φύλο, Ηλικία }, < QID → Aids, 10% > σημαίνει πως η πιθανότητα αποκάλυψης εγγραφών με την τιμή Aids δεν θα πρέπει να ξεπερνάει το 10%. Στα δεδομένα του πίνακα 5.8 αυτό το πρότυπο ιδιωτικότητας παραβιάζεται επειδή στην κλάση { Καλλιτέχνης, Θηλυκό, [30,35) } η πιθανότητα αποκάλυψης εγγραφής με την τιμή Aids είναι 75%.

Το μοντέλο οριοθέτηση δύναμης έχει δύο πλεονεκτήματα έναντι των μοντέλων της εντροπίας l-diversity και της αναδρομικής (c, l)-diverse. Πρώτον ο κάτοχος δεδομένων μπορεί να καθορίσει την δυνατότητα του αντιπάλου στην διεξαγωγή συμπερασμάτων γιατί είναι ένα μετρούμενο μέγεθος. Τέλος επιτρέπει στον κάτοχο δεδομένων να έχει την απαραίτητη ευελιξία που χρειάζεται ώστε να ορίζει σε κάθε κλάση ισοδυναμίας διαφορετικό κατώφλι h.

## *(X-Y)-Συνδεσιμότητα*

Στο μοντέλο (X,Y)-ανωνυμία είδαμε πως κάθε κλάση της X έχει τουλάχιστον k διακριτές τιμές στην Y (για παράδειγμα ασθένειες). Ωστόσο αν κάθε κλάση απεικονίζει k ανθρώπους, αυτό δεν σημαίνει πως η πιθανότητα αποκάλυψης καθενός από αυτούς θα είναι 1/k. Αν τιμές του Y παρουσιάζονται συχνότερα από άλλες τότε η πιθανότητα αποκάλυψης αυτών των τιμών είναι μεγαλύτερη του 1/k.

Ορισμός (X-Y)-Συνδεσιμότητα: Έστω x μια τιμή της X και y της Y. Η συνδεσιμότητα από το x στο y, ορίζεται ως $l_y(x)$ και είναι το ποσοστό των εγγραφών που περιλαμβάνουν τις τιμές x και y προς αυτές που περιλαμβάνουν μόνο την τιμή x (για παράδειγμα α(x, y)/α(x), όπου α(x, y) είναι το πλήθος των εγγραφών που περιλαμβάνουν τις τιμές x, y και α(x) μόνο την τιμή x.). Έστω $L_y(X) = \max\{l_y(\chi) \mid x \in X\}$ και $L_Y(X) = \max\{l_y(X) \mid y \in Y\}$. Λέμε πως ο πίνακας T ικανοποιεί την (X-Y)-συνδεσιμότητα [42] για $0 < k \leq 1$ αν $L_Y(X) \leq h$.

Οι καθηγητές Wang και Fung προτείνουν ένα γενικό μοντέλο ιδιωτικότητας που καλείται **(X-Y) ιδιωτικότητα** και το οποίο συνδυάζει τα μοντέλα (X,Y)-ανωνυμία και (X,Y)-συνδεσιμότητα ((X-Y)-linkability). Η γενική ιδέα είναι κάθε κλάση ισοδυναμίας x στο X να περιλαμβάνει τουλάχιστον k εγγραφές και η δυνατότητα αποκάλυψης οποιαδήποτε τιμής y ∈ Y από οποιαδήποτε τιμή x ∈ X να περιορίζεται σε ένα κατώφλι

h. Το μοντέλο τέλος μπορεί να εφαρμοστεί και σε πίνακες με πολλαπλές εκδόσεις.

## *(a, k)-ανωνυμία*

Ένα παρόμοιο μοντέλο με την (X-Y) ιδιωτικότητα είναι η **(a, k)-ανωνυμία ((a, k)-anonymity)** [47], η οποία απαιτεί κάθε κλάση ισοδυναμίας ενός πίνακα δεδομένων T να έχει τουλάχιστον k εγγραφές και conf(qid $\rightarrow$ s) $\leq$ a για κάθε ευαίσθητη τιμή s, όπου k και a είναι τα προκαθορισμένα από τον κάτοχο δεδομένων κατώφλια. Τα μοντέλα, (X-Y) ιδιωτικότητα και (a, k)-ανωνυμία προκαλούν μεγάλη παραμόρφωση των δεδομένων.

## *LKC-ιδιωτικότητα*

Στα περισσότερα μοντέλα που εξετάσαμε προηγουμένως η συνήθης τακτική είναι να γενικεύονται οι εγγραφές σε κλάσεις ισοδυναμίας μεγέθους τουλάχιστον k, με τέτοιο τρόπο ώστε σε κάθε μία κλάση οι τιμές των ψευδο-αναγνωριστικών να είναι ίδιες και να υπάρχει πλήθος από διαφορετικές ευαίσθητες τιμές. Όταν ένας πίνακας έχει πολύ μεγάλες διαστάσεις, νομοτελειακά το πλήθος των ψευδο-αναγνωριστικών θα είναι εξίσου μεγάλο. Για να εφαρμοστεί το μοντέλο της k-ανωνυμίας σε μεγάλο πίνακα θα πρέπει αναγκαστικά να διαγραφούν δεδομένα, γεγονός που υποβαθμίζει την χρησιμότητα και ποιότητα των πληροφοριών.

Για να αντιμετωπιστεί το παραπάνω πρόβλημα οι καθηγητές Mohammed, Fung και Debbabi προτείνουν το μοντέλο **LKC-ιδιωτικότητα (LKC-privacy)** [48] για την ανωνυμοποίηση υψηλών διαστάσεων πινάκων. Το μοντέλο εκμεταλλεύεται τις περιορισμένες γνώσεις του αντιπάλου. Σε πραγματικές συνθήκες, οι επιθέσεις ιδιωτικότητας είναι δύσκολο να εφαρμοστούν γιατί απαιτούν από τον επιτιθέμενο να αποκτήσει όλες τις τιμές των ψευδο-αναγνωριστικών. Είναι λογικό επομένως να θεωρήσουμε πως ένας αντίπαλος έχει στην κατοχή του το πολύ L τιμές των ψευδο-αναγνωριστικών του θύματος.

Η γενική ιδέα της LKC-privacy είναι να εξασφαλίσει πως όλοι οι συνδυασμοί τιμών στο $QID_j \subseteq QID$ του πίνακα δεδομένων T, με μέγιστο μήκος L να είναι ίδιοι με τουλάχιστον K εγγραφές και η δυνατότητα αποκάλυψης μια ευαίσθητης τιμής με

εγγραφή να είναι το πολύ ίση με C. Τα L, K, C είναι επομένως κατώφλια τα οποία καθορίζονται εξαρχής από τον κάτοχο δεδομένων. Η πιθανότητα στο μοντέλο LKC-privacy να πετύχει η αποκάλυψη εγγραφής είναι το πολύ 1/K ενώ η πιθανότητα αποκάλυψη χαρακτηριστικού γνωρίσματος το πολύ 1/C, με δεδομένο ότι οι προηγούμενες γνώσεις του αντιπάλου δεν ξεπερνούν τις L τιμές.

Ορισμός LKC-privacy: Έστω L ο μέγιστος αριθμός τιμών που κατέχει ο αντίπαλος. Έστω S να είναι το σύνολο των ευαίσθητων τιμών. Ένας πίνακας δεδομένων T ικανοποιεί την LKC-privacy αν και μόνο αν για οποιοδήποτε qid ισχύει $|qid| \leq L$,

- $|T[qid]| \geq K$, όπου $K > 0$ ένα ακέραιο κατώφλι ανωνυμίας και T[qid] το σύνολο των εγγραφών που περιλαμβάνουν τα qid μέσα στον πίνακα T και
- $conf(qid \rightarrow s) \leq C$ για οποιοδήποτε $s \in S$ , όπου $0 < C \leq 1$ ένας πραγματικός αριθμός που λειτουργεί ως κατώφλι περιορισμού των δυνατοτήτων του αντιπάλου.

Η LKC-ιδιωτικότητα χαρακτηρίζεται από τέσσερις σημαντικές ιδιότητες που την κάνουν κατάλληλη για την ανωνυμοποίηση μεγάλων πινάκων:

1. Απαιτεί, σε αντίθεση με το μοντέλο της k-ανωνυμίας, ένα μόνο υποσύνολο των ψευδο-αναγνωριστικών να είναι ίδιο με K εγγραφές. Η διαφορά αυτή με την οφείλεται στην παραδοχή του μοντέλου LKC-ιδιωτικότητα ότι η δύναμη του αντιπάλου είναι περιορισμένη.
2. Αποτελεί μια γενίκευση των μοντέλων που είδαμε παραπάνω:
   - Η k-ανωνυμία είναι ειδική περίπτωση της LKC-ιδιωτικότητας με: L = |QID|, K = k, και C = 100%, όπου |QID| είναι το πλήθος των ψευδο-αναγνωριστικών του πίνακα.
   - Το μοντέλο Οριοθέτηση δύναμης ομοίως με L = |QID|, K = 1.
   - Η (a, k)-ανωνυμίας ομοίως με L = |QID|, K = k, και C = a.
   - Η l-ποικιλομορφία ομοίως με L = |QID|, K = 1, και C = 1/ $\ell$.
3. Είναι ευέλικτη καθώς έχει παραμέτρους που ρυθμίζουν την σχέση μεταξύ του επιπέδου προστασίας της ιδιωτικότητας και της χρησιμότητας των δεδομένων. Με την αύξηση των L και K ή την μείωση της C μπορούμε να βελτιώσουμε την ιδιωτικότητα με κόστος την απώλεια δεδομένων.

4. Είναι ένα γενικό μοντέλο και ικανό να αποκρούσει τις συνδέσεις εγγραφής και χαρακτηριστικού γνωρίσματος. Μπορεί να ανωνυμοποιήσει δεδομένα με την ύπαρξη ή μη ευαίσθητων χαρακτηριστικών γνωρισμάτων.

## *(k, e)-ανωνυμία*

Η **(k, e)-ανωνυμία ((k, e)-anonymity)** είναι έργο των Zhang, Koudas, Srivastava και Yu [35]. Αφορά πίνακες δεδομένων στους οποίους τα πεδία των ευαίσθητων χαρακτηριστικών γνωρισμάτων είναι αριθμητικά. Οι μισθοί των εργαζόμενων μιας εταιρείας είναι ένα τέτοιο παράδειγμα. Η γενική ιδέα είναι να διαμελιστούν οι εγγραφές σε κλάσεις ώστε κάθε κλάση να περιέχει k διαφορετικές ευαίσθητες τιμές σε ένα εύρος τιμών τουλάχιστον e. Βέβαια η (k, e)-anonymity αγνοεί τον τρόπο με τον οποίο είναι διανεμημένες οι ευαίσθητες τιμές μέσα σε εύρος λ. Αν δηλαδή κάποιες ευαίσθητες τιμές εμφανίζονται συχνά σε αυτό το εύρος, τότε ο επιτιθέμενος μπορεί να βγάλει από μία κλάση ισοδυναμίας ασφαλή συμπεράσματα. Αυτός ο τύπος της αποκάλυψης χαρακτηριστικού γνωρίσματος καλείται *επίθεση εγγύτητας*. Ας φανταστούμε μια κλάση ισοδυναμίας με 10 εγγραφές, 7 ευαίσθητες ξεχωριστές τιμές, όπου οι 9 εγγραφές είναι μεταξύ του 30 και 35 και μια εγγραφή με τιμή 80. Ας δούμε τον πίνακα 5.12:

| (7, 50)-ανώνυμος πίνακας | | |
|---|---|---|
| Εργασία | Φύλο | Μισθός |
| Καλλιτέχνης | Θηλυκό | 30 |
| Καλλιτέχνης | Θηλυκό | 31 |
| Καλλιτέχνης | Θηλυκό | 30 |
| Καλλιτέχνης | Θηλυκό | 32 |
| Καλλιτέχνης | Θηλυκό | 35 |
| Καλλιτέχνης | Θηλυκό | 34 |
| Καλλιτέχνης | Θηλυκό | 33 |
| Καλλιτέχνης | Θηλυκό | 32 |
| Καλλιτέχνης | Θηλυκό | 35 |
| Καλλιτέχνης | Θηλυκό | 80 |

Πίνακας 5.12: Παράδειγμα (7, 50)-anonymous πίνακα

Τα γνωρίσματα { Εργασία, Φύλο } είναι QID και ο μισθός η ευαίσθητη πληροφορία. Ο πίνακας είναι (7, 50)-anonymous (η τιμή 50 προκύπτει από το εύρος των ακραίων τιμών 80-30). Ακόμα όμως ο αντίπαλος μπορεί να συμπεράνει με 90% πιθανότητα πως το θύμα έχει μισθό που ανήκει στην κλάση [30-35] αφού εκεί ανήκουν οι 9 από τις 10 εγγραφές.

## *(ε, m)-ανωνυμία*

Οι καθηγητές Li, Tao, Xiao προτείνουν ένα εναλλακτικό μοντέλο ιδιωτικότητας, την **(ε, m)-ανωνυμία ((ε, m)-anonymity)** [49]. Το μοντέλο αυτό είναι όμοιο με την (k, e)-ανωνυμία. Για οποιοδήποτε πίνακα δεδομένων T, με ευαίσθητες τιμές s, το μοντέλο περιορίζει την δύναμη του αντιπάλου να άρει την ανωνυμία σε ένα εύρος [s-ε, s+ε] το πολύ κατά 1/m.

## *t-εγγύτητα*

Η μέθοδος ανωνυμοποίησης της l-ποικιλομορφίας όπως είδαμε παραπάνω δεν αντιμετωπίζει αποδοτικά την επίθεση ανομοιόμορφων δεδομένων. Στην περίπτωση δηλαδή που σε μια κλάση ισοδυναμίας η απόσταση μεταξύ της κατανομής ενός ευαίσθητου γνωρίσματος και της κατανομής του γνωρίσματος αυτού στον συνολικό πίνακα είναι μεγάλη τότε η l-ποικιλομορφία δεν μπορεί να αποτρέψει τις συνδέσεις χαρακτηριστικού γνωρίσματος. Ας θεωρήσουμε σαν παράδειγμα έναν πίνακα ιατρικών δεδομένων όπου το 95% των εγγραφών έχει γρίπη, ενώ το υπόλοιπο 5% έχει καρκίνο. Μια κλάση ισοδυναμίας ικανοποιεί την 2-ποικιλομορφία αν υποθέσουμε πως το 50% των εγγραφών έχει γρίπη και το υπόλοιπο 50% καρκίνο. Σε αυτή την κλάση επομένως η πιθανότητα κάποιος να έχει καρκίνο είναι πολύ μεγαλύτερη από την αρχική της τάξεως του 5% και συνεπώς, ο επιτιθέμενος έχει αποκτήσει πληροφορίες που δεν είχε προηγουμένως. Αυτό συμβαίνει όταν η συνολική κατανομή είναι ανομοιόμορφη. Στις περιπτώσεις αυτές λέμε ότι έχουμε επίθεση ανομοιόμορφων δεδομένων (skewness attack). Για να αποτρέπεται η παραπάνω επίθεση προτάθηκε το μοντέλο της **t-εγγύτητας (t-closeness)** [21].

Ορισμός t-εγγύτητα: Μια κλάση ισοδυναμίας ικανοποιεί την t-εγγύτητα αν η απόσταση μεταξύ της κατανομής ενός ευαίσθητου γνωρίσματος στην κλάση αυτή και της κατανομής του γνωρίσματος αυτού στον συνολικό πίνακα δεν είναι μεγαλύτερη από ένα κατώφλι t. Ένας πίνακας δεδομένων ικανοποιεί την t-εγγύτητα αν όλες οι κλάσεις ισοδυναμίας του ικανοποιούν την παραπάνω παραδοχή.

Το μοντέλο χρησιμοποιεί για την μέτρηση της απόστασης μεταξύ δύο κατανομών στις ευαίσθητες τιμές την μετρική EMD (απόσταση μετακίνησης γης - Earth Mover Distance). Είναι απαραίτητη η χρήση της μετρικής στις τιμές των γνωρισμάτων έτσι ώστε να ορίζεται μεταξύ οποιουδήποτε ζεύγους τιμών η απόσταση της βάσης. Η EMD βασίζεται στο ελάχιστο ποσό έργου που χρειάζεται για να μετασχηματιστεί μια κατανομή σε μια άλλη μετακινώντας μάζα κατανομής μεταξύ αυτών.

- Για αριθμητικά γνωρίσματα χρησιμοποιείται η *διατεταγμένη απόσταση (ordered distance)*. Η διατεταγμένη απόσταση μεταξύ δύο τιμών υπολογίζεται με βάση το πλήθος των τιμών ανάμεσά τους.
- Για κατηγορικά δεδομένα υπάρχουν δύο μετρικές: της *ίσης απόστασης (equal distance)*, όπου η απόσταση μεταξύ οποιωνδήποτε κατηγορικών ορισμάτων ορίζεται ίση με ένα και της *ιεραρχικής απόστασης (hierarchical distance)* η οποία βασίζεται στο ελάχιστο επίπεδο στο οποίο μπορούν οι τιμές αυτές να γενικευθούν από κοινού σύμφωνα με την ιεραρχία.

## *Εξατομικευμένη Ιδιωτικότητα*

Οι καθηγητές Xiao και Tao προτείνουν την **εξατομικευμένη ιδιωτικότητα (Personalized Privacy)** [50] που επιτρέπει σε κάθε ιδιοκτήτη εγγραφής να ορίσει το δικό του επίπεδο προστασίας. Το μοντέλο υποθέτει πως κάθε ευαίσθητο γνώρισμα ανήκει σε ένα *ταξινομημένο δέντρο* και πως κάθε ιδιοκτήτης εγγραφής καθορίζει ο ίδιος τον *κόμβο προστασίας* εκείνου του δέντρου. Η ιδιωτικότητα αυτής της εγγραφής παραβιάζεται, εάν ο αντίπαλος καταφέρει να εντοπίσει την ευαίσθητη πληροφορία του υποδέντρου ενός κόμβου προστασίας με μία πιθανότητα μεγαλύτερη από ένα συγκεκριμένο κατώφλι. Η επίθεση αυτή καλείται *πιθανότητα παραβίασης (breach probability)*. Ας υποθέσουμε πως η Γρίπη και η Βρογχίτιδα είναι παιδιά της κορυφής *Λοίμωξη* σε ένα ταξινομημένο δέντρο. Αν η Αλίκη είναι φορέας της Βρογχίτιδας,

μπορεί να θέσει τον κόμβο προστασίας ίσο με Λοίμωξη, που σημαίνει πως επιτρέπει στους παραλήπτες των δεδομένων να ξέρουν ότι έχει μια λοίμωξη, αλλά όχι ποια συγκεκριμένα. Ο Βασίλης όμως που έχει Γρίπη μπορεί να μην τον απασχολεί αν μαθευτεί το ακριβές ιατρικό του ιστορικό και να μην θέσει κανέναν κόμβο προστασίας.

Τα μοντέλα *οριοθέτηση δύναμης* και *εξατομικευμένη ιδιωτικότητα* παρουσιάζουν ομοιότητες αφού και στα δύο περιορίζεται η δύναμη και η πιθανότητα για αντιστοίχιση ανθρώπων με ευαίσθητες τιμές. Παρόλα αυτά υπάρχουν διαφορές. Στο πρώτο μοντέλο ο κάτοχος δεδομένων εφαρμόζει μία καθολική επεξεργασία δεδομένων με τέτοιο τρόπο ώστε η κάθε εγγραφή να έχει το ίδιο επίπεδο προστασίας. Στην εξατομικευμένη ιδιωτικότητα ο εκάστοτε κόμβος προστασίας καθορίζεται από τον ιδιοκτήτη της εγγραφής. Πειράματα μάλιστα έδειξαν πως οι εξατομικευμένες επιλογές στην προστασία της ιδιωτικής ζωής μπορούν να βελτιώσουν σημαντικά την ποιότητα των δεδομένων του ανωνυμοποιημένου πίνακα σε σύγκριση με την καθολική εφαρμογή της ιδιωτικότητας. Στην πράξη όμως δεν μπορούμε να ξέρουμε εκ των προτέρων τι επιλογές θα κάνουν οι ιδιοκτήτες των εγγραφών στους κόμβους προστασίας. Μια λογική επιλογή κόμβου προστασίας εξαρτάται και από τον τρόπο με τον οποίο είναι μοιρασμένες οι ευαίσθητες τιμές στον πίνακα. Οι ιδιοκτήτες όμως συχνά δεν έχουν πρόσβαση στα δεδομένα του πίνακα, ούτε καν στην κλάση ισοδυναμίας που οι ίδιοι ανήκουν. Ως εκ τούτου θα έχουν την τάση να προτιμήσουν υψηλότερη προστασία, κάτι που επηρεάζει αρνητικά την χρησιμότητα των δεδομένων του πίνακα.

## *FF-ανωνυμία*

Τα μοντέλα που έχουμε ήδη αναλύσει θεωρούν δεδομένο πως τα χαρακτηριστικά γνωρίσματα ενός πίνακα μπορούν εξ' ολοκλήρου να κατηγοριοποιηθούν σε ψευδο-αναγνωριστικά και ευαίσθητα γνωρίσματα. Ωστόσο αυτή η υπόθεση δεν είναι ορθή όταν ένα γνώρισμα περιλαμβάνει ταυτόχρονα αναγνωριστικές και ευαίσθητες τιμές. Οι καθηγητές Wang, Xu, Fu και Wong μελέτησαν την παραπάνω περίπτωση καθώς και το πώς αυτά τα γνωρίσματα επηρεάζουν τα μοντέλα ιδιωτικότητας και την ανωνυμοποίηση δεδομένων. Τις επιθέσεις σε αυτά τα γνωρίσματα τα χαρακτήρισαν ως *ελευθέρας μορφής* και το μοντέλο αντιμετώπισης του ζητήματος **FF-ανωνυμία (FF-anonymity)** [51].

Ο παρακάτω πίνακας και τα συμπεράσματα που θα εξάγουμε από αυτόν θα μας κατατοπίσουν κατάλληλα για την ανάδειξη των επιθέσεων ελευθέρας μορφής:

| Ανεπεξέργαστος πίνακας δεδομένων T | | |
|---|---|---|
| Φύλο | Εισόδημα | Ασθένεια |
| Θηλυκό | 1900 | Γρίπη |
| Αρσενικό | 1900 | Aids |
| Θηλυκό | 1600 | Aids |
| Αρσενικό | 1600 | Aids |
| Αρσενικό | 1000 | E. Κολι |
| Αρσενικό | 800 | E. Κολι |
| Θηλυκό | 800 | Φυματίωση |
| Θηλυκό | 1000 | Φυματίωση |

Πίνακας 5.13: Ανεπεξέργαστος πίνακας δεδομένων T

Στον πίνακα T{ Φύλο, Εισόδημα, Ασθένεια } κάνουμε τις εξής διαπιστώσεις:

- **Ασθένεια:** Οι τιμές Γρίπη και Εσερίχια Κόλι είναι μη-ευαίσθητες και παρατηρήσιμες ενώ οι τιμές Aids και Φυματίωση είναι ευαίσθητες και μη-παρατηρήσιμες.
- **Εισόδημα:** Η παράθεση του ακριβούς εισοδήματος είναι ευαίσθητη πληροφορία και δεν μπορεί να παρατηρηθεί. Εάν το εισόδημα δηλωθεί κατηγορικά δηλαδή Υψηλό, Μεσαίο ή Χαμηλό τότε αφορά μη-ευαίσθητες και παρατηρήσιμες τιμές.
- **Φύλο:** Οι τιμές Θηλυκό και Αρσενικό είναι μη-ευαίσθητες και παρατηρήσιμες .

Επειδή τα γνωρίσματα { Ασθένεια, Εισόδημα } περιλαμβάνουν παρατηρήσιμες και ευαίσθητες τιμές προκύπτουν οι περιπτώσεις:

1. Ψευδο-αναγνωριστικό = { Φύλο } και Ευαίσθητο γνώρισμα = { Ασθένεια }
2. Ψευδο-αναγνωριστικό = { Φύλο } και Ευαίσθητο γνώρισμα = { Εισόδημα }

Παρατηρούμε πως και στις δυο περιπτώσεις ο πίνακας 5.13 ικανοποιεί την 2-ποικιλομορφία αφού καμία ευαίσθητη τιμή στις κλάσεις ισοδυναμίας δεν παρουσιάζεται σε ποσοστό μεγαλύτερο του 50%. Ωστόσο από τις διαπιστώσεις που κάναμε παραπάνω μπορεί να αρθεί η ανωνυμία. Καταρχάς με τις τιμές Αρσενικό και Υψηλό που είναι παρατηρήσιμες ο αντίπαλος συνδέει έναν άνθρωπο με δύο εγγραφές: ( Αρσενικό, 1900, Aids ) και ( Αρσενικό, 1600, Aids ) και συμπεραίνει με 100% πιθανότητα πως έχει Aids. Συνοπτικά όπως είδαμε και σε προηγούμενα μοντέλα η επίθεση διατυπώνεται ως { Αρσενικό, Υψηλό } → Aids. Μια άλλη περίπτωση είναι να χρησιμοποιήσει ο αντίπαλος τις παρατηρήσιμες τιμές Θηλυκό και Γρίπη και να συμπεράνει με 100% πιθανότητα πως ο άνθρωπος έχει εισόδημα 1900 ευρώ δηλαδή { Θηλυκό, Γρίπη } → 1900 ευρώ.

Το μοντέλο ιδιωτικότητας FF-ανωνυμία περιορίζει, κάτω από ένα καθορισμένο κατώφλι, την πιθανότητα παραβίασης της ιδιωτικότητας της μορφής X → s, όπου X παρατηρήσιμες τιμές και s η ευαίσθητη πληροφορία. Η κύρια ιδέα είναι ότι η κατηγοριοποίηση σε ψευδο-αναγνωριστικά και ευαίσθητα δεν γίνεται στο επίπεδο χαρακτηριστικού γνωρίσματος αλλά στο επίπεδο των τιμών του.

## *m-αμεταβλητότητα*

Ένα συχνό φαινόμενο στις πραγματικές δημοσιεύσεις συλλογών προσωπικών δεδομένων είναι η ανανέωση των περιεχομένων τους. Τα μοντέλα που έχουμε μελετήσει ως τώρα είναι εφαρμόσιμα στην περίπτωση της απλής και μοναδικής δημοσίευσης ενός πίνακα δεδομένων. Υπάρχει πιθανότητα ο επιτιθέμενος να συσχετίσει τις δύο ή και παραπάνω εκδόσεις και να αποκομίσει πληροφορία σχετική με κάποια εγγραφή παραβιάζοντας έτσι το απόρρητο των προσωπικών δεδομένων του συγκεκριμένου ατόμου. Με την πραγματικότητα αυτή ασχολήθηκαν οι Xiao και Tao και ανέπτυξαν το μοντέλο **m-αμεταβλητότητα (m-invariance)** [52] η οποία είναι μια επέκταση της l-ποικιλομορφίας.

Ορισμός m-αμεταβλητότητα: Μια ακολουθία δημοσιευμένων πινάκων T*(1),….,T*(n) ικανοποιεί την m-αμεταβλητότητα αν ισχύουν οι συνθήκες:

i. Ο πίνακας είναι m-μοναδικός για όλα τα j∈[ 1, n ] και

ii. Για κάθε εγγραφή t ∈ U(n) με διάρκεια ζωής [x, y] , οι ομάδες t.QI∗(x), t.QI∗(x + 1), ..., t.QI∗(y) στις οποίες εμφανίζεται η εγγραφή t στην αντίστοιχη έκδοση έχουν την ίδια υπογραφή.

Πιο συγκεκριμένα,

- ένας πίνακας είναι m-μοναδικός αν κάθε κλάση ισοδυναμίας σε αυτόν περιέχει το λιγότερο m εγγραφές και όλες οι εγγραφές στην κλάση έχουν διαφορετικές τιμές στα ευαίσθητα γνωρίσματα,

- το U(n) αναφέρεται στην ιστορική ένωση. Η U(n) για n≥1 περιέχει όλες τις πλειάδες που βρίσκονται στον πίνακα T πριν από κάθε δημοσίευση,

- ο όρος διάρκεια ζωής αναφέρεται σε κάθε εγγραφή t ∈ U(n) και ορίζεται ως το διάστημα [x,y], με x τον μικρότερο ακέραιο j και y τον αντίστοιχα μεγαλύτερο α-κέραιο j για τον οποίο η εγγραφή να εμφανίζεται στον πίνακα $T_j$ ,

- ως υπογραφή της QI* ορίζεται το σύνολο των διακριτών ευαίσθητων τιμών για κά-ποια δημοσίευση j που εμφανίζονται σε αυτήν.

Ας δούμε ένα παράδειγμα για να κατανοήσουμε καλύτερα το μοντέλο. Ας θεωρήσουμε πως έχουμε έναν δημοσιευμένο πίνακα ενός νοσοκομείου που ανανεώνει τα δεδομένα των ασθενών του τόσο με την προσθήκη νέων εγγραφών όσο και με την διαγραφή κά-ποιων παλαιότερων κάθε εξάμηνο. Στον πίνακα 5.14 παρουσιάζεται ο T(1) που είναι ο αρχικός πίνακας ο οποίος χρησιμοποιείται σαν βάση για να ανωνυμοποιηθεί και να δημοσιευθεί ο T*(1). Στον πίνακα 5.15 δημοσιεύεται ο πίνακας T*(2) με βάση τα δε-δομένα του τελικού πίνακα T(2), που διαμόρφωσε το νοσοκομείο στην διάρκεια έξι μηνών.

| Αρχικός πίνακας T(1) | | | |
|---|---|---|---|
| Όνομα | Ηλικία | Ταχυδρομικός Κώδικας | Ασθένεια |
| Αλίκη | 21 | 12000 | Δυσπεψία |
| Βασίλης | 22 | 14000 | Βρογχίτιδα |
| Γεωργία | 24 | 18000 | Γρίπη |
| Δημήτρης | 23 | 25000 | Γαστρίτιδα |
| Ελευθερία | 41 | 20000 | Γρίπη |
| Ζήσης | 36 | 27000 | Γαστρίτιδα |
| Ήβη | 37 | 33000 | Δυσπεψία |
| Θαλής | 40 | 35000 | Γρίπη |
| Ιωάννα | 43 | 26000 | Γαστρίτιδα |
| Κυριακή | 52 | 33000 | Δυσπεψία |
| Λεωνίδας | 56 | 34000 | Γαστρίτιδα |

| Γενικευμένος πίνακας T*(1) | | | |
|---|---|---|---|
| Όνομα | Ηλικία | Ταχυδρομικός Κώδικας | Ασθένεια |
| 1 | [21,22] | [12K-14K] | Δυσπεψία |
| 1 | [21,22] | [12K-14K] | Βρογχίτιδα |
| 2 | [23,24] | [18K-25K] | Γρίπη |
| 2 | [23,24] | [18K-25K] | Γαστρίτιδα |
| 3 | [36,41] | [20K-27K] | Γρίπη |
| 3 | [36,41] | [20K-27K] | Γαστρίτιδα |
| 4 | [37,43] | [26K-35K] | Δυσπεψία |
| 4 | [37,43] | [26K-35K] | Γρίπη |
| 4 | [37,43] | [26K-35K] | Γαστρίτιδα |
| 5 | [52,56] | [33K-34K] | Δυσπεψία |
| 5 | [52,56] | [33K-34K] | Γαστρίτιδα |

Πίνακας 5.14: Αρχικός T(1) και γενικευμένος T*(1) πίνακας

| Αρχικός πίνακας T(2) | | | |
|---|---|---|---|
| Όνομα | Ηλικία | Ταχυδρομικός Κώδικας | Ασθένεια |
| Αλίκη | 21 | 12000 | Δυσπεψία |
| Δημήτρης | 23 | 25000 | Γαστρίτιδα |
| Κατερινα | 24 | 21000 | Γρίπη |
| Ήβη | 37 | 33000 | Δυσπεψία |
| Ιωάννα | 43 | 26000 | Γαστρίτιδα |
| Ελευθερία | 41 | 20000 | Γρίπη |
| Ήρα | 46 | 30000 | Γαστρίτιδα |
| Θανάσης | 54 | 31000 | Δυσπεψία |
| Λεωνίδας | 56 | 34000 | Γαστρίτιδα |
| Νικολέτα | 60 | 44000 | Γαστρίτιδα |
| Αλέξιος | 65 | 36000 | Γρίπη |

| Γενικευμένος πίνακας T*(2) | | | |
|---|---|---|---|
| Κλάση | Ηλικία | Ταχυδρομικόq Κώδικας | Ασθένεια |
| 1 | [21,23] | [12K-25K] | Δυσπεψία |
| 1 | [21,23] | [12K-25K] | Γαστρίτιδα |
| 2 | [25,43] | [21K-33K] | Γρίπη |
| 2 | [25,43] | [21K-33K] | Δυσπεψία |
| 2 | [25,43] | [21K-33K] | Γαστρίτιδα |
| 3 | [41,46] | [20K-30K] | Γρίπη |
| 3 | [41,46] | [20K-30K] | Γαστρίτιδα |
| 4 | [54,56] | [31K-34K] | Δυσπεψία |
| 4 | [54,56] | [31K-34K] | Γαστρίτιδα |
| 5 | [60,65] | [36K-44K] | Γαστρίτιδα |
| 5 | [60,65] | [36K-44K] | Γρίπη |

Πίνακας 5.15: Αρχικός T(2) και γενικευμένος T*(2) πίνακας

Οι εγγραφές Βασίλης, Γεωργία, Ζήσης, Θαλής και Κυριακή έχουν διαγραφεί από τη βάση δεδομένων και έχουν προστεθεί αντίστοιχα οι ασθενείς Κατερίνα, Ήρα, Θανάσης, Νικολέτα και Αλέξιος. Παρόλο που και οι δύο δημοσιευμένοι πίνακες ικανοποιούν το 2-anonymity και το 2-diversity, ο επιτιθέμενος μπορεί να προσδιορίσει μοναδικά την ταυτότητα ενός ασθενή, με τη σύνδεση των T*(1) και T*(2). Για παράδειγμα έστω ότι,

ο αντίπαλος γνωρίζει την ηλικία και τον ταχυδρομικό κώδικα της Αλίκης και ταυτόχρονα ξέρει ότι τα στοιχεία της είναι δημοσιευμένα και στους δύο πίνακες επειδή η θεραπεία της κράτησε παραπάνω από έξι μήνες. Από τον πίνακα T*(1) ο επιτιθέμενος είναι βέβαιος ότι η Αλίκη πάσχει είτε από δυσπεψία είτε από βρογχίτιδα. Με βάση τον πίνακα T*(2) ο αντίπαλος μπορεί να βρει ότι η Αλίκη πάσχει είτε από δυσπεψία είτε από γαστρίτιδα. Με τη σύνδεση των δυο αυτών γνώσεων ο επιτιθέμενος είναι σίγουρος πλέον ότι η Αλίκη πάσχει από δυσπεψία.

| Γενικευμένος πίνακας T(3) με πλαστές εγγραφές | | | | |
|---|---|---|---|---|
| Όνομα | Κλάση | Ηλικία | Ταχυδρομικός Κώδικας | Ασθένεια |
| Αλίκη | 1 | [21,22] | [12K-14K] | Δυσπεψία |
| c1 | 1 | [21,22] | [12K-14K] | Βρογχίτιδα |
| Δημήτρης | 2 | [23,25] | [21K-25K] | Γαστρίτιδα |
| Κατερίνα | 2 | [23,25] | [21K-25K] | Γρίπη |
| Ήβη | 3 | [37,43] | [26K-33K] | Δυσπεψία |
| c2 | 3 | [37,43] | [26K-33K] | Γρίπη |
| Ιωάννα | 3 | [37,43] | [26K-33K] | Γαστρίτιδα |
| Ελευθερία | 4 | [41,46] | [20K-30K] | Γρίπη |
| Ήρα | 4 | [41,46] | [20K-30K] | Γαστρίτιδα |
| Θανάσης | 5 | [54,56] | [31K-34K] | Δυσπεψία |
| Λεωνίδας | 5 | [54,56] | [31K-34K] | Γαστρίτιδα |
| Νικολέτα | 6 | [60,65] | [36K-44K] | Γαστρίτιδα |
| Αλέξιος | 6 | [60,65] | [36K-44K] | Γρίπη |

Πίνακας 5.16: Γενικευμένος πίνακας T(3) με πλαστές εγγραφές

Σύμφωνα με το μοντέλο της m-αμεταβλητότητα, ο πίνακας T*(2) αντικαθίσταται από τον πίνακα T(3), όπως φαίνεται παραπάνω. Συγκεκριμένα ο πίνακας T(3) περιλαμβάνει τις γενικευμένες τιμές του πίνακα T(2) μαζί με δύο πλαστές εγγραφές c1 και c2. Οι 13 εγγραφές κατανέμονται σε 6 κλάσεις ισοδυναμίας.

Από την πλευρά του επιτιθέμενου, μια πλαστή εγγραφή δεν ξεχωρίζει από τις υπόλοιπες εγγραφές στην ίδια κλάση ισοδυναμίας. Για παράδειγμα, οι κλάσεις ισοδυναμίας στους πίνακες T*(1) και T(3) έχουν πλέον το ίδιο σύνολο ευαίσθητων εγγραφών {δυσπεψία, βρογχίτιδα}, οπότε ο επιτιθέμενος δεν μπορεί να προσδιορίσει με βεβαιότητα πάνω από 50% την ασθένεια της Αλίκης. Οι δύο δημοσιεύσεις των πινάκων T*(1) και T(3), έχουν μια πάρα πολύ σημαντική ιδιότητα και σε αυτήν στηρίζεται και η μέθοδος της μ-αμεταβλητότητας. Εάν μια εγγραφή εμφανίζεται και στις δύο δημοσιεύσεις, θα γενικευτεί σε δύο κλάσεις ισοδυναμίας με τα ίδια ευαίσθητα χαρακτηριστικά και για τις δύο δημοσιεύσεις. Για παράδειγμα η εγγραφή { Ήβη, 37, 33K, δυσπεψία }, εμφανίζεται και στους δύο πίνακες T(1) και T(2). Μετά από τη γενίκευση των πινάκων, η εγγραφή ανήκει στις κλάσεις ισοδυναμίας 4 και 3 στους πίνακες T*(1) και T(3), αντίστοιχα. Οι δύο αυτές κλάσεις ισοδυναμίας έχουν το ίδιο σύνολο ευαίσθητων γνωρισμάτων {δυσπεψία, γρίπη, γαστρίτιδα}. Αυτό οφείλεται στην προσθήκη της πλαστής εγγραφής c2 στον πίνακα T(3). Για την ακρίβεια το m-αμεταβλητότητα απαιτεί την ικανοποίηση του m-ποικιλομορφία και ταυτόχρονα μια εγγραφή να ανήκει πάντα σε μια κλάση ισοδυναμίας, η οποία έχει το ίδιο σύνολο ευαίσθητων ιδιοτήτων, για όλες τις δημοσιεύσεις.

### *5.3.3 Αποκάλυψη παρουσίας στον πίνακα*

Στις συνδέσεις εγγραφής και χαρακτηριστικού γνωρίσματος υποθέτουμε πως ο αντίπαλος γνωρίζει πως η εγγραφή του θύματος υπάρχει στον δημοσιευμένο πίνακα. Σε ορισμένες περιπτώσεις βέβαια η διαπίστωση της παρουσίας (ή της απουσίας) μιας εγγραφής στον πίνακα αποκαλύπτει ευαίσθητες πληροφορίες του θύματος. Αν υποθέσουμε πως ένα νοσοκομείο δημοσιεύει έναν πίνακα για μία ασθένεια, ο αντίπαλος δεν θα πρέπει να είναι βέβαιος πως το θύμα περιλαμβάνεται σε αυτό τον πίνακα, γιατί τότε θα ξέρει με βεβαιότητα ότι πάσχει από την συγκεκριμένη ασθένεια. Αυτή η μέθοδος επίθεσης καλείται **αποκάλυψη παρουσίας πίνακα**.

## *δ-παρουσία*

Ας υποθέσουμε πως ο κάτοχος δεδομένων δημοσιεύει τον πίνακα 5.8 που όπως έχουμε δει είναι 3-ανώνυμος. Θεωρούμε πως ο επιτιθέμενος έχει πρόσβαση στον πίνακα 5.9

και πως η Αλίκη είναι το υποψήφιο θύμα. Ο πίνακας 5.8 είναι υποσύνολο του 5.9. Η πιθανότητα η Αλίκη να υπάρχει στον πίνακα 5.8 είναι (4/5) = 0.8 αφού στην κλάση ισοδυναμίας < Καλλιτέχνης, Θηλυκό, [30, 35) > ο πίνακας 5.8 έχει 5 εγγραφές ενώ ο 5.9 έχει 4. Ομοίως η πιθανότητα για τον Βασίλη είναι (3/4) = 0.75.

Το μοντέλο **δ-παρουσία (δ-presence)** [53] εξετάζει δηλαδή κατά πόσο τα δεδομένα θεωρούνται επαρκώς ανώνυμα μέσω της ανάλυσης των ανωνυμοποιημένων πινάκων για τον κίνδυνο επιβεβαίωσης της συμμετοχής ή όχι ενός φυσικού προσώπου. Στην συνέχεια θεωρούμε πως η δ-παρουσία ικανοποιείται αν η πιθανότητα να ανήκει οποιαδήποτε εγγραφή ενός πίνακα και σε άλλον είναι μεταξύ δύο τιμών δ=( $\delta_{min}$, $\delta_{max}$).

Ορισμός δ-παρουσία: Αν δίνεται ένας εξωτερικός δημόσιος πίνακας E και ένας ιδιωτικός T, όπου $T \subseteq E$, ο γενικευμένος πίνακας T΄ ικανοποιεί την ($\delta_{min}$, $\delta_{max}$)-παρουσία εάν $\delta_{min} \leq P(t \in T|T') \leq \delta_{max}$ για όλα τα $t \in E$.

Η δ-παρουσία μπορεί εμμέσως να αποτρέψει τις επιθέσεις αποκάλυψης εγγραφής και χαρακτηριστικού γνωρίσματος επειδή ο αντίπαλος έχει το πολύ δ% πιθανότητα να συνδέσει ένα πρόσωπο με έναν πίνακα και κατ' επέκταση η πιθανότητα να τον συνδέσουν με την εγγραφή του ή ένα ευαίσθητο γνώρισμα του επίσης δεν ξεπερνάει σε πιθανότητα το δ%. Τέλος να σημειώσουμε πως το μοντέλο είναι ιδιαίτερα αξιόπιστο, θεωρεί όμως δεδομένο πως ο κάτοχος και ο αντίπαλος έχουν πρόσβαση στον ίδιο εξωτερικό πίνακα E, κάτι που πρακτικά δεν συμβαίνει πάντα.

## *ε-Διαφορική Ιδιωτικότητα*

Για λογαριασμό της εταιρίας Microsoft η ερευνήτρια Cynthia Dwork ανέπτυξε το μοντέλο **ε-Διαφορική Ιδιωτικότητα (ε-Differential Privacy)** [54]. Βασίζεται στην ιδέα ότι, οποιαδήποτε προσθήκη προσωπικής εγγραφής σε μία στατιστική βάση δεδομένων κι αν γίνει, δεν θα πρέπει να αυξάνει σημαντικά το ρίσκο της παραβίασης της ανωνυμίας. Αντί να συγκρίνει τις προηγούμενες και τις μετέπειτα γνώσεις του αντιπάλου από την δημοσίευση των πινάκων, η Dwork προτείνει να συγκρίνουμε τον κίνδυνο να παραβιαστεί η ιδιωτικότητα πριν και μετά την είσοδο της κάθε εγγραφής στην βάση. Κατά συνέπεια, το μοντέλο εξασφαλίζει πως η αφαίρεση ή η πρόσθεση μιας εγγραφής δεν επηρεάζει σημαντικά το επίπεδο προστασίας του πίνακα.

Ορισμός ε-Διαφορικής ιδιωτικότητα: Μια συνάρτηση F τυχαιότητας εξασφαλίζει την

ε-διαφορική ιδιωτικότητα για όλες τις ομάδες δεδομένων T1 και T2, με διαφορά μίας το πολύ εγγραφής , $\frac{|\ln(P[F(t1) = S]}{|P[F(T2) = S]|} \leq \varepsilon$ για όλα τα S ∈ Εύρος(F), όπου Εύρος(F) είναι το σύνολο όλων των πιθανών εξόδων της συνάρτησης F.

Τέλος το μοντέλο δεν είναι ικανό να αποτρέψει συνδέσεις εγγραφής και χαρακτηριστικού γνωρίσματος αλλά εξασφαλίζει στους κατόχους των εγγραφών πως καμία δική τους προσωπική πληροφορία δεν θα συμβάλει στην αποκάλυψη στοιχείων και ευαίσθητων πληροφοριών του πίνακα.

### *(d, γ)-ιδιωτικότητα*

Οι Rastogi, Suciu και Hong παρουσίασαν το μοντέλο **(d, γ)-ιδιωτικότητα (d, γ)-privacy** [55]. Ορίζεται η P(r) ως η προηγούμενη πιθανότητα να εξακριβωθεί εάν η εγγραφή του θύματος υπάρχει στον πίνακα ή όχι και η P(r|T) ως η επόμενη πιθανότητα μετά την εξέταση του δημοσιευμένου πίνακα T. Το μοντέλο περιορίζει τις δύο αυτές πιθανότητες και παρέχει σε αντίθεση με τα περισσότερα μοντέλα εγγυήσεις ιδιωτικότητας και ποιότητα πληροφοριών. Αποδείχθηκε μάλιστα πως η καλύτερη σχέση μεταξύ ιδιωτικότητας και ποιότητας πληροφοριών επιτυγχάνεται μόνο όταν η P(r) είναι μικρή. Μια επίθεση καλείται d-ανεξάρτητη όταν η προηγούμενη πιθανότητα P(r) ικανοποιεί τις συνθήκες P(r) = 1 ή P(r) ≤ d για όλες τις εγγραφές r, όπου P(r) = 1 σημαίνει ότι ο αντίπαλος γνωρίζει πως το θύμα έχει εγγραφή στον πίνακα και επομένως η r του δεν χρειάζεται προστασία. Η (d, γ)-ιδιωτικότητα μπορεί να προστατεύσει τους πίνακες μόνο από τις d-ανεξάρτητες επιθέσεις.

## *5.4 Μοντελοποίηση πληροφοριών αντιπάλου*

Παρά τις διάφορες τεχνικές που έχουν αναπτυχθεί, και τις διάφορες εγγυήσεις που προτείνονται, ελλοχεύουν για την ιδιωτικότητα πολλοί κίνδυνοι για την ιδιωτικότητα χωρίς αποτελεσματική αντιμετώπιση. Η διαθέσιμη πληροφορία που μπορεί να κατέχει ο επιτιθέμενος μπορεί να έχει πολλές μορφές. Παράλληλα, τα μοντέλα των δημοσιευμένων δεδομένων μπορεί να διαφέρουν κάθε φορά, με αποτέλεσμα η κάθε περίπτωση να απαιτεί διαφορετική επεξεργασία προκειμένου να εξασφαλίζεται η ιδιωτικότητα των βά-

σεων δεδομένων. Συγκεκριμένα ένας μεγάλος αριθμός μοντέλων υποθέτει πως οι γνώσεις του αντιπάλου περιορίζονται στα ψευδο-αναγνωριστικά. Όλες οι μελέτες έχουν δείξει πως είναι ιδιαίτερα σημαντικό για τον κάτοχο των δεδομένων να γνωρίζει τι γνώσεις έχει ο αντίπαλος ώστε να επεξεργαστεί κατάλληλα την βάση δεδομένων. Οι γνώσεις των αντιπάλων προέρχονται από την κοινή λογική από εκλογικούς καταλόγους, δημογραφικές βάσεις, κοινωνικά δίκτυα και άλλες προσωπικές πληροφορίες.

Αναπτύχθηκε το μοντέλο skyline ιδιωτικότητα [56] το οποίο υποστηρίζει ότι, αφού είναι ανέφικτο για έναν εκδότη δεδομένων να προβλέψει το ακριβές γνωστικό υπόβαθρο που κατέχει ο αντίπαλος, θα πρέπει να μελετήσει τις πληροφορίες που είναι φυσικό να έχει και που μπορεί να διαχειριστεί. Ειδικότερα ορίζει τρεις τύπους γνωστικού υπόβαθρου, γνωστούς και ως «γνώσεις τριών-διαστάσεων». Πρόκειται για τις:

- γνώσεις για το θύμα
- γνώσεις για άλλους κατόχους εγγραφών του πίνακα
- γνώσεις για άλλους κατόχους εγγραφών στην κλάση με το ίδιο ευαίσθητο στοιχείο του θύματος

Το γνωστικό υπόβαθρο εκφράζεται ποσοτικά από τα γράμματα l, k, m που υποδεικνύουν πως ο αντίπαλος ξέρει:

i. l ευαίσθητες τιμές που το θύμα t δεν έχει,

ii. τις ευαίσθητες τιμές από k ανθρώπους και

iii. m ανθρώπους που έχουν την ίδια ευαίσθητη τιμή όπως ο t

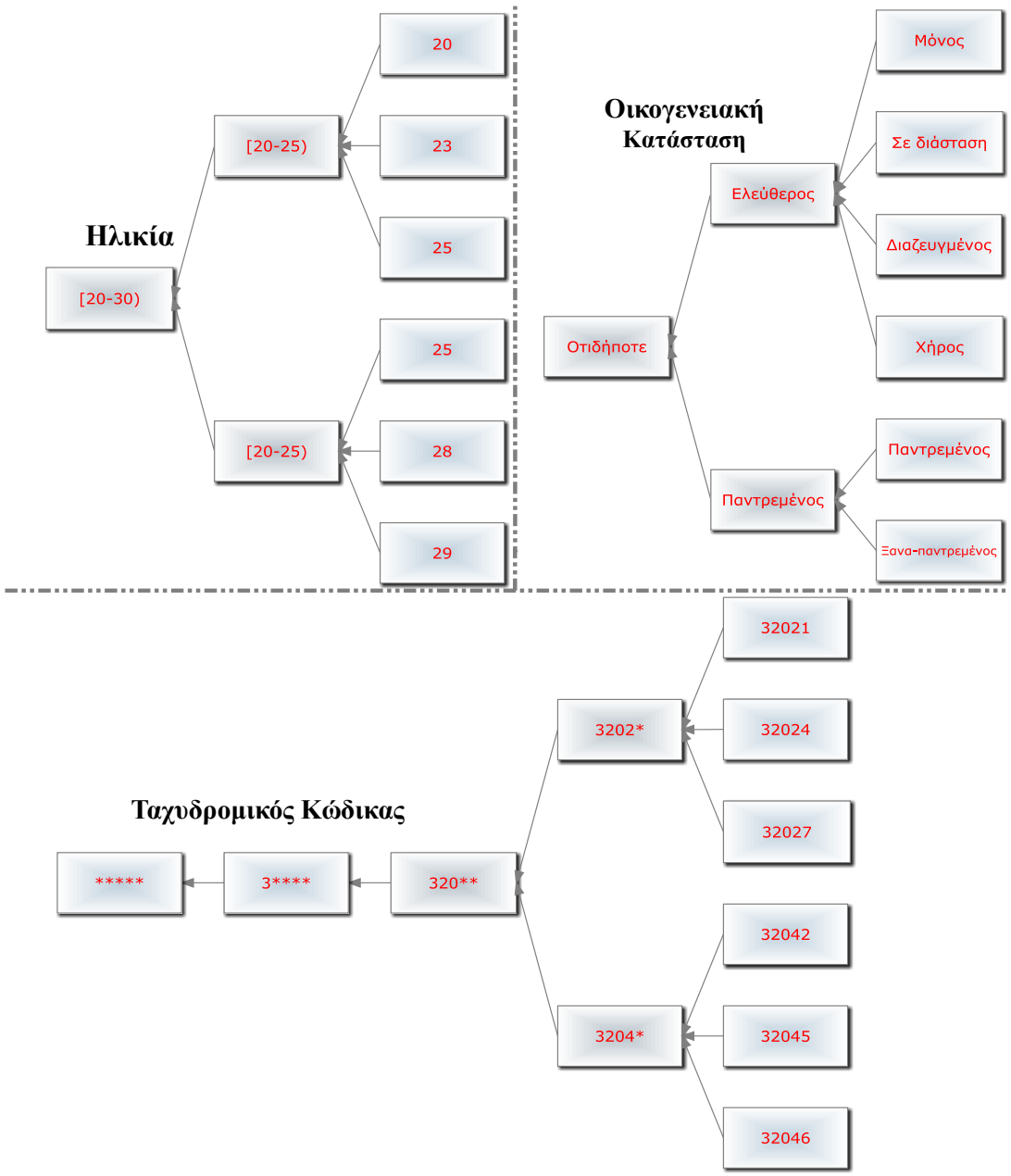
Έπειτα το μοντέλο αξιοποιεί τις παραπάνω τιμές και δέχεται κάποιες επιπλέον παραμέτρους που εισάγει ο κάτοχος των δεδομένων για μεγαλύτερη ακρίβεια και ευελιξία στην ανωνυμοποίηση. Η επιτυχής ανωνυμοποίηση των δεδομένων είναι εφικτή εάν οριστούν οι τιμές l, k, m σωστά.

## *5.5 Αλγόριθμοι Ανωνυμοποίησης*

Σε αυτό το κεφάλαιο θα εξετάσουμε κάποιους χαρακτηριστικούς αλγορίθμους ανωνυμοποίησης. Η επιλογή τους έγινε με τέτοιο τρόπο ώστε να παρουσιαστούν όσο το δυνατόν περισσότερες υλοποιήσεις μοντέλων αλλά και τεχνικές ανωνυμοποίησης. Συγκεκριμένα θα παρουσιαστούν και θα συγκριθούν οι:

| | **Αλγόριθμοι ανωνυμοποίησης** |
|---|---|
| **1** | Datafly |
| **2** | μ-Argus |
| **3** | Βελτιωμένη έκδοσή του μ-Argus |
| **4** | Mondrian |
| **5** | Bottom-up |
| **6** | Top-Down |
| **7** | Incognito |
| **8** | Apriori |
| **9** | Anatomy |

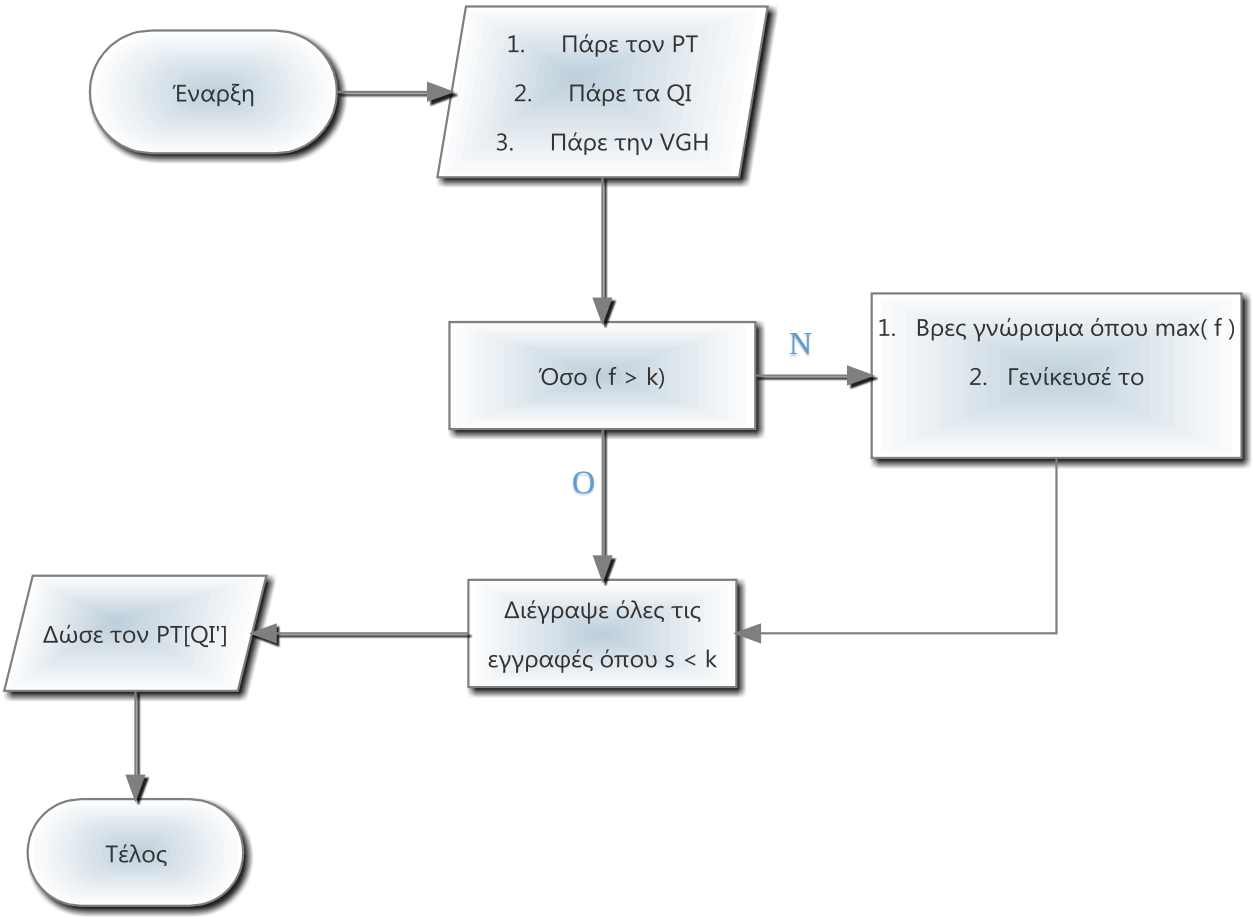
Παρακάτω απεικονίζονται τρεις ιεραρχίες γενίκευσης που θα μας χρησιμοποιηθούν ως αναφορά στα παραδείγματα των αλγορίθμων:




Σχήμα 5.11: Ιεραρχία γενίκευσης τιμών σε { Ηλικία, Οικογενειακή κατάσταση, Τ.Κ}

## *Datafly*

Ο **Datafly** [29, 57] είναι ένας ευριστικός και άπληστος αλγόριθμος. Είναι δημιούργημα της καθηγήτριας Sweeney και ανωνυμοποιεί πίνακες με την εφαρμογή *ολικής γενίκευσης πεδίων* και *κατάπνιξης εγγραφών*. Είναι από τους πρώτους αλγορίθμους που δημοσιεύτηκαν και ικανός να χειριστεί μεγάλων διαστάσεων πίνακες. Ο αλγόριθμος σε διάγραμμα ροής:




Σχήμα 5.12: Διάγραμμα ροής Datafly

Αρχικά ο αλγόριθμος παίρνει ως είσοδο τον πίνακα PT, την λίστα με τα ψευδο-αναγνωριστικά, την παράμετρο ανωνυμοποίησης k και την ιεραρχία γενικευμένης τιμής στα δεδομένα που την χρειάζεται. Υποθέτει πως το μέγεθος του πίνακα είναι μεγαλύτερο του k αφού διαφορετικά δεν θα είχε νόημα η διαδικασία. Έπειτα υπολογίζει την τιμή f, όπου f είναι το πλήθος των μοναδικών τιμών του κάθε γνωρίσματος των QID. Όσο η f είναι μεγαλύτερη της παραμέτρου k τότε βρίσκει το γνώρισμα με το μεγαλύτερο f και αντιστοιχίζει κάθε τιμή του γνωρίσματος αυτού στο αμέσως υψηλότερο επίπεδο ιεραρχίας. Έπειτα διαγράφει όποια εγγραφή ή κλάση έχει μέγεθος s μικρότερο του k. Τέλος επιστρέφει τον ανωνυμοποιημένο πίνακα.

Παράδειγμα Datafly:

Έστω πως έχουμε τον πίνακα 5.17 όπου το γνώρισμα { Όνομα } είναι αναγνωριστικό

ταυτότητας και διαγράφεται, τα γνωρίσματα { Ηλικία, Οικογενειακή κατάσταση, Ταχυδρομικός κώδικας } είναι ψευδο-αναγνωριστικά και το γνώρισμα { Αδίκημα} είναι ευαίσθητο:

| ΕΓΓΡΑΦΗ | ΌΝΟΜΑ | ΟΙΚΟΓΕΝΕΙΑΚΗ ΚΑΤΑΣΤΑΣΗ | ΗΛΙΚΙΑ | ΤΑΧΥΔΡΟΜΙΚΟΣ ΚΩΔΙΚΑΣ | ΑΔΙΚΗΜΑ |
|---|---|---|---|---|---|
| **1** | Κώστας | Διαζευγμένος | 29 | 32042 | Δολοφονία |
| **2** | Μαντώ | Μόνος | 20 | 32021 | Κλοπή |
| **3** | Μαρίνα | Χήρος | 24 | 32024 | Τρομοκρατία |
| **4** | Δημήτρης | Διαζευγμένος | 28 | 32046 | Απαγωγή |
| **5** | Βαλάντης | Χήρος | 25 | 32045 | Λαθρεμπόριο |
| **6** | Άννα | Μόνος | 23 | 32027 | Εμπρησμός |

Πίνακας 5.17: Αρχικός πίνακας δεδομένων

1. Ο πίνακας 1 δεν ικανοποιεί την k = 3 ανωνυμία και έχει 6 εγγραφές. Αρχικά υπολογίζει την τιμή f του κάθε χαρακτηριστικού γνωρίσματος και γενικεύει την στήλη { Ηλικία}:

| **Οικογενειακή κατάσταση** | **Ηλικία** | **Ταχυδρομικός κώδικας** | **s** |
|---|---|---|---|
| **Διαζευγμένος** | 29 | 32042 | 1 |
| **Μόνος** | 20 | 32021 | 1 |
| **Χήρος** | 24 | 32024 | 1 |
| **Διαζευγμένος** | 28 | 32046 | 1 |
| **Χήρος** | 25 | 32045 | 1 |
| **Μόνος** | 23 | 32027 | 1 |
| **f=3** | f=6 | f=6 | |

Πίνακας 5.18: Πίνακας δεδομένων στο βήμα 1

2. Ο πίνακας δεν ικανοποιεί την k-ανωνυμία και γενικεύει την στήλη { Ταχυδρομικός κώδικας}:

| Οικογενειακή κατάσταση | Ηλικία | Ταχυδρομικός κώδικας | s |
|---|---|---|---|
| **Διαζευγμένος** | [25-30) | 32042 | 1 |
| **Μόνος** | [20-25) | 32021 | 1 |
| **Χήρος** | [20-25) | 32024 | 1 |
| **Διαζευγμένος** | [25-30) | 32046 | 1 |
| **Χήρος** | [25-30) | 32045 | 1 |
| **Μόνος** | [20-25) | 32027 | 1 |
| **f=3** | f=2 | f=6 | |

Πίνακας 5.19: Πίνακας δεδομένων στο βήμα 2

3. Ο πίνακας δεν ικανοποιεί την k-ανωνυμία και γενικεύει την στήλη { Οικογενειακή κατάσταση }

| Οικογενειακή κατάσταση | Ηλικία | Ταχυδρομικός κώδικας | s |
|---|---|---|---|
| **Διαζευγμένος** | [25-30) | 3204* | 2 |
| **Μόνος** | [20-25) | 3202* | 2 |
| **Χήρος** | [20-25) | 3202* | 1 |
| **Χήρος** | [25-30) | 3204* | 1 |
| **f=3** | f=2 | f=2 | |

Πίνακας 5.20: Πίνακας δεδομένων στο βήμα 3

4. Ο πίνακας πλέον ικανοποιεί την k-ανωνυμία:

| **Οικογενειακή κατάσταση** | **Ηλικία** | **Ταχυδρομικός κώδικας** | **s** |
|---|---|---|---|
| **Ελεύθερος** | [25-30) | 3204* | 3 |
| **Ελεύθερος** | [20-25) | 3202* | 3 |
| **f=2** | f=2 | f=2 | |

Πίνακας 5.21: Πίνακας δεδομένων στο βήμα 4

5. Ο αλγόριθμος τέλος ανακατασκευάζει και επιστρέφει τον ανωνυμοποιημένο πίνακα:

| | | **Ψευδο-αναγνωριστικά** | | | **Ευαίσθητο γνώρισμα** |
|---|---|---|---|---|---|
| Εγγραφή | Κλάση ισοδυναμίας | Οικογενειακή κατάσταση | Ηλικία | Ταχυδρομικός κώδικας | Αδίκημα |
| 1 | 1 | Ελεύθερος | [25-30) | 3204* | Δολοφονία |
| 2 | | Ελεύθερος | [25-30) | 3204* | Κλοπή |
| 3 | | Ελεύθερος | [25-30) | 3204* | Τρομοκρατία |
| 4 | 2 | Ελεύθερος | [20-25) | 3202* | Απαγωγή |
| 5 | | Ελεύθερος | [20-25) | 3202* | Λαθρεμπόριο |
| 6 | | Ελεύθερος | [20-25) | 3202* | Εμπρησμός |

Πίνακας 5.22: Τελικός πίνακας δεδομένων

Τα μειονεκτήματα του αλγορίθμου Datafly είναι ότι πραγματοποιεί ολική γενίκευση χαρακτηριστικού γνωρίσματος, ολική διαγραφή εγγραφών και ότι η επιλογή γνωρίσματος για γενίκευση με την επιλογή μέγιστου f πιθανόν να προκαλέσει υπέρ-γενίκευση.

## *μ-Argus*

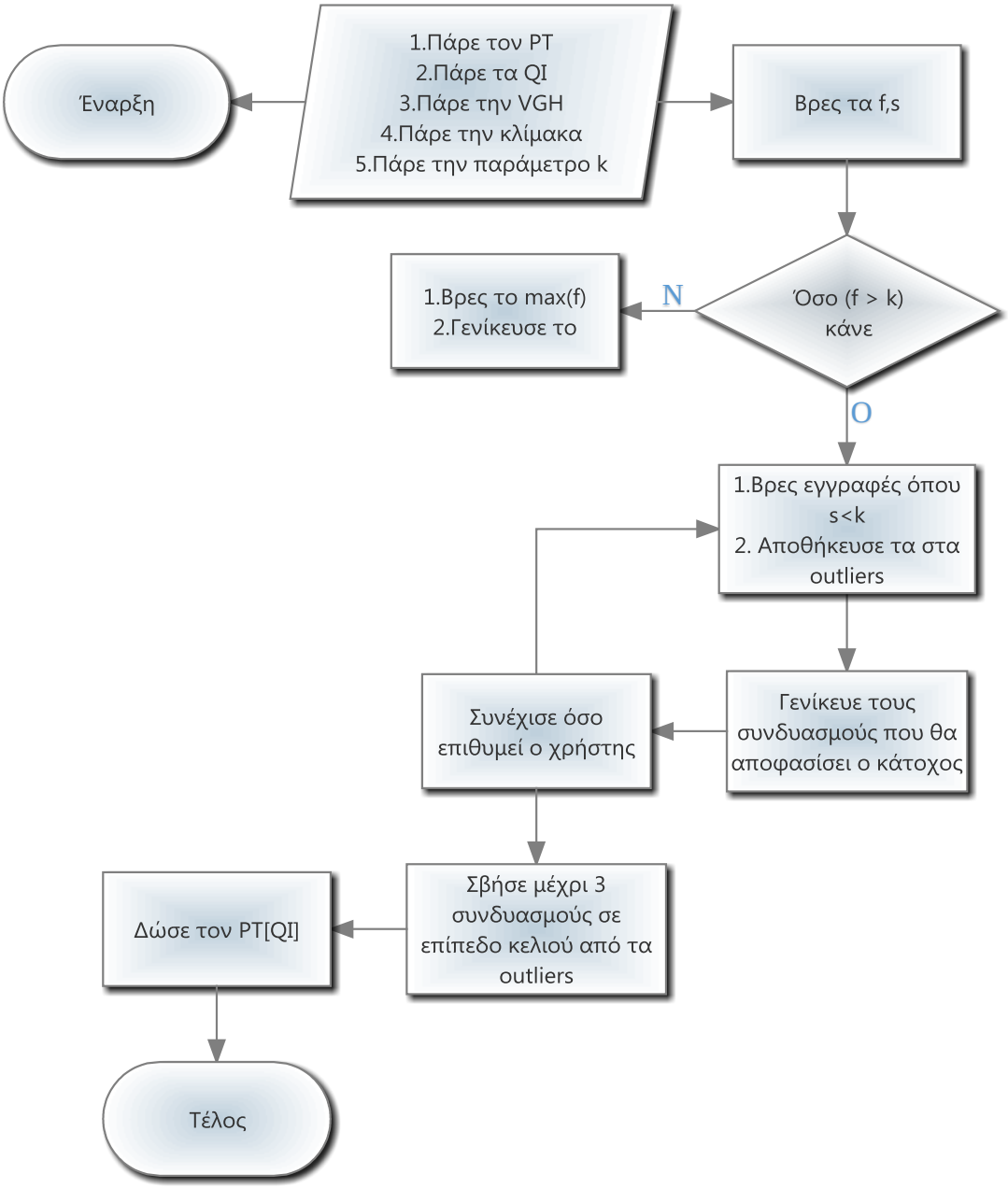
Η επόμενη προσπάθεια για πρακτική εφαρμογή της k-ανωνυμίας έρχεται από τους Hundpool και Willenborg το 1996. Ο άπληστος αλγόριθμος μ-Argus [9, 29] χρησιμοποιεί τις τεχνικές της γενίκευσης και κατάπνιξης σε επίπεδο κελιού και λειτουργεί με διαφορετικό τρόπο από τον Datafly. Αρχικά ο χρήστης κατηγοριοποιεί τα ψευδο-αναγνωριστικά σε μία κλίμακα επικινδυνότητας. Οι τιμές των γνωρισμάτων ανάλογα την κατηγοριοποίηση παίρνουν τις τιμές 0 έως 3 με βάση την κλίμακα:


Σχήμα 5.13: Κλίμακα αξιολόγησης Ψευδο-αναγνωριστικών

Η είσοδος του αλγορίθμου περιλαμβάνει τον πίνακα PT, την λίστα με τα ψευδο-αναγνωριστικά, την παράμετρο ανωνυμοποίησης k, την ιεραρχία γενικευμένης τιμής στα δεδομένα που την χρειάζεται και την κλίμακα αξιολόγησης. Στην συνέχεια βρίσκει τις τιμές f, s και γενικεύει τα γνωρίσματα όσο υπάρχουν f > k όπως και στον Datafly. Έπειτα αποθηκεύει στην βάση outliers δυο με τρεις επικίνδυνους συνδυασμούς βασισμένους στην κλίμακα αξιολόγησης και ο χρήστης αποφασίζει εάν θα τα γενικεύσει ή όχι. Η διαδικασία αυτή διαρκεί μέχρι να αποφασίσει ο χρήστης να διακοπεί. Αυτόματα στην συνέχεια ο αλγόριθμος διαγράφει από την βάση outliers μέχρι τρείς συνδυασμούς, τους οποίους θεωρεί πιο επικίνδυνους.

Το μεγάλο μειονέκτημα του μ-Argus είναι πως μπορεί τελικά να μην επιτευχθεί η k-ανωνυμία αφού στην βάση outliers μπορεί να υπάρχουν τέσσερις ή και παραπάνω μοναδικοί συνδυασμοί. Μπορεί επίσης να έχουμε υπεργενίκευση δεδομένων.

Ο αλγόριθμος σε διάγραμμα ροής:

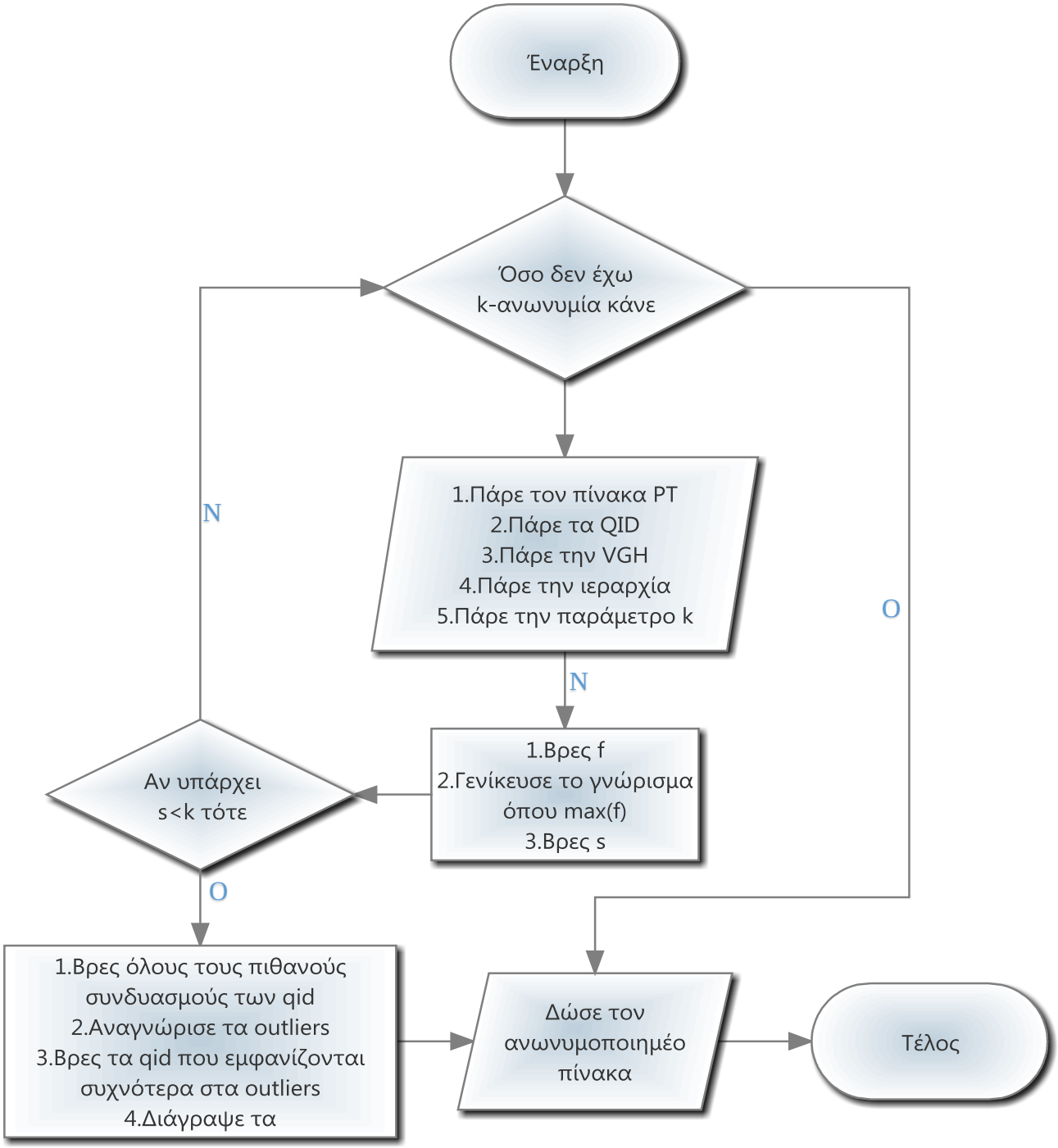



Σχήμα 5.14: Διάγραμμα ροής μ-Argus

## *Βελτιωμένη μ-Argus*

Για να ξεπεράσει την μεγάλη ευπάθεια της μ-Argus, μια και υπάρχει πιθανότητα ο

αλγόριθμός των Hundpool και Willenborg να μην πετύχει k-ανωνυμία η εθνική στατιστική υπηρεσία της Ολλανδίας δημοσίευσε μια βελτιωμένη έκδοσή της [58]. Ο βελτιωμένος αλγόριθμος σε διάγραμμα ροής:




Σχήμα 5.15: Διάγραμμα ροής βελτιωμένου μ-Argus

## *Mondrian*

Ο Mondrian [31, 39] είναι ένας άπληστος αλγόριθμος ο οποίος στηρίζεται στην συνεχή διαμέριση του χώρου. Επιτελεί γενικεύσεις n-διαστάσεων, πετυχαίνει την ελάχιστη k-ανωνυμία. Βάσει του πολυδιάστατου μοντέλου τοπικής ανακωδικοποίησης προσφέρει υψηλότερης ποιότητας ανωνυμοποίηση και λιγότερη απώλεια πληροφοριών συγκριτικά με τους παραπάνω αλγορίθμους.

Αρχικά ο αλγόριθμος λαμβάνει τον πίνακα, τα ψευδο-αναγνωριστικά και την παράμετρο ανωνυμοποίησης k. Επιλέγει μία διάσταση του δηλαδή ένα από τα χαρακτηριστικά γνωρίσματα των ψευδο-αναγνωριστικών. Συνήθως επιλέγει την διάσταση η οποία έχει το μέγιστο κανονικοποιημένο εύρος, αλλά αυτό είναι μια επιλογή του χρήστη Εάν όλες οι διαστάσεις παίρνουν την ίδια τιμή τότε επιλέγεται αυτή που έχει το μικρότερο απόλυτο εύρος. Έπειτα εξετάζει όλες τις τιμές του γνωρίσματος και επιλέγει την ενδιάμεση τιμή. Με βάση την ενδιάμεση τιμή και το επιλεγμένο γνώρισμα οι εγγραφές χωρίζονται σε δύο υποχώρους. Οι εγγραφές που έχουν μικρότερη τιμή από την επιλεγμένη αποθηκεύονται στην lpart ενώ οι μεγαλύτερες στην rpart. Η διαδικασία αυτή συνεχίζεται αναδρομικά μέχρι να όλοι οι υποχώροι να έχουν μέγεθος μικρότερο του 2k ώστε να ικανοποιείται η k-ανωνυμία. Τέλος επιστρέφει τον ανωνυμοποιημένο πίνακα.

Ακολουθώντας τα παραπάνω βήματα ο αλγόριθμος πετυχαίνει την βέλτιστη πολυδιάστατη διαμέριση, και k-ανωνυμία αφού σε κάθε τελικό υποχώρο έχει διαμορφωθεί και μία κλάση ισοδυναμίας μεγέθους τουλάχιστον k.

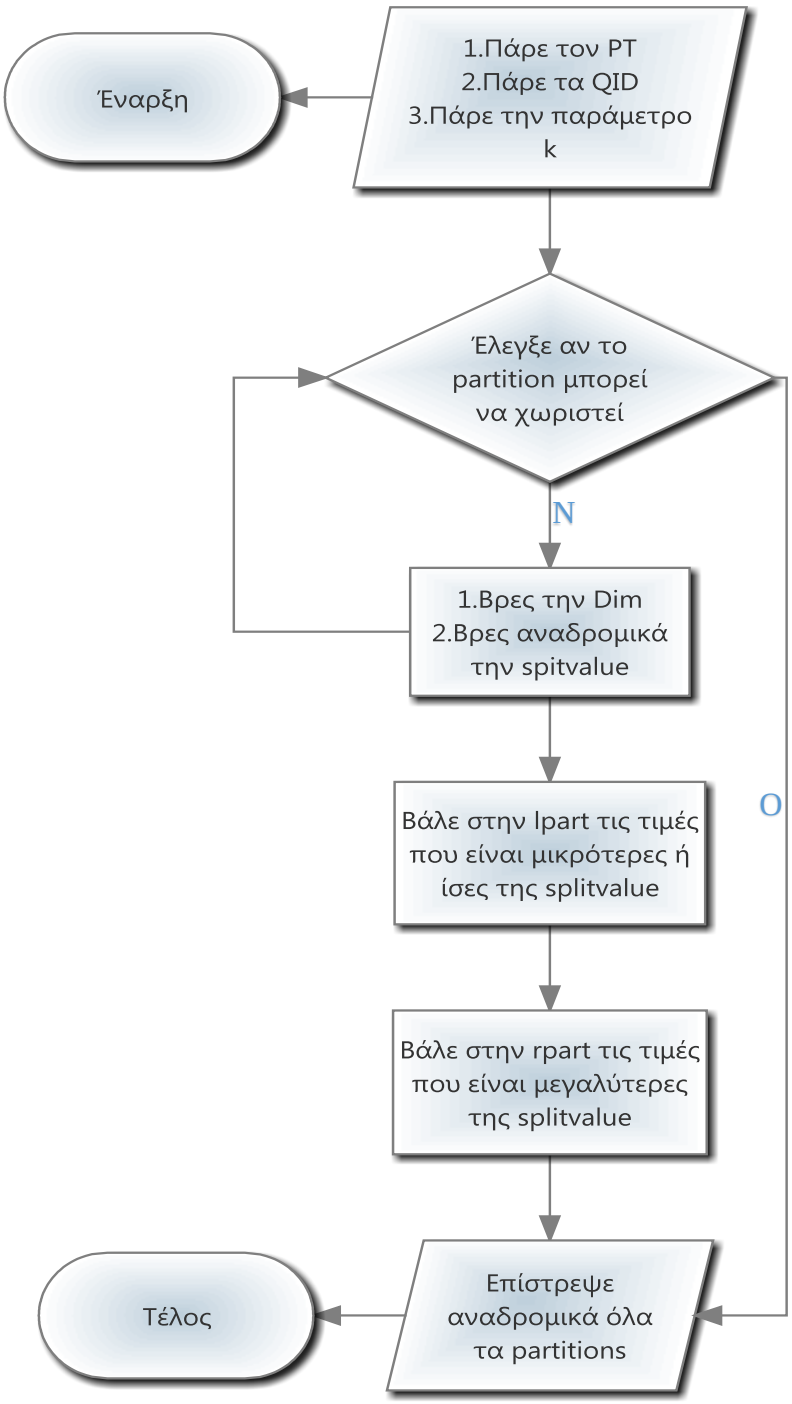
Ο αλγόριθμος Mondrian έχει πολυπλοκότητα O(nlogn) και το πρόβλημα εύρεσης της βέλτιστης διαμέρισης ανήκει στην κατηγορία NP-hard. Πρέπει να σημειωθεί πως το βασικό μειονέκτημα του αλγορίθμου είναι ότι, για να επεξεργαστεί τα κατηγορικά δεδομένα χρειάζεται την διάταξή τους, ενώ στα αριθμητικά δεδομένα η ταχύτητα εκτέλεσής του είναι ιδιαίτερα ικανοποιητική.

Για τον αλγόριθμο Mondrian σε διάγραμμα ροής θεωρώ ως:

- partition την διαμέριση του χώρου
- Dim την επιλογή της διάστασης
- splitvalue την ενδιάμεση τιμή

- lpart και rpart τον αριστερά και δεξιά υποχώρο αντίστοιχα

Αλγόριθμος Mondrian σε διάγραμμα ροής:

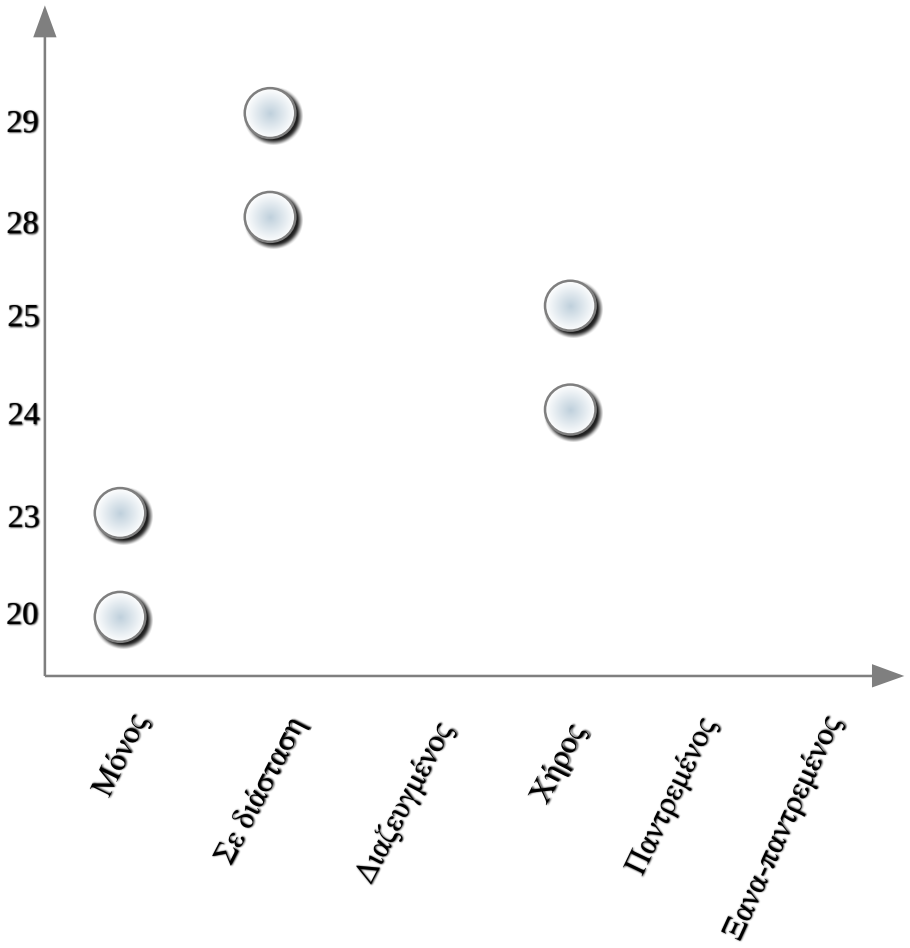



Σχήμα 5.16: Διάγραμμα ροής Mondrian

Παράδειγμα Mondrian

Έστω ο πίνακας 5.17 όπου το γνώρισμα { Όνομα } είναι αναγνωριστικό ταυτότητας και διαγράφεται, τα γνωρίσματα { Ηλικία, Οικογενειακή κατάσταση } είναι ψευδο-αναγνωριστικά, το γνώρισμα { Αδίκημα} είναι ευαίσθητο και ο { Ταχυδρομικός κώδικας } μη ευαίσθητο-γνώρισμα.

1. Η αρχική χωρική αναπαράσταση των τιμών των ψευδο-αναγνωριστικών { Ηλικία, Οικογενειακή κατάσταση } είναι:




Σχήμα 5.17: Χωρική αναπαράσταση αρχικών δεδομένων

Εξετάζονται οι περιοχές: [20 - 29] και [Μόνος - Ξανα-παντρεμένος].

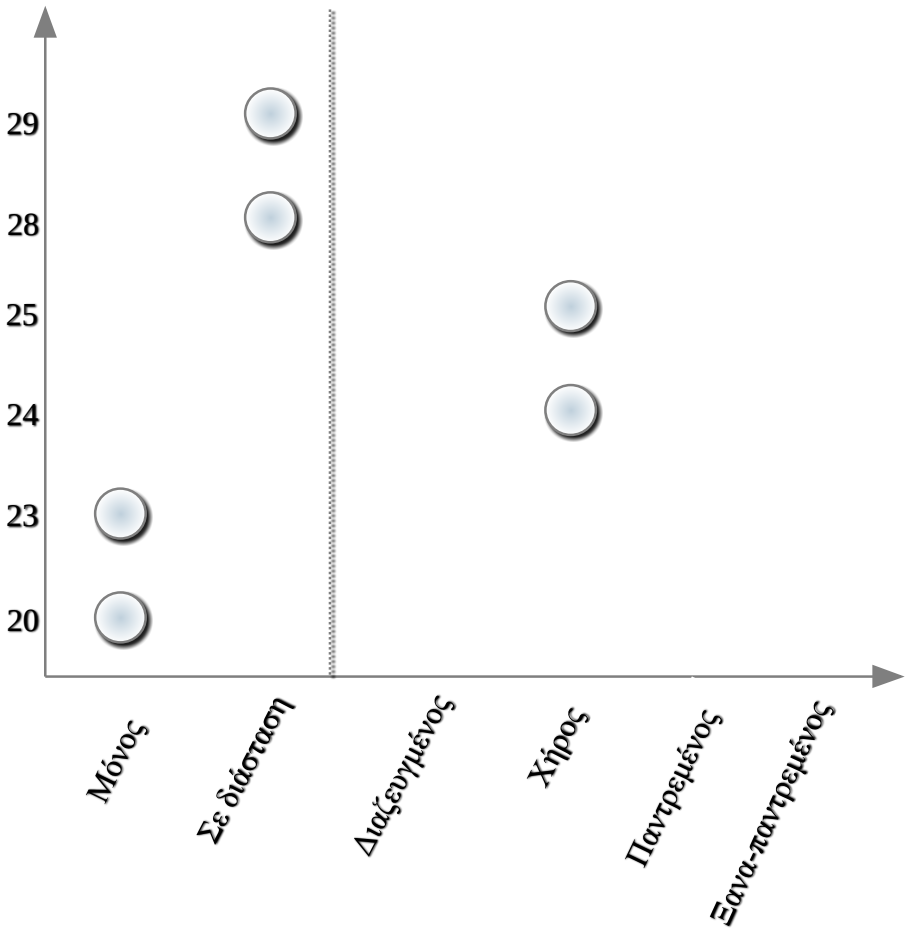
Το κανονικοποιημένο εύρος και των δύο γνωρισμάτων είναι 1.

Ο αλγόριθμος επιλέγει:

- dim = { Οικογενειακή κατάσταση }

- splitvalue = { Σε διάσταση }
- lpart = [ Μόνος – Σε διάσταση ]
- rpart = ( Σε διάσταση – Ξανα-παντρεμένος ]

2. Η χωρική αναπαράσταση των τιμών μετά την πρώτη επανάληψη:




Σχήμα 5.18: Χωρική αναπαράσταση δεδομένων στο βήμα 1

Εξετάζονται οι περιοχές: [20 - 29] και [Μόνος – Σε διάσταση].

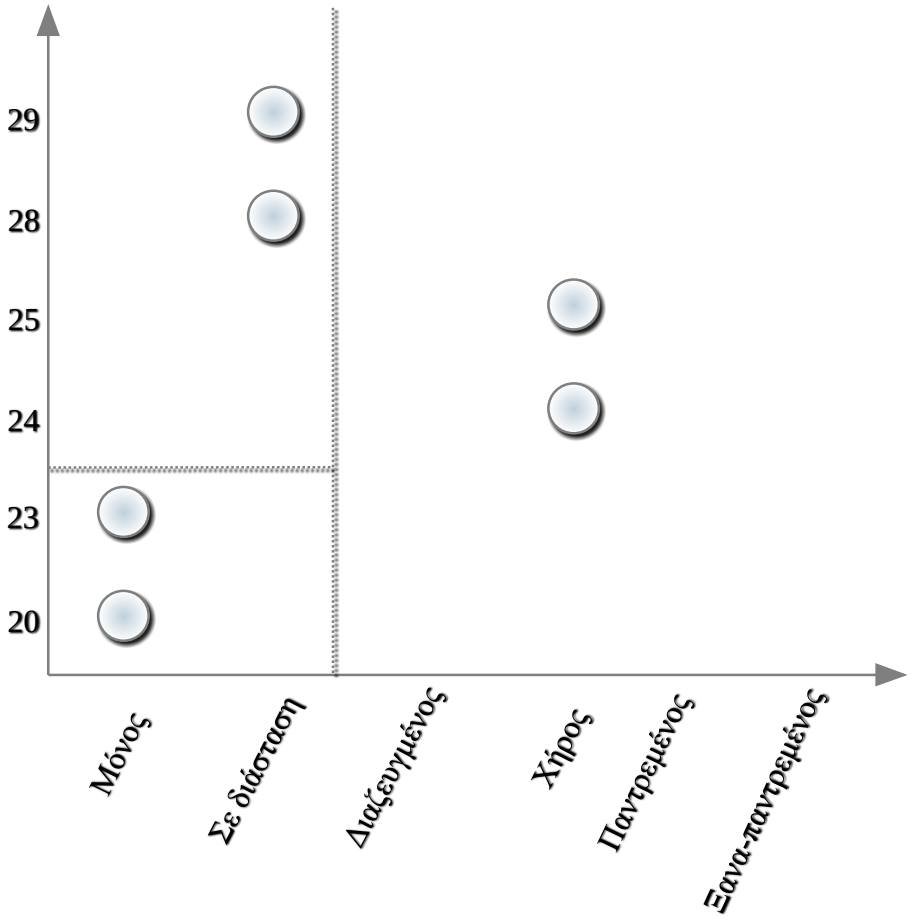
Το κανονικοποιημένο εύρος του γνωρίσματος { Ηλικία } είναι 1 και της { Οικογενειακή κατάσταση } είναι 0.2

Ο αλγόριθμος επιλέγει:

- dim = { Ηλικία }

- splitvalue = { 23 }
- lpart = [ 20 - 23 ]
- rpart = ( 23 – 29]

3. Η χωρική αναπαράσταση των τιμών μετά την δεύτερη επανάληψη:




Σχήμα 5.19: Χωρική αναπαράσταση δεδομένων στο βήμα 2

Πλέον ο αλγόριθμος δεν μπορεί να δημιουργήσει καινούρια partitions και επιστρέφει τον ανωνυμοποιημένο πίνακα:

| | | **Ψευδο-αναγνωριστικά** | | **Μη ευαίσθητο γνώρισμα** | **Ευαίσθητο γνώρισμα** |
|---|---|---|---|---|---|
| **Εγγραφή** | **Κλάση ισοδυναμίας** | **Οικογενειακή κατάσταση** | **Ηλικία** | **Ταχυδρομικός κώδικας** | **Αδίκημα** |
| **1** | **1** | Ελεύθερος | (23-30) | 32042 | Δολοφονία |
| **2** | | Ελεύθερος | (23-30) | 32046 | Κλοπή |
| **3** | | Ελεύθερος | [20-23] | 32041 | Τρομοκρατία |
| **4** | **2** | Ελεύθερος | [20-23] | 32027 | Απαγωγή |
| **5** | | Ελεύθερος | [20-30) | 32024 | Λαθρεμπόριο |
| **6** | | Ελεύθερος | [20-30) | 32025 | Εμπρησμός |

Πίνακας 5.23: Ανωνυμοποιημένος πίνακας δεδομένων από τον αλγόριθμο mondrian

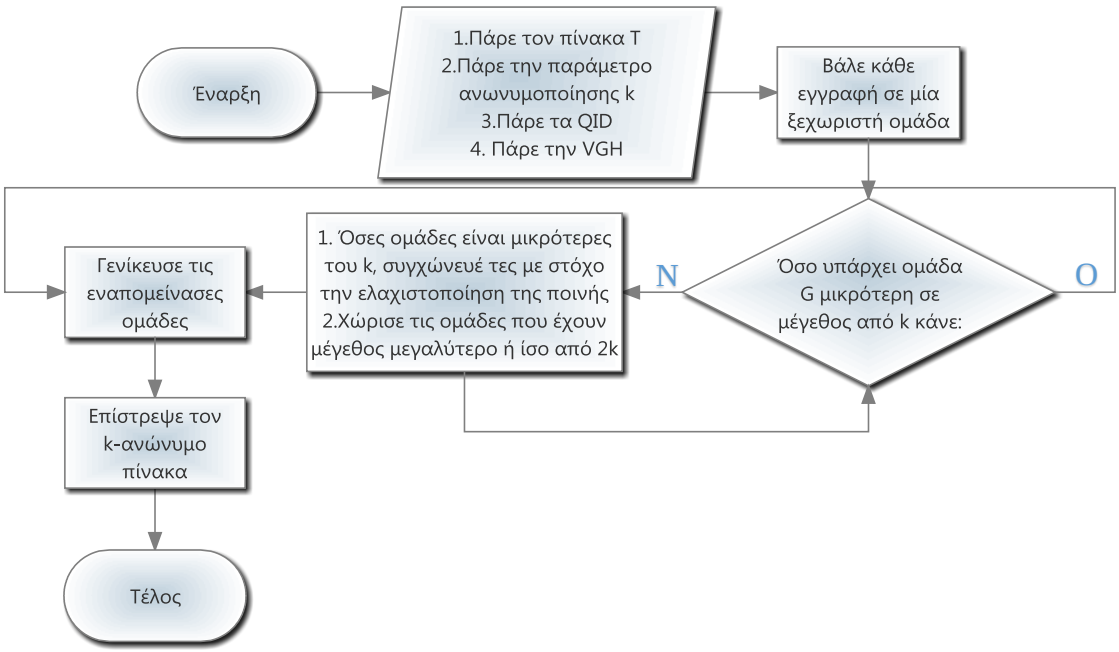
### *Bottom up*

Ο αλγόριθμος Bottom up [59] είναι της ίδια φιλοσοφίας με τον Mondrian, αλλά ενεργεί με διαφορετική λογική. Ο στόχος του είναι η επίτευξη της μέγιστης ποιότητας των πληροφοριών με την ελάχιστη δυνατή ποινή. Είναι και αυτός άπληστος αφού επιλέγει τοπικά την βέλτιστη λύση.

Αρχικά η κάθε εγγραφή αποτελεί μία ομάδα. Σταδιακά και όσο τρέχει ο αλγόριθμος οι ομάδες συγχωνεύονται με την μικρότερη ποινή, ώστε κάποια στιγμή κάθε ομάδα να έχει τουλάχιστον k εγγραφές. Επειδή υπάρχει η πιθανότητα στο στάδιο της συγχώνευσης να προκύψουν ομάδες μεγαλύτερες από 2k σε μέγεθος, ο αλγόριθμος φροντίζει και τις σπάει στην μέση.

Σε αντίθεση με τους προηγούμενους αλγορίθμους ο Bottom up δεν στοχεύει στην διαμόρφωση ομάδων με το μικρότερο δυνατό μέγεθος αλλά στην ελαχιστοποίηση της ποινής. Το βασικό μειονέκτημα του είναι η μεγάλη του πολυπλοκότητα, $O(\log(k)|T|^2)$.

Ο αλγόριθμος σε διάγραμμα ροής:




Σχήμα 5.20: Διάγραμμα ροής Bottom up

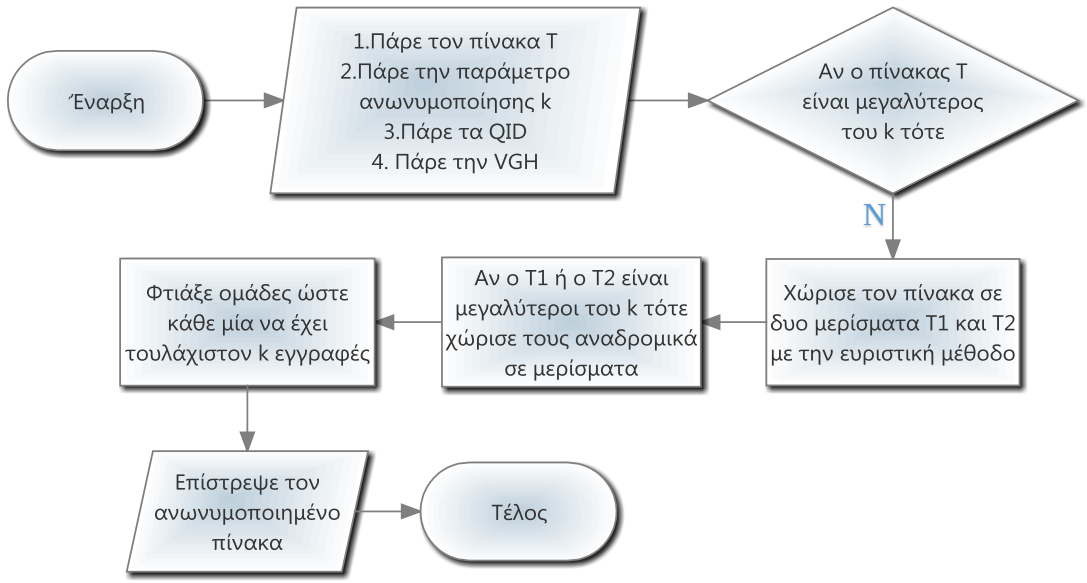
### *Top Down*

Βασικό μειονέκτημα του Bottom up είναι η ταχύτητά του. Ο top down [59] αντίθετα είναι γρηγορότερος αλλά παρουσιάζει χειρότερα αποτελέσματα στην ποιότητα των ανωνυμοποιημένων πληροφοριών.

Αρχικά ο αλγόριθμος παίρνει τον πίνακα και τον χωρίζει σε μερίσματα ώστε ο καθένας τοπικά να διατηρεί μεγαλύτερη πληροφορία. Εμείς θέλουμε κάθε φορά που σπάει ο πίνακας η ποινή να μειώνεται. Τελικά οι ομάδες που περισσεύουν και έχουν μέγεθος μικρότερο του k συγχωνεύονται ώστε να πετύχουμε k-ανωνυμία.

Το βασικό πρόβλημα του αλγορίθμου είναι πώς θα χωριστούν τα μερίσματα με τρόπο ώστε να μειώνετε η ποινή. Η ευριστική μέθοδος που το πετυχαίνει αρχικά επιλέγει δύο εγγραφές ώστε η ποινή να είναι η μέγιστη δυνατή και στην συνέχεια αυτές θα ανήκουν σε διαφορετικά μερίσματα. Οι υπόλοιπες εγγραφές ταξινομούνται στο μέρισμα εκείνο που θα επιβάλλει την μικρότερη ποινή.

Ο αλγόριθμος σε διάγραμμα ροής:




Σχήμα 5.21: Διάγραμμα ροής Top Down

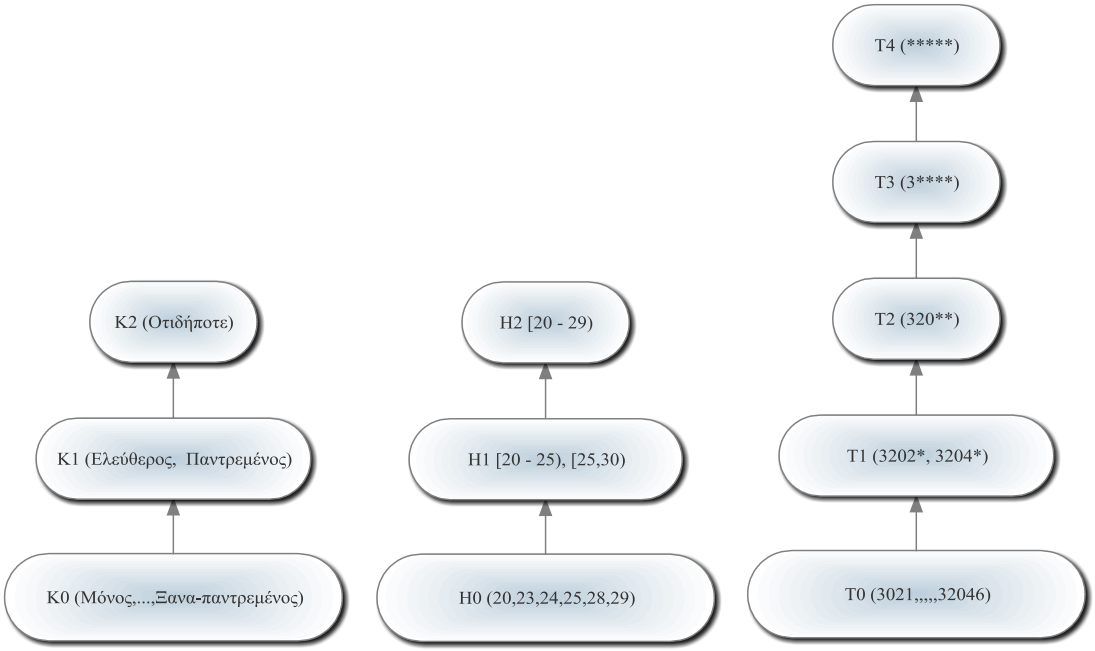
## *Incognito*

Ο αλγόριθμος Incognito είναι ένας ορθός και πλήρης αλγόριθμος [30]. Χρησιμοποιεί γενίκευση πλήρους πεδίου και μπορεί να θέσει ένα όριο στις καταπνίξεις δεδομένων. Στόχος του είναι η επίτευξη της k-ανωνυμίας με τον ελάχιστο συνδυασμό γενικεύσεων.

Για την επίτευξη της ανωνυμίας κάνει χρήση δύο πολύ σημαντικών ιδιοτήτων:

- *Ιδιότητα γενίκευσης*: εάν σε κάποιο κόμβο του πλέγματος γενίκευσης ισχύει η k-ανωνυμία τότε ισχύει και για οποιοδήποτε πρόγονο του κόμβου και
- *Ιδιότητα υποσυνόλου*: εάν ένα σύνολο ψευδο-αναγνωριστικών δεν πληρεί την k-ανωνυμία τότε δεν την πληρεί ούτε ένα υπερσύνολό του

Σε αντίθεση με τους παραπάνω αλγορίθμους ο Incognito χρειάζεται την ιεραρχία γενικευμένου πεδίου και λαμβάνοντας τα δεδομένα του σχήματος 5.11, σχηματικά είναι:




Σχήμα 5.22: Ιεραρχία γενικευμένου πεδίου

Ως είσοδο ο αλγόριθμος λαμβάνει τον πίνακα PT, την λίστα με τα ψευδο-αναγνωριστικά, την παράμετρο ανωνυμοποίησης k και την ιεραρχία γενικευμένου πεδίου για κάθε QID. Ξεκινά και χτίζει ένα πλέγμα γενίκευσης που το διασχίζει με την μέθοδο της αναζήτησης κατά πλάτους. Πρώτα γίνεται η αξιολόγηση ενός γνωρίσματος κάθε φορά και στην συνέχεια κάνοντας χρήση των ιδιοτήτων της γενίκευσης και του υποσυνόλου προχωρά προοδευτικά στην αξιολόγηση μεγαλύτερων συνόλων. Οι επαναλήψεις σταματούν όταν ο αλγόριθμος έχει συμπεριλάβει όλα τα ψευδο-αναγνωριστικά, έχει επιλέξει το βέλτιστο σύνολο και τα έχει γενικεύσει επαρκώς.

Παράδειγμα Incognito

Έστω πως έχουμε όπως και στα προηγούμενα παραδείγματα τον πίνακα 1 όπου τα γνωρίσματα { Οικογενειακή κατάσταση, Ηλικία, Ταχυδρομικός κώδικας } είναι ψευδο-αναγνωριστικά και το γνώρισμα { Αδίκημα} είναι ευαίσθητο. Η παράμετρος ανωνυμοποίησης είναι k = 2.

Στην πρώτη επανάληψη ο αλγόριθμος ελέγχει αν ο πίνακας 1 είναι k-ανώνυμος για γενικεύσεις ενός γνωρίσματος. Έστω ότι αρχικά ελέγχει την k-ανωνυμία αν κρατήσει το γνώρισμα { Οικογενειακή κατάσταση } και απομακρύνει τα { Ηλικία, Ταχυδρομικός κώδικας }. Βρίσκει ότι ο πίνακας είναι k-ανώνυμος με βάση το υποσύνολο K0 και άρα με όλες τις γενικευμένες τιμές που ορίζονται από την ιεραρχία του γενικευμένου πεδίου.

Επαναλαμβάνει την διαδικασία και για τα υπόλοιπα γνωρίσματα. Το αποτέλεσμα στα υπόλοιπα γνωρίσματα είναι H1 και T1.

Στη δεύτερη επανάληψη ο αλγόριθμος θα ελέγξει αν ισχύει η k-ανωνυμία για τα υποσύνολα δύο γνωρισμάτων δηλαδή τα: { Οικογενειακή κατάσταση, Ηλικία } , { Οικογενειακή κατάσταση, Ταχυδρομικός Κώδικας } και { Ηλικία, Ταχυδρομικός κώδικας }. Για παράδειγμα ο αλγόριθμος ελέγχει αρχικά το σύνολο [ K0, H0 ] και βλέπει ότι δεν ικανοποιείται η k-ανωνυμία, οπότε περνά στον έλεγχο του [ K1, H0 ] και [ K0, H1 ]. Βλέπει ότι πάλι δεν ικανοποιείται η k-ανωνυμία με κανένα από τα δύο σύνολα οπότε και απορρίπτονται. Στην συνέχεια ελέγχει και βρίσκει πως συνδυασμός [ K1, H1 ] ικανοποιεί την k-ανωνυμία. Ως εκ τούτου οι υψηλότερες στο πλέγμα διευθύνσεις γενίκευσης των συνόλων [ K, H ] δεν ελέγχονται γιατί αποτελούν γενικεύσεις του [ K1, H1 ]. Η ίδια διαδικασία ακολουθείται για όλα τα υποσύνολα δύο γνωρισμάτων.

Στην τρίτη επανάληψη ο αλγόριθμος θα ελέγξει αν ισχύει η k-ανωνυμία και για τα τρία γνωρίσματα [ K, H, T]. Ακολουθώντας την ίδια διαδικασία με τις παραπάνω επαναλήψεις τα σύνολα που ικανοποιούν την k-ανωνυμία είναι τα [ K0, H2, T2] και [ K1, H1, T1]. Με βάση αυτό το αποτέλεσμα θα επιλεγεί το πιο αποδοτικό σύνολο με βάση το μετρικό σύστημα που έχει ήδη επιλέξει ο χρήστης. Το ανωνυμοποιημένο σύνολο [ K0, H2, T2] που έδωσε ως έξοδο ο αλγόριθμος Incognito είναι:

| | | **Ψευδο-αναγνωριστικά** | | | **Ευαίσθητο γνώρισμα** |
|---|---|---|---|---|---|
| **Εγγραφή** | **Κλάση ισοδυναμίας** | **Οικογενειακή κατάσταση** | **Ηλικία** | **Ταχυδρομικός κώδικας** | **Αδίκημα** |
| **1** | **1** | Σε διάσταση | [20-30) | 320** | Δολοφονία |
| **2** | | Σε διάσταση | [20-30) | 320** | Κλοπή |
| **3** | | Μόνος | [20-30) | 320** | Τρομοκρατία |
| **4** | **2** | Μόνος | [20-30) | 320** | Απαγωγή |
| **5** | | Χήρος | [20-30) | 320** | Λαθρεμπόριο |
| **6** | | Χήρος | [20-30) | 320** | Εμπρησμός |

Πίνακας 5.24: Ανωνυμοποιημένος πίνακας δεδομένων από τον αλγόριθμο Incognito

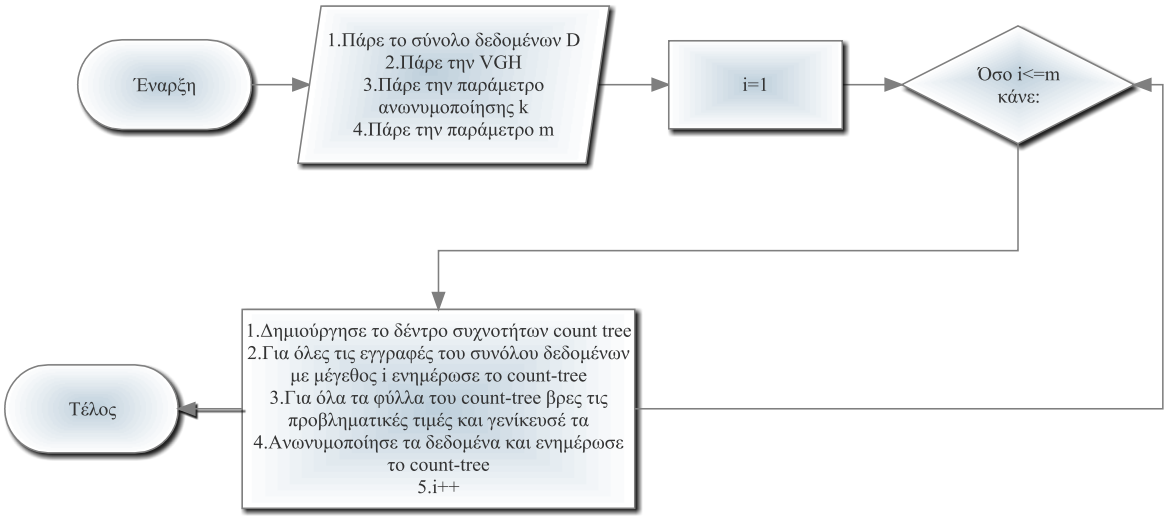
## *Apriori*

Οι καθηγητές Τεροβίτης, Μαμούλης και Καλνής ανέπτυξαν τον apriori [45] αλγόριθμο. Είναι ένας αποδοτικός και ευριστικός αλγόριθμος που εφαρμόζει $k^m$-ανωνυμία σε σύνολα δεδομένων. Εκμεταλλεύεται την αρχή της apriori ιδιότητας σύμφωνα με την οποία εάν ένα σύνολο D παραβιάζει την ιδιωτικότητα της βάσης, τότε το ίδιο θα συμβαίνει και για οποιοδήποτε υπερσύνολο του D.

Ως είσοδο λαμβάνει το σύνολο των δεδομένων D, την ιεραρχία γενίκευσης I, την παράμετρο ανωνυμοποίησης k και την παράμετρο γνώσης του αντιπάλου m. Αρχικά ελέγχει για παραβιάσεις ιδιωτικότητας υποθέτοντας ότι ο επιτιθέμενος γνωρίζει μόνο μια τιμή (i=1) από το σύνολο του ψευδο-αναγνωριστικού. Στη συνέχεια επαναλαμβάνει τον έλεγχο για 2 τιμές και συνεχίζει μέχρι να ελέγξει για m τιμές ( i=m ). Το πλεονέκτημα του συγκεκριμένου αλγόριθμου είναι ότι εκμεταλλεύεται τις γενικεύσεις που έγιναν στο βήμα i με αποτέλεσμα να μειώνεται ο αριθμός των γενικεύσεων στο επόμενο 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 στην ιεραρχία γενίκευσης τιμής, 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.

Ο αλγόριθμος σε διάγραμμα ροής:




Σχήμα 5.23: Διάγραμμα ροής Apriori

## *Anatomize*

Ο αλγόριθμος anatomize [34] ικανοποιεί τις απαιτήσεις της l-ποικιλομορφίας. Η πολυπλοκότητά του είναι ιδιαίτερα χαμηλή τόσο χρονικά, όσο και χωρικά.

Ο αλγόριθμος λαμβάνει ως είσοδο το πίνακα T, τα ψευδο-αναγνωριστικά και την παράμετρο $\ell$. Σε πρώτη φάση κατηγοριοποιεί τις εγγραφές με βάση το ευαίσθητο γνώρισμά τους. Η κάθε μια ομάδα δηλαδή θα ξεχωρίζει για την ευαίσθητη τιμή της και ονομάζεται $A^s$ ή αλλιώς κουτί. Στην συνέχεια κατασκευάζονται οι ομάδες που θα αποτελέσουν τις κλάσεις ισοδυναμίας. Όσο υπάρχουν τουλάχιστον $\ell$ στο πλήθος $A^s$ που δεν είναι κενά, επιλέγει τα $\ell$ μεγαλύτερα και εξάγοντας από καθένα από αυτά μια εγγραφή, κατασκευάζει την κλάση. Η διαδικασία επαναλαμβάνεται και κατασκευάζονται συνέχεια νέες κλάσεις μέχρι να έχουμε λιγότερα από $\ell$ $A^s$ μη κενά.

Όταν τελειώσει και αυτό το στάδιο θα έχουν μείνει το πολύ $\ell$ -1 $A^s$ μη κενά και αποδεικνύεται πως:

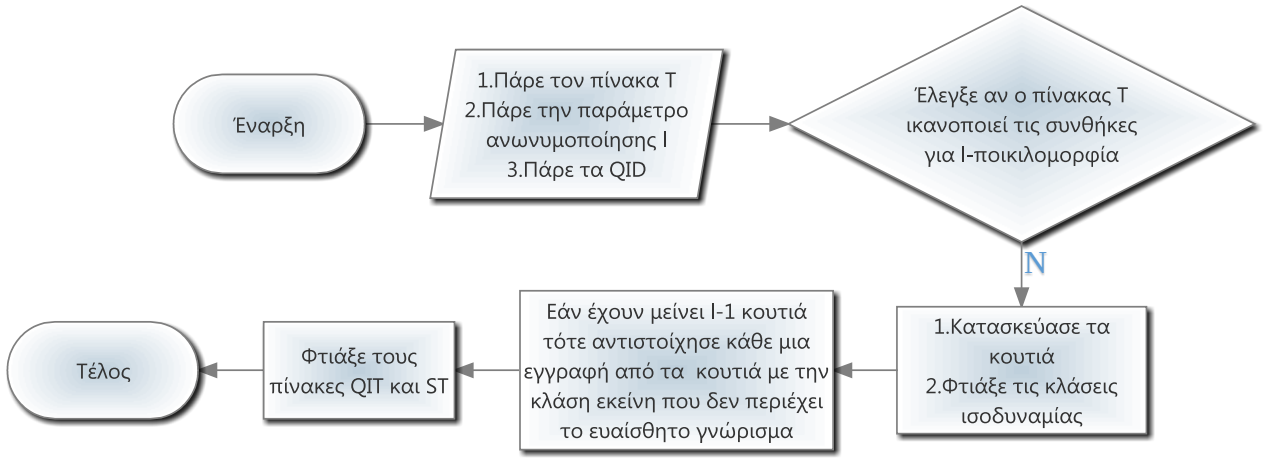
- κάθε ένα από τα $A^s$ έχει ακριβώς μια εγγραφή και
- για κάθε ένα από τα $A^s$ υπάρχει μία τουλάχιστον κλάση ώστε να μην περιέχει την ευαίσθητη πληροφορία η οποία αντιστοιχεί στο $A^s$.

Επομένως στο επόμενο στάδιο για κάθε μη κενό $A^s$ επιλέγουμε μία τυχαία κλάση που δεν έχει αυτή την τιμή και αντιστοιχούμε αυτήν την εγγραφή.

Στο τρίτο και τελευταίο στάδιο απλά κατασκευάζονται τα QIT και ST.

Για να εξάγει σωστό αποτέλεσμα και να επιτευχθεί η προστασία της ιδιωτικότητας θα πρέπει και ο πίνακας που θα λάβει ως είσοδο να είναι ικανός να γίνει l-ποικιλόμορφος. Για να ελεγχθεί αυτό αρκεί να στο τέλος του πρώτου σταδίου να υπάρχει $A^s$ με παραπάνω της μίας εγγραφής.

Ο αλγόριθμος σε διάγραμμα ροής:




Σχήμα 5.24: Διάγραμμα ροής Anatomize

# *Κεφάλαιο 6. Εργαλείο εύρεσης μοντέλου ανωνυμοποίησης*

Στα πλαίσια της παρούσας διπλωματικής αναπτύχθηκε εργαλείο για την εύρεση ενός κατάλληλου μοντέλου ανωνυμοποίησης μιας σχεσιακής βάσης δεδομένων. Το εργαλείο κάνει χρήση προσεκτικά επιλεγμένων ερωτήσεων και με βάση τις απαντήσεις που θα του δώσει ο χρήστης, επιστρέφει το βέλτιστο μοντέλο ανωνυμοποίησης. Η σωστή επιλογή μοντέλου είναι ιδιαίτερα σημαντική για να πετύχει ο ιδιοκτήτης των δεδομένων αρχικά την προστασία του απορρήτου των εγγραφών της βάσης και έπειτα να διασφαλίσει την ποιότητα των επεξεργασμένων πληροφοριών. Συγκεκριμένα το εργαλείο απευθύνεται σε τρεις κατηγορίες ανθρώπων:

i. Πρώτον σε εκείνους οι οποίοι δεν έχουν ούτε το γνωσιακό υπόβαθρο που απαιτείται αλλά ούτε και την πολυτέλεια του χρόνου να ασχοληθούν. Ως εκ τούτου τους παρέχεται η δυνατότητα να αναζητήσουν και να βρουν στον ελάχιστο δυνατό χρόνο ένα σωστό μοντέλο ανωνυμοποίησης που να εξυπηρετεί τις ανάγκες τους.

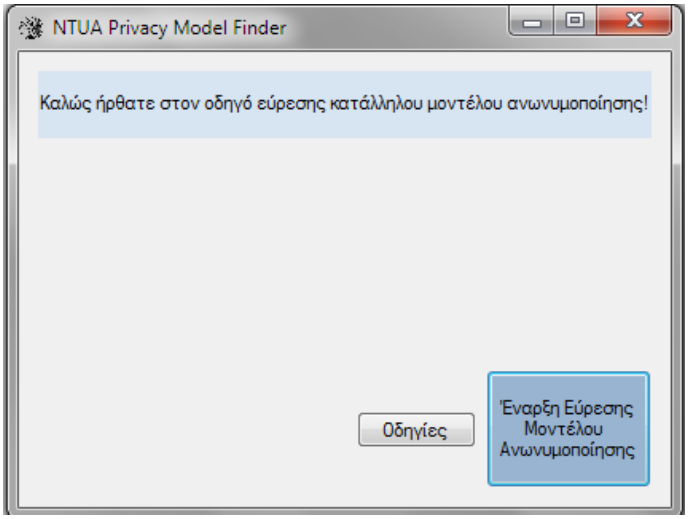
ii. Στους ιδιοκτήτες δεδομένων οι οποίοι έχουν ήδη από μόνοι τους επιλέξει ένα μοντέλο και επιθυμούν να πάρουν μία δεύτερη γνώμη για την επιλογή τους. Ανάλογα με το αποτέλεσμα που θα δώσει το εργαλείο, οι ιδιοκτήτες θα ξέρουν εάν είναι συνετό να συνεχίσουν ή να διερευνήσουν μια καλύτερη λύση.

iii. Στον οποιονδήποτε επιθυμεί να πάρει πληροφορίες και να εμπλουτίσει τις γνώσεις του στην ανωνυμοποίηση των προσωπικών δεδομένων.

Στην συνέχεια θα αναλυθούν αναλυτικά οι λειτουργίες του εργαλείου καθώς και οι τεχνικές λεπτομέρειες του.

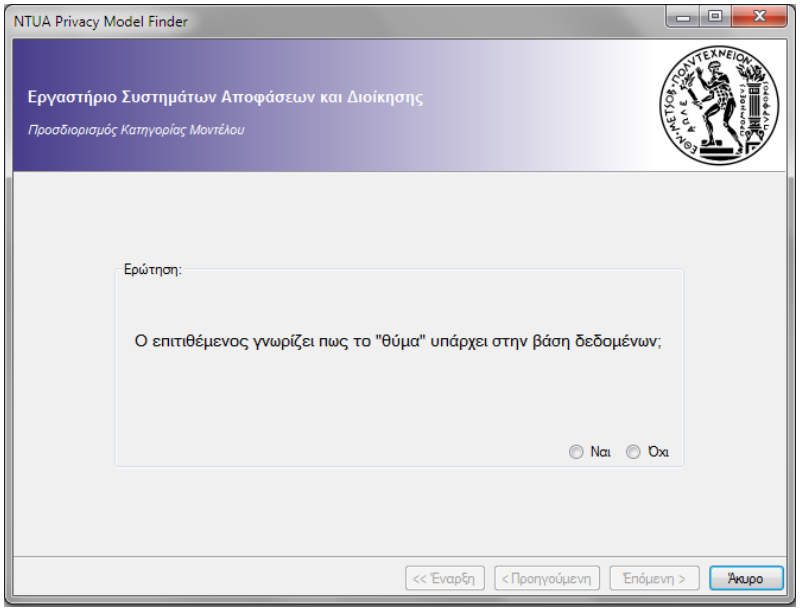
## *6.1 Περιγραφή λειτουργιών*

Στην έναρξη της εφαρμογής παρουσιάζεται ένα εύχρηστο, γραφικό περιβάλλον, το οποίο καλωσορίζει τον χρήστη:




Σχήμα 6.1: Περιβάλλον στην έναρξη εργαλείου

Για όσους χρήστες χρησιμοποιούν για πρώτη φορά την εφαρμογή, ή δεν έχουν προηγούμενη εμπειρία στον κλάδο, υπάρχει πλήκτρο με το οποίο ανοίγει pdf αρχείο που περιλαμβάνει συνοπτικές αλλά πλήρεις οδηγίες. Οι οδηγίες δεν αναφέρονται μόνο στο εργαλείο αλλά συνολικά στις σχεσιακές βάσεις δεδομένων και την ανωνυμοποίηση τους. Στην συνέχεια οι ερωτήσεις ξεκινούν με το πάτημα του διπλανού κουμπιού "Έναρξη Εύρεσης Μοντέλου ανωνυμοποίησης". Ενδεικτικά η πρώτη φόρμα που απεικονίζεται είναι η:




Σχήμα 6.2: Έναρξη ερωτήσεων

Παρακάτω γίνεται η αναφορά των ερωτήσεων του εργαλείου και οι ονομασίες page1,…,page15 προκύπτουν από τον τίτλο της κάθε κλάσης στο περιβάλλον υλοποίησης του εργαλείου και οι spage01,…spage18 αντίστοιχα για το μοντέλο ανωνυμοποίησης που προτείνεται στον χρήστη:

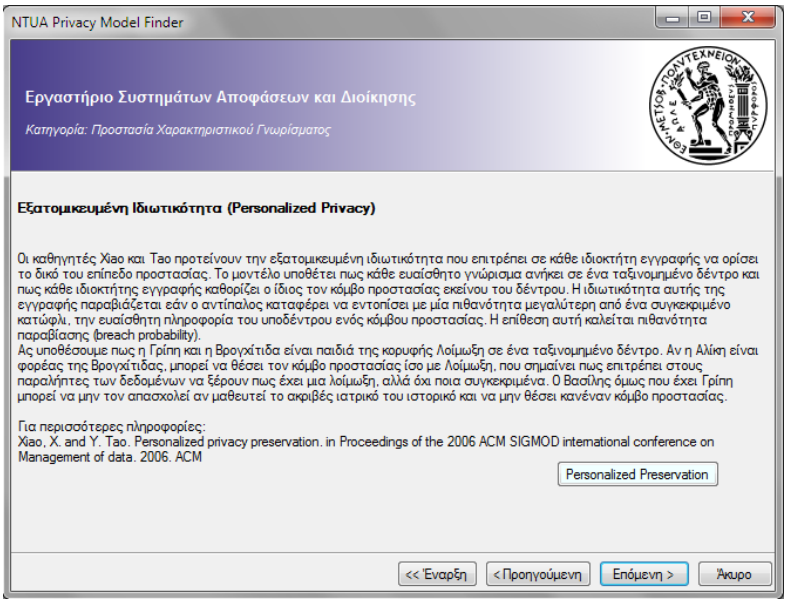
- **page01:** Ο επιτιθέμενος γνωρίζει πως το "θύμα" υπάρχει στην βάση δεδομένων;
  - Από την απάντηση θα ξέρουμε εάν το μοντέλο ανωνυμοποίησης ανήκει στην κατηγορία της προστασίας της παρουσίας της εγγραφής στον πίνακα ή όχι. Οι σελίδες page2α και page15 αφορούν τα μοντέλα αυτής της κατηγορίας.
- **page02α:** Η σχεσιακή βάση δεδομένων είναι στατιστική;
  - Εάν ναι τότε το προτεινόμενο μοντέλο είναι η ε-Διαφορική Ιδιωτικότητα **(spage01)**.
- **page15:** Έχουν όλες οι εγγραφές πιθανότητα να αποκαλυφθεί η παρουσία τους μικρότερη της σταθεράς d;
  - Εάν ναι τότε θα επιστρέψει το μοντέλο (d, g)-privacy **(spage03)** διαφορετικά το δ-Presence **(spage02).**
- **page02b:** Τα ευαίσθητα γνωρίσματα της βάσης περιλαμβάνουν πολλά διακριτά στοιχεία;
  - Εάν ναι τότε αρκεί να συνεχίσει με ερωτήματα για τα μοντέλα προστασίας εγγραφής. Οι σελίδες page3a και page{4 έως 10} ανήκουν στην κατηγορία αυτή. Διαφορετικά συνεχίζει στα μοντέλα προστασίας χαρακτηριστικού γνωρίσματος. Οι σελίδες page3b και page{11 έως 14} ανήκουν στην κατηγορία αυτή. Θα συνεχίσουμε με την κατηγορία προστασίας εγγραφής.
- **page03a:** Στον κάθε άνθρωπο που αφορά η βάση αντιστοιχεί μία εγγραφή;
  - Εάν η απάντηση είναι αρνητική τότε αναγκαστικά το βέλτιστο μοντέλο είναι η (X- Y)-Privacy **(spage05).**
- **page04:** Η δημοσίευση περιλαμβάνει πολλαπλούς πίνακες;
  - Εάν ναι τότε επιστρέφετε η MultiRelational k-Anonymity **(spage06).**
- **page05:** Τα περιεχόμενα του πίνακα θα ανανεώνονται συχνά;
  - Εάν ναι τότε επιστρέφετε η m-invariance **(spage09).**

- **page06:** Τα επιλεγμένα ψευδο-αναγνωριστικά περιλαμβάνουν ταυτόχρονα και ευαίσθητες πληροφορίες;
  - Εάν ναι τότε επιστρέφετε η FF-Anonymity **(spage10).**
- **page07:** Ο αντίπαλος γνωρίζει m τιμές της εγγραφής και είναι αδύνατο να προσδιοριστούν ποιες;
  - Εάν ναι τότε επιστρέφετε η $k^m$-Anonymity **(spage08).**
- **page08:** Τα ψευδο-αναγνωριστικά είναι με ακρίβεια γνωστά;
  - Εάν ναι τότε επιστρέφετε η k-Anonymity **(spage04).**
- **page09:** Τα γνωρίσματα της βάσης δεδομένων είναι όλα ή η πλειοψηφία τους αριθμητικά;
  - Εάν ναι τότε επιστρέφετε η (c, t)-isolation **(spage07).**
- **page10:** Τα ευαίσθητα γνωρίσματα είναι μόνο αριθμητικά και θέλετε να σας δίνεται η δυνατότητα να καθορίζεται την μέγιστη πιθανότητα παραβίασης προσωπικών πληροφοριών;
  - Εάν ναι τότε επιστρέφετε η (e, m)-anonymity **(spage11)** διαφορετικά επιστρέφεται το μοντέλο (k, e)-anonymity **(spage12).**
- **page03b:** Ο πίνακας δεδομένων είναι υπερβολικά μεγάλων διαστάσεων;
  - Εάν ναι τότε επιστρέφετε η LKC-privacy **(spage13).**
- **page11:** Υπάρχει η δυνατότητα για την κάθε εγγραφή του πίνακα να έρθετε σε επικοινωνία με τον ιδιοκτήτη της ώστε να καθορίσετε το επίπεδο της προστασίας του;
  - Εάν ναι τότε επιστρέφετε το μοντέλο Personalized Privacy Preservation **(spage14).**
- **page12:** Θέλετε για συγκεκριμένα ευαίσθητα γνωρίσματα να προσδιορίσετε το πόσο ασφαλή να είναι;
  - Εάν ναι τότε επιστρέφετε το μοντέλο Confidence bounding **(spage15).**
- **page13:** Θέλετε καθολικά να μπορείτε να προσδιορίσετε το μέγεθος των κλάσεων ισοδυναμίας και την πιθανότητα παραβίασης ευαίσθητων πληροφοριών;

- Εάν ναι τότε επιστρέφετε το μοντέλο (a, k)-anonymity **(spage16)**

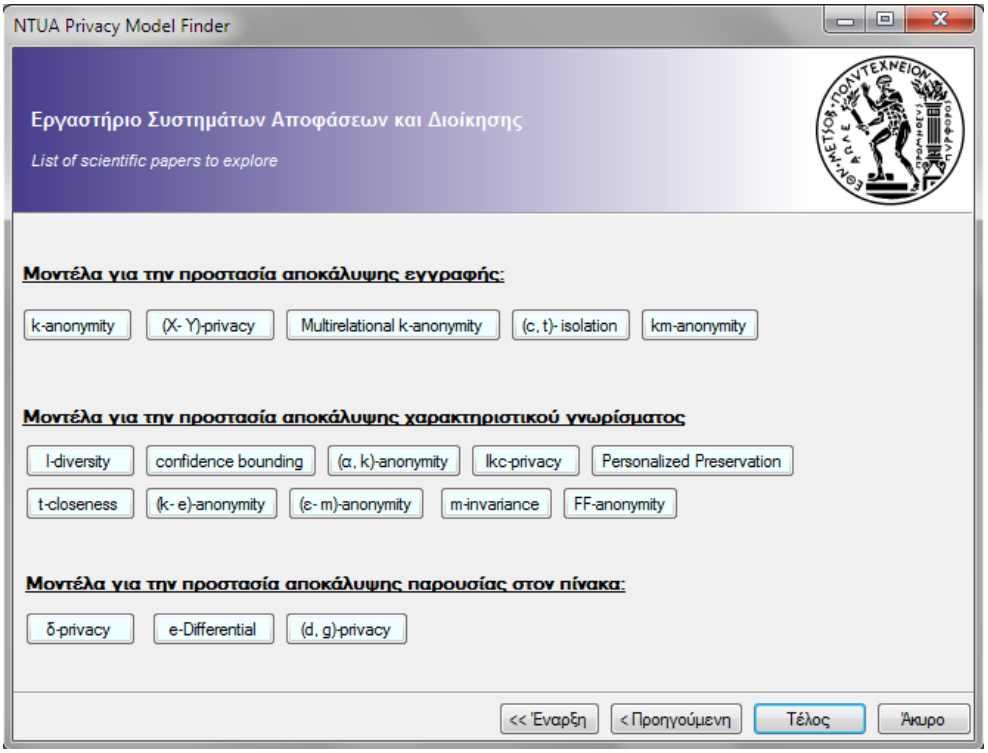
- **page14:** Οι συχνότητες εμφάνισης των ευαίσθητων τιμών έχουν μεγάλη απόκλιση;
  - Εάν ναι τότε επιστρέφετε η t-closeness **(spage17)** διαφορετικά επιστρέφεται το μοντέλο l-diversity **(spage18).**

Κάθε σελίδα spage{1-18} παρουσιάζει τον τίτλο του μοντέλου και μερικά λόγια για αυτόν. Εάν επιθυμεί ο χρήστης να αναζητήσει περισσότερες πληροφορίες η σελίδα περιλαμβάνει κουμπί το οποίο θα τρέξει pdf αρχείο του ερευνητικού paper που περιγράφει αναλυτικά το μοντέλο. Ένα ενδεικτικό παράδειγμα μιας σελίδας spage αποτελεί η παρακάτω:




Σχήμα 6.3: Παράδειγμα σελίδας αποτελέσματος

Στην τελευταία σελίδα της εφαρμογής και ανεξάρτητα από το αποτέλεσμα που θα προκύψει στον χρήστη εμφανίζεται η σελίδα finalpage. Εκεί περιλαμβάνονται και τα 18 μοντέλα που αναλύονται στην παρούσα διπλωματική μαζί με την βιβλιογραφία που τα συνοδεύει. Απευθύνεται στους χρήστες, οι οποίοι επιθυμούν να εξερευνήσουν τα υπόλοιπα μοντέλα ανωνυμοποίησης. Η σελίδα finalpage όπως παρουσιάστηκε:




Σχήμα 6.4: Σελίδα finalpage

## *6.2 Τεχνικές Λεπτομέρειες*

Για την ανάπτυξη του το εν λόγω εργαλείου χρησιμοποιήθηκε το Microsoft Visual Studio Community 2015 και η γλώσσα C-Sharp. Η εφαρμογή απαρτίζεται κυρίως από κλάσεις που εξυπηρετούν την διαδραστικότητα με τον χρήστη μέσω του γραφικού περιβάλλοντος. Συγκεκριμένα οι κλάσεις περιλαμβάνονται σε δύο projects, τα WizardFormLib και NtuaPrivacysolution. Η εισαγωγή του project WizardFormLib έγινε από την ιστοσελίδα codeproject [60]. Χάρη σε αυτήν την προσθήκη το εργαλείο μπορεί να επεκταθεί άμεσα και εύκολα έπειτα από οποιαδήποτε νέα δημοσίευση ή βελτίωση ενός μοντέλου ανωνυμοποίησης.

Συνολικά το project NtuaPrivacysolution περιλαμβάνει μία κλάση για την εισαγωγή του εργαλείου, 17 κλάσεις για τις ερωτήσεις, 18 κλάσεις για κάθε αποτέλεσμα ανωνυμοποίησης, μια κλάση για την τελευταία σελίδα καθώς επίσης και 19 αρχεία pdf.

Παρακάτω γίνεται μία αναλυτικότερη αναφορά της λειτουργίας της κάθε κλάσης:

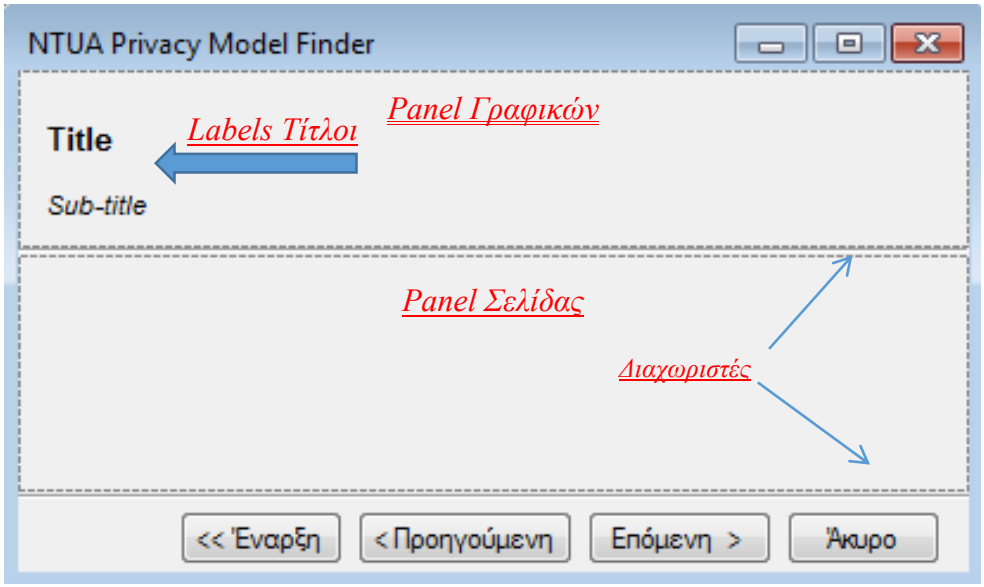
- **Κλάση WizardPageChain**: Διατηρεί την λίστα των σελίδων που έχει επισκεφτεί ο χρήστης, τις διαχειρίζεται και ορίζει την ορατότητά τους. Λειτουργεί όπως ακριβώς μια στοίβα και μας είναι απαραίτητη αφού το εργαλείο παρέχει με βάση την επιλογή των απαντήσεων πολλαπλούς δρόμους μέχρι το τελικό αποτέλεσμα. Όταν ο χρήστης επισκέπτεται μια σελίδα (όταν δηλαδή ο χρήστης ενεργοποιήσει το οδηγό ή όταν πατήσει το κουμπί *Επόμενο* >) τότε η σελίδα προστίθεται στο τέλος της λίστας και θεωρείται τρέχουσα. Εάν ο χρήστης πατήσει το κουμπί < *Προηγούμενη* τότε η τελευταία σελίδα αφαιρείται και την θέση της παίρνει η προηγούμενη σελίδα της λίστας. Στο πάτημα του κουμπιού << *Έναρξη* η κλάση έχει λειτουργία για την εμφάνιση της πρώτης σελίδας που βρίσκεται στην λίστα. Τελικά στην λίστα θα υπάρχουν όλες οι σελίδες που παρουσιάστηκαν και απεικονίζουν το μονοπάτι που ακολουθήθηκε.

- **Κλάση EventArgs:** Για την διευκόλυνση της αλληλεπίδρασης των σελίδων χρησιμοποιούνται προσαρμοσμένα events. Αφορούν τρεις βασικές λειτουργίες: την ενεργοποίηση, την αλλαγή και την δημιουργία μιας σελίδας.

- **Κλάση Enums:** Απλός ορισμός χαρακτηριστικών των σελίδων.

- **Κλάση WizardExceptions:** Για τον προσδιορισμό των προβλημάτων μέσα στον κώδικα και την γρήγορη διόρθωσή του.

- **Κλάση WizardFormBase:** Είναι η βασική κλάση που περιλαμβάνει και τον κορμό του κώδικα για την βάση των σελίδων. Συγκεκριμένα στην παρακάτω φόρμα απεικονίζεται η βάση στην οποία στηρίζονται οι υπόλοιπες σελίδες που δημιουργούνται. Περιλαμβάνει δύο Panels, δύο διαχωριστές, ένα ζευγάρι Label Τίτλων και τα τέσσερα κουμπιά για την περιήγηση. Το μέγεθός της εδώ δεν μας ενδιαφέρει τόσο, αφού στο τέλος το μέγεθος της θα εξαρτηθεί από τις υπόλοιπες σελίδες και συγκεκριμένα με αυτήν που διατηρεί το μεγαλύτερο. Για τον λόγο αυτό ορίζουμε το πάνω Panel από τις ιδιότητές του να είναι ‘docked’ στο πάνω μέρος του με την φόρμα και στο κάτω μέρος με τον πάνω διαχωριστή. Επίσης το Panel Σελίδας και τα κουμπιά είναι ‘docked’ στο δεξί μέρος.




Σχήμα 6.5: Πατέρας Σελίδων

**Panel Γραφικών:**


Σχήμα 6.6: Panel γραφικών

Στα πλαίσια της δυνατότητας το εργαλείο αυτό να επεκταθεί και οι σελίδες του να επεξεργαστούν η κλάση WizardFormBase παρέχει τις εξής επιλογές για την καθολική εφαρμογή τους:

- Ορισμός στερεού ή βαθμιδωτού χρώματος παρασκηνίου
- Εικόνα με δυνατότητα επιλογής θέσης (αριστερά, κέντρο, δεξιά). Η θέση των Label Τίτλων θα εξαρτηθεί από την τοποθέτηση της εικόνας με εξαίρεση την περίπτωση της τοποθέτησης της εικόνας στο κέντρο όπου και αποκρύπτονται.
- Αυτόματη εισαγωγή των Label Τίτλων από κάθε σελίδα για τον προσδιορισμό τους

Οι κυριότερες μέθοδοί της είναι:

i. PageCreated, είναι υπεύθυνη για την εμφάνιση των τεσσάρων κουμπιών για κάθε σελίδα που προστίθεται.

ii. DiscoverPanelSize, για την εύρεση της μεγαλύτερης σε μέγεθος σελίδας και ορισμός του μεγέθους σε default.

iii. StartWizard, για μερικούς ελέγχους ορθότητας και την διαμόρφωση μεγέθους του Panel με βάση το αποτέλεσμα της μεθόδου DiscoverPanelSize.

iv. UpdateWizardForm, για να ενημερώνει την κατάσταση των κουμπιών στη φόρμα με βάση την καθορισμένη σελίδα, το Panel Γραφικών και το κείμενο στο κουμπί Επόμενο.

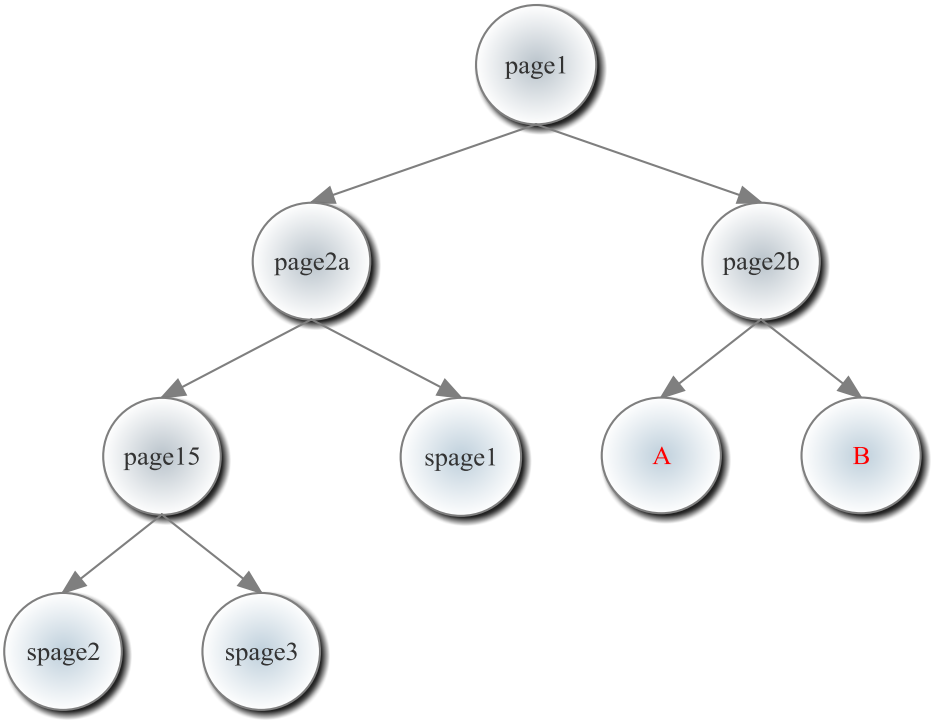
- **Κλάση WizardPage:** Η κλάση αρχικά περιλαμβάνει τρεις υπερφορτωμένους κατασκευαστές. Ο πρώτος είναι ο υπεύθυνος στην υποστήριξη του γραφικού Panel στις παράγωγες σελίδες. Οι άλλοι δύο κατασκευαστές παρέχουν υποστήριξη για τον καθορισμό και τον τύπο της σελίδας. Καλούν την μέθοδο Init η οποία είναι υπεύθυνη για την διαμόρφωση του στυλ της σελίδας, της αρχικής ορατότητας των στοιχείων, τον τύπο της σελίδας και προσθέτει έναν handler για το γεγονός της αλλαγής σελίδας. Οι υπόλοιπες μέθοδοι είναι ήσσονος σημασίας και αφορούν την λειτουργικότητα των σελίδων.

- **Κλάση questionBase:** Η κλάση είναι υπεύθυνη για την αρχικοποίηση των σελίδων-παράγωγων. Σε πρώτη φάση προσδιορίζονται επακριβώς οι κύριες ιδιότητες του Panel Γραφικών. Έπειτα προστίθενται handlers για τα event του πατήματος κουμπιού και της αρχικοποίησης σελίδων. Τέλος προσδιορίζεται για την κάθε σελίδα, ποιος είναι ο πατέρας, ποια τα παιδιά του και ξεκινάει το εργαλείο.

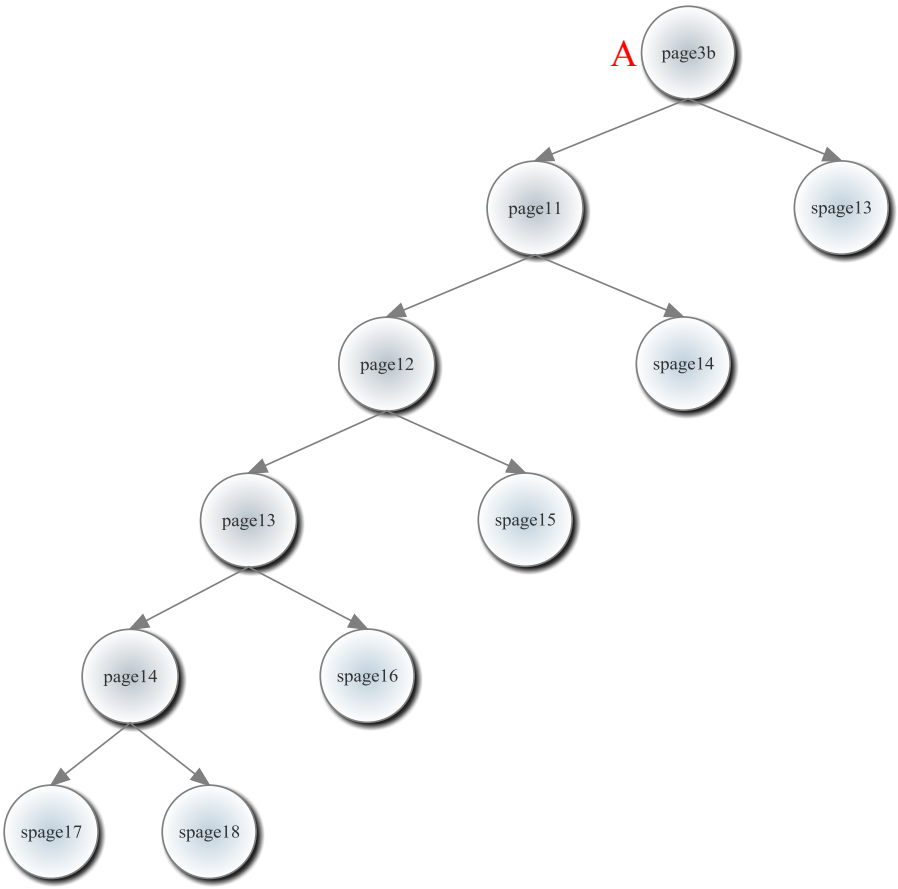
- **Κλάσεις σελίδων spage και page:**

  κάθε κλάση spage και page περιλαμβάνει οδηγίες για τον καθορισμό των Panel Τίτλων, τον ορισμό του κουμπιού *Επόμενο* > σε 'Enabled' (μέχρι να πατήσει ο χρήστης ένα radiobutton). Όλα τα γραφικά στοιχεία που βρίσκονται μέσα στο Panel Σελίδας της κάθε σελίδας προστίθενται μέσα σε groupbox και τοποθετούνται προγραμματιστικά στην μέση ακριβώς. Τέλος οι σελίδες που έχει επισκεφτεί ο χρήστης αποθηκεύονται ώστε να μπορεί να επιστρέψει σε αυτήν εάν το επιθυμεί.

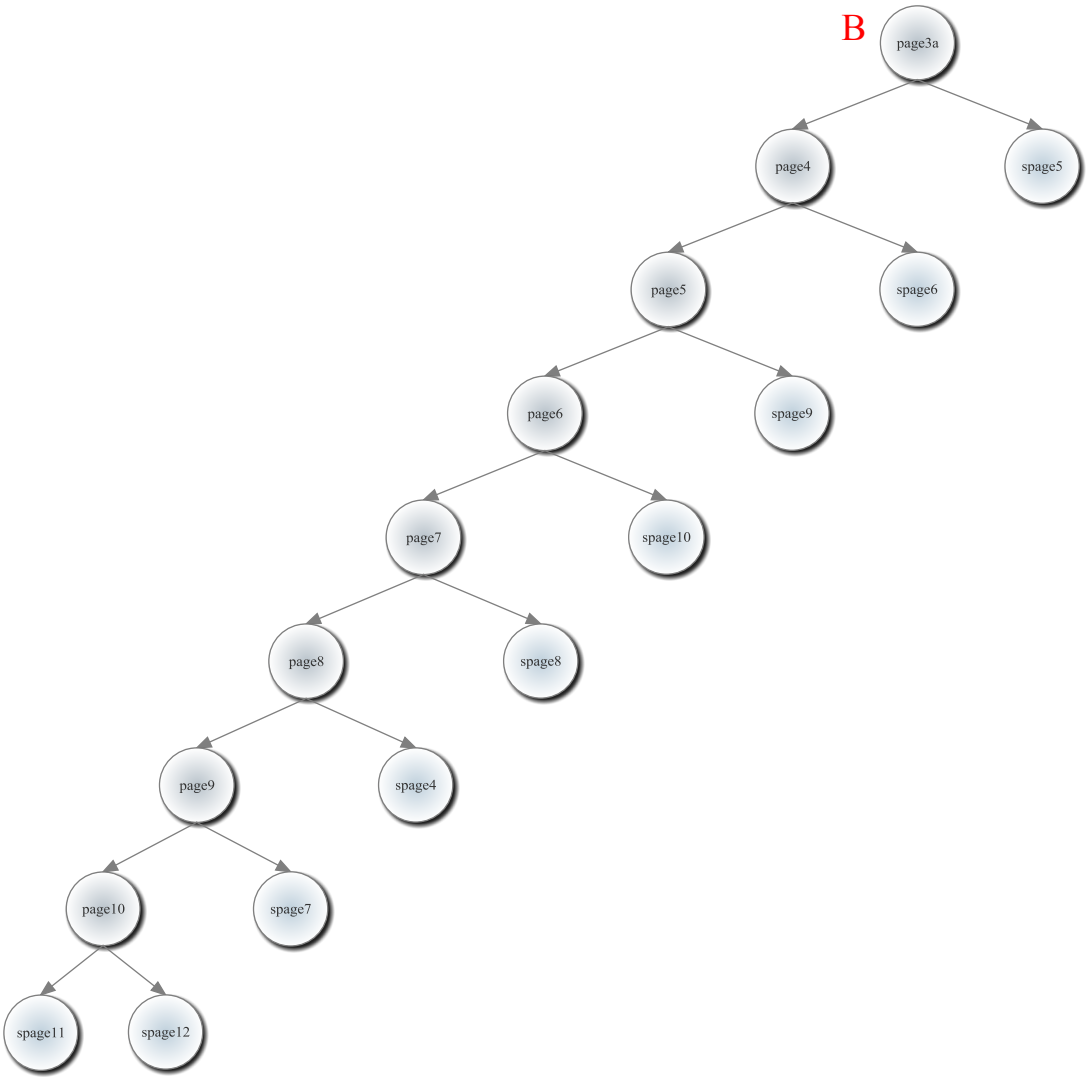
Η δομή των κλάσεων βάση τις ερωτήσεις που προβάλλει το εργαλείο απεικονίζεται γραφικά παρακάτω (λόγω της έκτασης των ερωτήσεων η δομή χωρίζεται σε 3 σχήματα):




Σχήμα 6.7: Δομή κλάσεων (Προστασία παρουσίας πίνακα)




Σχήμα 6.8: Δομή κλάσεων (Προστασία χαρακτηριστικού γνωρίσματος)




Σχήμα 6.9: Δομή κλάσεων (Προστασία εγγραφής)

# *Κεφάλαιο 7. Επίλογος*

## *7.1 Σύνοψη και συμπεράσματα*

Η παρούσα διπλωματική εργασία ασχολήθηκε με το πρόβλημα της διασφάλισης απορρήτου σε δεδομένα γράφων και σχεσιακών βάσεων δεδομένων. Η ανάγκη για δημοσίευση προσωπικών πληροφοριών με τρόπο ώστε να εξασφαλίζεται απόλυτα η ανωνυμία παραμένει ακόμα ανοιχτό ζήτημα. Ταυτόχρονα οι μέθοδοι απεικόνισης και αποθήκευσης προσωπικών δεδομένων συνεχώς εξελίσσονται.

Η αποθήκευση των δεδομένων στα κοινωνικά δίκτυα πραγματοποιείται με τη χρήση γράφων. Η έρευνα πάνω στην ανωνυμοποίησή τους βρίσκεται ακόμα σε πρώιμο στάδιο εξαιτίας της πολυπλοκότητας της δομής τους. Η ενημέρωση και επαγρύπνηση των χρηστών μπορεί να συμβάλει καθοριστικά στη διασφάλιση του απορρήτου.

Στις σχεσιακές βάσεις δεδομένων έχει προταθεί ένας μεγάλος αριθμός μοντέλων ανωνυμοποίησης. Καθένα από αυτά έχει μοναδικά χαρακτηριστικά, πλεονεκτήματα και μειονεκτήματα. Ο διαχωρισμός των μοντέλων που ακολουθήθηκε στην εργασία, προκύπτει από τους διαφορετικούς τρόπους αποκάλυψης των ευαίσθητων προσωπικών πληροφοριών. Αναλύθηκαν τέλος οι κυριότερες τεχνικές ανωνυμοποίησης, οι μετρικές απώλειας πληροφορίας και μερικοί από τους σημαντικότερους αλγορίθμους ανωνυμοποίησης.

Από τη θεωρία και τα παραδείγματα καταλήγουμε στο συμπέρασμα πως είναι ιδιαίτερα δύσκολο να αναπτυχθεί μια καθολική εφαρμογή ανωνυμοποίησης δεδομένων, η οποία να εφαρμόζεται αποδοτικά στον οποιονδήποτε πίνακα. Η επιλογή ενός μοντέλου ανωνυμοποίησης που ανταποκρίνεται στις απαιτήσεις και στις ιδιαιτερότητες του πίνακα, εξασφαλίζει τη χρησιμότητα πληροφοριών και την προστασία απορρήτου. Για το λόγο αυτό αναπτύχθηκε εργαλείο, το οποίο επιστρέφει το βέλτιστο μοντέλο ανωνυμοποίησης κάνοντας χρήση προσεκτικά επιλεγμένων ερωτήσεων.

## *7.2 Μελλοντικές επεκτάσεις*

Η κυριότερη και χρησιμότερη επέκταση που μπορεί να πραγματοποιηθεί είναι η ανάπτυξη ενός εργαλείου, το οποίο θα συγκεντρώνει όλα τα υπάρχοντα μοντέλα ανωνυμοποίησης. Το εργαλείο που αναλύθηκε στο κεφάλαιο 6 μπορεί να αποτελέσει τη βάση πάνω στην οποία θα αναζητείται και θα επιλέγεται, δεδομένων των συνθηκών, το καταλληλότερο μοντέλο ανωνυμοποίησης. Στην συνέχεια θα δέχεται ως είσοδο τον αρχικό πίνακα και τις παραμέτρους που απαιτούνται και θα το επιστρέφει ανωνυμοποιημένο. Παράλληλα υπάρχει η δυνατότητα ανάπτυξης μιας βελτιωμένης και καθολικής μετρικής απώλειας πληροφορίας, η οποία θα παρουσιάζει με γραφικό και απλοποιημένο τρόπο τα αποτελέσματα ποιότητας και απώλειας των ανωνυμοποιημένων δεδομένων. Σημαντική θα είναι επίσης η οποιαδήποτε ανάπτυξη ή ακόμα και βελτίωση ενός μοντέλου ανωνυμοποίησης, είτε αυτό αφορά τους γράφους, είτε τις σχεσιακές βάσεις δεδομένων.

# *Κεφάλαιο 8. Βιβλιογραφία*


1. Acquisti, A., L.K. John, and G. Loewenstein, What is privacy worth? The Journal of Legal Studies, 2013. 42(2): p. 249-274.

2. Xu, Y., et al., A survey of privacy preserving data publishing using generalization and suppression. Appl. Math, 2014. 8(3): p. 1103-1116.

3. Gehrke, J. Models and methods for privacypreserving data publishing and analysis. in Proceedings of the 22nd International Conference on Data Engineering (ICDE). 2006.

4. Co-operation, O.f.E. and Development, Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. 1981: OECD.

5. Dpa.gr. Νομοθεσία για τα προσωπικά δεδομένα - Ευρώπη. 2016 18/1/2016]; Available from: http://www.dpa.gr/portal/page?_pageid=33,123482&_dad=portal&_schema=PORTAL.

6. Dpa.gr. Νόμος 2472/1997. 2016; Available from: http://www.dpa.gr/portal/page?_pageid=33,19052&_dad=portal&_schema=PORTAL#19.

7. Καργιώτου, Α.Ε. and A.E. Kargiotou, Μελέτη σχεδιασμού ολοκληρωμένου Πληροφοριακού Συστήματος Γεωχωρικών και Κτηματολογικών Πληροφορίων Απαλλοτριώσεων αρμοδιότητας ΓΓΔΕ του ΥΠΥΔΕΔΙ-Θεσμικό Τεχνολογικό Πλαίσιο. 2014.

8. Ingram, W. and W. Mahavier, Mathematics and Computer Science. 2004: p. 121-196.

9. Sweeney, L., k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002. 10(05): p. 557-570.

10. Fung, B.C., et al., Introduction to privacy-preserving data publishing: concepts and techniques. 2010: CRC Press.

11. Witten, I.H., E. Frank, and M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques. 2011.

12. Raynes-Goldie, K.S., Privacy in the age of Facebook: Discourse, architecture, consequences. 2012: Curtin University.

13. Consumerreports.org. Facebook Privacy - Consumer Reports. [Accessed 10 Jan. 2016]]; Available from: http://www.consumerreports.org/cro/magazine/2012/06/facebook-your-privacy/index.htm

14. Goel, V. Facebook to Update Privacy Policy, but Adjusting Settings Is No Easier. [online] Bits Blog. Accessed 18 Jan. 2016]; Available from: http://bits.blogs.nytimes.com/2013/08/29/facebook-to-update-privacy-policy-but-adjusting-settings-is-no-easier/?_r=1

15. Yamada, A., T.H.-J. Kim, and A. Perrig, Exploiting privacy policy conflicts in online social networks. CMU-CyLab-12-005, Carnegie Mellon University, 2012.

16. Lucas, M.M. and N. Borisov. Flybynight: mitigating the privacy risks of social networking. in Proceedings of the 7th ACM workshop on Privacy in the electronic society. 2008. ACM.

17. Könings, B., et al. PrivacyJudge: effective privacy controls for online published information. in Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. 2011. IEEE.

18. Beato, F., M. Kohlweiss, and K. Wouters. Scramble! your social network data. in Privacy Enhancing Technologies. 2011. Springer.

19. Dingledine, R., N. Mathewson, and P. Syverson, Tor: The second-generation onion router. 2004, DTIC Document.

20. Zhou, B. and J. Pei. Preserving privacy in social networks against neighborhood attacks. in Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. 2008. IEEE.

21. Li, N., T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. in Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. 2007. IEEE.

22. Wang, D.-W., C.-J. Liau, and T.-s. Hsu. Privacy protection in social network data disclosure based on granular computing. in Fuzzy Systems, 2006 IEEE International Conference on. 2006. IEEE.

23. Hay, M., et al., Resisting structural re-identification in anonymized social networks. Proceedings of the VLDB Endowment, 2008. 1(1): p. 102-114.

24. Backstrom, L., C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. in Proceedings of the 16th international conference on World Wide Web. 2007. ACM.

25. Narayanan, A. and V. Shmatikov. De-anonymizing social networks. in Security and Privacy, 2009 30th IEEE Symposium on. 2009. IEEE.

26. Liu, K. and E. Terzi. Towards identity anonymization on graphs. in Proceedings of the 2008 ACM SIGMOD international conference on Management of data. 2008. ACM.

27. Blitzstein, J. and P. Diaconis, A sequential importance sampling algorithm for generating random graphs with prescribed degrees. Internet Mathematics, 2011. 6(4): p. 489-522.

28. Soria-Comas, J., et al. Improving the utility of differentially private data releases via k-anonymity. in Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on. 2013. IEEE.

29. Sweeney, L., Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002. 10(05): p. 571-588.

30. LeFevre, K., D.J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. in Proceedings of the 2005 ACM SIGMOD international conference on Management of data. 2005. ACM.

31. LeFevre, K., D.J. DeWitt, and R. Ramakrishnan. Workload-aware

anonymization. in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006. ACM.

32. Wang, K., B.C. Fung, and S.Y. Philip, Handicapping attacker's confidence: an alternative to k-anonymization. Knowledge and Information Systems, 2007. 11(3): p. 345-368.

33. Bayardo, R.J. and R. Agrawal. Data privacy through optimal k-anonymization. in Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on. 2005. IEEE.

34. Xiao, X. and Y. Tao. Anatomy: Simple and effective privacy preservation. in Proceedings of the 32nd international conference on Very large data bases. 2006. VLDB Endowment.

35. Zhang, Q., et al. Aggregate query answering on anonymized tables. in Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. 2007. IEEE.

36. Adam, N.R. and J.C. Worthmann, Security-control methods for statistical databases: a comparative study. ACM Computing Surveys (CSUR), 1989. 21(4): p. 515-556.

37. Reiss, S.P. Practical data-swapping: The first steps. in null. 1980. IEEE.

38. Samarati, P., Protecting respondents identities in microdata release. Knowledge and Data Engineering, IEEE Transactions on, 2001. 13(6): p. 1010-1027.

39. LeFevre, K., D.J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. in Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. 2006. IEEE.

40. Nergiz, M.E. and C. Clifton, Thoughts on k-anonymization. Data & Knowledge Engineering, 2007. 63(3): p. 622-645.

41. Dalenius, T., Towards a methodology for statistical disclosure control. Statistik Tidskrift, 1977. 15(429-444): p. 2-1.

42. Wang, K. and B. Fung. Anonymizing sequential releases. in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006. ACM.

43. Nergu, M.E., C. Clifton, and A.E. Nergu, MULTIRELATIONAL K-ANONYMITY. 2008.

44. Chawla, S., et al., Toward privacy in public databases, in Theory of Cryptography. 2005, Springer. p. 363-385.

45. Terrovitis, M., N. Mamoulis, and P. Kalnis, Privacy-preserving anonymization of set-valued data. Proceedings of the VLDB Endowment, 2008. 1(1): p. 115-125.

46. Machanavajjhala, A., et al., l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. 1(1): p. 3.

47. Wong, R.C.-W., et al. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006. ACM.

48. Mohammed, N., et al. Anonymizing healthcare data: a case study on the blood transfusion service. in Proceedings of the 15th ACM SIGKDD international conference

on Knowledge discovery and data mining. 2009. ACM.

49. Li, J., Y. Tao, and X. Xiao. Preservation of proximity privacy in publishing numerical sensitive data. in Proceedings of the 2008 ACM SIGMOD international conference on Management of data. 2008. ACM.

50. Xiao, X. and Y. Tao. Personalized privacy preservation. in Proceedings of the 2006 ACM SIGMOD international conference on Management of data. 2006. ACM.

51. Wang, K., et al. FF-anonymity: When quasi-identifiers are missing. in Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on. 2009. IEEE.

52. Xiao, X. and Y. Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. in Proceedings of the 2007 ACM SIGMOD international conference on Management of data. 2007. ACM.

53. Nergiz, M.E., M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. in Proceedings of the 2007 ACM SIGMOD international conference on Management of data. 2007. ACM.

54. Dwork, C., Differential privacy, in Encyclopedia of Cryptography and Security. 2011, Springer. p. 338-340.

55. Rastogi, V., D. Suciu, and S. Hong. The boundary between privacy and utility in data publishing. in Proceedings of the 33rd international conference on Very large data bases. 2007. VLDB Endowment.

56. Chen, B.-C., K. LeFevre, and R. Ramakrishnan. Privacy skyline: Privacy with multidimensional adversarial knowledge. in Proceedings of the 33rd international conference on Very large data bases. 2007. VLDB Endowment.

57. Sweeney, L. Guaranteeing anonymity when sharing medical data, the Datafly System. in Proceedings of the AMIA Annual Fall Symposium. 1997. American Medical Informatics Association.

58. Acharjya, D. and N.C.S. Iyengar, Improved Anonymization Algorithms for Hiding Sensitive Information in Hybrid Information System. International Journal of Computer Network and Information Security (IJCNIS), 2014. 6(6): p. 9.

59. Xu, J., et al., Utility-based anonymization for privacy preservation with less information loss. ACM SIGKDD Explorations Newsletter, 2006. 8(2): p. 21-30.

60. John, L.K. Wizard Form Implementation. 2010; Available from: http://www.codeproject.com/Articles/31770/Wizard-Form-Implementation.