ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ ΤΜΗΜΑ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΤΕΧΝΙΚΕΣ ΠΡΟΣΤΑΣΙΑΣ ΙΔΙΩΤΙΚΟΤΗΤΑΣ ΕΥΑΙΣΘΗΤΩΝ ΔΕΔΟΜΕΝΩΝ

ΚΑΡΑΒΑΝΑΣ ΘΕΟΦΑΝΗΣ

ΜΕΡΔΗ ΑΓΑΘΗ

ΕΠΟΠΤΕΥΩΝ ΚΑΘΗΓΗΤΗΣ: **ΚΩΝΣΤΑΝΤΙΝΟΣ ΣΤΑΜΟΣ**

ΠΑΤΡΑ 2016

Περιεχόμενα

Περίληψη

Με την αύξηση του πληθυσμού στις μέρες μας και την ποικιλία των προσωπικών πληροφοριών που διατίθενται σε διάφορες βάσεις δεδομένων, η ιδιωτικότητα των ανθρώπων έχει αρχίσει να παραβιάζεται.

Καθημερινά επιχειρήσεις, οργανισμοί και διάφοροι φορείς όπως νοσοκομεία, τράπεζες χρησιμοποιούν την ιδιωτική τους συλλογή από προσωπικά δεδομένα πελατών για ερευνητικούς σκοπούς, για στατιστικές μελέτες ακόμα και για προσωπική τους χρήση. Η διακίνηση τέτοιων επώνυμων πληροφοριών κλονίζει την ιδιωτικότητα των ατόμων αφήνοντας εκτεθειμένα ευαίσθητα προσωπικά δεδομένα.

Τον μεγαλύτερο ωστόσο κίνδυνο τον εγκυμονεί η δυνατότητα επεξεργασίας, σύγκρισης και ταυτοποίησης των πληροφοριών αυτών, που η σύγχρονη πληροφοριακή τεχνολογία διευκολύνει.

Για την διαφύλαξη της ιδιωτικής ζωής και των ευαίσθητων δεδομένων έχουν προταθεί μέθοδοι και τεχνικές που χρησιμοποιούν αλγορίθμους με σκοπό να γενικεύσουν και να ανωνυμοποιήσουν τα δεδομένα.

Η παρούσα πτυχιακή εργασία έχει ως θέμα τις τεχνικές προστασίας των ευαίσθητων δεδομένων, εστιάζεται στις τεχνικές και στους αλγορίθμους που έχουν δημιουργηθεί για το σκοπό αυτό και κυρίως στην πιο βασική μέθοδο ανωνυμοποίησης, στην k-ανωνυμία.

Λέξεις Κλειδιά:

k-ανωνυμία

l-διαφορετικότητα

αλγόριθμος ανωνυμοποίησης

προστασία ιδιωτικότητας

Abstract

Nowadays due to the increase of population and the variety of personal information which is available in various data, the privacy of people is being encroached.

Every day companies, organizations and various institutions such as hospitals and banks use their private collection of customers' personal data for research purposes, statical studies even for personal use. The distributions of this personal information undermine the privacy of individuals by exposing sensitive personal data.

However, the biggest risk is posed by the possibility of processing, comparison and identification of this data and modern information technology makes it easier.

Methods and techniques, using algorithms in order to generalize and anonymize data, have been proposed to protect the privacy of individuals and the sensitive data.

The issue of this thesis is about the techniques of data protection, focusing on the techniques and algorithms that have been created for this purpose especially in the most basic anonymization method, the k-anonymity.

Keywords:

k-anonymity

l-diversity

anonymization algorithm

privacy protection

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά

Ο ιδιωτικός βίος αποτελεί θεμελιώδες ανθρώπινο δικαίωμα, προϋποθέτει την ατομική ελευθερία και προστατεύεται νομικά από όλα τα δημοκρατικά συντάγματα.

Όμως σύμφωνα με το άρθρο της Χ. Ακριβοπούλου, το δικαίωμα στην προστασία των προσωπικών δεδομένων δεν ταυτίζεται με το δικαίωμα στην ιδιωτική ζωή του προσώπου. Παρόλα αυτά, η σχέση μεταξύ τους εμφανίζεται ως εγγενής στο βαθμό που το δικαίωμα στην προστασία των προσωπικών δεδομένων διαπλάθεται προκειμένου να αντιμετωπίσει τις καινοφανείς διακινδυνεύσεις που γεννά για την προστασία της ιδιωτικής ζωής, η σύγχρονη κοινωνία της πληροφορίας. (Ακριβοπούλου, 2011)

Τι ακριβώς όμως εννοούμε με τον όρο ιδιωτικότητα; Κατά καιρούς έχουν διατυπωθεί αρκετοί ορισμοί από ειδικούς που όλοι έχουν ένα κοινό, την έννοια της πρόσβασης, όπου πρόσβαση είναι η φυσική εγγύτητα σ' ένα πρόσωπο ή σε μια γνώση γι' αυτό το πρόσωπο.

Ζούμε σε μια εποχή της πληροφορίας και τα δεδομένα είναι ένα από τα νομίσματα της. Σχεδόν για κάθε άνθρωπο πάνω στη γη υπάρχει τουλάχιστον μια πληροφορία καταγεγραμμένη σε μία βάση δεδομένων από την στιγμή που γεννιέται. Με το πέρασμα των χρόνων, την ανάπτυξη της τεχνολογίας και με την αύξηση του πληθυσμού ολοένα και περισσότερα προσωπικά δεδομένα συλλέγονται σε ποικίλες βάσεις δεδομένων. Περισσότερες από 15.000 ειδικευμένες διαφημιστικές βάσεις δεδομένων περιέχουν δύο δισεκατομμύρια ονόματα καταναλωτών, μαζί με μια εκπληκτική ποσότητα ατομικών πληροφοριών. Ο μέσος αμερικάνος καταναλωτής βρίσκεται σε τουλάχιστον 25 διαφημιστικές λίστες. Οι διαφημιστικές βάσεις δεδομένων είναι μόνο η αρχή του παγόβουνου. Πολλές απ' αυτές τις λίστες οργανώνονται σύμφωνα με χαρακτηριστικά όπως εισόδημα, ηλικία, πολιτικές και θρησκευτικές απόψεις ακόμα και σεξουαλικές προτιμήσεις και όλες αυτές οι λίστες

αγοράζονται και πωλούνται καθημερινά. (Beekman & Quinn, 2010) Σε διάφορους, λοιπόν, τομείς από ιδιωτικούς και κοινωνικούς φορείς όπως τράπεζες, εταιρείες κινητής τηλεφωνίας, ακόμα και επιχειρήσεις, συλλέγονται και αποθηκεύονται αναλυτικά προσωπικές πληροφορίες των πελατών τους για διάφορους λόγους. Πληροφορίες πίστωσης και τραπεζικών συναλλαγών, φορολογικά, υγεία, ασφάλειες, πολιτικές συνεισφορές, ψηφοφορίες, αγορές με πιστωτική κάρτα, εγγραφές εγγύησης, συνδρομές σε περιοδικά, τηλεφωνικές κλήσεις, διαβατήρια, αεροπορικές πτήσεις, αγορές αυτοκινήτων, συλλήψεις και εξερευνήσεις στο Internet, όλα καταγράφονται σε υπολογιστές και έχουνε πολύ λίγο ή καθόλου έλεγχο επί των πραγμάτων που συμβαίνουν με τις περισσότερες απ' αυτές τις εγγραφές αφού συλλεχθούν.

Μερικές αντιπροσωπευτικές ιστορίες κατάχρησης τέτοιων βάσεων δεδομένων είναι οι εξής. Όταν τα μέλη του αμερικανικού Κογκρέσου ερεύνησαν τους δεσμούς μεταξύ του αδελφού του προέδρου Τζίμυ Κάρτερ, Μπίλυ, και της κυβέρνησης της Λιβύης, κατέληξαν σε μια αναφορά η οποία ανέφερε μεταξύ άλλων την ακριβή ώρα και θέση τηλεφωνικών κλήσεων μεταξύ του Billy Carter σε τρεις διαφορετικές πολιτείες. Οι τηλεφωνικές εγγραφές, οι οποίες αποκάλυψαν πολλά για τις δραστηριότητες του Μπίλυ Κάρτερ, αποκτήθηκαν από το τεράστιο δίκτυο υπολογιστών συλλογής δεδομένων της εταιρίας τηλεπικοινωνιών AT&T. Παρόμοιες πληροφορίες διατίθενται για όλους τους πελάτες εταιριών τηλεφωνίας. Ένα πιο πρόσφατο και πιο τυπικό παράδειγμα κλοπής ταυτότητας είναι όταν ένας απατεώνας κατάφερε να προωθήσει την αλληλογραφία ενός αθώου ανθρώπου σε δική του ταχυδρομική θυρίδα, ώστε να μπορεί εύκολα να συλλέγει αριθμούς πιστωτικών καρτών και άλλα προσωπικά δεδομένα. Μέχρι το θύμα να ανακαλύψει έναν εκκρεμή λογαριασμό της κάρτας Visa, ο κλέφτης είχε ήδη χρεώσει 42.000 δολάρια την κάρτα. Το θύμα δεν ήταν υπεύθυνο για τις χρεώσεις, αλλά χρειάστηκε πολύ χρόνο για να διορθώσει όλα τα λάθη της υπηρεσίας πίστωσης (Beekman & Quinn, 2010).

Όπως δείχνουν αυτά τα παραδείγματα, υπάρχουν πολλοί τρόποι για να χρησιμοποιήσει κάποιος τις βάσεις δεδομένων, ώστε να αφαιρέσει την ιδιωτικότητα κάποιου. Μερικές φορές οι παραβιάσεις ιδιωτικότητας οφείλονται σε κυβερνητικές δραστηριότητες παρακολούθησης. Άλλες φορές είναι το αποτέλεσμα του έργου ιδιωτικών επιχειρήσεων. Οι παραβιάσεις της ιδιωτικότητας μπορεί να είναι αθώα σφάλματα, στρατηγικές ενέργειες ή απατεωνιά. Η μεγάλη αύξηση των περιστατικών κλοπής ταυτότητας, η οποία έχει πολλά θύματα κάθε έτος, δείχνει ξεκάθαρα ότι η

τεχνολογία των βάσεων δεδομένων μπορεί να γίνει ένα πολύ ισχυρό εγκληματικό εργαλείο.

Όσο «παράνομη» και αν ακούγεται η διαδικασία δημοσίευσης πληροφοριών , κάποιες φορές είναι απαραίτητη η εξόρυξη τους καθώς και συμπερασμάτων που προκύπτουν από τη συλλογή και επεξεργασία επώνυμων πληροφοριών, που αφορούν συγκεκριμένο πληθυσμό ώστε να πραγματοποιηθούν έρευνες τόσο δημογραφικές όσο οικονομικές. Για παράδειγμα το Αμερικανικό Κέντρο Πληροφοριών Εγκλήματος (NCIC), το οποίο διοικείται από το FBI, περιέχει περισσότερες από 39 εκατομμύρια εγγραφές που σχετίζονται με κλεμμένα αυτοκίνητα, κλεμμένα ή απολεσθέντα όπλα, αγνοούμενους, καταζητούμενους, κατάδικους, ύποπτους τρομοκράτες κ.α.. Το NCIC παραχώρησε στο FBI τις πληροφορίες που ήθελε για να αναγνωρίσει τον James Earl Ray ως τον δολοφόνο του Δρ. Martin Luther King, Jr. Βοήθησε το FBI να εντοπίσει τον Timothy McVeigh, τον άνθρωπο που καταδικάστηκε για τον βομβαρδισμό του ομοσπονδιακού κτιρίου Alfred P. Murrah στην Οκλαχόμα. Επειδή το NCIC είναι προσβάσιμο από κρατικές και τοπικές υπηρεσίες επιβολής του νόμου, διευκολύνει περισσότερες από 100.000 συλλήψεις και την ανακάλυψη περισσοτέρων από 100.000 κλεμμένων αυτοκινήτων κάθε χρόνο (Beekman & Quinn, 2010).

Γι αυτό το λόγο δημιουργήθηκε η ανάγκη της προστασίας της ιδιωτικότητας, ώστε να γίνεται ορθή χρήση των προσωπικών πληροφοριών και να αποφευχθεί ταυτόχρονα η παραβίαση τους. Συνεπώς, με το πέρασμα των χρόνων έχουν διατυπωθεί ποικίλες έννοιες και τεχνικές προστασίας της ιδιωτικότητας των ευαίσθητων δεδομένων με σκοπό την διατήρηση της ισορροπίας ανάμεσα στην εκμετάλλευση των προσωπικών δεδομένων και το σεβασμό προς τα άτομα.

1.2 Αντικείμενο πτυχιακής

1.2.1 **Συνεισφορά**

Με σκοπό να αντιμετωπιστεί το πρόβλημα της παραβίασης της ιδιωτικότητας, που παρουσιάστηκε στην αρχή της διπλωματικής εργασίας, καταγράφηκαν και αναπτύχθηκαν διάφορες τεχνικές προστασίας ευαίσθητων προσωπικών δεδομένων.

Στη παρούσα εργασία θα αναπτυχθεί κυρίως η τεχνικής της $k-$ ανωνυμίας και $l-$ ποικιλομορφίας, καθώς όλοι οι μέθοδοι και οι αλγόριθμοι που χρησιμοποιούν για να επιτύχουν τη βέλτιστη ανωνυμοποίηση.

Η συνεισφορά της συγκεκριμένης εργασίας συνοψίζεται ως εξής:

1. Ερευνήθηκε και μελετήθηκε η κατάλληλη βιβλιογραφία με αντικείμενο την προστασία της ιδιωτικότητας των δημοσιευμένων δεδομένων με σκοπό την εύρεση των τεχνικών ανωνυμοποίησης για την του προβλήματος της εργασίας.
2. Καταγραφή και ανάπτυξη όλων των τεχνικών προστασίας ευαίσθητων δεδομένων.
3. Εκτενείς αναφορά στις τεχνικές της k-ανωνυμίας και της l-Diversity, παρουσίαση των ψευδοκωδικών των αλγορίθμων που χρησιμοποιούν καθώς και τις επιμέρους συμπληρωματικές τεχνικές. Ανάπτυξη αντίστοιχων παραδειγμάτων.
4. Καταγραφή συμπερασμάτων, λογισμικών ανωνυμοποίησης και παρουσίαση πιθανών μελλοντικών επεκτάσεων.

1.3 Οργάνωση κειμένου

Η δομή του κειμένου της παρούσας εργασίας παρουσιάζεται σύμφωνα με τα παρακάτω κεφάλαια:

Στο **δεύτερο** κεφάλαιο παρουσιάζεται και αναλύεται βιβλιογραφία που μελετήθηκε και έχει σχέση με την προστασία της ιδιωτικής ζωής πάνω σε μία βάση δεδομένων ,την ανωνυμοποίηση των πληροφοριών . Καθώς επίσης παρουσιάζονται οι τεχνικές προστασίας ευαίσθητων δεδομένων και οι αλγόριθμοι τους.

Στο **τρίτο** κεφάλαιο ορίζεται και γίνεται εκτενείς περιγραφή της ιδιότητας της k-ανωνυμίας . Διατυπώνεται με παραδείγματα η έννοια της k-ανωνυμίας, παρουσιάζονται οι μέθοδοι που χρησιμοποιεί, οι αλγόριθμοι που βοηθούν στην ανωνυμοποίηση καθώς και οι αδυναμίες που την χαρακτηρίζουν.

Στο **τέταρτο** κεφάλαιο παρουσιάζονται τεχνικές που διατυπώθηκαν με σκοπό να καλύψουν τα κενά της k-ανωνυμίας, όπως οι l-διαφορετικότητα (l-diversity), τ-εγγύτητα (t-closeness), δ-παρουσία (δ-presence). Διατυπώνονται οι μεθολογίες και οι αλγόριθμοι που χρησιμοποιούν καθώς επίσης οι αδυναμίες που παρουσιάζουν.

Στο **πέμπτο** κεφάλαιο γίνεται περιγραφή της πιο αποτελεσματικής και «μοντέρνας» τεχνικής, της km-ανωνυμίας, παρουσιάζονται οι ιδιότητες και οι αλγόριθμοι της και δίνονται παραδείγματα για την καλύτερη κατανόηση της μεθόδου.

Στο **έκτο** κεφάλαιο συνοψίζονται τα αποτελέσματα της πτυχιακής εργασίας γύρω από τις τεχνικές που έχουν διατυπωθεί ώστε να προστατεύονται οι επώνυμες πληροφορίες. Αναφέρονται τα διαθέσιμα λογισμικά ανωνυμοποίησης και προτείνονται πιθανές μελλοντικές επεκτάσεις της παρούσας εργασίας.

Κεφάλαιο 2

Προστασία προσωπικών δεδομένων-Εισαγωγικές έννοιες

2.1 Θεωρητικό υπόβαθρο

Με την πληροφορία να είναι η τροφή των νέων τεχνολογιών, η ανάγκη για προστασία και ασφάλεια των δεδομένων είναι πιο αναγκαία από ποτέ. Ένα από τα βασικά προβλήματα που απασχολούν ιδιαίτερα τους επιστήμονες στο χώρο της πληροφορικής είναι αυτό της παραβίασης της ιδιωτικότητας. Οι βάσεις δεδομένων ολοένα και γεμίζουν με πληροφορίες που αφορούν ανθρώπινες δραστηριότητες και χρησιμοποιούνται από κακόβουλους είτε για προσωπικούς είτε για ερευνητικούς λόγους.

2.2 Οδηγός προστασίας των δεδομένων για τον πολίτη

Μερικές φορές οι παραβιάσεις της ιδιωτικότητας που γίνονται με την βοήθεια των υπολογιστών είναι απλές ενοχλήσεις, μερικές φορές είναι απειλές κατά της ζωής, της ελευθερίας και της επιδίωξης της ευτυχίας. (Beekman & Quinn, 2010)

Ο καθένας μπορεί να λάβει κάποια μέτρα ώστε να προστατέψει εν μέρει τα ευαίσθητα προσωπικά του δεδομένα συνεπώς και την ιδιωτικότητα του. Παρακάτω παρουσιάζεται ένας σύντομος οδηγός προστασίας και ασφάλειας των δεδομένων και της ιδιωτικότητας. (Βασιλοπούλου, και συν., 2015)

1. Ο αριθμός ταυτότητάς σας είναι δικός σας. Μην τον μοιράζεστε. Επειδή ο αριθμός ταυτότητας είναι στοιχείο που σας αναγνωρίζει μοναδικά, μπορεί να χρησιμοποιηθεί για την συλλογή πληροφοριών για εσάς χωρίς τη δική σας άδεια ή γνώση.

2. Μην δίνετε πληροφορίες για τον εαυτό σας, μην απαντάτε σε προσωπικές ερωτήσεις επειδή ένα ερωτηματολόγιο ή κάποιος αντιπρόσωπος εταιρίας σας το ζητάει. Όταν συμπληρώνετε οποιοδήποτε έγγραφο (εγγύηση, έρευνα, λαχείο) σκεφτείτε αν θέλετε αυτές οι πληροφορίες να αποθηκευτούν στον υπολογιστή κάποιου άλλου.
3. Να ενημερώνεστε συχνά και λεπτομερώς για τα θέματα του διαδικτύου αλλά και γενικά της τεχνολογίας σε όποιο περιβάλλον και αν λειτουργούν.
4. Χρησιμοποιείτε πάντα πολύ ισχυρούς κωδικούς, αποτελούμενοι από σύμβολα, γράμματα(κεφαλαία και μικρά) και αριθμούς.
5. Χρησιμοποιείτε πάντα διαφορετικούς λογαριασμούς ηλεκτρονικού ταχυδρομείου για διαφορετικές εργασίες καθώς επίσης και ταυτοποίηση του χρήστη σε δύο επίπεδα.
6. Επιλέξτε το κατάλληλο anti-virus (προστασία από ιούς) καθώς μπορεί να είναι σωτήριο για τους υπολογιστές και τα δεδομένα τους.
7. Αποφεύγεται την αποθήκευση μεγάλου ιστορικού δεδομένων καθώς και πολλών δεδομένων σε υπηρεσίες cloud.
8. Ενδυνάμωση της ασφάλειας του wi-fi router (WPA2 με strong encryption και firewall).

2.3 **Εισαγωγή στις τεχνικές προστασίας προσωπικών δεδομένων**

Τα τελευταία χρόνια, η μετάβαση στην πληροφοριακή εποχή και η εξόρυξη δεδομένων δημιουργεί πλήθος κινδύνων για τον ιδιώτη, τον δημόσιο και τον ιδιωτικό τομέα λόγω της ταχείας διάδοσης των προσωπικών δεδομένων είτε σε ιστοσελίδες είτε σε περιορισμένο κύκλο. Αυτό έχει ως αποτέλεσμα την αύξηση της ανησυχίας σχετικά με την προστασία της ιδιωτικής ζωής. Τα τελευταία χρόνια, έχει προταθεί ένας αριθμός τεχνικών με σκοπό την τροποποίηση των δεδομένων ώστε να διατηρείτε το απόρρητο (V.S., Elmagarmid, Bertino, Saygin , & Dasseni, 2004).

Παρακάτω θα εξετάσουμε μια σειρά από τεχνικές, αφού κατά καιρούς έχουν γίνει προσπάθειες για την μεγιστοποίηση της ασφάλειας. Κάποιες από αυτές τις μεθόδους συνεχίζουν μέχρι και σήμερα να είναι αποτελεσματικές. Οι περισσότερες

τεχνικές για την προστασία της ιδιωτικής ζωής χρησιμοποιούν κάποια μορφή μετασχηματισμού των δεδομένων, προκειμένου να γενικεύσουν ή να αποκρύψουν πληροφορίες και να διαφυλάξουν την ιδιωτικότητα. Σε κάποιες περιπτώσεις αυτή η απόκρυψη δεδομένων είχε ως αποτέλεσμα την αλλοίωση της πληροφορίας και την απώλεια της αποτελεσματικότητας τους. Αυτή είναι εξάλλου η φυσική εξισορρόπηση μεταξύ της απώλειας των πληροφοριών και της προστασίας της ιδιωτικής ζωής. Μερικά παραδείγματα τέτοιων τεχνικών είναι τα ακόλουθα:

1. **Μέθοδος τυχαίας επιλογής:** Η *μέθοδος τυχαίας επιλογής* (randomization method) είναι μία τεχνική εξόρυξης δεδομένων για τη διατήρηση της ιδιωτικότητας στην οποία ο θόρυβος προστίθεται στα δεδομένα με σκοπό την συγκάλυψη των τιμών των χαρακτηριστικών των εγγραφών (R. & Srikant, 2000) (D. & Aggarwal , 2001). Ο θόρυβος που προστίθεται είναι αρκετά μεγάλος έτσι ώστε οι τιμές των εγγραφών να μην μπορούν να ανακτηθούν. Επομένως, οι τεχνικές έχουν δημιουργηθεί με σκοπό την άντληση ενός συνονθυλεύματος κατανεμημένων δεδομένων από τις διαταραγμένες εγγραφές, στα οποία μπορούν στη συνέχεια να εφαρμοστούν οι τεχνικές εξόρυξης γνώσης.
2. **Το μοντέλο της *k*-ανωνυμίας και της *l*-ποικιλομορφίας:** Το *μοντέλο της k-ανωνυμίας* (*k*-anonymity model) δημιουργήθηκε για τη πιθανότητα της μη άμεσης ταυτοποίησης των εγγραφών σε δημόσιες βάσεις δεδομένων. Πιο συγκεκριμένα, το συγκεκριμένο μοντέλο εξασφαλίζει πως ο επιτιθέμενος δεν θα καταφέρει να προσδιορίσει μοναδικά μια πληροφορία σε μία βάση δεδομένων αφού θα υπάρχει άλλη μία εγγραφή τουλάχιστον με τις ίδιες ακριβώς τιμές. Στη *μέθοδο της k-ανωνυμίας* μειώνουμε τη διακριτότητα των δεδομένων με τη χρήση τεχνικών όπως η γενίκευση (generalization), η απόκρυψη (suppression). Μ’ αυτόν τον τρόπο η διακριτότητα μειώνεται αρκετά έτσι ώστε κάθε εγγραφή να αντιστοιχεί σε τουλάχιστον k-1 άλλες εγγραφές στο σύνολο δεδομένων. Το μοντέλο της *l*-ποικιλομορφίας σχεδιάστηκε με σκοπό να χειρίζεται τις αδυναμίες του μοντέλου της *k*-ανωνυμίας εφόσον η σημαντικότητα της προστασία των ταυτοτήτων των *k*-ατόμων δεν είναι η ίδια με την προστασία των ευαίσθητων τιμών, ιδιαίτερα όταν υπάρχει ομοιογένεια μεταξύ των αντίστοιχων ευαίσθητων τιμών μέσα στο ίδιο σύνολο. (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2007).

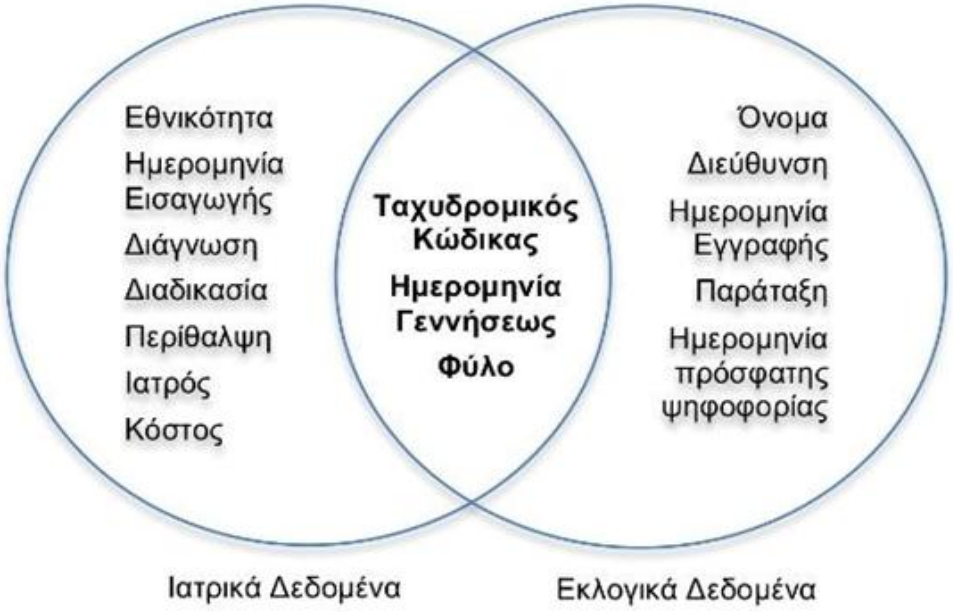
3. **Κατανεμημένη διατήρηση της ιδιωτικότητας (distributed privacy preservation):** Σε πολλές περιπτώσεις, οι ατομικές οντότητες μπορεί να επιθυμούν την εξαγωγή συγκεντρωτικών αποτελεσμάτων από τα σύνολα δεδομένων, τα οποία κατηγοριοποιούνται μέσω αυτών των οντοτήτων. Αυτή η κατηγοριοποίηση μπορεί να είναι οριζόντια (αν οι εγγραφές κατανέμονται σε πολλαπλές οντότητες) ή κάθετη (αν οι εγγραφές διανέμονται σε πολλαπλά χαρακτηριστικά). Όταν οι ατομικές οντότητες δεν επιθυμούν να μοιραστούν ολόκληρα τα σύνολα δεδομένων τους, διαμοιράζουν περιορισμένες πληροφορίες με τη χρήση μίας ποικιλίας πρωτοκόλλων. Το συνολικό αποτέλεσμα μίας τέτοιας μεθόδου είναι η διατήρηση της ιδιωτικότητας για κάθε ατομική οντότητα ενώ διατηρούνται και τα συγκεντρωτικά αποτελέσματα ολόκληρου του συνόλου δεδομένων.

4. **Υποβαθμισμός της εφαρμογής της αποτελεσματικότητας (downgrading Application Effectiveness):** Σε πολλές περιπτώσεις, ακόμα και διαμέσου των δεδομένων, τα οποία δεν είναι διαθέσιμα, το αποτέλεσμα των εφαρμογών όπως η εξόρυξη κανόνων συσχέτισης (association rule mining) ή η ταξινόμηση (classification) μπορεί να οδηγήσει στη παραβίαση της ιδιωτικότητας. Για αυτό το σκοπό, η έρευνα έχει επικεντρωθεί στην υποβάθμιση της αποτελεσματικότητας των εφαρμογών αυτών με την τροποποίηση των δεδομένων ή ακόμα και των ίδιων εφαρμογών. Παραδείγματα τέτοιων εφαρμογών περιλαμβάνουν τον ταξινομητή υποβάθμισης (classifier downgrafing) (I. & Chang , 2000) ή την απόκρυψη κανόνων συσχετίσεων (association rule hiding) (V.S., Elmagarmid, Bertino, Saygin , & Dasseni, 2004).

Στη παρούσα εργασία θα αναπτυχθεί εκτενέστερα η τεχνικής της $k-$ ανωνυμίας και $l-$ποικιλομορφίας, δύο από τις πιο διαδεδομένες και αποτελεσματικές μεθόδους. Θα γίνει αναφορά και σε μεταγενέστερες τεχνικές και μεθόδους καθώς και στην km-ανωνυμία μία εξίσου αποδοτική και δημοφιλή τεχνική ανωνυμοποίησης.

2.4 **Εισαγωγή στη προστασία της ιδιωτικότητας**

Σύμφωνα με μελέτη που έγινε στο πανεπιστήμιο του Harvard το 2000, το 87% του πληθυσμού των Ηνωμένων Πολιτειών της Αμερικής, (Sweeney, Uniqueness of Simple Demographics in the U.S. Population, 2000) μπορεί να προσδιοριστεί εύκολα με την χρήση κάποιας επώνυμης πληροφορίας όπως ηλικία, ταχυδρομικός κώδικας, φύλο. Επιπλέον στην ίδια μελέτη αναφέρεται πως το 53% του πληθυσμού μπορεί να προσδιοριστεί και από πιο ασήμαντες και γενικές πληροφορίες, τα ψευδοαναγνωριστικά.



Εικόνα 1: Εισαγωγή δεδομένων από διαφορετικά σύνολα εγγραφών

Στο παραπάνω παράδειγμα που δημοσιεύτηκε σε άρθρο της Sweeney το 2002,παρατηρούμε πόσο εύκολο είναι για κάποιον εισβολέα να εκτελέσει μια επίθεση γνωρίζοντας τα στοιχεία από δυο βάσεις δεδομένων. (L.Sweeney, 2002) Έστω ότι ο κακόβουλος τρίτος συνδυάσει τις πληροφορίες από μία βάση που αφορούσε ιατρικά δεδομένα που είχαν καταγραφεί σ ένα πρόγραμμα υγείας μιας ασφαλιστικής και μια δεύτερη βάση δεδομένων με εκλογικά στοιχεία των πολιτών από δημόσιους καταλόγους, μπορεί εύκολα να εντοπίσει τον ιατρικό φάκελο της διευθύντριας του στην δουλειά γνωρίζοντας μόνο την ημερομηνία γέννησης, το φύλο και το Τ.Κ. της περιοχής της. Πιο συγκεκριμένα η διευθύντρια ζούσε στην Αθήνα, σύμφωνα με τους

καταλόγους ψηφοφορίας 7 άτομα είχαν το ίδιο ταχυδρομικό κώδικα, μόνο τέσσερα ήταν γεννημένα την ίδια ημερομηνία και μόνο η διευθύντρια ήταν γυναίκα.

Ακόμα και όταν δημοσιεύονται βάσεις δεδομένων με ελλιπής στοιχεία πολιτών για διάφορους λόγους μπορεί εύκολα να προσδιοριστούν μοναδικά τα άτομα αφού σίγουρα υπάρχουν άλλες πληροφορίες τους δημοσιευμένες σε κάποια άλλη βάση. Έτσι ο επιτιθέμενος συνδυάζοντας και απορρίπτοντας πληροφορίες εντοπίζει εύκολα τον στόχο του.

Κεφάλαιο 3

Η τεχνική της k-ανωνυμίας

3.1 $k-$ ανωνυμία

Η $k-$ ανωνυμία αποτελεί μία από τις πιο γνωστές λύσεις για το πρόβλημα της παραβίασης της ευαίσθητης προσωπικής ζωής. Είναι μια ιδιότητα που κατέχεται από ορισμένα ανώνυμα στοιχεία. Έχε προταθεί ως μια προσέγγιση για την προστασία των ταυτοτήτων αποβάλλοντας αληθείς πληροφορίες. Η έννοια της $k-$ ανωνυμίας διατυπώθηκε για πρώτη φορά από την Latanya Sweeney σ' ένα έγγραφο που δημοσιεύτηκε το 2002, σε μία προσπάθεια επίλυσης προβλήματος παραβίασης ευαίσθητων προσωπικών δεδομένων (L.Sweeney, 2002). Η μέθοδος $k-$ ανωνυμίας έχει την ιδιότητα ότι κάθε εγγραφή είναι δυσδιάκριτη από τουλάχιστον $k-1$ εγγραφές όπου η τιμή k αντανακλά το βαθμό του επιπέδου προστασίας των προσωπικών δεδομένων. Με άλλα λόγια μια βάση δεδομένων έχει ανωνυμοποιηθεί εάν οι πληροφορίες κάθε εγγραφής δεν μπορούν να διακριθούν από τουλάχιστον k-1 άτομα των οποίων οι πληροφορίες είναι επίσης καταγεγραμμένες.

Οι διάφορες διαδικασίες και τα προγράμματα για τη δημιουργία ανώνυμων δεδομένων με την χρήση της k-ανωνυμίας έχουν καταχωρηθεί με δίπλωμα ευρεσιτεχνίας στις Ηνωμένες Πολιτείες (Sweeney, Systems and methods for de-identifying entries in a data source, retrieved 19 january 2014) .

Ορισμός. (Ψευδοαναγνωριστικό σύνολο). Το ελάχιστο σύνολο από γνωρίσματα $Q = Q_1, Q_2, \dots, Q_N$ με το οποίο ένας πίνακας T μπορεί να συζευχθεί με κάποιες εξωτερικές πληροφορίες για να προσδιοριστούν ατομικές εγγραφές ονομάζεται *ψευδό-αναγνωριστικό σύνολο*.

Ορισμός ($k-$ ανώνυμος πίνακας). Ένας πίνακας Τ θα λέμε ότι είναι k-ανώνυμος με βάση ένα σύνολο γνωρισμάτων $Q = Q_1, Q_2, \dots, Q_N$ αν το μέγεθος κάθε κλάσης ισοδυναμίας του πίνακα T με βάση τα $Q_1, Q_2, \dots, Q_N$έχει πληθάριθμο τουλάχιστον k.

3.2 Μέθοδος $k-$ ανωνυμίας

Στο πλαίσιο των προβλημάτων k-ανωνυμίας, μια βάση δεδομένων είναι ένας πίνακας με n γραμμές και m στήλες. Κάθε γραμμή του πίνακα αντιπροσωπεύει μια εγγραφή που σχετίζεται με έναν συγκεκριμένο αριθμό του πληθυσμού. Οι εγγραφές των διαφόρων γραμμών δεν χρειάζεται να είναι μοναδικές και οι τιμές στις διάφορες στήλες είναι οι τιμές των χαρακτηριστικών που σχετίζονται με τα μέλη του πληθυσμού.

Σύμφωνα με τους παραπάνω ορισμούς , ορίζεται ότι ένας προς δημοσίευση πίνακας RT(A_1,A_2,…,A_n) είναι k-ανώνυμος, αν κάθε εγγραφή του πίνακα είναι ίδια ως προς τα ψευδο-αναγνωριστικά QI_{RT} = (A_1,A_2,…,A_i) πεδία του, με k-1 άλλες εγγραφές. το σύνολο των γνωρισμάτων A_1,A_2,…,A_i ικανοποιεί την k- ανωνυμία (k-anonymity) αν κάθε ακολουθία τιμών στον πίνακα RT[QI_{RT}] του ψευδο-αναγνωριστικού εμφανίζεται τουλάχιστον k φορές (L.Sweeney, 2002). Με αυτό τον τρόπο το έργο ενός κακόβουλου γίνεται αρκετά δύσκολο και δεν θα καταφέρει να προσδιορίσει εύκολα ένα άτομο. Όταν τα δημοσιευμένα δεδομένα ικανοποιούν την k-ανωνυμία, για κάθε συνδυασμό τιμών στα γνωρίσματα του ψευδοαναγνωριστικού θα υπάρχουν το λιγότερο k εγγραφές που θα τον περιέχουν. Ένας πίνακας που ικανοποιεί την k-ανωνυμία αποτελείται από κλάσεις ισοδυναμίας (equivalence class). Αυτό σημαίνει πως σ' ένα σύνολο ατόμων με συγκεκριμένα γνωρίσματα, μετά την ανωνυμοποίηση εμφανίζονται ταυτόσημες τιμές στα ψευδοαναγνωριστικά. Επομένως, ο επιτιθέμενος πάνω σε τέτοια δεδομένα δεν μπορεί με βεβαιότητα να προσδιορίσει το θύμα αφού θα εμφανίζονται άλλες τουλάχιστον k εγγραφές με τα ίδια χαρακτηριστικά.

Οι περιπτώσεις που τα δεδομένα σ' έναν πίνακα ικανοποιούν την k-ανωνυμία στην αρχική μορφή είναι σπάνιες επομένως ο τομέας της πληροφορικής για να λύσει

το πρόβλημα της παραβίασης της ιδιωτικής ζωής έχει αναπτύξει τεχνικές και αλγορίθμους με σκοπό να διαφοροποιούνται τα οι τιμές των δεδομένων με τέτοιο τρόπο ώστε να ικανοποιείται η k-ανωνυμία. Από αυτές τις διαδικασίες συνήθως προκύπτει ένας νέος διαφοροποιημένος πίνακας ο RT'(A1,A2,…,An) στον οποίο αποκρύπτονται ή αλλάζουν δεδομένα και ονομάζεται k-ανωνυμοποίηση του αρχικού RT(A1,A2,…,An).

Ο παρακάτω πίνακας είναι μια βάση δεδομένων που αποτελείται από φανταστικά στοιχεία ασθενών κάποιου νοσοκομείου.

Ιατρικά Δεδομένα Ασθενών			
ΦΥΛΛΟ	ΧΡΟΝΟΛΟΓΙΑ ΓΕΝΝΗΣΗΣ	Τ.Κ.	ΑΣΘΕΝΕΙΑ
ΑΡΡΕΝ	1976	30115	ΚΑΡΚΙΝΟΣ
ΘΥΛΗ	1986	30115	ΑΜΥΓΔΑΛΙΤΙΔΑ
ΑΡΡΕΝ	1976	30103	ΓΡΙΠΗ
ΑΡΡΕΝ	1976	30103	ΗΠΑΤΙΤΙΔΑ
ΘΥΛΗ	1986	30106	ΟΙΔΗΜΑ
ΑΡΡΕΝ	1986	30106	ΒΡΟΓΧΙΤΙΔΑ

Πίνακας 1: Στοιχεία ιατρικού οργανισμού

2-ΑΝΩΝΥΜΟΣ ΠΙΝΑΚΑΣ			
Ιατρικά Δεδομένα Ασθενών			
ΦΥΛΛΟ	*ΧΡΟΝ. ΓΕΝΝΗΣΗΣ*	*Τ.Κ.*	*ΑΣΘΕΝΕΙΑ*
ΑΡΡΕΝ	197*	30115	ΚΑΡΚΙΝΟΣ
ΘΥΛΗ	198*	30115	ΑΜΥΓΔΑΛΙΤΙΔΑ
ΑΡΡΕΝ	197*	30103	ΓΡΙΠΗ
ΑΡΡΕΝ	1976	3010*	ΗΠΑΤΙΤΙΔΑ
ΘΗΛΥ	198*	30106	ΟΙΔΗΜΑ
ΑΡΡΕΝ	1986	3010*	ΒΡΟΓΧΙΤΙΔΑ

Πίνακας 2: Στοιχεία ιατρικού οργανισμού σε 2-Ανωνυμία

Για την καλύτερη κατανόηση της μεθόδου της k-Ανωνυμίας, χρησιμοποιείται το παράδειγμα του Πίνακα 1. Από αυτόν προκύπτει ο πίνακας 2, οποίος έχει επεξεργαστεί για να ικανοποιεί την k-Ανωνυμία (k=2).

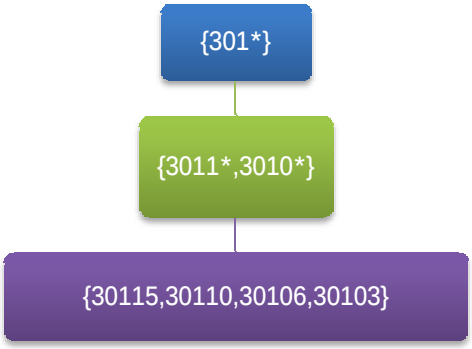
Ο πίνακας ικανοποιεί την 2-Ανωνυμία γιατί για κάθε δύο εγγραφές του συνόλου δεδομένων υπάρχουν ισοτιμίες στα γνωρίσματα. Ο λόγος είναι γιατί ο πίνακας έχει χωριστεί σε διάφορες κλάσεις ισοδυναμίας, έτσι ώστε σε κάθε υποσύνολο να υπάρχουν τουλάχιστον δύο εγγραφές (k=2) οι οποίες να έχουν τις ίδιες τιμές για τα ψευδο-αναγνωριστικά.

3.2.1. Γενίκευση

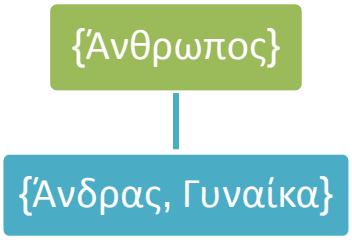
Υπάρχουν πολλές τεχνικές για την επίτευξη της k-ανωνυμίας μία από αυτές είναι η γενίκευση που χρησιμοποιείται ευρέως στον τομέα της ασφάλειας της ιδιωτικότητας. Με τον όρο γενίκευση (generalization) ορίζουμε την διαδικασία κατά την οποία οι τιμές των ψευδο-αναγνωριστικών (Quasi Identifiers) αντικαθίστανται με γενικότερες ή με βάση συγκεκριμένες ιεραρχίες γενίκευσης. Στόχος της γενίκευσης είναι η διατήρηση μέρους της πληροφορίας της αρχικής τιμής χωρίς αυτή να αλλάζει και να αλλοιώνεται πλήρως.

Επιπρόσθετα, κάθε τιμή μπορεί να γενικευθεί σε πολλά στάδια και σε πιο γενικές σημασιολογικές τιμές. Τα διαφορετικά επίπεδα γενίκευσης του πεδίου τιμών ενός γνωρίσματος, στα οποία οδηγούνται οι αρχικές τιμές με κάθε γενίκευση συνήθως αποτυπώνονται με τη μορφή δένδρου, το οποίο ονομάζεται ιεραρχία γενίκευσης του πεδίου τιμών, (Domain Generalization Hierarchy).

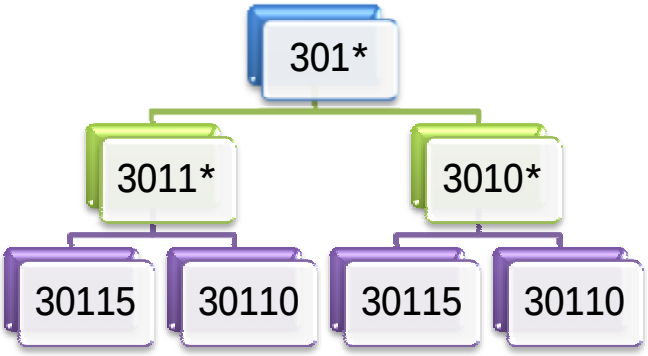
Για παράδειγμα, η τιμή 30115 βρίσκεται στο κύριο πεδίο Τ.Κ. του πίνακα 1. Για να επιτύχουμε k-ανωνυμία θα τροποποιήσουμε τις τιμές του πεδίου. Αυτό θα το καταφέρουμε με το να αντικαταστήσουμε την τιμή του πεδίου με κάποια πιο γενική, λιγότερο συγκεκριμένη τιμή που μπορεί να χρησιμοποιηθεί για να περιγράψουμε το πεδίο που θέλουμε να τροποποιήσουμε, δηλαδή θα αντικαταστήσουμε το τελευταίο ψηφίο με * (30115->3011*). Αυτή σχεδίαση ξεκίνησε με σκοπό την γενίκευση των δεδομένων όμως για να εφαρμοστεί πρέπει να τηρούνται οι εξής συνθήκες: (1) κάθε πεδίο έχει το πολύ ένα κύριο γενικευμένο πεδίο και (2) όλα τα μέγιστα στοιχεία είναι μοναδικά.



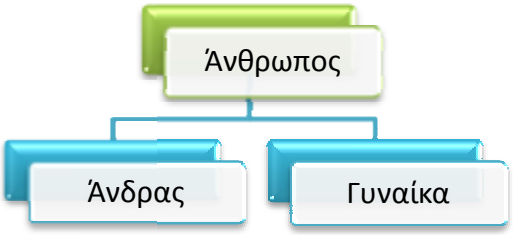
Σχήμα α: Τ.Κ.



Σχήμα β: Φύλο



Σχήμα γ: Τ.Κ. ιεραρχία



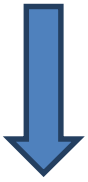
Σχήμα δ: Φύλο ιεραρχία

Όπως βλέπουμε στα σχήματα (α) και (γ) χρησιμοποιείται γενίκευση σε αριθμητικά πεδία τιμών σε αντίθεση με τα (β) και (δ) που γενικεύονται κατηγορικά πεδία. Στην πρώτη περίπτωση, στα αριθμητικά η αρχική τιμή γενικεύεται σ' ένα πιο ευρύ διάστημα τιμών και στη συνέχεια σ' ένα ακόμα πιο γενικό διάστημα.

Στην περίπτωση της γενίκευσης σε κατηγορικά δεδομένα, τα πεδία γενικεύονται σε μια πιο γενική τιμή με βάση της σημασιολογία τους.

ΑΡΡΕΝ	1976	30103	ΓΡΙΠΗ
ΑΡΡΕΝ	1976	30106	ΒΡΟΓΧΙΤΙΔΑ

ΓΕΝΙΚΕΥΣΗ (generalization)



ΑΡΡΕΝ	1976	3010*	ΓΡΙΠΗ
ΑΡΡΕΝ	1976	3010*	ΒΡΟΓΧΙΤΙΔΑ

Πίνακας 3: διαδικασία γενίκευσης για δημιουργία κλάσης

Στο παραπάνω πίνακα φαίνεται πως επιτυγχάνεται η γενίκευση και η δημιουργία κλάσης ισοδυναμίας για δύο τιμές του πίνακα, έτσι ώστε να ικανοποιείται η 2-ανωνυμία. Σε αυτή την περίπτωση, όσο πιο ψηλά βρίσκεται η γενίκευση στο δέντρο ιεραρχίας, τόσο πιο μεγάλη είναι και η απώλεια πληροφορίας (information loss).

Υπάρχουν δύο τρόποι για την εφαρμογή την γενίκευσης σε μία βάση δεδομένων, η τοπική και η ολική γενίκευση. Στην πρώτη περίπτωση η γενίκευση γίνεται σ ένα μέρος των αρχικών τιμών ενώ στην ολική, αντικαθίστανται όλες οι αρχικές τιμές της βάσης δεδομένων με πιο γενικευμένες. (Terrovitis, Mamoulis, & Kalnis, 2008)

3.2.2. Απόκρυψη εγγραφών

Μία άλλη μέθοδος που χρησιμοποιείται για την επίτευξη της k-ανωνυμίας είναι η μέθοδος της απόκρυψης εγγραφών (suppression). Σε αυτή την περίπτωση αφαιρούνται ολοκληρωτικά τιμές από διάφορα πεδία τιμών προκειμένου να ελαχιστοποιηθεί το επίπεδο γενίκευσης και να χάνονται πληροφορίες στα δεδομένα. (Cox, 1980)

ΑΡΡΕΝ	1976	3010*	ΓΡΙΠΗ
ΑΡΡΕΝ	1976	3010*	ΒΡΟΓΧΙΤΙΔΑ

ΓΕΝΙΚΕΥΣΗ (GENERALIZATION)

ΑΠΟΚΡΥΨΗ ΕΓΓΡΑΦΩΝ(SUPPRESSION)

*	1976	3010*	ΓΡΙΠΗ
*	1976	3010*	ΒΡΟΓΧΙΤΙΔΑ

Πίνακας 4: Απόκρυψη εγγραφών κατά τη διαδικασία ανωνυμοποίησης

Συνήθως η απόκρυψη εγγραφών γίνεται σε πεδία που δεν επιτρέπουν μεγάλη γενίκευση. Πρακτικά η απόκρυψη χρησιμοποιείται για να μετριάσει την διαδικασία της γενίκευσης όταν υπάρχουν χαρακτηριστικά πλήθους λιγότερα από k δηλαδή γνωρίσματα που έχουν εμφανιστεί λιγότερο από k φορές. Για παράδειγμα, στον πίνακα 4 έχει χρησιμοποιηθεί η απόκρυψη στο πεδίο φύλλο καθώς δεν ήταν δυνατή η γενίκευση σε επιμέρους μεγαλύτερα πεδία.

3.2.3. Γενίκευση (generalization) vs απόκρυψη (suppression)

Η γενίκευση καθώς και η συμπίεση δεδομένων είναι δύο διαφορετικές τεχνικές που χρησιμοποιούνται για την επίτευξη της k-ανωνυμοποίησης, δηλαδή της μεθόδου τροποποίησης των δεδομένων σε μια βάση δεδομένων. Και οι δύο τεχνικές είναι εξίσου διαδεδομένες και ισχυρές. Το ποιά τεχνική θα εφαρμοστεί εξαρτάται από τα δεδομένα του πίνακα και πως αυτά μπορούν να τροποποιηθούν καλύτερα. Η βασική διαφορά τους είναι πως με την συμπίεση των δεδομένων έχουμε λιγότερη πληρότητα ενώ με την γενίκευση χάνουμε μέρος της ακρίβειας. Κάποια από τα μοντέλα χρησιμοποιούν μόνο την γενίκευση, γενικεύοντας δεδομένα ώστε να επιτύχουν k-ανωνυμοποιήσεις, ενώ κάποια άλλα μόνο την απόκρυψη, αποκρύπτοντας όσα δεδομένα επηρεάζουν *k*- ανωνυμία. Η επιλογή της τεχνικής που θα χρησιμοποιηθεί μπορεί να είναι και ένας συνδυασμός αυτών των δύο όπως φαίνεται στον πίνακα 4. Παρόλα αυτά σύμφωνα με έρευνες που έχουν γίνει πάνω στην εφαρμογή των δύο τεχνικών καταλήγουμε στο συμπέρασμα πως προτιμάται η συμπίεση από την γενίκευση χωρίς αυτό να σημαίνει πως και η γενίκευση δεν είναι εξίσου ισχυρή και αποτελεσματική. Ο λόγος είναι πως η γενίκευση επηρεάζει όλες τις πλειάδες ενός πίνακα ενώ η συμπίεση αποκρύπτει συγκεκριμένες εγγραφές.

3.2.4 Απώλεια πληροφορίας

Ο λόγος που δημοσιεύουμε μία βάση δεδομένων είναι η εκμετάλλευση των χρήσιμων πληροφοριών που διαθέτουν τα δεδομένα για διάφορους λόγους. Για να προστατέψουμε την ιδιωτικότητα των εγγραφών από διάφορους επιτήδειους γενικεύουμε τα δεδομένα, αντικαθιστώντας τα από πιο γενικευμένες τιμές ή τα αποκρύπτουμε, όπως έχουμε ήδη αναφέρει και πιο πάνω.

Μια τέτοια τροποποίηση έχει σαν αποτέλεσμα την απώλεια χρήσιμης πληροφορίας που διαθέτουν τα δεδομένα στην αρχική τους μορφή, με αποτέλεσμα να υπάρχει μεγάλη πιθανότητα να βγάλουμε ελλειπή ή λανθασμένα συμπεράσματα. Γι΄αυτό το λόγο όλοι οι αλγόριθμοι που έχουν δημιουργηθεί για την ανωνυμοποίηση εξετάζονται πολύ προσεκτικά ώστε να μην χάνεται μεγάλος όγκος χρήσιμων

πληροφοριών αλλά μόνο όσες πληροφορίες δυσκολεύουν το έργο κάποιου κακόβουλου.

3.3 Αλγόριθμοι εύρεσης k-ανώνυμων πινάκων

Δύο από τους πιο διαδεδομένους αλγορίθμους που υλοποιούν την k-ανωνυμοποίηση και μελετήθηκαν στη συγκεκριμένη πτυχιακή εργασία είναι ο Incognito και ο Mondrian. Και οι δύο αλγόριθμοι είναι ευρέως γνωστοί, με υψηλά επίπεδα αποτελεσματικότητας και έχουν ως βασικό στόχο την βέλτιστη ανωνυμοποίηση μιας βάσης δεδομένων με την μικρότερη απώλεια χρήσιμης πληροφορίας.

Χρησιμοποιούν και οι δύο αλγόριθμοι την τεχνική της γενίκευσης, δέχονται ως είσοδο το σύνολο των αρχικών δεδομένων και επιστρέφουν πίσω τις γενικευμένες τιμές που προκύπτουν κατά την ανωνυμοποίηση.

Η κύρια διαφορά τους σχετίζεται με το μοντέλο ανακωδικοποίησης αφού ο Incognito είναι αλγόριθμος μονοδιάστατης καθολικής ανακωδικοποίησης, ενώ ο Mondrian εφαρμόζει πολυδιάστατη ανακωδικοποίηση. Και οι δύο αλγόριθμοι στοχεύουν στην εφαρμογή της k-ανωνυμίας στα δεδομένα, όμως παρουσιάζουν διαφορετικής μορφής ανωνυμοποιημένα δεδομένα λόγω των διαφορετικών μεθόδων που ακολουθούν.

3.3.1. Αλγόριθμος Incognito

Ο αλγόριθμος Incognito δημιουργήθηκε με σκοπό την επίλυση προβλημάτων ιδιωτικότητας με *k*-ανωνυμία. Όπως αναφέραμε ο συγκεκριμένος αλγόριθμος βασίζεται στην τεχνική της πλήρους γενίκευσης αντιστοιχίζοντας κάθε τιμή ενός γνωρίσματος, με την ίδια γενικευμένη τιμή σε όλες τις τιμές του πίνακα (Agrawal & Srikant, 1994), (K. Le Fevre, D. J. De Witt, R. Ramakrishnan 1995).

Αλγόριθμος: Incognito

Είσοδος: Ένας πίνακας T προς k-ανωνυμοποίηση, ένα σύνολο Q του ψευδο-αναγνωριστικού και μία ιεραρχία για κάθε γνώρισμα του ψευδο-αναγνωριστικού.

Έξοδος: Ένα σύνολο από k-ανώνυμων γενικεύσεων γενικού συνόλου.

C_1= {Κόμβοι της ιεραρχίας γενίκευσης συνόλου τιμών των γνωρισμάτων του Q}

E_1= {Ακμές της ιεραρχίας γενίκευσης συνόλου τιμών των γνωρισμάτων του Q}

queue= μία άδεια ουρά

for i=1 μέχρι n do

//C_i και E_i ορίζουν ένα γράφο γενικεύσεων

S_i = ένα αντίγραφο του C_i

{roots} = {όλοι οι κόμβοι στο C_i χωρίς κάποια άκρη στο E_i κατευθυνόμενοι προς αυτούς.}

Εισαγωγή των roots στην queue, διατηρώντας την queue ταξινομημένη ως προς το ύψος.

while queue δεν είναι άδεια do

node =αφαίρεση του πρώτου στοιχείου της queue

if node δεν είναι μαρκαρισμένος then

if node είναι root then

frequencySet = υπολόγισε την συχνότητα του συνόλου T σε σχέση με τα χαρακτηριστικά του node χρησιμοποιώντας το T

else

frequencySet = υπολόγισε την συχνότητα του συνόλου T σε σχέση με τα χαρακτηριστικά του node χρησιμοποιώντας την καθορισμένη συχνότητα του γονέα.

end if

χρησιμοποίησε το frequencySet για να τσεκάρεις την k-ανωνυμία σε σχέση με τα χαρακτηριστικά του node

if το T είναι k-ανώνυμο με βάση τα γνωρίσματα του κόμβου then

μάρκαρε όλες τις άμεσες γενικεύσεις του κόμβου

else

διέγραψε τον κόμβο από το S_i

εισήγαγε τις άμεσες γενικεύσεις του κόμβου στην ουρά , κρατώντας την ουρά ταξινομημένη ως προς το ύψος.

end if

end if

end while

C_i+1,E_i+1=κατασκευή Γράφους (S_i,E_i)

end for

return προβολή όλων των γνωρισμάτων του Sn στο T και στις ιεραρχίες

Χρησιμοποιώντας την προκαθορισμένη ιεραρχία γενίκευσης δημιουργεί ένα πλέγμα γενίκευσης πολλαπλών γνωρισμάτων. Στο πλέγμα παρουσιάζονται σχηματικά όλοι οι δυνατοί συνδυασμοί μεταξύ των επιπέδων των ιεραρχιών γενίκευσης των γνωρισμάτων του ψευδοαναγνωριστικού, όπου εκφράζονται ουσιαστικά όλες οι δυνατές γενικεύσεις των πλειάδων. Οι συνδυασμοί αυτοί, ελέγχονται για την ικανοποίηση της k-ανωνυμίας. Στόχος του αλγορίθμου είναι η εύρεση της ελάχιστης γενίκευσης πλήρους πεδίου, προκειμένου να υπάρχει η λιγότερο δυνατή απώλεια πληροφορίας.

Ο αλγόριθμος Incognito χρησιμοποιεί την ιδιότητα του υποσυνόλου (subset property), σύμφωνα με την οποία, αν ένας πίνακας T είναι k-ανώνυμος ως προς ένα σύνολο γνωρισμάτων Q της βάσης δεδομένων, τότε είναι k-ανώνυμος και ως προς οποιοδήποτε υποσύνολο γνωρισμάτων $P \subseteq Q$.

Στη συνέχεια, παρουσιάζεται ένα παράδειγμα της εφαρμογής του αλγορίθμου Incognito στα ιατρικά δεδομένα ενός οργανισμού και δίνεται έμφαση στη διαδικασία

που ακολουθεί ο αλγόριθμος προκειμένου να ελέγξει όλες τις δυνατές γενικεύσεις στο πλέγμα.

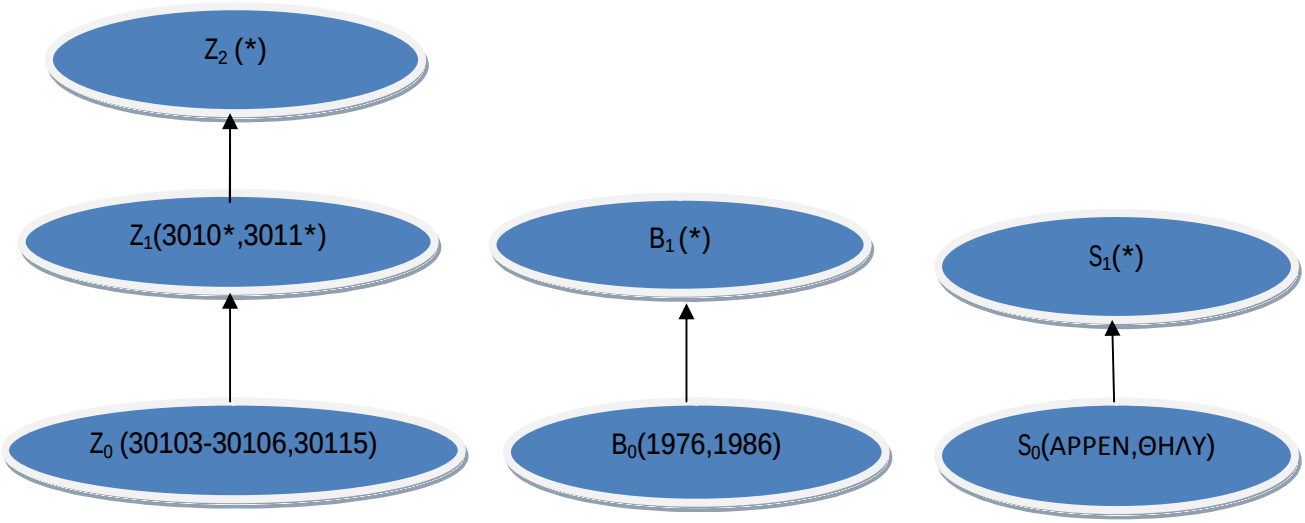
Ιατρικά Δεδομένα Ασθενών			
ΦΥΛΛΟ	ΧΡΟΝΟΛΟΓΙΑ ΓΕΝΝΗΣΗΣ	Τ.Κ.	ΑΣΘΕΝΕΙΑ
ΑΡΡΕΝ	1976	30115	ΚΑΡΚΙΝΟΣ
ΘΥΛΗ	1986	30115	ΑΜΥΓΔΑΛΙΤΙΔΑ
ΑΡΡΕΝ	1976	30103	ΓΡΙΠΗ
ΑΡΡΕΝ	1976	30103	ΗΠΑΤΙΤΙΔΑ
ΘΥΛΗ	1986	30106	ΟΙΔΗΜΑ
ΑΡΡΕΝ	1986	30106	ΒΡΟΓΧΙΤΙΔΑ

Πίνακας 5: Ιατρικά Δεδομένα Οργανισμού

$1^{η}$ Επανάληψη $(i = 1)$

Ο αλγόριθμος ελέγχει αν ο πίνακας T είναι k-ανώνυμος για γενικεύσεις ενός συνόλου γνωρισμάτων με μέγεθος $i = 1$. Πιο συγκεκριμένα ξεκινάει να ελέγχει την ανωνυμία του πίνακα αν αφαιρέσει τα πεδία <ταχυδρομικός κώδικας> και <φύλλο> και κρατήσει μόνο το πεδίο <χρονολογία γέννησης>. Με την διαδικασία αυτή βγαίνει το αποτέλεσμα πως ο πίνακας είναι k-ανώνυμος με βάση αυτό το υποσύνολο και άρα με όλες τις γενικευμένες τιμές που ορίζονται από την ιδιότητα της γενίκευσης.

Στη συνέχεια επαναλαμβάνεται η διαδικασία και για τα υπόλοιπα γνωρίσματα του ψευδοαναγνωριστικού. (Εικόνα 1)



Εικόνα 1: Αλγόριθμος Incognito, Επανάληψη 1: Ιεραρχίες γενίκευσης πεδίων {Ημερομηνία Γεννήσεως, Ταχυδρομικός Κώδικας, Φύλο}

$2^{η}$ Επανάληψη $(i = 2)$

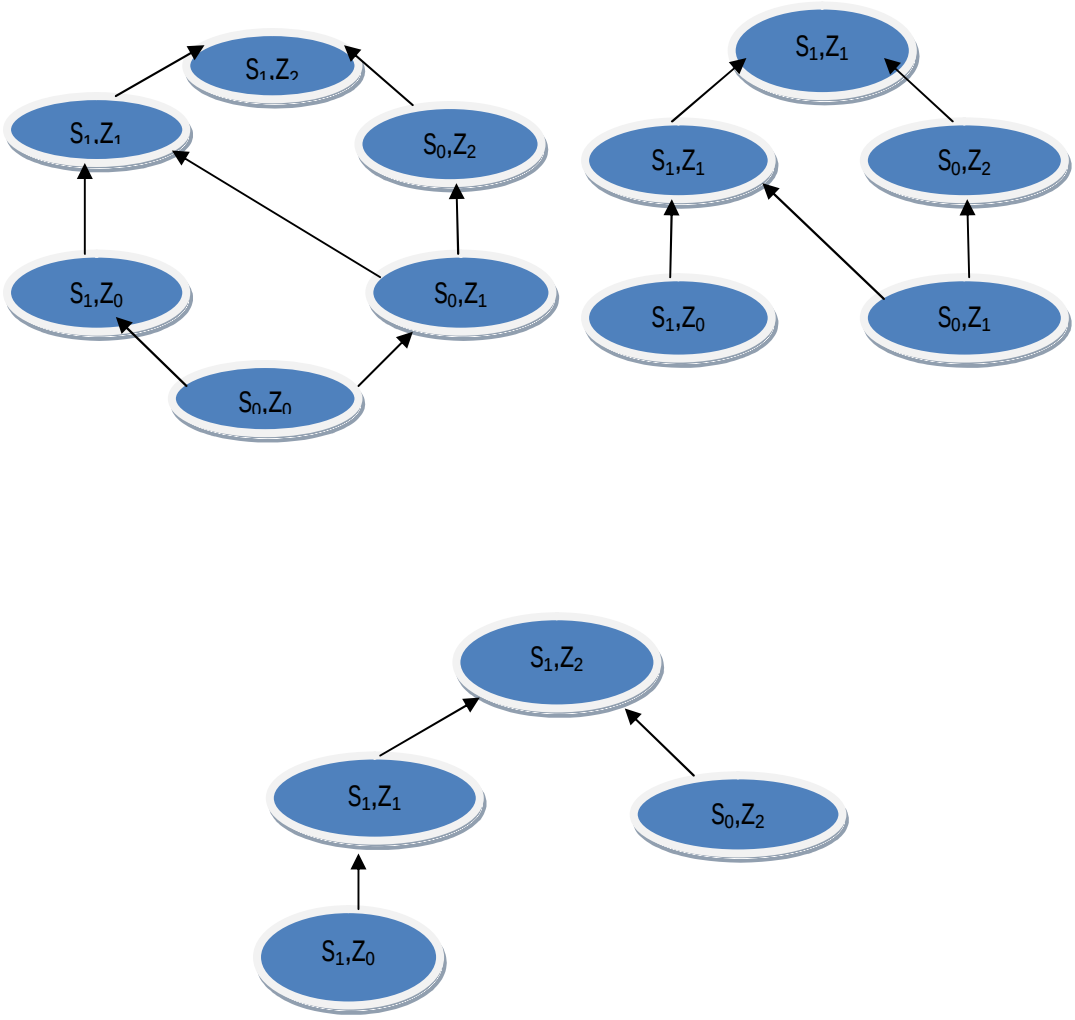
Στη δεύτερη επανάληψη ο αλγόριθμος θα κάνει ακριβώς το ίδιο με πιο πάνω όμως για $i = 2$, δηλαδή

<Χρονολογία Γέννησης, Φύλο>

<Χρονολογία Γέννησης, Ταχυδρομικός Κώδικας>

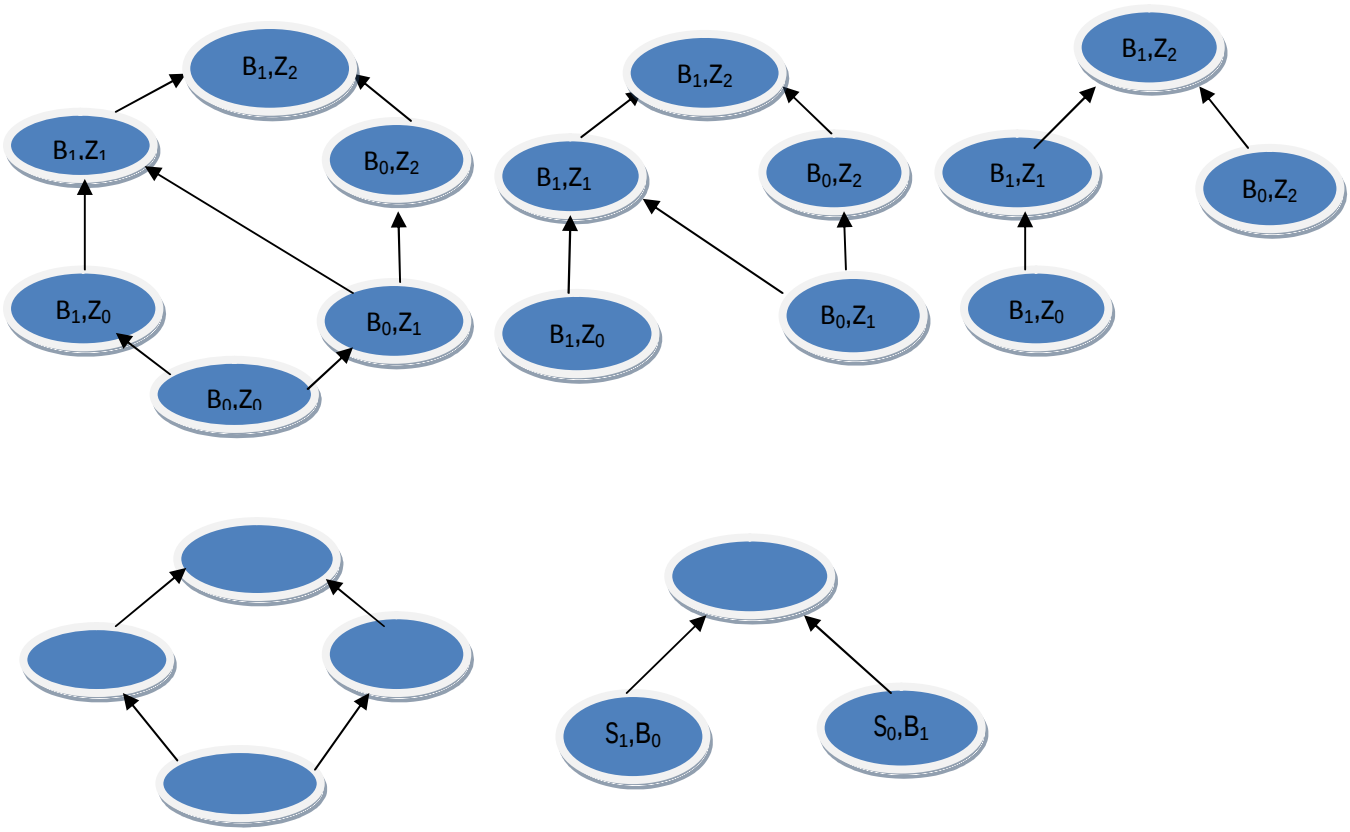
<Φύλο, Ταχυδρομικός Κώδικας>

Πιο συγκεκριμένα, ο αλγόριθμος ελέγχει αρχικά το frequency set του συνόλου $< S0, Z0 >$ και βλέπει ότι δεν ικανοποιείται η k-ανωνυμία, οπότε περνά στον έλεγχο του $< S1, Z0 >$ και $< S0, Z1 >$. Με βάση το $< S1, Z0 >$ ο πίνακας ικανοποιεί το k-anonymity και κατ' επέκταση όλες οι γενικεύσεις του το επιτυγχάνουν (ιδιότητα γενίκευσης). Στη συνέχεια ελέγχει το frequency set του $< S0, Z1 >$ και βλέπει ότι δεν ικανοποιείται το k-ανωνυμία οπότε και απορρίπτεται. Ο συνδυασμός $< S1, Z1 >$ δεν ελέγχεται γιατί είναι γενίκευση του <S1, Z0>. Ο επόμενος έλεγχος είναι το σύνολο $< S0, Z2 >$ με βάση το οποίο ικανοποιείται η k-ανωνυμία, οπότε και σταματά ο έλεγχος όπως παρουσιάζεται στην Εικόνα 2.



Εικόνα 2: Αλγόριθμος Incognito, Επανάληψη 2: Απόρριψη κόμβων στο υποσύνολο γνωρισμάτων <Φύλο, Ταχυδρομικός Κώδικας>

Η ίδια διαδικασία ακολουθείται για όλα τα υποσύνολα γνωρισμάτων. Όπως παρουσιάζεται στην Εικόνα 3, φαίνονται ποιοι κόμβοι απορρίπτονται στην 2η επανάληψη του αλγόριθμου για τα υπόλοιπα σύνολα γνωρισμάτων μεγέθους $i = 2$.

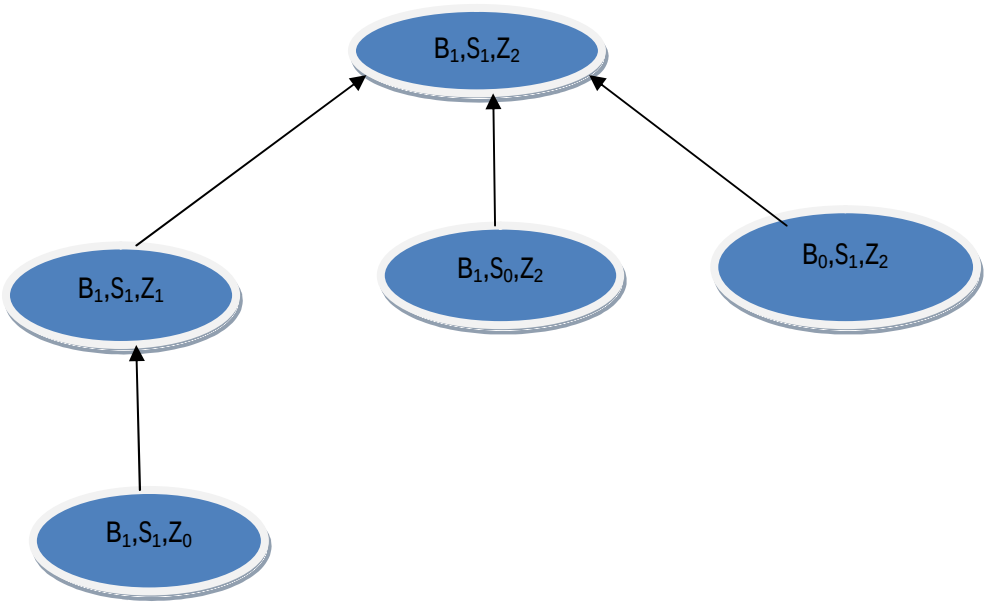


Εικόνα 3: Αλγόριθμος Incognito, Επανάληψη 2: Απόρριψη κόμβων στο υποσύνολο γνωρισμάτων <Φύλο, Ημερ. Γεννήσεως>

Τελευταία επανάληψη

Στο τελευταίο βήμα, ο αλγόριθμος αντιστρέφει την ιδιότητα του υποσυνόλου. Συμπεραίνεται πως αν ένα υποσύνολο δεν ικανοποιεί την k-ανωνυμία τότε το ίδιο θα συμβαίνει για κάθε σύνολο που το περιέχει.

Ύστερα από πολλές επαναλήψεις δημιουργείται το τελικό δέντρο με όλους τους k-ανώνυμους συνδυασμούς των διαφορετικών επιπέδων γενίκευσης και στη συνέχεια επιλέγεται ο πιο αποδοτικός συνδυασμός, όπως φαίνεται και στην εικόνα 4.



Εικόνα 4: Αλγόριθμος Incognito: Τελικό πλέγμα γενίκευσης του αλγόριθμου

Όπως συμπεραίνεται από τα παραπάνω πρόκειται για έναν αλγόριθμο με πολυπλοκότητα εκθετική ως προς το μέγεθος του συνόλου των γνωρισμάτων του ψευδοαναγνωριστικού. Είναι ένας ουσιαστικός, αποτελεσματικός και πλήρης αλγόριθμος με βασικό μειονέκτημα ότι εκτελεί όλες τις ανωνυμοποιήσεις που μπορούν να γίνουν σ' έναν πίνακα με αποτέλεσμα να είναι χρονοβόρος, με μεγάλη έκταση αλλά δίνει το πιο αποδοτικό αποτέλεσμα.

3.3.2. Αλγόριθμος Mondrian

Όπως έχει ήδη αναφερθεί πιο πάνω, ο αλγόριθμος Incognito μπορεί να επιστρέφει αποδοτικά αποτελέσματα αλλά σε μία πλήρης γενικευμένη μορφή. Αυτό έχει ως αποτέλεσμα την υπεργενίκευση των δεδομένων μιας βάσης πληροφοριών, τα οποία πιθανόν χάνουν τις χρήσιμες πληροφορίες τους και γίνονται κάποιες φορές άχρηστα για εκείνους που απευθύνονται.

Πράγματι, για μία βάση με αριθμητικά δεδομένα, μία γενίκευση πλήρους πεδίου σημαίνει αντικατάσταση όλων των αρχικών τιμών με σταθερά μη επικαλυπτόμενα μεταξύ τους διαστήματα ή πλήρη απόκρυψή τους.

Ο αλγόριθμος Mondrian όπως ορίζεται και παρουσιάζεται από (Fevre, witt, & Ramakrishnan, 2006) έχει ως σκοπό να επιλύσει τέτοιου είδους προβλήματα, προσφέροντας υψηλές τιμές ανωνυμοποίησης, χάρη στο πολυδιάστατο μοντέλο τοπικής ανακωδικοποίησης με το οποίο εφαρμόζεται.

Σύμφωνα με αυτό το μοντέλο, ορίζεται ένας χώρος π-διαστάσεων, όπου π το πλήθος των ψευδοαναγνωριστικών. Αναζητούμε μια k-ανώνυμη λύση, χωρίζοντας αυτό το χώρο σε διαμερίσεις.

Διαδικασία που ακολουθεί ο αλγόριθμος:

1. Αρχικά ορίζονται οι πολυδιάστατες περιοχές και επιλέγεται η διάσταση κατά την οποία θα γίνει η διαμέριση.
2. Υλοποιείται η διαμέριση κατά την πιο πάνω διάσταση και προκύπτουν δύο υποχώροι. Στους οποίους οι τιμές που είναι ίσες ή μικρότερες με τον μέσο να βρίσκονται αριστερά (R1) και οι υπόλοιπες να βρίσκονται στην δεξιά κλάση (R2).
3. Επαναλαμβάνεται η διαδικασία και για τους δύο υποχώρους (R1) και (R2) μέχρι να μην υπάρχει άλλη επιτρεπόμενη πολυδιάστατη τομή για διαμέρισμα σε καμία διάσταση.
4. Τέλος προκύπτει η βέλτιστη πολυδιάστατη διαμέριση, με την πιο κατάλληλη γενίκευση που θα χρησιμοποιηθεί για την ανακωδικοποίηση.

.

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

Για την ευκολότερη κατανόηση του αλγορίθμου πρέπει να μελετηθεί το παρακάτω παράδειγμα στο οποίο ορίζονται τα δύο μοντέλα ανωνυμοποίησης, το μονοδιάστατο και το πολυδιάστατο για τον πίνακα των ιατρικών δεδομένων ενός οργανισμού.

Σύμφωνα με το (Fevre, witt, & Ramakrishnan, 2006), όταν πρόκειται για μονοδιάστατη ανωνυμοποίηση σε κάθε βήμα της γενίκευσης υπάρχουν τιμές από συγκεκριμένα μη επικαλυπτόμενα διαστήματα ενώ στην πολυδιάστατη ανωνυμοποίηση σε κάθε της γενίκευσης επιτρέπονται τα επικαλυπτόμενα διαστήματα.

Ιατρικά Δεδομένα			
Ηλικία	**Φύλο**	**Κώδικας**	**Ασθένεια**
35	Άρρεν	30511	Καρκίνος
35	Θήλυ	30512	Αμυγδαλίτιδα
36	Άρρεν	30511	Γρίπη
37	Άρρεν	30510	Ηπατίτιδα
37	Θήλυ	30512	Οίδημα
38	Άρρεν	30511	Βρογχίτιδα

Πίνακας 6: Ιατρικά Δεδομένα Οργανισμού

Μονοδιάστατη Ανωνυμοποίηση			
Ηλικία	**Φύλο**	**Κώδικας**	**Ασθένεια**
[35-38]	Άρρεν	[30510-30511]	Καρκίνος
[35-38]	Θήλυ	30512	Αμυγδαλίτιδα
[35-38]	Άρρεν	[30510-30511]	Γρίπη
[35-38[	Άρρεν	[30510-30511]	Ηπατίτιδα
[35-38]	Θήλυ	30512	Οίδημα
[35-38]	Άρρεν	[30510-30511]	Βρογχίτιδα

Πίνακας 7: Μονοδιάστατη ανωνυμοποίηση πίνακα ασθενών

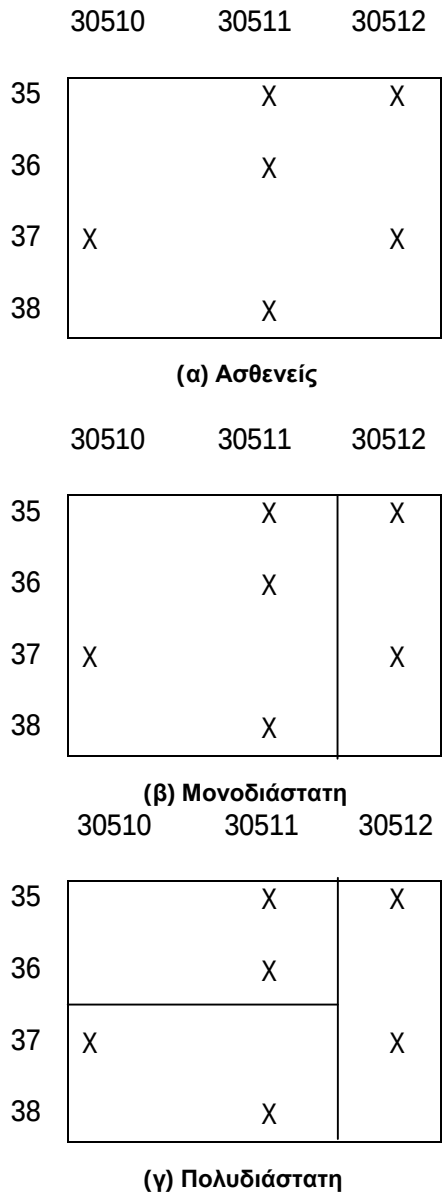
Πολυδιάστατη Ανωνυμοποίηση			
Ηλικία	**Φύλο**	**Κώδικας**	**Ασθένεια**
[35-36]	Άρρεν	30511	Καρκίνος
[35-37]	Θήλυ	30512	Αμυγδαλίτιδα
[35-36]	Άρρεν	30511	Γρίπη
[37-38]	Άρρεν	[30510-30511]	Ηπατίτιδα
[35-37]	Θήλυ	30512	Οίδημα
[37-38]	Άρρεν	[30510-30511]	Βρογχίτιδα

Πίνακας 8: Πολυδιάστατη ανωνυμοποίηση πίνακα ασθενών

Όπως φαίνεται στο πιο κάτω σχήμα, στην ουσία στο μονοδιάστατο μοντέλο για να ορίσουμε σωστά τις διαμερίσεις, τραβάμε παράλληλες ευθείες ως προς τους άξονες οι οποίες διαπερνάνε από όλο τον χώρο. Από την άλλη μεριά στο πολυδιάστατο μοντέλο, αρχικά τραβάμε μια ευθεία παράλληλη ως τον έναν άξονα και

στη συνέχεια τραβάμε επιπλέον ευθείες ως προς οποιονδήποτε άξονα για να χωρίσουμε τους δύο υποχώρους σε άλλους.

Για να επιλυθεί η k-ανωνυμία αρκεί σε κάθε υποχώρο να υπάρχουν τουλάχιστον k εγγραφές.



Σχήμα: Χωρική αναπαράσταση ασθενών και διαμερίσεων (QI: Ηλικία, Ταχυδρομικός κώδικας)

Παρόλα αυτά το πρόβλημα του αλγόριθμου είναι πως δεν μπορεί να υπολογίσει την απώλεια χρήσιμης πληροφορίας ακόμα και αν δίνει ικανοποιητικά αποτελέσματα σε σχέση με άλλα μοντέλα που έχουν διατυπωθεί για την επίλυση του

ίδιου προβλήματος, και σε μικρότερο χρόνο μιας και έχει πολυπλοκότητα O(nlogn), όπου n ο αριθμός των εγγραφών του πίνακα.

3.4 Αδυναμίες k-ανωνυμίας

Το αρχικό πρόβλημα έχει να κάνει με τις πληροφορίες που διαθέτει ο επιτιθέμενος στο προσωπικό του αρχείο ή εξωτερικές γνώσεις, και με τις τελικές πληροφορίες των δεδομένων έπειτα από την γενίκευση. Έστω ότι μία ανωνυμοποιημένη βάση δεδομένων παρουσιάζει ελλειπή ή γενικευμένες πληροφορίες και έστω ότι ο ενδιαφερόμενος γνωρίζει προσωπικές πληροφορίες για το θύμα, συνδυάζοντας τις πληροφορίες που έχει με αυτές του πίνακα είναι αρκετά εύκολο να εντοπίσει το πρόσωπο που τον ενδιαφέρει.

Επιπλέον είναι πιθανόν κάποιος επιτιθέμενος να γνωρίζει μέρος του ψευδό-αναγνωριστικού ενός ατόμου. Με αυτό τον τρόπο μπορεί να υπολογίσει τα προσωπικά δεδομένα του αν αυτά έχουν μεγάλη συχνότητα εμφάνισης.

Το δεύτερο πρόβλημα έχει να κάνει με τη σειρά που εμφανίζονται οι πλειάδες αφού ανωνυμοποιηθεί ο πίνακας. Πιο κάτω βλέπουμε ένα παράδειγμα δύο ανωνυμοποιημένων πινάκων ύστερα από γενίκευση του αρχικού και ικανοποιούν την k-Ανωνυμία για k=2. Παρατηρούμε πως στους δυο τελικούς πίνακες όσο και στον αρχικό οι σειρά των δεδομένων είναι η ίδια. Επομένως αν είχε δημοσιευτεί πρώτα ο Πίνακας 1 και έπειτα ο Πίνακας 2, ή αντίστροφα θα ήταν δυνατό με μια απλή σύνδεση να εντοπιστούν όλες οι εγγραφές του αρχικού.

ΑΡΧΙΚΟΣ ΠΙΝΑΚΑΣ	
ΟΝΟΜΑ	**ΕΤΟΣ ΓΕΝΝΗΣΗΣ**
ΓΙΑΝΝΗΣ	1986
ΚΩΣΤΑΣ	1991
ΜΑΡΙΑ	1987
ΕΛΕΝΗ	1993
ΓΕΩΡΓΙΟΣ	1986

ΠΙΝΑΚΑΣ 1	
ΟΝΟΜΑ	**ΕΤΟΣ ΓΕΝΝΗΣΗΣ**
ΑΡΡΕΝ	1986
ΑΡΡΕΝ	1991
ΘΗΛΥ	1987
ΘΗΛΥ	1993
ΑΡΡΕΝ	1986

ΠΙΝΑΚΑΣ 2	
ΟΝΟΜΑ	**ΕΤΟΣ ΓΕΝΝΗΣΗΣ**
ΓΙΑΝΝΗΣ	198*
ΚΩΣΤΑΣ	199*
ΜΑΡΙΑ	198*
ΕΛΕΝΗ	199*
ΓΕΩΡΓΙΟΣ	198*

Πίνακας 9: Αδυναμία k-anonymity με πίνακες ίδιας σειράς πλειάδων

3.4.1. Επιθέσεις κατά της k-ανωνυμίας

Για τους παραπάνω λόγους δεν εξασφαλίζεται πλήρως η ιδιωτικότητα και έτσι δίνεται η δυνατότητα σε κάποιον κακόβουλο εξωτερικό χρήστη να την παραβίαση εντοπίζοντας διάφορα ευαίσθητα γνωρίσματα εγγραφών.

Οι δύο πιο σύνηθες επιθέσεις κατά την k-ανωνυμία είναι οι επιθέσεις ομοιογένειας (homogeneity attack) και οι επιθέσεις με πρότερη γνώση (background knowledge attack) (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2007).

Στην πρώτη περίπτωση, δηλαδή στην επίθεση ομοιογένειας, σε ένα σύνολο από ψευδό-αναγνωριστικά υπάρχουν k-1 ίδιες εγγραφές από το σύνολο δεδομένων. Έτσι μερικές φορές η k-ανωνυμία οδηγεί στη διαρροή των πληροφοριών και, όταν αυτές οι πληροφορίες αποκαλύπτονται, μία επίθεση ομοιογένειας συμβαίνει. Δηλαδή, εάν κάποιος γνωρίζει ορισμένες τιμές για κάποια γνωρίσματα του ψευδό-

αναγνωριστικού μπορεί εύκολα να προσδιορίσει την τιμή κάποιας εγγραφής από την κλάση ισοδυναμίας, χωρίς να έχει γίνει η ταυτοποίηση της. Για να γίνει πιο κατανοητό, εξετάζοντας τους πίνακες πιο κάτω παρατηρούμε πως παρόλο που στον δεύτερο έχει εφαρμοστεί η k-ανωνυμία αν κάποιος ανήκει στην τελευταία κλάση τότε είναι εύκολο να εντοπίσει πως κάποιος πάσχει σίγουρα από αμυγδαλίτιδα.

Στην δεύτερη περίπτωση, δηλαδή στην επίθεση με πρότερη γνώση, ένας επιτιθέμενος κατέχει βαθιά γνώση για ένα συγκεκριμένο άτομο το οποίο γνωρίζει πως με σιγουριά βρίσκεται στη βάση δεδομένων. Έτσι αποκλείοντας διάφορες τιμές από την κλάση ισοδυναμίας στην οποία ανήκει το συγκεκριμένο άτομο, μπορεί να φτάσει στην εύρεση της τιμής ενός γνωρίσματος για το συγκεκριμένο άτομο.

ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ ΑΣΘΕΝΩΝ				
Α/Α	**Τ.Κ.**	**ΗΛΙΚΙΑ**	**ΕΘΝΙΚΟΤΗΤΑ**	**ΑΘΕΝΕΙΑ**
1	30013	20	ΕΛΛΗΝΑΣ	ΓΡΙΠΗ
2	30059	27	ΙΑΠΩΝΑΣ	ΚΑΡΚΙΝΟΣ
3	30046	23	ΕΛΛΗΝΑΣ	ΚΑΡΔΙΟΠΑΘΕΙΑ
4	30017	29	ΑΛΒΑΝΟΣ	ΚΑΡΚΙΝΟΣ
5	34145	49	ΑΜΕΡΙΚΑΝΟΣ	ΑΜΥΓΔΑΛΙΤΙΔΑ
6	34149	62	ΓΕΡΜΑΝΟΣ	ΓΡΙΠΗ
7	34143	53	ΓΑΛΛΟΣ	ΗΠΑΤΙΤΙΔΑ
8	34140	42	ΓΕΡΜΑΝΟΣ	ΚΑΡΚΙΝΟΣ
9	30058	38	ΕΛΛΗΝΑΣ	ΑΜΥΔΓΑΛΙΤΙΔΑ
10	30049	34	ΣΟΥΗΔΟΣ	ΑΜΥΓΔΑΛΙΤΙΔΑ
11	30012	36	ΙΤΑΛΟΣ	ΑΜΥΓΔΑΛΙΤΙΔΑ
12	30013	38	ΙΑΠΩΝΑΣ	ΑΜΥΓΔΑΛΙΤΙΔΑ

Πίνακας 10: ιατρικά δεδομένα ασθενών

ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ ΑΣΘΕΝΩΝ				
Α/Α	**Τ.Κ.**	**ΗΛΙΚΙΑ**	**ΕΘΝΙΚΟΤΗΤΑ**	**ΑΘΕΝΕΙΑ**
1	300**	<30	*	ΓΡΙΠΗ
2	300**	<30	*	ΚΑΡΚΙΝΟΣ
3	300**	<30	*	ΚΑΡΔΙΟΠΑΘΕΙΑ
4	300**	<30	*	ΚΑΡΚΙΝΟΣ
5	341**	>=40	*	ΑΜΥΓΔΑΛΙΤΙΔΑ
6	341**	>=40	*	ΓΡΙΠΗ
7	341**	>=40	*	ΗΠΑΤΙΤΙΔΑ
8	341**	>=40	*	ΚΑΡΚΙΝΟΣ
9	300**	3*	*	ΑΜΥΔΓΑΛΙΤΙΔΑ
10	300**	3*	*	ΑΜΥΓΔΑΛΙΤΙΔΑ
11	30012	3*	*	ΑΜΥΓΔΑΛΙΤΙΔΑ
12	30013	3*	*	ΑΜΥΓΔΑΛΙΤΙΔΑ

Πίνακας 11: Πρόβλημα Homogeneity Attack

Κεφάλαιο 4

Παραλλαγές της k-ανωνυμίας

4.1 *l*-Διαφορετικότητα *(l*-Diversity*)*

Η l-διαφορετικότητα ή l-ποικιλομορφία είναι μια νέα τεχνική που ανήκει στην ομάδα της ανωνυμοποίησης και χρησιμοποιείται για τη διατήρηση της ιδιωτικής ζωής σε σύνολα δεδομένων, μειώνοντας τη διασπορά μιας αναπαράστασης δεδομένων. Διατυπώθηκε για πρώτη φορά το 2007 με βασικό σκοπό να επιλύσει τα προβλήματα του εμφανίζει η k-ανωνυμία και αναφέραμε λίγο πιο πάνω. Στην ουσία το μοντέλο l-ποικιλομορφία είναι μια επέκταση του μοντέλου k-ανωνυμίας που έχει ως στόχο την διασφάλιση της ανωνυμίας των ευαίσθητων προσωπικών δεδομένων και όχι την ταυτότητα της εγγραφής.

Εξάλλου σκοπός της προστασίας της ιδιωτικής ζωής δεν είναι μόνο να διασφαλίσουμε την ταυτότητα μίας εγγραφής αλλά ταυτόχρονα να είμαστε βέβαιοι πως επιτιθέμενος δεν θα μπορέσει να βρει προσωπικά στοιχεία για ένα άτομο.

Ένας πίνακας ικανοποιεί την l-diversity εάν σε κάθε κλάση υπάρχουν τουλάχιστον l διαφορετικές τιμές για το σύνολο των δεδομένων του.

Ορισμός :Ενας πίνακας T είναι l-ποικιλόμορφος εάν σε κάθε κλάση ισοδυναμίας QI η συχνότερη τιμή που εμφανίζεται στο ευαίσθητο γνώρισμα *S* δεν εμφανίζεται πάνω από 1/l ×|QI| φορές.

Ο παρακάτω πίνακας ικανοποιεί την l-ποικιλομορφία με l=3 και βασίζεται στον αρχικό πίνακα 10. Υπάρχει μία διαφορετική κατάταξη των εγγραφών καθώς επίσης μία διαφορετική γενίκευση με αποτέλεσμα να επιλύεται το πρόβλημα του πίνακα 11 το Homogeneity Attack. Όπως παρατηρούμε σε κάθε κλάση υπάρχουν τουλάχιστον τρεις διαφορετικές τιμές στο γνώρισμα ασθένεια στην πέμπτη στήλη.

ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ ΑΣΘΕΝΩΝ				
Α/Α	Τ.Κ.	ΗΛΙΚΙΑ	ΕΘΝΙΚΟΤΗΤΑ	ΑΘΕΝΕΙΑ
1	3001*	<=40	*	ΓΡΙΠΗ
9	3005*	<=40	*	ΑΜΥΔΓΑΛΙΤΙΔΑ
5	3414*	>40	*	ΑΜΥΓΔΑΛΙΤΙΔΑ
2	3005*	<=40	*	ΚΑΡΚΙΝΟΣ
7	3414*	>40	*	ΗΠΑΤΙΤΙΔΑ
3	3004*	<=40	*	ΚΑΡΔΙΟΠΑΘΕΙΑ
6	3414*	>40	*	ΓΡΙΠΗ
11	3001*	<=40	*	ΑΜΥΓΔΑΛΙΤΙΔΑ
10	3004*	<=40	*	ΑΜΥΓΔΑΛΙΤΙΔΑ
4	3001*	<=40	*	ΚΑΡΚΙΝΟΣ
12	3001*	<=40	*	ΑΜΥΓΔΑΛΙΤΙΔΑ
8	3414*	>=40	*	ΚΑΡΚΙΝΟΣ

Πίνακας 12: 3-Διαφορετικός πίνακας

4.1.1 Αδυναμίες l-diversity

Το πρώτο και ένα από τα πιο βασικά μειονεκτήματα της l-ποικιλομορφίας είναι πως ο επιτιθέμενος αν ανακαλύψει το στόχο του θα είναι βέβαιος γι’ αυτόν. Για παράδειγμα έστω ότι έχουμε μία βάση δεδομένων με ιατρικά δεδομένα ασθενών. Εάν στο πεδίο ασθένεια από τις δέκα εγγραφές του πίνακα η μία είναι γρίπη η άλλη καρκίνος και οι άλλες 8 αμυγδαλίτιδα παρόλο που ικανοποιεί την l-ποικιλομορφία για l=3 είναι εύκολο ο θύτης να συμπεράνει με ποσοστό 80% πως η νόσος του ατόμου που ψάχνει είναι αμυγδαλίτιδα.

Επιπλέον, με την μέθοδο l-diversity δεν μπορούμε να εγγυηθούμε με σιγουριά πως ο επιτιθέμενος δεν θα μπορέσει αν βρει τουλάχιστον προσωπικά δεδομένα ενός ατόμου. Για να γίνει αυτό πιο κατανοητό, ας εξετάσουμε τον παρακάτω πίνακα 12 που ικανοποιεί την l-διαφορετικότητα για l=3. Εάν γνωρίζουμε πως ο Γιώργος είναι 31 χρονών και μένω σε περιοχή με ταχυδρομικό κώδικα 30100 τότε είναι εύκολο να βρούμε πως ο Γιώργος πάσχει από κάποια καρδιολογική νόσο και πως παίρνει έναν σχετικά μικρό μισθό.

ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ ΑΣΘΕΝΩΝ			
Τ.Κ.	ΜΙΣΘΟΣ	ΗΛΙΚΙΑ	ΑΣΘΕΝΕΙΑ
301**	3,000	3*	καρδιακή ανεπάρκεια
301**	4,000	3*	αρρυθμίες
301**	5,000	3*	περικαρδίτιδα
3090*	10,000	>40	ηπατίτιδα
3090*	25,000	>40	γρίπη
3090*	30,000	>40	καρκίνος
301*	40,000	2*	αμυγδαλίτιδα
301*	50,000	2*	πυρετός
301*	60,000	2*	αρρυθμίες

Πίνακας 13: 3-Διαφορετικός πίνακας

Εκτός αυτού μια άλλη βασική αδυναμία της l-ποικιλομορφίας είναι πως δεν γίνεται να εφαρμοστεί εύκολα σε μεγάλες βάσεις δεδομένων. Για παράδειγμα, σε μία βάση δεδομένων με 10000 εγγραφές χρειάζονται 100 ισοδύναμες κλάσεις για να ικανοποιηθεί η l-diversity για l=2.

4.2 Ανατομία (Anatomy)

Προκειμένου να προστατέψουμε τα ευαίσθητα προσωπικά μας δεδομένα τα ανωνυμοποιούμε με διάφορους τρόπους όπως αναφέραμε και πιο πάνω. Αυτές οι τροποποιήσεις όσο και να κάνουν πιο δύσκολο το έργο του επιτιθέμενου έχουν και σαν αποτέλεσμα την απώλεια χρήσιμης πληροφορίας, με αποτέλεσμα τα τελικά δεδομένα να μην είναι αξιόπιστα.

Η τεχνική της Ανατομίας έχει ως στόχο την μείωση της απώλειας πληροφορίας (Xiao & Tao, 2006b). Πράγμα που επιτυγχάνει καθώς στα δημοσιευμένα δεδομένα παραμένουν οι αρχικές τιμές των πλειάδων, ενώ αποκρύπτεται η συσχέτιση κάθε εγγραφής με την ευαίσθητη τιμή τους.

Αλγόριθμος της Ανατομίας

QIT=∅; ST=∅; gcnt=0

Κατακερμάτισε τις εγγραφές στο T με βάση τις τιμές του A^s (ένα καλάθι για κάθε τιμή)

/*Πρώτη Φάση: Δημιουργία μιας νέας κλάσης ισοδυναμίας*/

while υπάρχουν τουλάχιστον ℓ μη κενά καλάθια do

gcnt = gcnt + 1

QI_{gnct}=∅;

S=Το σύνολο των l μεγαλύτερων καλαθιών

for κάθε καλάθι στο BC do

αφαίρεση μίας τυχαίας εγγραφής t από το S

$QI_{gnct=}$ QI_{gnct}∪{t}

end for

end while

/*Δεύτερη Φάση: Διαμοιρασμός των υπολοίπων εγγράφων*/

for κάθε μη κενό καλάθι do

/*Αυτό το καλάθι έχει ακριβώς μία εγγραφή*/

t = η μοναδική εγγραφή στο καλάθι

S’ = το σύνολο των κλάσεων ισοδυναμίας του προηγούμενου σταδίου που δεν περιέχουνε την τιμή *t*[d+1] στην S

ανάθεση της *t* σε μία τυχαία κλάση ισοδυναμίας στο BC’

end for

/*Δημιουργία του QIT και του ST*/

for j = 1 to gcnt do

for κάθε εγγραφή $t \in QI_j$ do

εισαγωγή της εγγραφής (t[1], …, t[d], j) στο QIT

end for

for κάθε μοναδική τιμή ν της A^s στο QI_j do

c_j(ν) = το πλήθος των εγγραφών στο QI_j με τιμή ν στην A^s

εισαγωγή της (j, ν, cj(ν)) στο ST

end for

end for

return QIT και ST

Μεθοδολογία της ανατομίας

Με την χρήση της μεθόδου ανατομίας κατασκευάζονται δύο τελικοί νέοι πίνακες, ο QIT και ο ST ώστε να ικανοποιηθεί η l-ποικιλομορφία. Αρχικά για κάθε εγγραφή t του T συμβολίζουμε με *[i] (1 <=i <= d)* την τιμή του γνωρίσματος Qi και με t[d+1] την τιμή του S.

Οι πίνακες που παράγονται από την ανατομία είναι ο ψευδοαναγνωριστικός QIT και ο ευαίσθητων τιμών ST με βάση τις ακόλουθες ιδιότητες:

1. Το QIT έχει το σχήμα (Q1,...,Qd,Σύνολο-ΑΑ).
2. Για κάθε κλάση ισοδυναμίας QIj($1 \leq j \leq m$) και για κάθε εγγραφή t του QIj, το QIT έχει μία εγγραφή της μορφής: (t[1],t[2],...,t[d],j).
3. Το ST έχει σχήμα (Σύνολο-ΑΑ,S,Πλήθος).
4. Για κάθε κλάση ισοδυναμίας Qij ($1 \leq j \leq m$) και για κάθε τιμή ν του ευαίσθητου γνωρίσματος S στο QIj, το ST έχει μία εγγραφή της μορφής: (j,ν,cj(ν)), όπου cj(u) ο αριθμός των εγγραφών t του QIj έτσι ώστε t[d+1]=u.

4.3 Μ-αμεταβλητοτητα (m-Invariance)

Η μ-αμεταβλητότητα (m-Invariance) είναι μια μεθοδολογία της l-diversity η οποία κρατά συνεχώς ενημερωμένη τη βάση δεδομένων καθώς υποστηρίζει την εκ νέου δημοσίευση των δεδομένων, στη περίπτωση που υπάρξουν αλλαγές στη βάση, όπως η προσθήκη και η διαγραφή δεδομένων.

Για παράδειγμα έστω ότι ένα νοσοκομείο δημοσιεύει τα δεδομένα των ασθενών του κάθε χρόνο. Έχοντας σαν βάση τον αρχικό πίνακα 14, ανωνυμοποιείται και δημοσιεύεται ο γενικευμένος πίνακας 15. Στη συνέχεια μετά από ένα χρόνο με βάση τον αρχικό πίνακα 16, δημοσιεύεται ο πίνακας 17.

ΑΡΧΙΚΟΣ ΠΙΝΑΚΑΣ			
ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ			
ΟΝΟΜΑ	**ΗΛΙΚΙΑ**	**ΤΑΧ. ΚΩΔΙΚΑΣ**	**ΑΣΘΕΝΕΙΑ**
Κώστας	31	22000	Αμυγδαλίτιδα
Βασίλης	32	24000	Βρογχίτιδα
Γιώργος	34	28000	Γρίπη
Σταύρος	33	35000	Οίδημα
Μαρία	51	30000	Γαστρίτιδα
Κατερίνα	46	37000	Αμυγδαλίτιδα
Βάσω	47	43000	Οίδημα
Γιάννης	50	45000	Γαστρίτιδα
Έλενα	53	36000	Γρίπη
Σοφία	62	43000	Γρίπη
Παύλος	66	44000	Βρογχίτιδα

Πίνακας 14: Αρχικός πίνακας κατά την πρώτη δημοσίευση

ΓΕΝΙΚΕΥΜΕΝΟΣ ΠΙΝΑΚΑΣ			
ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ			
ΚΛΑΣΗ	**ΗΛΙΚΙΑ**	**ΤΑΧ. ΚΩΔΙΚΑΣ**	**ΑΣΘΕΝΕΙΑ**
1	[31,32]	[22000-24000]	Αμυγδαλίτιδα
1	[31,32]	[22000-24000]	Βρογχίτιδα
2	[33,34]	[28000-35000]	Γρίπη
2	[33,34]	[28000-35000]	Οίδημα
3	[46,51]	[30000-37000]	Γαστρίτιδα
3	[46,51]	[30000-37000]	Αμυγδαλίτιδα
4	[47,53]	[36000-45000]	Οίδημα
4	[47,53]	[36000-45000]	Γαστρίτιδα
4	[47,53]	[36000-45000]	Γρίπη
5	[62,66]	[43000-44000]	Γρίπη
5	[62,66]	[43000-44000]	Βρογχίτιδα

Πίνακας 15: Γενικευμένος πίνακας κατά την πρώτη δημοσίευση

ΑΡΧΙΚΟΣ ΠΙΝΑΚΑΣ			
ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ			
ΟΝΟΜΑ	**ΗΛΙΚΙΑ**	**ΤΑΧ. ΚΩΔΙΚΑΣ**	**ΑΣΘΕΝΕΙΑ**
Κώστας	31	22000	Αμυγδαλίτιδα
Σταύρος	33	35000	Οίδημα
Σταυρούλα	35	31000	Γρίπη
Βάσω	47	43000	Οίδημα
Έλενα	53	36000	Γρίπη
Μαρία	51	30000	Γαστρίτιδα
Χρήστος	56	40000	Γαστρίτιδα
Θωμάς	64	41000	Δυσπεψία
Παύλος	66	44000	Βρογχίτιδα
Μαρία	70	54000	Γαστρίτιδα
Θάνος	75	46000	Βρογχίτιδα

Πίνακας 16: Αρχικός πίνακας κατά τη δεύτερη δημοσίευση

ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ			
ΚΛΑΣΗ	**ΗΛΙΚΙΑ**	**ΤΑΧ. ΚΩΔΙΚΑΣ**	**ΑΣΘΕΝΕΙΑ**
1	[31-33]	[22000-35000]	Αμυγδαλίτιδα
1	[31-33]	[22000-35000]	Οίδημα
2	[35-53]	[31000-43000]	Γρίπη
2	[35-53]	[31000-43000]	Οίδημα
2	[35-53]	[31000-43000]	Γρίπη
3	[51-56]	[30000-40000]	Γαστρίτιδα
3	[51-56]	[30000-40000]	Γαστρίτιδα
4	[64-66]	[41000-44000]	Δυσπεψία
4	[64-66]	[41000-44000]	Βρογχίτιδα
5	[70-75]	[46000-54000]	Γαστρίτιδα
5	[70-75]	[46000-54000]	Βρογχίτιδα

Πίνακας 17: Γενικευμένος πίνακας κατά τη δεύτερη δημοσίευση

Οι ασθενείς Βασίλης, Γιώργος, Κατερίνα, Γιάννης και Σοφία έχουν διαγραφεί από τη βάση δεδομένων. Στη θέση τους έχουν προστεθεί οι ασθενείς Σταυρούλα, Χρήστος, Θωμάς, Μαρία και Θάνος. Το πρόβλημα είναι ότι ο επιτιθέμενος μπορεί να αναγνωρίσει τη ταυτότητα ενός ασθενή συνδέοντας τους πίνακες 15 και 17, παρόλο που και οι δύο δημοσιευμένοι πίνακες ικανοποιούν το 2-anonymity και το 2-diversity.

Για παράδειγμα, έστω ότι για τον επιτιθέμενο είναι γνωστά, η ηλικία και ο ταχυδρομικός κώδικας του Κώστα, και ότι αυτά τα δεδομένα έχουν δημοσιευτεί και στους δύο πίνακες. Από τον πίνακα 15 ο επιτιθέμενος συμπεραίνει ότι ο Κώστας έχει αμυγδαλίτιδα ή βρογχίτιδα. Αντίστοιχα από τον πίνακα 17 ο αντίπαλος βρίσκει ότι ο Κώστας πάσχει είτε από αμυγδαλίτιδα είτε από οίδημα. Επομένως με τη σύνδεση αυτών των πινάκων, ο αντίπαλος εύκολα μπορεί να συμπεράνει ότι ο Κώστας νοσηλεύεται λόγω αμυγδαλίτιδας.

Με τη μέθοδο όμως της μ-αμεταβλητότητας ο πίνακας 17 αντικαθιστάται με τον πίνακα 18, ο οποίος περιλαμβάνει τις πλαστές εγγραφές Φ1 και Φ2, καθώς επίσης τις γενικευμένες τιμές του πίνακα 16.

ΠΙΝΑΚΑΣ ΜΕ ΠΛΑΣΤΕΣ ΕΓΓΡΑΦΕΣ				
ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ				
ΟΝΟΜΑ	**ΚΛΑΣΗ**	**ΗΛΙΚΙΑ**	**ΤΑΧΥΔ.ΚΩΔΙΚΑΣ**	**ΑΣΘΕΝΕΙΑ**
Κώστας	1	[31-32]	[22000-24000]	Αμυγδαλίτιδα
Φ1	1	[31-32]	[22000-24000]	Βρογχίτιδα
Σταύρος	2	[33-35]	[31000-35000]	Οίδημα
Σταυρούλα	2	[33-35]	[31000-35000]	Γρίπη
Βάσω	3	[47-53]	[36000-43000]	Οίδημα
Φ2	3	[47-53]	[36000-43000]	Οίδημα
Έλενα	3	[47-53]	[36000-43000]	Γρίπη
Μαρία	4	[51-56]	[30000-40000]	Γαστρίτιδα
Χρήστος	4	[51-56]	[30000-40000]	Γαστρίτιδα
Θωμάς	5	[64-66]	[41000-44000]	Δυσπεψία
Παύλος	5	[64-66]	[41000-44000]	Βρογχίτιδα
Μαρία	6	[70-75]	[46000-54000]	Γαστρίτιδα
Θάνος	6	[70-75]	[46000-54000]	Βρογχίτιδα

Πίνακας 18: Πίνακας με πλαστές εγγραφές

Το πλεονέκτημα είναι ότι ο επιτιθέμενος δεν μπορεί να ξεχωρίσει τις πλαστές εγγραφές από τις υπόλοιπες στην ίδια κλάση ισοδυναμίας. Για παράδειγμα, στη περίπτωση του Κώστα, ο επιτιθέμενος δεν μπορεί να καταλάβει από τι ασθένεια πάσχει ο Κώστας, καθώς οι κλάσεις ισοδυναμίας στους πίνακες 15 και 18 έχουν πλέον το ίδιο σύνολο ευαίσθητων εγγραφών.

Δηλαδή η μ-αμεταβλητότητα απαιτεί μια εγγραφή να ανήκει πάντα σε μια κλάση ισοδυναμίας, η οποία έχει το ίδιο σύνολο ευαίσθητων ιδιοτήτων, για όλες τις δημοσιεύσεις, καθώς επίσης να ικανοποιείται ταυτόχρονα η μ-διαφορετικότητα.

4.4 Τ-εγγύτητα (t-closeness)

Η τ-εγγυτητα είναι μια έννοια ιδιωτικότητας που μπορεί να αναπαραστήσει το γενικό γνωστικό υπόβαθρο του επιτιθέμενου πάνω στη κατανομή των τιμών του ευαίσθητου γνωρίσματος. Δηλαδή δεν έχει τη δυνατότητα της προστασίας από επιθέσεις που έχουν σκοπό την αποκάλυψη της ταυτότητας μιας εγγραφής από τη βάση δεδομένων. Για την επίτευξη αυτού θα πρέπει να χρησιμοποιούνται ταυτόχρονα και η k-ανωνυμία και η τ-εγγύτητα. Στην ουσία σκοπός της τ-εγγυτητας είναι να

περιορίσει τη διαφοροποίηση στη γνώση του επιτιθέμενου μεταξύ της γνώσης που αποκτά από το δημοσιευμένο σύνολο των δεδομένων αναφορικά με τη κατανομή των τιμών του ευαίσθητου γνωρίσματος και της γνώσης που αποκτά για τη κατανομή των ευαίσθητων τιμών στη κλάση ισοδυναμίας που βρίσκεται η εγγραφή που αναζητά. Μία κλάση ισοδυναμίας ικανοποιεί την τ-εγγυτητα αν η απόσταση της κατανομής των τιμών του ευαίσθητου γνωρίσματος μέσα στην κλάση ισοδυναμίας από την κατανομή των ευαίσθητων τιμών του γνωρίσματος στο σύνολο των δεδομένων δεν υπερβαίνει το άνω όριο ***t.***

Ένας πίνακας ικανοποιεί την τ-εγγύτητα αν όλες οι κλάσεις ισοδυναμίας του την ικανοποιούν. Ο καλύτερος τρόπος μέτρησης της απόστασης μεταξύ των δύο κατανομών πιθανότητας και των τιμών του ευαίσθητου γνωρίσματος είναι η EMD(Earth Mover's Distance) σύμφωνα με τους (Li, T., & S., 2007) .

4.4.1 Μεθολογία τ-Εγγύτητας

Έστω P=(p1, p2, ...pm), Q=(q1, q2, ...qm) και dij η σταθερή απόσταση μεταξύ των τιμών pi και qj. Θέλουμε να βρούμε την ροή F = [fij], όπου fij είναι η ροή της μάζας από το pi στο qj,έτσι ώστε να ελαχιστοποιηθεί το συνολικό έργο.

$$\text{WORK}(P,Q,F)=\sum_{i=1}^{m}\sum_{j=1}^{m} dij\, fij$$

Ισχύουν οι παρακάτω περιορισμοί:

$$fij \geq 0 \quad , \quad 1\leq i \leq m,\ 1 \leq j \leq m$$

$$pi - \sum_{j=1}^{m} fij + \sum_{j=1}^{m} fji = qi \ , \ 1 \leq i \leq m$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} fij = \sum_{=1}^{m} pi = \sum_{i=1}^{m} qi = 1$$

Το συνολικό έργο το οποίο είναι ίσο με την ***EMD*** και δίνεται από τη σχέση:

$$D[P,Q]= \text{WORK}(P,Q,F)=\sum_{i=1}^{m}\sum_{j=1}^{m} dij\, fij$$

4.5 δ-παρουσία (δ-Presence)

Μία άλλη τεχνική που έχει αναπτυχθεί για τη προστασίας της ιδιωτικότητας είναι η δ-παρουσία. Συγκεκριμένα εμποδίζει τον επιτιθέμενο να συμπεράνει αν κάποιο άτομο βρίσκεται σε μια συγκεκριμένη βάση δεδομένων. Αυτό το επιτυγχάνει με την ανωνυμοποίηση της βάσης δεδομένων καθώς έτσι ο επιτιθέμενος δεν μπορεί να είναι σίγουρος αν κάποιο άτομο βρίσκεται στη συγκεκριμένη βάση δεδομένων με πιθανότητα μεγαλύτερη από δ. Ωστόσο δεν μπορεί να αντιμετωπίσει τη πρόκληση που ένας επιτιθέμενος γνωρίζει ήδη ότι ένα άτομο βρίσκεται σίγουρα σε μία βάση δεδομένων. Αυτό έχει εφαρμογή σε πολλές περιπτώσεις. Για παράδειγμα, ας υποθέσουμε ότι έχουμε μία φορολογική βάση δεδομένων, τότε ένας φίλος του Γιάννη που ξέρει ότι ο Γιάννης δουλεύει εκεί, θα είναι σίγουρος ότι ο Γιάννης θα βρίσκεται στη συγκεκριμένη βάση δεδομένων.

Λαμβάνοντας υπόψη ένα εξωτερικό γνωστικό υπόβαθρο P και έναν πίνακα T έχουμε:

δ = (δmin, δmax)-της υπάρχουσας παρουσίας για μία γενίκευση T* του T αν $\delta min \leq Pr(t \in T \mid T^*,P) \leq \delta max$

για κάθε $t \in P$

4.6 de-identification

Η de-identification (επαναταυτοποίηση) είναι μία διαδικασία που χρησιμοποιείται στην ιατρική πληροφορική, για να αποτρέψει τη σύνδεση της ταυτότητας ενός ατόμου με περαιτέρω πληροφορίες. Συχνές χρήσεις της de-identification περιλαμβάνουν έρευνες ανθρώπινων ζητημάτων για λόγους προστασίας της ιδιωτικής ζωής για τους συμμετέχοντες της έρευνας. Είναι μια κοινή στρατηγική για τη διαγραφή ή συγκάλυψη προσωπικών στοιχείων, όπως το όνομα, ο αριθμός κοινωνικής ασφάλισης και για τη γενίκευση αναγνωριστικών όπως η ημερομηνία γέννησης και ο ταχυδρομικός κώδικας. (Privacy Analytics, 2015)

Παράδειγμα:

Μία έρευνα που διεξάγεται, (απογραφή) για τη συλλογή πληροφοριών σχετικά με μια ομάδα ανθρώπων. Για να ενθαρρυνθεί η συμμετοχή και η προστασία της ιδιωτικής ζωής των συμμετεχόντων στην έρευνα, οι ερευνητές προσπαθούν να σχεδιάσουν την έρευνα με τέτοιο τρόπο έτσι ώστε οι άνθρωποι να μπορούν να συμμετέχουν στην έρευνα και όταν το αποτέλεσμα δημοσιεύεται να μην είναι δυνατόν να ταιριαστούν τα στοιχεία του οποιουδήποτε συμμετέχοντα με οποιαδήποτε στοιχεία που δημοσιεύονται στο αποτέλεσμα.

Εφαρμογή:

Η de-identification χρησιμοποιείται κυρίως για τη προστασία πληροφοριών που αφορούν την υγεία. Ορισμένες βιβλιοθήκες έχουν υιοθετήσει τις μεθόδους που χρησιμοποιούνται στο κλάδο της υγείας για να διαφυλαχθεί η ιδιωτική ζωή των αναγνωστών τους.

Κεφάλαιο 5

Τεχνική της km-ανωνυμίας

5.1 km-Ανωνυμία (km-Anonymity)

Παρά τους τρόπους προστασίας των ευαίσθητων προσωπικών μας δεδομένων, όπως αναφέραμε παραπάνω, ο κίνδυνος για την παραβίαση τους παραμένει. Αυτό συμβαίνει γιατί ο επιτιθέμενος μπορεί να έχει διάφορες πληροφορίες για το θύμα του σε διάφορες μορφές. Επίσης, είναι πιθανόν το κάθε δημοσιευμένο μοντέλο να είναι με τέτοιο τρόπο κωδικοποιημένο ώστε κάθε φορά να χρήζει διαφορετικής επεξεργασίας ώστε να εξασφαλιστεί η ιδιωτικότητά τους.

Οι M. Terrovitis, N. Mamoulis, P. Kanlis αντιμετωπίζοντας τα δεδομένα ως πιθανά ψευδοαναγνωριστικά και ως πιθανά ευαίσθητα δεδομένα, προσπάθησαν να δώσουν λύση στο πρόβλημα, διατυπώνοντας την μέθοδο της km-ανωνυμία. (Terrovitis, Mamoulis, & Kalnis, 2008)

Έτσι για παράδειγμα, υποθέτουμε πως κάποιος πραγματοποίησε αγορά κάποιων ενδυμάτων σε μια επίσκεψή του σε συγκεκριμένο πολυκατάστημα, τα οποία ήταν γούνες, παλτό, τζιν παντελόνια, βερμούδες παντελόνια, φορέματα και πυτζάμες. Βγαίνοντας στο δρόμο, κάποιος γνωστός του είδε τα προϊόντα που βρίσκονταν σε μια από τις σακούλες, δηλαδή το παλτό, τις γούνες και το τζιν. Εάν στο δημοσιευμένο πίνακα με τις αγορές εκείνης της μέρας, υπήρχε μόνο μια εγγραφή που να περιείχε το παλτό, τις γούνες και το τζιν, αμέσως γνωστός αποκτά γνώση όλων των προϊόντων τα οποία είχε αγοράσει το πρόσωπο.

Σύμφωνα με τον ορισμό της km-ανωνυμίας (Terrovitis, Mamoulis, & Kalnis, 2008) , εάν ο επιτιθέμενος γνωρίζει το πολύ m στοιχεία , για κάθε υποσύνολο m ή λιγότερων στοιχείων, θα πρέπει να υπάρχουν στον πίνακα τουλάχιστον k άλλες για το υποσύνολο αυτό. Κατά συνέπεια, ο επιτιθέμενος του παραδείγματός μας που έχει γνώση τριών στοιχείων θα ήταν δύσκολο να αναγνωρίσει τα ψώνια του γνωστού του μεταξύ πέντε άλλων αγορών εάν ο δημοσιοποιημένος πίνακας ήταν 5^3-ανώνυμος.

Παρακάτω βλέπουμε ένα παράδειγμα ενός πίνακα οποίος δεν ικανοποιεί την 2^2-ανωνυμία καθώς εάν ο επιτιθέμενος γνωρίζει δύο στοιχεία, τζιν παντελόνι και γούνες, τότε μπορεί εύκολα να καταλάβει τι άλλο έχει αγοράσει ο γνωστός του, στην συγκεκριμένη περίπτωση τις βερμούδες.

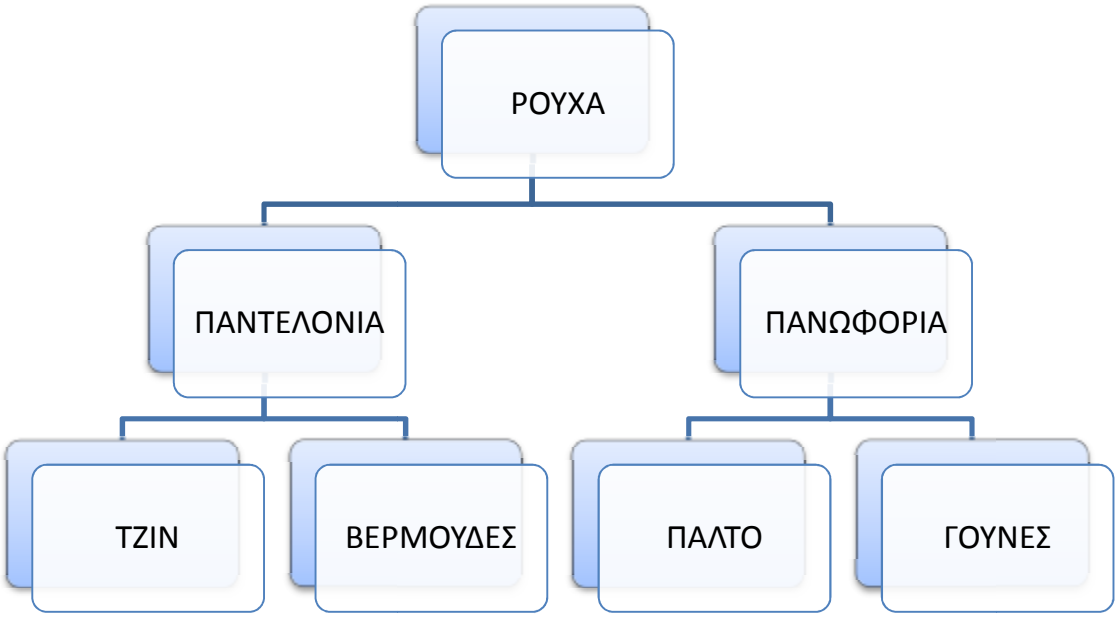
Α/Α	**ΠΡΟΙΟΝΤΑ**
1	Φορέματα, τζιν, βερμούδες
2	Βερμούδες, γούνες, τζιν
3	Γούνες, πυτζάμες, φορέματα
4	Τζιν, πυτζάμες, φορέματα

5.1.2 Μοντέλο γενίκευσης

Όπως αναφέρθηκε στην αρχή της διπλωματικής εργασίας, με τον όρο γενίκευση (generalization) ορίζουμε την διαδικασία κατά την οποία οι τιμές των ψευδο-αναγνωριστικών (Quasi Identifiers) αντικαθίστανται με γενικότερες ή με βάση συγκεκριμένες ιεραρχίες γενίκευσης. Στόχος της γενίκευσης είναι η διατήρηση μέρους της πληροφορίας της αρχικής τιμής χωρίς αυτή να αλλάζει και να αλλοιώνεται πλήρως.

Συγκεκριμένα η μέθοδος της km-ανωνυμίας χρησιμοποιεί την τεχνική της γενίκευσης για να επιτευχθεί η ανωνυμοποίηση των δεδομένων.

Έτσι στο παράδειγμά μας, αξιοποιείται η ιεραρχία της γενίκευσης, γενικεύοντας τα τζιν παντελόνια και τα βερμούδες σε παντελόνια και τα παλτό και τις γούνες σε πανωφόρια.



Από την παραπάνω γενίκευση προκύπτει ο ακόλουθος πίνακας:

Α/Α	**ΠΡΟΙΟΝΤΑ**
1	Παντελόνια, φορέματα, παλτό
2	Παντελόνια, φορέματα
3	Παντελόνια, φορέματα, παλτό
4	Παντελόνια, παλτό

Διαπιστώνουμε λοιπόν πως ο παραπάνω πίνακας ικανοποιεί την 2^2-ανωνυμία καθώς κάποιος που θέλει να μάθει τις συναλλαγές ενός θύματος και γνωρίζει το πολύ 2 στοιχεία από τον πίνακα, δεν θα μπορέσει να αναγνωρίσει τι ψώνισε.

Η τεχνική της γενίκευσης μπορεί να βοηθάει ώστε να επιτευγχθεί η ανωνυμοποίηση των δεδομένων αποτελεσματικά, δεν παύει όμως να παρουσιάζει κάποια μειονεκτήματα. Αρχικά, όταν γενικεύεται μία τιμή για να ικανοποιηθεί η km-ανωνυμία, αυτόματα γενικεύονται όλες οι τιμές που βρίσκονται στον ίδιο κλάδο. Αυτό έχει ως συνέπεια να γίνονται άσκοπες γενικεύσεις σε στοιχεία που υπό άλλες συνθήκες δεν θα χρειάζονταν να γενικευτούν καθώς δεν βοηθάνε το έργο του επιτιθέμενο. Επιπλέον, με τις ανούσιες γενικεύσεις χάνεται αρκετός χρόνος μιας και η μέθοδος της γενίκευσης είναι χρονοβόρα διαδικασία.

Ένα άλλο εξίσου σημαντικό πρόβλημα που προκύπτει από την εφαρμογή της διαδικασίας της γενίκευσης είναι ότι χάνεται πολύτιμη πληροφορία από την γενίκευση όλων των στοιχείων ενός κλάδου. Καθώς επίσης η km-ανωνυμία δεν επιτυγχάνεται με τον ίδιο τρόπο σε όλα τα δεδομένα αλλά εξαρτάται κάθε φορά από την διάταξη της ιεραρχίας γενίκευσης. Για παράδειγμα, ένα σύνολο δεδομένων μπορεί να ανωνυνομοποιηθεί ακολουθώντας την γενίκευση και να τηρεί την km-ανωνυμία με έναν τρόπο γενικεύοντας συγκεκριμένα στοιχεία σε άλλες κατηγορίες πιο γενικές. Ωστόσο, το ίδιο σύνολο δεδομένων θα μπορούσε να γενικευτεί μ' έναν διαφορετικό τρόπο, γενικεύοντας άλλα στοιχεία σε διαφορετικές γενικεύσεις.

5.1.3 Apriori αλγόριθμος ανωνυμοποίησης

Ο αλγόριθμος Apriori διατυπώθηκε το 1994 από τους Agarwal και Srikant με σκοπό να λειτουργήσει σε βάσεις δεδομένων που περιλάμβαναν συναλλαγές (π.χ. προϊόντα που αγοράστηκαν από καταναλωτές) (Apriori algorithm, n.d.) .

Η km-ανωνυμία εκμεταλλεύτηκε την αρχή της apriori ιδιότητας σύμφωνα με την οποία όταν ένα σύνολο J παραβιάζει την ιδιωτικότητα μιας βάσης, τότε όλα τα υποσύνολα του J θα παραβιάζουν αντίστοιχα την ιδιωτικότητα της βάσης (Terrovitis, Mamoulis, & Kalnis, 2008). Ο αλγόριθμος Apriori λειτουργεί με βάση ότι αρχικά ο επιτιθέμενος γνωρίζει μια τιμή από το σύνολο του ψευδοαναγνωριστικού και αρχίζει να ελέγχει αν υπάρχουν παραβιάσεις στην ιδιωτικότητα. Η παραπάνω διαδικασία ελέγχου επαναλαμβάνεται για δυο έως m τιμές. Το θετικό με τον συγκεκριμένο αλγόριθμο είναι ότι χρησιμοποιεί τις γενικεύσεις που έχει κάνει στα προηγούμενα βήματα με αποτέλεσμα να μειώνεται ο αριθμός των γενικεύσεων στο επόμενο βήμα i+1.

Ο apriori για να υλοποιηθεί χρησιμοποιεί την διαδικασία της γενίκευσης σε όλους τους πιθανούς συνδυασμούς τιμών για μεγέθη από i={1,2,3,…,m}. Σε κάθε επανάληψη i ο αλγόριθμος καταγράφει σ' ένα δέντρο συχνοτήτων (count-tree) τα αποτελέσματα κάθε συνδυασμού. Έπειτα καταγράφει στο count tree τις τιμές που παίρνουν οι κόμβοι-φύλλα που εμφανίζονται λιγότερες φορές από k. Για κάθε μία από τις τιμές που βρίσκει επιστρέφει στο δέντρο γενίκευσης και αντικαθιστά τις

προβληματικές τιμές με άλλες πιο γενικευμένες, ώστε να εμφανίζεται συχνότερα κάθε τιμή σε πλήθος μεγαλύτερο από k. Η διαδικασία επαναλαμβάνεται για κάθε προβληματική τιμή που έχει εντοπιστεί στο δέντρο συχνοτήτων.

Αλγόριθμος Apriori Ανωνυμοποίησης

AA (D,i,k,m,c)

όπου:
k: παράμετρος ανωνυμίας
m: μέγιστη γνώση επιτιθέμενου
c: γενίκευση

Αρχικοποίηση δέντρου ιεραρχίας γενίκευσης

για *i*=1 μέχρι m

δημιουργία νέου δέντρου συχνοτήτων (count-tree) για όλες τις εγγραφές *t*∈*D*

ενημέρωση του δέντρου με όλους τους συνδυασμούς μεγέθους *i* της εγγραφής
για όλα τα φύλλα *V* του δέντρου συχνοτήτων
εάν το *support(V)* < *k*

εύρεση γενίκευσης στο δέντρο ιεραρχίας τέτοια ώστε *support(V)* >= *k*
ανωνυμοποίηση των δεδομένων και ενημέρωση του δέντρου συχνοτήτων

επιστροφή *c*=γενικεύσεις

Κεφάλαιο 6

Συμπεράσματα και μελλοντικές επεκτάσεις

6.1 Ασφαλές δίκτυα και τέλεια ανωνυμοποίηση

Αρχικά , η έννοια του ασφαλές δικτύου είναι πολύ γενική και καθόλου απόλυτη. Παρόλο που γοητεύει τους περισσότερους χρήστες δεν μπορούμε να ταξινομήσουμε ένα δίκτυο σε ασφαλή ή μη ασφαλή, καθώς εξαρτάται καθαρά από τον ιδιοκτήτη του δικτύου μέχρι που επιτρέπει την πρόσβαση σ' αυτό.

Για παράδειγμα, ένας οργανισμός που διατηρεί πολύτιμα εμπορικά μυστικά θα θέλει να εμποδίζει τους ξένους από το να αποκτούν πρόσβαση στους υπολογιστές του. Από την άλλη, μια εταιρεία που διαθέτει μια τοποθεσία Ιστού η οποία κάνει κάποιες πληροφορίες διαθέσιμες στο κοινό, μπορεί να ορίζει ως ασφαλές ένα δίκτυο που επιτρέπει την πρόσβαση στα δεδομένα, αλλά απαγορεύει την αλλαγή των δεδομένων αυτών από τρίτους. Άλλοι φορείς, πάλι εστιάζουν την προσοχή τους στο να διατηρούν τις επικοινωνίες εμπιστευτικές. Αυτοί ορίζουν ως ασφαλές ένα δίκτυο στο οποίο κανένας άλλος εκτός από τον αποστολέα και τον τελικό αποδέκτη δεν μπορεί να υποκλέψει και να διαβάσει ένα μήνυμα. Πολλοί μεγάλοι οργανισμοί χρειάζονται ένα σύνθετο ορισμό της ασφάλειας, που να επιτρέπει την πρόσβαση σε κάποια επιλεγμένα δεδομένα, ενώ εμποδίζει την πρόσβαση ή την τροποποίηση των ευαίσθητων δεδομένων τα οποία διατηρούνται εμπιστευτικά.

Επειδή δεν υπάρχει απόλυτος ορισμός του ασφαλούς δικτύου, το πρώτο βήμα που πρέπει να κάνει ένας οργανισμός για να επιτύχει ένα ασφαλές σύστημα είναι να ορίσει την πολιτική ασφαλείας του (security policy). Η πολιτική αυτή δεν καθορίζει πως θα επιτευχθεί η προστασία. Καθορίζει όμως ρητά και με σαφήνεια τα στοιχεία που πρέπει να προστατεύονται. (Comer, 2004)

Έτσι δεν μπορούμε να μιλάμε για τέλεια ανωνυμοποίηση. Αφενός γιατί δεν ξέρουμε μέχρι που θέλει κάποιος να προστατέψει μια βάση δεδομένων και πως ορίζεται η τέλεια ανωνυμοποίηση για τον καθένα και αφετέρου γιατί σε όλες τις τεχνικές και τις μεθόδους ανωνυμοποίησης που έχουν διατυπωθεί μέχρι στιγμής εντοπίζονται κενά που επιτρέπει στον εκάστοτε επιτιθέμενο να αποσπάσει στοιχεία.

6.2 Λογισμικά Ανωνυμοποίησης

Τα λογισμικά ανωνυμοποίησης είναι λογισμικά ανοιχτού κώδικα έχουν ως σκοπό να ανωνυμοποιήσουν τα ευαίσθητα προσωπικά δεδομένα και να συμβάλλουν στο έργο των τεχνικών ανωνυμοποίησης. Συνήθως είναι εύκολα στη χρήση και υποστηρίζουν ανωνυμίες με βάση το ρίσκο, επιπλέον μεθόδους με την ανάλυση της χρησιμότητας δεδομένων και την επαναταυτοποίηση του κινδύνου. Μερικά από τα κύρια χαρακτηριστικά τους είναι ότι παρέχουν βοήθεια στις τεχνικές προστασίας ευαίσθητων προσωπικών δεδομένων όπως τις k-ανωνυμία, l-diversity, t-closeness ή δ-presence. Μερικά από τέτοια λογισμικά ανωνυμοποίησης είναι τα ακόλουθα:

- ARX open source data anonymization tool

  Link: https://github.com/arx-deidentifier/arx

- ARX Data Anonymization Tool

  Link: http://arx.deidentifier.org/

- UTD Anonymization ToolBox
  Link: http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php

- FDiversity: a software package for the integrated analysis of functional diversity
  Link: http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2010.00082.x/full

- SECRETA: System for Evaluating and Comparing RElational and Transaction Anonymization algorithms
  Link http://secreta.uop.gr/

- InfoSphere Optim Data Privacy

  Link: http://www-03.ibm.com/software/products/en/infosphere-optim-data-privacy/

6.3 Σύνοψη και συμπεράσματα

Ο πολίτης είναι ο πρώτος κερδισμένος από τις προσπάθειες διασφάλισης των επώνυμων πληροφοριών μιας και οι επιθέσεις αυξάνονται ολοένα με την εξέλιξη της τεχνολογίας.

Η συγκεκριμένη πτυχιακή εργασία ασχολήθηκε με το πρόβλημα της διασφάλισης της ιδιωτικότητας των εγγράφων σε συλλογές δεδομένων με συνεχή γνωρίσματα. Θεωρήθηκε η περίπτωση επίθεσης στην οποία ο επιτιθέμενος έχει γνώση μερικών τιμών μιας εγγραφής του συνόλου δεδομένων.

Αναπτύχθηκαν όλες οι τεχνικές για την επίτευξη της διασφάλισης της ιδιωτικότητας σε μια βάση δεδομένων με έμφαση στη μεθοδολογία της k-ανωνυμίας.

Η k-ανωνυμία είναι η πρώτη και η πιο διαδεδομένη τεχνική ανωνυμοποίησης . Αναπτύχθηκαν αλγόριθμοι με στόχο την εγγύηση της ανωνυμίας των εγγραφών χρησιμοποιώντας τη γενίκευση και την απόκρυψη εγγραφών.

Οι αδυναμίες της k-ανωνυμίας οδήγησαν στη δημιουργία νέων μεθόδων.

Μια από αυτές είναι η l-diversity, η οποία διατυπώθηκε για πρώτη φορά το 2007 και έχει στόχο την διασφάλιση της ανωνυμίας των δεδομένων και την ταυτότητα της εγγραφής. Χρησιμοποιεί την μεθοδολογία της μ-αμεταβλητότητας και του αλγορίθμου της ανατομίας.

Στη συνέχεια διατυπώθηκαν και καταγράφηκαν αρκετές άλλες μέθοδοι με σκοπό να καλύψουν τα κενά των προηγούμενων καθώς επίσης και σε εξειδικευμένες τεχνικές πάνω σε συγκεκριμένους τομείς, όπως ο d-identification που χρησιμοποιείται στην υγεία και στην ιατρική πληροφορική.

Συμπεραίνεται πως για το πρόβλημα της προστασίας της ιδιωτικότητας σε σύνολα δεδομένων, από επιθέσεις με μερική γνώση σε κάποιες τιμές των γνωρισμάτων του ψευδο-αναγνωριστικού μιας εγγραφής, ο αλγόριθμος που επιτυγχάνει μικρότερη απώλεια πληροφορίας είναι αυτός της k-ανωνυμίας με χρήση ιεραρχιών γενίκευσης. Επιπρόσθετα, αξίζει να σημειωθεί ότι μπορεί να γίνει συνδυασμός περισσότερων του ενός αλγορίθμου για ην επίτευξη της βέλτιστης ανωνυμοποίησης.

6.4 Μελλοντικές επεκτάσεις

Οι αλγόριθμοι και οι τεχνικές οι οποίες αναπτύχθηκαν στην παρούσα πτυχιακή εργασία, αναζητούν και εφαρμόζουν αποκρύψεις και γενικεύσεις οι οποίες είναι χρήσιμες ώστε να επιτευχθεί η ανωνυμοποίηση των δεδομένων σε μια βάση.

Πρωταρχικό και βασικό σημείο στο οποίο υστερούν οι παραπάνω αλγόριθμοι αποτελεί ο χρόνος εκτέλεσής τους, που κάποιες φορές είναι πιθανόν να καθιστά αδύνατη την πρακτική εφαρμογή τους. Αξίζει λοιπόν μελλοντικά η ανωτέρω τεχνικές να διερευνηθούν ως προς τη δυνατότητα χρήσης συμπληρωματικών μεθόδων ώστε να καταστούν αποδοτικότερη σε σχέση με την χρονική τους εκτέλεση, κάνοντας τους έτσι φιλικότερους προς τη χρήση σε πρακτικές εφαρμογές κατά την διαδικασία ανωνυμοποίησης των δεδομένων.

Τέλος, ως μια νέα τεχνική θα μπορούσε να θεωρηθεί ο συνδυασμός περισσότερων του ενός αλγορίθμου, ώστε ο ένας να καλύπτει τα κενά και τις αδυναμίες του άλλου με σκοπό τη βέλτιστη ανωνυμοποίηση. Ένας τέτοιος πιθανός συνδυασμός είναι η ικανοποίηση της l-διαφορετικότητας για τα ευαίσθητα γνωρίσματα με την km-ανωνυμία.

6.5 Γλωσσάριο

ΕΛΛΗΝΙΚΑ	**ΑΓΓΛΙΚΑ**
k-ανωνυμία	k-anonymity
l-διαφορετικότητα,l-ποικιλομορφία	l-diversity
επαναταυτοποίηση	de-identification
μ-αμεταβλητότητα	m-invariance
τ-εγγύτητα	t-closeness
δ-παρουσία	δ-presence
πίνακας	table
ιδιότητες κλειδιά	key attributes

στήλη-γνωρίσματα	attribute
γενίκευση	generalization
απόκρυψη εγγραφών	suppression
ταξινόμηση	classification
κλάσεις ισοδυναμίας	equivalence class
ψευδο-αναγνωριστικό	quasi identifiers
ουρά	queue
επιθέσεις ομοιογένειας	homogeneity attack
ανατομία	anatomy

Κεφάλαιο 7

Βιβλιογραφία

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules ιn Very Large Databases. *Proc. of the 20th Int'l Conference on Very Large Databases.* Santiago.

Apriori algorithm. (n.d.). Ανάκτηση από Wikipedia: https://en.wikipedia.org/wiki/Apriori_algorithm

Beekman, G., & Quinn, M. J. (2010). Στο *Εισαγωγή στην Πληροφοριακή* (σσ. 245-246). Αθήνα : Μόσχος Γκιούρδας.

Comer, D. E. (2004). Στο *Δίκτυα και διαδίκτυα υπολογιστών και εφαρμογές τους στο internet* (σσ. 721-722). εκδόσεις Κλειδάριθμος.

Cox, L. (1980). Suppression methodology in statistical discourse analysis. *Journal of American Statistical Association*, 377-385.

D., A., & Aggarwal , C. (2001). On the Design and Quantification of Privacy-Preserving Data Mining Algorithms. *PODS '01 Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, (σσ. 247-255).

Fevre, K. L., witt, D. J., & Ramakrishnan, R. (2006). *Mondrian Multidimendional k-Anonymity* (Conference on Data Engineering εκδ.). In Proc. Intl.

I., M., & Chang , L. (2000). A decision theoretic system for information downgrading. joint conference on information sciences.

L.Sweeney. (2002). k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 557-570.

Li, N., T., L., & S., V. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-diversity. *IEEE 23rd International Conference on Data Engineering* (σσ. 106-115). Instabul: IEEE.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2007). l-Diversity: privacy beyondk-Anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1-47.

Privacy Analytics. (2015). Ανάκτηση από http://www.privacy-analytics.com/de-id-university/white-papers/de-identification-201/

R., A., & Srikant, R. (2000). Privacy-preserving Data Mining. *ACM SIGMOD Record*, 439-450.

Sweeney, L. (2000). Uniqueness of Simple Demographics in the U.S. Population. *Data Privacy Working Paper, carnegie mellon university*.

Sweeney, L. (2002). Database Security:k-anonymity. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 557-570.

Sweeney, L. (retrieved 19 january 2014). *Systems and methods for de-identifying entries in a data source* (united states patents and trademarks office εκδ.). united states.

Terrovitis, M., Mamoulis, N., & Kalnis, P. (2008). Privacy-preserving Anonymization of Set-valued Data. *PVLDB '08*, 115-125.

V.S., V., Elmagarmid, A., Bertino, E., Saygin , Y., & Dasseni, E. (2004, April). Association Rule Hiding. *IEEE Transactions on Knowledge and Data Engineering, 16*(4), σσ. 434-447.

Xiao, X., & Tao, Y. (2006b). Anatomy: Simple and effective Privacy Preservation. (σσ. 139-150). In proc 32nd Int Conf on VLDB.

Ακριβοπούλου, Χ. (2011). *Το δικαίωμα στην προστασία των προσωπικών δεδομένων μέσα από το φακό του δικαιώματος στην ιδιωτική ζωή*. Ανάκτηση από Νομικό Σπουδαστήριο Α. Προυσανίδη: http://www.nomikospoudastirio.gr/%CE%B4%CE%B7%CE%BC%CE%BF%CF%83%CE%B9%CE%B5%CF%8D%CF%83%CE%B5%CE%B9%CF%82-%E2%80%93%CE%BD%CE%BF%CE%BC%CE%B9%CE%BA%CE%AD%CF%82-%CE%BC%CE%B5%CE%BB%CE%AD%CF%84%CE%B5%CF%82/1300-%CF%87-%CE%B1%CE%BA%CF%81%CE%B9%CE%B2%CE%BF%CF%80

Βασιλοπούλου, Κ., Καραγεώργος, Γ., Κική, Ι., Τσαγκαλίδου, Ε., Σταίου, Ε. Ρ., Ξυτάκη, Μ., . . . Χιώνη, Β. (2015). *Άποψη: Προστασία δεδομένων, ιδιωτικότητα και ασφάλεια στο Διαδίκτυο*. Ανάκτηση από Η Καθημερινή : http://www.kathimerini.gr/836171/article/oikonomia/ellhnikh-oikonomia/apoyh-prostasia-dedomenwn-idiwtikothta-kai-asfaleia-sto-diadiktyo