ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. «Τεχνοοικονομική Διοίκηση και Ασφάλεια Ψηφιακών Συστημάτων»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

k-Ανωνυμία, Μια προσέγγιση στην προστασία των δεδομένων και της ιδιωτικότητας

Όνομα: Θρασύβουλος

Επώνυμο: Πιστοφίδης

Α.Μ. ΜΤΕ 1532

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:

Λαμπρινουδάκης Κωνσταντίνος

Τριμελής Εξεταστική Επιτροπή:

Κωνσταντίνος Λαμπρινουδάκης, Καθηγητής

Χρήστος Ξενάκης, Καθηγητής

Χριστόφορος Νταντογιάν, Καθηγητής

Πειραιάς Φεβρουάριος 2018

Ευχαριστίες

Κατ' αρχήν, θα ήθελα να εκφράσω τις ευχαριστίες μου στον επιβλέποντα καθηγητή μου
Δρ. Κωνσταντίνο Λαμπρινουδάκη, τόσο για την ανάθεση της εν λόγω εργασίας όσο για την υποστήριξη, υπομονή και καθοδήγηση του κατα την διάρκεια της σύνταξης της. Η συνεργασία μας ήταν άψογη τόσο κατα την διάρκεια του Μεταπτυχιακού όσο και στο τέλος. Εν συνεχεία, θα ήθελα να ευχαριστήσω τον Δρ. Χρήστο Ξενάκη, Δρ. Χριστόφορο Νταντογιάν και τον Δρ. Παναγιώτη Ριζομυλιώτη για τις παραγωγικές συζητήσεις και το ενδιαφέρον τους.

Άξια αναφοράς, και η άριστη συνεργασία μου με την γραμματεία κα. Σκούντζου Σοφία.

Τέλος, ένα ευχαριστώ στην οικογένεια μου που συνεχίζει να με στηρίζει σε όλες μου τις επιλογές και αποφάσεις.

Περίληψη

Με την καλπάζουσα αύξηση της τεχνολογίας, αυξάνεται η σχέση και η εξάρτηση μας απο αυτή. Το διαδίκτυο, ως ο απόλυτος πλέον εκφραστής αυτής, γίνεται ο δρόμος της ανταλλαγής, επεξεργασίας και διαχείρισης της πληφοροφορίας τόσο σε ατομικό επίπεδο ως χρήστες, όσο σε μαζικό ως εταιρείες, οργανισμοί και ιδρύματα. Η πληροφορία διαχέεται ακατάπαυστα και αποθηκεύεται παντού. Τόσο οι καταχωρήσεις όσο και τα μεταδεδομένα που αφήνουμε πίσω. Έτσι η πληροφορία, και κατ επέκταση τα Big Data αποκτούν το χαρακτηρισμό του "μαύρου χρυσού" του 21 Αιώνα.

Σε ένα τέτοιο αμυδρό περιβάλλον χρειάζονται κανόνες και μέτρα ασφάλειας, προκειμένου να διασφαλιστεί αρχικώς η ιδιωτικότητα των ατόμων, και κατ επέκταση η διασφάλιση των εταιρείων απο επιθέσεις. Σε Ευρωπαϊκό επίπεδο η θέσπιση του GDPR έχει θετικό αντίκτυπο τόσο στην ενημέρωση - συνειδητοποίηση μας, όσον αφορά το εμπόριο των δεδομένων αλλά και την συμόρφωση των εταιρείων, οργανισμών ως προς την διαχείριση τους. Ένας απο τους τρόπους που μπορούν να διασφαλίσουν τα παράπανω είναι η ανωνυμία.

Στην εν λόγω διπλωματική θα ασχοληθούμε αναφορικά με τους αλγόριθμους ανωνυμίας, και θα επικεντρωθούμε σε έναν απο τους πρώτους -αν όχι, πρώτο- την k-ανωνυμία όπως επίσης και θα αναλύσουμε τον τρόπο λειτουργίας ενός απο πιο σημαντικά πρόγράμματα εφαρμογής της, σε μεγάλες βάσεις δεδομένων, το ARX.

Λέξεις κλειδιά: k-ανωνυμία, l-ποικιλομορφία, t-εγγύτητα, ιδιωτικότητα δεδομένων, ανωνυμία, απόρρητο, ανωνυμοποίηση, γενίκευση, απόκρυψη, ψευδοαναγνωριστικά

Πίνακας Περιεχομένων

Κεφάλαιο 1. Εισαγωγή

Στην ψηφιακή εποχή, η ραγδαία αύξηση της δυνατότητας συλλογής, διαχείρισης και ανταλλαγής της πληροφορίας αυξάνει τις ανησυχίες περί προστασίας της ιδιωτικότητας Οι υπολογιστές μπορούν πλέον να χρησιμοποιηθούν για να συλλέξουν και επεξεργαστούν συνήθειες, τρόπο ζωής, τοποθεσίας, όπως της συσχέτισης όλων αυτων απο δεδομένα που συλλέγονται σε καθημερινή βάση. Αυτό δημιουργεί έναν αμφιλεγόμενο συμβιβασμό μεταξύ της αξίας (τόσο για την κοινωνία όσο για τα άτομα της) τής διαθέσιμης,διαχέουσας γνώσης, με τον κίνδυνο που θα προκαλέσει για τα άτομα, η αποκάλυψη ή κακή χρήση των ιδιωτικών τους δεδομένων. Μια λύση σε αυτό το πρόβλημα είναι η ανωνυμία.

Ο όρος ανωνυμία αναφέρεται στην απουσία των αναγνωριστικών πληροφοριών ενός ατόμου. Ωστόσο, μια απλή κατάργηση των πιο ξεκάθαρων – σαφών αναγνωριστικών των ατόμων μπορεί να μην αποτελεί επαρκή προστασία.

Ο κύριος λόγος είναι οτι όταν οι πληροφοριές βγούν στην επιφάνεια, και συνδυαστούν με τις ήδη υπάρχουσες δημοσιευμένες, μπορούν να αποκαλύψουν την ταυτότητα ενός ατόμου.

Ένα χαρακτηριστικό και αρκετά γνωστό παράδειγμα ειναι με τον Crowdsourcing διαγωνισμό του Netflix. Το 2012, το Netflix δημοσιευσε μια λίστα με τους χρήστες και τις βαθμολογίες τους. Ο κόσμος μπορούσε απλα να κατεβάσει αυτή την βάση και να ψάξει για μοτίβα - patterns. Τα δεδομένα περιλάμβαναν ψεύτικα αναγνωριστικά δεδομένα (customer ID) , μαζι με τις ταινίες, τις βαθμολογίες τους και την ημέρα της βαθμολογίας. Υποστηρίζεται οτι απο την στιγμή που τα αναγνωριστικά των πελατών έχουν αφαιρεθεί, οι εν λόγω πληροφορίες δεν παραβιάζουν το απόρρητο των χρηστών. Ωστόσο, οι Narayanan και Shmatikov (2008) έδειξαν πως οι πελάτες μπορούν να ταυτοποιηθούν όταν αυτή η -Netflix- λίστα συνδυαστεί με κάποια βοηθητικά δεδομένα (όπως π.χ απο το IMDB). [1]

Ένα άλλο τρανταχτό παράδειγμα ειναι αυτό της AOL με ουσιαστική πληροφορία αυτή την φορά τα log των αναζητήσεων των μηχανών αναζήτησης.

Τα logs αναζητήσεων των μηχανών αναζήτησης ειναι τεράστια πηγή πληροφορίας τόσο για τους ερευνητές όσο και για τις εταιρείες marketing, ομως παράλληλα η δημοσίευση τους ενδέχεται να εκθέσει την ιδιωτικότητα των χρηστών εκ των οποιων τα logs δημιουργούνται. Υπάρχει το λιγότερο μια τρανταχτή περίπτωση τέτοιας εξέθεσης των προσωπικών logs αναζητήσεων χρηστών οπου εφαρμόστηκε πολύ απλή ανωνυμοποίηση, η οποία ειχε σαν αποτέλεσμα να αποκαλύψει αρκετές πληροφορίες για την αναγνώριση τους. Αυτη η αποκάλυψη έγινε απο την AOL σε μια προσπάθεια να βοηθήσει την “Ερευνητική κοινότητα ανάκτησης πληροφοριών” που έφερε σαν αποτέλεσμα να προκαλέσει ισχυρό πλήγμα τόσο στην πολιτική απορρήτου των χρηστών της AOL όσο και στην ίδια την AOL, με αγωγές και ενστάσεις εναντίον της.

Ιδανικά τα logs αναζήτησης θα πρέπει να ανωνυμοποιούνται πριν κοινοιποιηθούν δημοσιως. Το πρόβλημα ειναι οτι η επιτευξη του επιθυμητου επιπέδου ιδιωτικότητας στα logs καθίσταται δύσκολη, παρουσιάζοντας παράλληλα αντιστάθμισμα μεταξύ χρηστικότητας και ιδιωτικότητας των δεδομένων. Υπάρχουν αρκετές προσεγγίσεις για την ανωνυμοποιηση τέτοιων δεδομένων, αλλα συνήθως περιορίζονται στην διαγραφή συγκεκριμένων queries ή logs. Επιπλέον, συνηθισμένες τεχνικές που χρησιμοποιούνται σε στατιστικούς ελέγχους αποκάλυψης δεν χρησιμοποιηθηκαν ποτε, παρα μονο μέχρι προσφατα για τέτοιου ειδους περιπτώσεις.

Ιδια λογική ισχύει και σε κάθε ειδους δεδομένα, όπως τα χρονικής σήμανσης(timestamped) που παρατηρούνται σε περιπτώσεις οπως αποθήκευση τραπεζικών συναλλαγων, ιατρικές εξετάσεις, λογιστικά βιβλία αρχεία ασφάλισης. Δεδομένα δηλαδή προς διακίνηση, οπου ενδέχεται να ζητηθούν προς δημόσια κοινοποιηση

Τα προηγούμενα παραδείγματα εγείρουν την ερώτηση: *τι είδους πληροφορία επιθυμούμε να προστατεύσουμε* όταν μιλάμε για προστασία της ιδιωτικότητας και πως κατατάσσονται οι απειλές της

1.1 Τα είδη πληροφορίας

Με την αναφορά των ειδών πληροφορίας που επιθυμούμε να προστατεύσουμε αναφερόμαστε συνήθως στα δεδομένα χρηστών/πελατών. Αυτά δομούνται σε σχέση με τον τύπο τους, την ευαισθησία και τις συνέπειες τους στην ασφάλεια, ιδιωτικότητα και ανωνυμία. Κατηγοριοποιούνται σε γενικότερο πλαίσιο στους εξής ακόλουθους 3 τύπους:

i) Δεδομένα αναγνώρισης χρήστη: Πρόκειται για τα δεδομένα και τις παραμέτρους που προσδιορίζουν τον χρήστη, όπως το όνομα/επώνυμο, το όνομα εισόδου (αναγνωριστικό/username), ο κωδικός, το e-mail του και η IP του

ii) Μεταδεδομένα: Πρόκειται για δεδομένα που ορίζουν κάποια ιδιότητα του χρήστη, όπως τοποθεσία, ημερομηνία γεννήσεως, φύλο, κριτήρια αναζήτησης, και log files

iii) Δεδομένα περιεχομένου: Πρόκειται για δεδομένα που περιέχονται σε μια συναλλαγή, όπως πληροφορίες πληρωμής, το περιεχόμενο ενός e-mail και μιας ιστοσελίδας

Στο ειδικότερο πλαίσιο της βασικής ανωνυμοποίησης, και ομαδοποίησης τους σε πίνακες, η διασύνδεση θα γίνει κρίτήριο τού τι θα ορίζεται ως ευαίσθητο, μη ευαίσθητο και ψευδοαναγνωρίστικό.

1.2 Βάσεις Δεδομένων : Ο Τρόπος παρουσίασης και διαχείρισης της πληροφορίας

1.2.1 Σχεσιακές βάσεις δεδομένων.

Με τον όρο σχεσιακή βάση δεδομένων εννοείται μία συλλογή δεδομένων οργανωμένη σε συσχετισμένους πίνακες που παρέχει ταυτόχρονα ένα μηχανισμό για ανάγνωση, εγγραφή, τροποποίηση ή και πιο πολύπλοκες διαδικασίες πάνω στα δεδομένα. [14]

Κάθε γραμμή ονομάζεται tuple και περιλαμβάνει ενα σύνολο πληροφοριων που σχετίζονται με ενα άτομο.Οι στήλες χωρίζουν τα δεδομένα σε σημασιολογικές κατηγορίες που ονομάζονται γνωρίσματα. Ενα σύνολο δεδομένων – dataset αναφέρεται σε ένα μονο tuple σε εναν συγκεκριμένο πίνακα.

Για να ειμαστε πιο συγκεκριμένοι βάση αναφοράς της εκθεσης της Sweeney ενας πίνακας T αναφέρεται σαν $T(A_1,\dots ,A_n)$ με τα γνωρίσματα του $\{A_1,\dots ,A_n\}$. Ενας διατεταγμένος n-tuple $[d_1,d_2,\dots,d_n]$ περιλαμβάνει τις τιμές που συσχετίζονται με τα γνωρίσματα του πίνακα. Για κάθε $j=1,2,\dots,n$ η τιμή του d_j ανατίθεται στο γνώρισμα A_j.

Καθε $T[A_1,\dots ,A_j]$ σημαινει την προβολή του T, συμπεριλαμβάνει μονο τις τιμές $A_1,\dots ,A_j$

1.3 Απειλές στην ιδιωτικότητα

Απειλή στην ιδιωτικότητα, κατα γενική προσέγγιση, έχουμε όταν ο επιτιθέμενος είναι ικανός να συσχετίσει την ταυτότητα ενός χρήστη με την πληροφορία που ο χρήστης θεωρεί ιδιωτική.

Ακολουθούν, οι τρείς πιο κοινοί τύποι απειλών της ιδιωτικότητας[2]:

1. **Αποκάλυψη της ιδιότητας μέλους**: σημαίνει ότι η διασύνδεση των δεδομένων επιτρέπει σε έναν επιτιθέμενο να καθορίσει το κατα πόσο τα δεδομένα ενός συγκεκριμένου ατόμου περιέχονται σε ένα σύνολο δεδομένων ή όχι. Ενώ αυτό άμεσα δεν αποκαλύπτει καμία πληροφορία από το ίδιο το σύνολο δεδομένων, μπορεί να επιτρέψει στον επιτιθέμενο να συναγάγει πορίσματα και μεταδεδομένα.

Στο παράδειγμα μας (πίνακας 1), τα δεδομένα προέρχονται από μητρώο καρκινοπαθών και μπορεί να συναχθεί ότι ένα άτομο έχει ή είχε καρκίνο. Αυτό αφορά τα έμμεσα ευαίσθητα χαρακτηριστικά (εννοώντας χαρακτηριστικά ενός ατόμου που δεν περιλαμβάνεται στο σύνολο δεδομένων), άλλα μοντέλα γνωστοποίησης, αφορούν απόλυτα, ευαίσθητα χαρακτηριστικά.

2. **Αποκάλυψη χαρακτηριστικών**: μπορεί να επιτευχθεί χωρίς απαραίτητης της σύνδεσης ενός ατόμου με ένα συγκεκριμένο στοιχείο σε ένα σύνολο δεδομένων. Προστατεύει ευαίσθητα χαρακτηριστικά, τέτοια τα οποία δεν θα ήθελε κανένα άτομο να συσχετίζεται. Ως εκ τούτου μπορεί να ενδιαφέρουν τον επιτιθέμενο, και αν αποκαλυφθούν θα προκαλέσουν κακό στα υποκείμενα των δεδομένων. Για παράδειγμα η σύνδεση σε ένα σύνολο καταχωρήσεων δεδομένων επιτρέπει την εξαγωγή πληροφοριών έαν όλα τα στοιχεία μοιράζονται μια ορισμένη τιμή ευαίσθητου χαρακτηριστικού (π.χ καρκίνος του μαστού)

3. **Αποκάλυψη ταυτότητας**: (ή επαναπροσδιορισμός) σημαίνει ότι ένα άτομο μπορεί να συνδεθεί με μια συγκεκριμένη καταχώρηση δεδομένων. Πρόκειται για πολύ σοβαρό είδος επίθεσης, καθώς έχει νομικές συνέπειες για τους κατόχους δεδομένων σύμφωνα με πολλούς νόμους και κανονισμούς παγκοσμίως. Από τον ορισμό προκύπτει επίσης ότι ένας επιτιθέμενος μπορεί να μάθει όλες τις ευαίσθητες πληροφορίες που εμπεριέχονται στην καταχώρηση δεδομένων για το άτομο.

	Directly identifying		Quasi-identifying		Insensitive	Sensitive	
	Firstname	Lastname	Age	Gender	State	Diagnosis	
Membership disclosure	Bradley	Rider	51	Male	NY	Colon cancer	Identity disclosure
	Michael	Harlow	45	Male	MS	Hodgkin disease	
	Adella	Bartram	63	Female	NY	Breast cancer	Attribute disclosure
	Freya	King	78	Female	TX	Breast cancer	
	Laurena	Milton	81	Female	AL	Breast cancer	

Πίνακας 1 : Τύποι χαρακτηριστικών και τύποι αποκάλυψης

Με την πάροδο των χρόνων, η ερευνητική κοινότητα στην προσπάθεια της να αντιμετωπίσει τις προαναφερόμενες απειλές ανέπτυξε διάφορα μοντέλα ιδιωτικότητας, συμπεριλαμβανομένης της k-ανωνυμίας (k-anonymity : Sweeney, 2002) και της διαφορικής ιδιωτικότητας (Differential Privacy : Dwork, 2006). Στην παρούσα διπλωματική θα δοθεί έμφαση στην k-ανωνυμία, στις αδυναμίες και στις εναλλακτικές προτάσεις που ήρθαν αργότερα για να προσφέρουν επίλυση στις επιμέρους αδυναμίες της.

Κεφάλαιο 2. Ο ορισμός της k-ανωνυμίας

Κατα Sweeney:
"Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful"

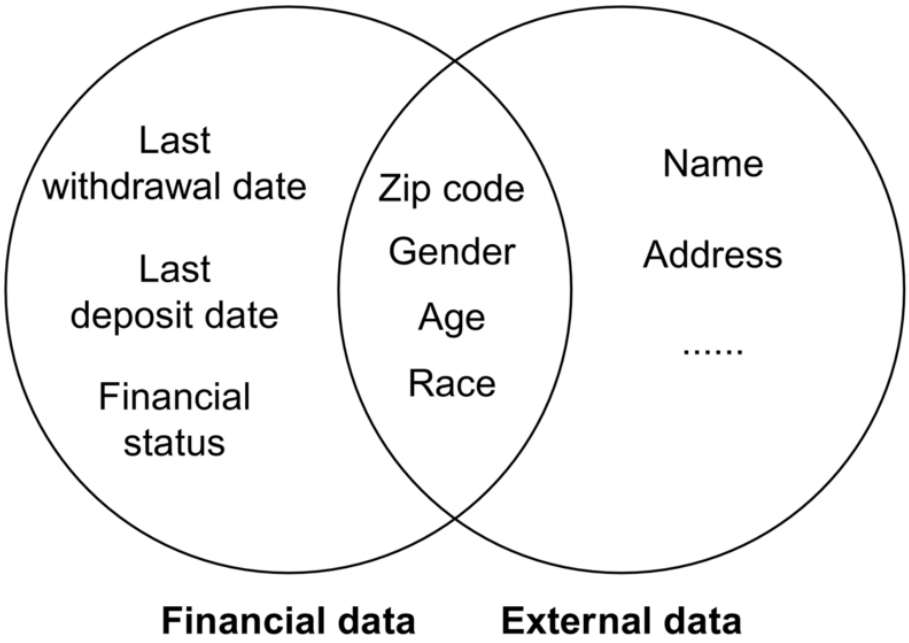
Η *k-ανωνυμία*, που προτάθηκε απο την Sweeney (2002), ειναι μια "αρχή προστασίας" (property) των δημοσιευμένων πληροφοριών απο την επαλήθευση (reidentification). Μπορεί να χρησιμοποιηθεί για παράδειγμα, όταν μια ιδιωτική εταιρεία, όπως μια Τράπεζα που επιθυμεί να εκδόσει – μοιραστεί μια εκδοχή μιας βάσης δεδομένων πελάτων που αφορά οικονομικές πληροφορίες σε κάποιους δημόσιους οργανισμούς για ερευνητικούς σκοπούς.

Σε μια τέτοια περίπτωση τα δεδομένα που θα δημοσιοποιηθούν θα πρέπει να έχουν την ιδιότητα της μη επαληθευσιμότητας, αναγνώρισης δηλαδή των μεμονομένων υποκειμένων τους προκειμένου να επιτευχθεί η προστασία της ιδιωτικότητας τους. Με άλλα λόγια, όλα τα δεδομένα που θα αναφέρονται στην βάση δεδομένων θα πρεπει να ειναι μη – συσχετιζόμενα, ανεξάρτητα των πελατών.

Τα αρχικά δεδομένα των πελατών μιας Τράπεζας συνήθως περιέχουν πληροφορίες όπως:
Όνομα, *Διεύθυνση*, και *Τηλεφωνικό αριθμό* όπου μπορούν άμεσα να ταυτοποιήσουν τους πελάτες. Ένας πιθανός τρόπος απόκρυψης της ταυτότητας τους ειναι η άμεση αφαίρεση των ευαίσθητων αυτών πληροφοριών απο την βάση. Ώστόσο, αυτό δεν αποτελεί εξασφάλιση της ιδιωτικότητας των πελατών.

Οι πληροφορίες, όπως ο ταχυδρομικός κώδικας, το φύλλο, η ηλικία και η φυλή των πελατών μπορούν να τους ταυτοποιήσουν. Ο Τ.Κ (ταχυδρομικός κώδικας) προσφέρει μια κατα προσέγγιση τοποθεσία. Αναζητώντας, βάση συγκεκριμένων κριτηρίων όπως της ηλικίας, του γένους και της φυλής, ειναι πολύ πιθανό να αποκαλυφθούν οι ταυτότητες των πελατών.

Ένας άλλος πιθανός τρόπος επίτευξης επαληθευσιμότητας ονομάζεται "επίθεση διασύνδεσης" (linking attack). Εκτός των γνωρίσματων όπως το όνομα και η διεύθυνση οπου μπορούν άμεσα να "σπάσουν" την ανωνυμία των δεδομένων, υπάρχουν επίσης γνωρίσματα που ονομάζονται ψευδοαναγνωριστικά – (quasi-identifier) QID που χρησιμοποιούνται για να συνδέσουν τα δημοσιευμένα δεδομένα με εξωτερικά δεδομένα. Το φύλο, η ηλικία, ο ταχυδρομικός κώδικας ειναι μια τυπική **tuple** (η πλείαδα θα χαρακτηρίζεται ώς tuple) και αυτή - η tuple - απο τα δημοσιευμένα δεδομένα έχει αυξημένη πιθανότητα εμφάνισης εξίσου σε κάποια εξωτερικά δεδομένα. Δηλαδή, έαν υπάρχουν εξωτερικοί πίνακες όπως, λίστες καταχώρησης ψηφοφόρων, τότε συνδέοντας τα με τα QID των δημοσιευμένων δεδομένων, μπορούν να αποκαλυφθούν οι ταυτότητες των πελατών (σχεδ. 1).[1]



σχεδ.1 *Επίθεση της διασύνδεσης μεταξύ των δημοσιευμένων δεδομένων και των εξωτερικών δεδομένων*

Η *k-ανωνυμία* ως πρόταση απαιτεί κάθε εγγραφή των δημοσιευμένων δεδομένων, να μπορεί να αντιστοιχιστεί σε τουλάχιστον k εγγραφές των αρχικών δεδομένων. Με άλλα λόγια, κάθε εγγραφή απο τα δημοσιευμένα δεδομένα θα έχει τουλάχιστον k-1 ίδιες εγγραφές στα ίδια δημοσιευμένα δεδομένα. Για παράδειγμα στον Πίνακα 1, ο (a) ειναι τα αυθεντικά δεδομένα και (b) ειναι τα δεδομένα που προκύπτουν απο τον (a). Ο (b) έχει k-ανωνυμία όπου k = 2.

Η Latanya Sweeney [3] παρουσίασε την αρχή της k-ανωνυμίας και απέδειξε ότι εαν τα δημοσιευμένα δεδομένα εφαρμόζουν την ιδιότητα της k-ανωνυμίας, τότε μπορούν να προστατευθούν απο την “Επίθεση της διασύνδεσης”. Αυτό γιατι κάθε εγγραφή των δημοσιευμένων δεδομένων θα έχει τουλάχιστον k-1 ίδιες εγγραφές.

Άρα, σκοπός της k-ανωνυμίας είναι να καταστήσει κάθε μια εγγραφή μη-διάκριτη ανάμεσα σε άλλες k-1.

Παράδειγμα k-ανωνυμίας (k=2)

Πίνακας 1

(a) Αρχικά Δεδομένα				
Όνομα	**Γένος**	**Φυλή**	**Ηλικία**	**Τ.Κ**
Αλίκη	Θηλυκό	Λευκή	17	21103
Su	Θηλυκό	Κίτρινη	22	21300
Denzel	Αρσενικό	Μαύρη	27	21110
Κατερίνα	Θηλυκό	Λευκή	15	21102
Rose	Θηλυκό	Μαύρη	29	21109
Andy	Αρσενικό	Κίτρινη	24	21304

(b) Διαμοιραζόμενα δεδομένα προερχόμενα απο (a)			
Γένος	**Φυλή**	**Ηλικία**	**Τ.Κ**
Θ ή Α	Λευκή	15 - 19	211*
Θ ή Α	Κίτρινη	20 - 24	213*
Θ ή Α	Μαύρη	25 - 29	211*
Θ ή Α	Λευκή	15 - 19	211*
Θ ή Α	Μαύρη	25 - 29	211*
Θ ή Α	Κίτρινη	20 - 24	213*

2.1 Quasi identifiers – Ψευδοαναγνωριστικά

Μια ομάδα απο γνωρίσματα που δεν ειναι δομικά μοναδικά, αλλα θα μπορούσαν να ειναι εμπειρικά μοναδικά, συνεπώς κατα βάση θα μπορούσαν να αναγνωρίζουν – προσδιορίζουν μια πληθυσμιακή μονάδα.

	Quasi – Identifier - Ψευδοαναγνωριστικά			**Ευαίσθητα Δεδομένα**
	Ηλικία	**Γένος**	**Τ.Κ**	**Έσοδα**
1	35	Αρσενικό	81243	300,000
2	48	Θηλυκό	83123	30,000
3	40	Αρσενικό	81205	1,000,000
4	60	Αρσενικό	73193	100,000
5	27	Θηλυκό	83123	60,000
6	60	Αρσενικό	71234	20,000
7	27	Θηλυκό	83981	25,000
8	35	Θηλυκό	83012	30,000
9	27	Αρσενικό	81021	40,000
10	46	Αρσενικό	73013	25,000
11	46	Θηλυκό	83561	70,000
12	40	Αρσενικό	81912	40,000
13	48	Αρσενικό	72231	1,500,000

Πίνακας 1α: Ιδιωτικός πίνακας P

Ο παραπάνω πινακας ειναι ιδιωτικός καθότι δεν συμπεριλαμβάνει προσωπικα δεδομένα οπως Ονοματεπώνυμο. Το εισόδημα κατατάσεται στα ευαίσθητα δεδομένα. Δεν θα πρέπει να ειναι εφικτή η αναγνώριση του εισοδήματος ενος ατομου χρησιμοποιώντας αυτον τον πίνακα.

Εαν συγκεκριμένα δεδομένα όπως η ηλικία, το γένος, και ο Τ.Κ ειναι μοναδικοι τότε μπορεί να χρησιμοποιηθούν για την αναγνώριση ενός ατόμου και ανάλογα του εισοδήματος του.
Μπορούν να χρησιμοποιηθούν για να ταυτοποιησουν τα δεδομένα σε σχέση με έναν πίνακα που περιλαμβάνει ηλικία, γένος, Τ.Κ οπως και το όνομα. Εαν υπάρχει και το ανάλογο ενδιαφέρον απο τον επιτιθέμενο σε συγκεκριμένο άτομο που ήδη γνωρίζει...και αρα ορισμένα απο τα πεδια, γινεται ακόμα πιο εύκολο.

Το ψευδοαναγνωριστικό για τον πίνακα Ρ απο το Πίνακα 1a μπορεί να έχει τον εξής τύπο
Q_P = {age, gender, zip}.

Το ελάχιστο σύνολο από γνωρίσματα Q=Q1,...,Q_d με το οποίο ένας πίνακας Τ μπορεί να συζευχθεί με κάποιες εξωτερικές πληροφορίες για να αναγνωριστούν ατομικές εγγραφές ονομάζεται ψευδόαναγνωριστικό σύνολο.

	Quasi – Identifier - Ψευδοαναγνωριστικά			**Ευαίσθητα Δεδομένα**
	Ηλικία	**Γένος**	**Τ.Κ**	**Έσοδα**
1	<45	Αρσενικό	81***	40,000
2	<45	Αρσενικό	81***	40,000
3	<45	Αρσενικό	81***	300,000
4	<45	Αρσενικό	81***	1,000,000
5	≥45	Αρσενικό	7****	20,000
6	≥45	Αρσενικό	7****	25,000
7	≥45	Αρσενικό	7****	100,000
8	≥45	Αρσενικό	7****	1,500,000
9	*	Θηλυκό	83***	25,000
10	*	Θηλυκό	83***	30,000
11	*	Θηλυκό	83***	30,000
12	*	Θηλυκό	83***	60,000
13	*	Θηλυκό	83***	70,000

Πίνακας 1b: Γενικευμένος πίνακας G1 οπου βασίζεται στον Ρ Εφαρμογή k-ανωνυμίας για k=4

Για να πετύχουμε k-ανωνυμία στον P, τα γνωρίσματα που ειναι στα πεδία των ψευδοαναγνωρίστικών θα πρέπει να γενικευτούν. Ένας * στο T.K. θα μπορούσε να σημαίνει οποιοδήποτε στοιχείο. Στην στήλη της ηλικίας το <45 σημαίνει οτι η ηλικία ειναι κάτω των 45, ενω >= σημαίνει οτι η ηλικία μπορεί να ειναι ίση ανω των 45 και * μπορεί να σημαινει οποιοδήποτε αριθμό. Ο G1 οπως φαινεται και απο τον πίνακα ειναι μια γενικευμένη έκδοχη του P που εκπληρώνει της k-ανωνυμία με k=4. Κάθε πεδίο με τον ίδιο quasi-identifier περιλαμβάνει τουλάχιστον 4 εισόδους. Έαν ο επιτιθέμενος ενδιαφέρεται για τα έσοδα ενος ατομου με ένα συγκεκριμενο ψευδοαναγνωριστικο, υπάρχουν τουλάχιστον k-1=3 αλλοι άνθρωποι με το ιδιο ψευδοαναγνωριστικό στον πίνακα.

Εαν υπάρχουν το λιγότερο k tuples με τα ίδια ψευδοαναγνωριστικά, δεν ειναι εφικτό να προσδιορίσεις ένα μονο που βασίζεται σε αυτά. Υπάρχουν k-1 tuples με το ίδιο ψευδοαναγνωριστικο που δεν ειναι διακριτα απο το tuple που ψάχνει ο επιτιθέμενος.

Κεφάλαιο 3. Μηχανισμοί που υποστηρίζουν την *k-ανωνυμία*

Μετα την πρόταση της k-ανωνυμίας, υλοποιήθηκαν δίαφορες προσπάθειες σχεδίασης ενος καλού αλγόριθμου που θα μετέτρεπε μια βάση σε μορφή ανάλογη της υλοποίησης του ορισμού. Οι δύο κύριες τεχνικές που χρησιμοποιήθηκαν για την επιβολή k-ανωνυμίας στα δημοσιευμένα δεδομένα είναι η Γενίκευση και η Απόκρυψη (suppression).

3.1. Γενίκευση (Generalization):

Η διαδικασία κατα την οποία οι τιμές των ψευδοαναγνωριστικών (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).

Στην περίπτωση του Πίνακα 1, οι τιμές του φύλου, της ηλικίας και του Τ.Κ. απο τον (a) αντικαθίστανται απο μια γενικευμένη έκδοση (b). Η γενίκευση μπορεί να εφαρμοστεί σε επίπεδα απο ένα και μόνο κελί σε ολόκληρο tuple (γραμμή) γνωρισμάτων για την επίτευξη k-ανωνυμίας.

3.2. Απόκρυψη (Suppression),

Συνίσταται στην αφαίρεση ευαίσθητων γνωρισμάτων για την μείωση της ποσότητας της γενίκευσης όταν επιτυγχάνεται k-ανωνυμία. Όπως και η γενίκευση, η απόκρυψη μπορεί να εφαρμοστεί τόσο σε κελιά όσο σε ολόκληρα γνωρίσματα. Ο συνδυασμός της Γενίκευσης και της Απόκρυψης χρησιμοποιήθηκε για την δημιουργία διαφορετικών αλγόριθμων με σκοπό την ικανοποίηση της k-ανωνυμίας.

Το συμβατικό πλαίσιο – η λογική, ενός τέτοιου αλγόριθμου πάντα ξεκινά με την απόκρυψη πολλών ευαίσθητων γνωρίσματων και εν συνέχεια χωρίζει tuples των εναπομήνοντων γνωρισμάτων σε ομάδες καθως υποκαθιστά τις ακριβής τιμές των QID με γενικευμένες μορφές τους για κάθε ομάδα, οπου ονομάζονται κλάσεις ισοδυναμίας. Αυτή η γενίκευση ειναι ομοιογενής γενίκευση και έχει χρησιμοποιήθει για την εφαρμογή της k-ανωνυμίας στις αναφορές των Iwuchukwu and Naughton (2007) [4] Ghinita et al. (2007) , και LeFevre et al. (2008). [5]

Η ιδιότητα μιας ομοιογενής γενίκευσης ορίζει ότι, αν μια αρχική εγγραφή t_i ταιριάζει με μια δημοσιευμένη εγγραφή t_j' της οποίας η αντίστοιχη εγγραφή είναι t_j , τότε η t_j θα ταιρίαζει με την t_i'. Αυτή η ιδιότητα ονομάζεται αμοιβαιότητα.

Το πιο σημαντικό στοιχείο της ομοιογενούς γενίκευσης ειναι ο τρόπος διαίρεσης των κλάσεων ισοδυναμίας. Η “στρατηγική” γενίκευσης θα επηρεάζει άμεσα την χρησιμότητα των δημοσιευμένων δεδομένων.

Υπάρχουν δύο τύποι μοντέλων γενίκευσης: Ολική και Τοπική Γενίκευση. Ακολουθεί η ανάλυση τους.

3.2.1. Ολική Γενίκευση - Global Recoding

(πλήρη ανωνυμοποίηση τομέα – domain) [7][8]

Στην Ολική Γενίκευση, σε μία στήλη εφαρμόζεται η ίδια στατηγική γενίκευσης στην ίδια τιμή. Έτσι αν δύο tuples στις αρχικές καταγραφές έχουν ίδιες QID τιμές, τότε θα πρέπει να έχουν την ίδια δημοσιευμένη τιμή.

Εν προκειμένω, τα σύνολα των δεδομένων που ανωνυμοποιήθηκαν με την εν λόγω μέθοδο έχουν όλα τα γνωρίσματα ισάξια γενικευμένα ή κρυμμένα σε όλες τις καταχωρίσεις. Εαν διαφορετικά tuples είχαν την ίδια τιμή γνωρίσματος, αυτή η τιμή θα ανιστοιχούσε στην ίδια γενικευμένη τιμή. (Πίνακας 2a)

Η Ολική Γενίκευση μπορεί να χαρακτηριστεί ως μονοδιάστατη γενίκευση ή πολυδιάστατη γενίκευση. Η μονοδιάστατη ορίζεται απο την συνάρτηση $f_i : D x_i \rightarrow D$ για κάθε γνώρισμα x_i του ψευδοαναγνωριστικού. Η τιμή της ανωνυμοποίησης μπορεί να ληφθεί εφαρμόζοντας την συνάρτηση f_I στις τιμές των ψευδοαναγνωριστικών x_i σε κάθε tuple του αρχικού συνόλου δεδομένων. Ενώ, η πολυδιάστατη γενίκευση ορίζεται απο μια μόνο συνάρτηση
$f : D x_i * \ldots * D x_n \rightarrow D$, η τιμή της ανωνυμοποίησης λαμβάνεται εφαρμόζοντας την συνάρτηση f στο διάνυσμα κάθε tuple του αρχικού συνόλου δεδομένων.

3.2.2 Τοπική Γενίκευση - Local Recoding

Στην περίπτωση της Τοπικής Γενίκευσης 2 tuples με ίδιες QID τιμές μπορεί να έχουν διαφορετικές γενικευμένες τιμές.

Σε αντίθεση με την Ολική Γενίκευση, η τοπική αποκρύπτει τα γνωρίσματα ανα στοιχείο κατάχωρησης. Αποτέλεσμα αυτού ένα σύνολο δεδομένων με λιγότερα κρυφά δεδομένα σε σχέση με της Ολικής Γενίκευσης. Ο χώρος των δεδομένων χωρίζεται σε διαφορετικές περιοχές και όλες οι εγγραφές της ίδιας περιοχές αντιστιχίζονται στην ίδια γενικευμένη ομάδα. Απο την στιγμή που τα όρια ποιότητας ειναι αρκετά μεγάλα δεν είναι εύκολη η απόδειξη της βέλτιστης λύσης που δόθηκε στην γενίκευση . [9]

Η ολική γενίκευση επιτυγχάνει την ανωνυμία χαρτογραφώντας/αντιστοιχίζοντας τους τομείς των γνωρισμάτων των ψευδοαναγνωριστικών σε γενικές ή τροποποιημένες τιμές. Ενώ τα μοντέλα τοπικής γενίκευσης αποτυπώνουν (μη διακριτά) μεμονωμένα στοιχεία δεδομένων σε γενικευμένες τιμές (Πίνακας 2b). Το πλεονέκτημα της ολικής γενίκευσης είναι ότι ο ανωνυμοποιημένος πίνακας θα έχει ένα ομοιογενές σύνολο τιμών ενώ το μειονέκτημα του είναι ότι υπερ-γενικεύει τον αρχικό πίνακα και με αποτέλεσμα να χάνεται μεγάλη πληροφορία κατα την διαδικασία. Σε αντίθεση η τοπική γενίκευση δίνει πολύ καλύτερα αποτελέσματα σε πραγματικό σενάριο και έχει πολύ μεγαλύτερη χρησιμότητα σε σχέση με την ολική γενίκευση

Πινακας 2: Αρχικός Πίνακας στοιχείων δεδομένων

Ηλικία	Φύλο	Τ.Κ.	Αρρώστια
23	Άντρας	34756	Βρογχίτιδα
29	Άντρας	34201	Καρκίνος
35	Γυναίκα	36020	Γρίπη
39	Άντρας	37013	Ιλαρά

Πίνακας 2a: Με χρήση ολικής γενίκευσης

Ηλικία	Φύλο	Τ.Κ.	Αρρώστια
[21 - 30]	Άνθρωποι	[34201 - 37013]	Βρογχίτιδα
[21 - 30]	Άνθρωποι	[34201 - 37013]	Καρκίνος
[31 - 40]	Άνθρωποι	[34201 - 37013]	Γρίπη
[31 - 40]	Άνθρωποι	[34201 - 37013]	Ιλαρά

Πίνακας 2b: Πίνακας με χρήση τοπικής γενίκευσης

Ηλικία	Φύλο	Τ.Κ.	Αρρώστια
[21 - 30]	A	[34201 - 34756]	Βρογχίτιδα
[21 - 30]	*	[34201 - 34756]	Καρκίνος
[31 - 40]	*	[36020 - 37013]	Γρίπη
[31 - 40]	A	[36020 - 37013]	Ιλαρά

*Οι αλγόριθμοι που χρησιμοποιούν Τοπική Γενίκευση μπορεί να εγγυηθούν περισσότερη ανωνυμία σε συσκεκριμένες περιπτώσεις.(Ninghui Li and Su, 2011).

Κεφάλαιο 4. Αλγόριθμοι k-Ανωνυμίας

Στην ολική γενίκευση οι τιμές των αρχικών δεδομένων γενικεύονται στο επίπεδο του τομέα. Υπάρχουν αρκέτες εργασίες που βασίζονται στην μέθοδο της ολικής γενίκευσης, κάποιες απο αυτές απόδωσαν τους κάτωθι καρπούς – αλγόριθμους που θα εξετάσουμε.

4.1 Αλγόριθμος Incognito

Ο Αλγόριθμος **Incognito** που προτάθηκε στη [7] χρησιμοποιεί δυναμικό προγραμματισμό και φαίνεται να υπερέχει απο προηγούμενους αλγόριθμους σε εφαρμογή του πάνω σε σε σενάριο με 2 βάσεις δεδομένων βασισμένων σε πραγματικά στοιχεια.
Η κύρια ιδέα του Incognito ειναι οτι κάθε υποσύνολο tuple των QID με k-ανωνυμία, θα πρέπει να εφαρμόζει και αυτό k-ανωνυμία. (subset property)

Άρα, 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
$P \subseteq Q$.

Ο Incognito παράγει ελάχιστη γενίκευση πλήρους τομέα. Στην ολική γενίκευση, αν ένας τομέας χαμηλότερου επιπέδου πρέπει να γενικευθεί σε τομέα υψηλότερου επιπέδου, όλες οι τιμές του χαμηλότερου γενικεύονται σε υψηλότερου. Αυτό μπορεί προκαλέσει την υπεργενίκευση του πίνακα, που έχει σαν αποτέλεσμα μεγάλη απώλεια πληροφορίας.

Στην πολυδιάσταση και τοπική γενίκευση, η γενίκευση γίνεται πάντα στο επίπεδο των στοιχείων, γεγονός που δεν προκαλεί υπεργενίκευση, οδηγώντας σε πιο ευέλικτη γενίκευση με μικρότερη απώλεια πληροφίας.

4.2 **Αλγόριθμος Mondrian**

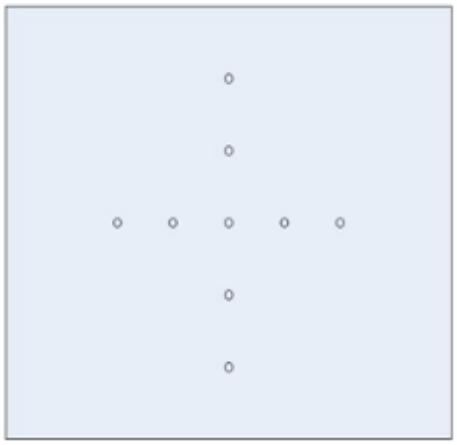
Ο Αλγόριθμος **Mondrian** [11] χρησιμοποιεί την στατηγική της πολυδιάστατης Ολικής Γενίκευσης. Πρόκειται για έναν πολυ γρήγορο και κλιμακωτό αλγόριθμο εύρεσης, ο οποίος παράγει καλύτερα αποτελέσματα.

Στον Mondrian, κάθε γνώρισμα στο σύνολο των δεδομένων αντιπροσωπεύει μια διάσταση και κάθε εγγραφή αντιπροσωπεύει ενα σημείο στον χώρο[12]. Αντί, να διαιρεί τις εγγραφές, ο Mondrian διαιρεί τον χώρο σε πολλές περιοχές και σε κάθε περιοχή υπάρχουν το λιγότερα k σημεία.
Χρησιμοποιεί μεθόδους αυστηρής και χαλαρής κατάτμισης, οι οποίες οδηγούν σε καλύτερα αποτελέσματα χρηστικότητας των δεδομένων.

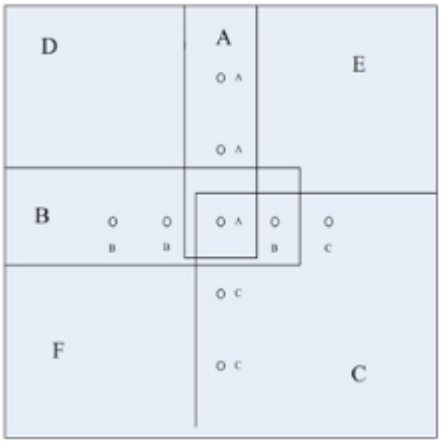
Η τεχνική της κατάτμισης αντιστοιχίζει κάθε tuple του συνόλου δεδομένων σε ένα πολυδιάστατο χώρο. Στην συνέχεια, η γενίκευση του συνόλου δεδομένων ισούται με την κατάτμιση του αντίστοιχου πολυδιάστατου χώρου. Ένα κομμάτι του πολυδιάστατου χώρου αντιστοιχεί σε μοναδικό αποτέλεσμα ανωνυμίας. Εάν τα κομμάτια δεν διασταυρώνονται ή επικαλύπτονται το ένα με το άλλο τότε είναι γνωστό ως Αυστηρή Κατάτμιση. Αντιθέτως, αν αλληλεπικαλύπτονται μεταξύ τους τότε είναι γνωστό ως Χαλαρή Κατάτμιση.[12]

Η χαλαρή κατάτμιση θεωρείται καλύτερη τεχνική σε σχέση με την αυστηρή, και παραθέτουμε μια περίπτωση που το αποδεικνύει.

Παράδειγμα: Αν υποθέσουμε ότι θα πρέπει να εφαρμόσουμε αυστηρή κατάτμιση στον πίνακα του σχεδίου 2, σε δύο το λιγότερο περιοχές, τότε θα πρέπει να υπάρχει μια περιοχή που να περιλαμβάνει όχι περισσότερα απο δύο tuples.
Επομένως, αν απαιτηθεί 3-ανωνυμία ή περισσότερες, τότε δεν θα υπάρξει κανένας “αυστηρής κατάτμισης” αλγόρθμος για να υποστηρίξει το σκοπό μας



Σχέδ 2: Δεδομένα σε 2-διάστατο επίπεδο



Σχέδ 3: Χαλαρή Κατάτμιση

Αλλά μπορούμε να χρησιμοποιήσουμε ενα αλγόριθμο χαλαρής κατάτμισης για το διαμοιρασμό του πίνακα σε 6 περιοχές, A,B,C,D,E,F όπως υποδεικνύει και το σχέδιο 3. Οι περιοχές A,B, και C περιλαμβάνουν 3 tuples ενώ οι περιοχές D,E και F δεν περιλαβάνουν tuples. Κάθε tuple στην διασταύρωση των περιοχών ανήκει σε μόνο μία περιοχή.[12][13]

Άρα, ο αλγόριθμος Mondrian επιστρέφει την βέλτιστη πολυδιάστατη διαίρεση σε κάθε περιοχή της οποίας ανήκουν περισσότερες από k-εγγραφές και συνεπώς ικανοποιείται η k-ανωνυμία.

4.3 Μη-ομοιογενής γενίκευση

Οι περισσότεροι αλγόριθμοι που χρησιμοποιούν την γενίκευση όπως και οι προαναφερόμενοι μοιράζονται μια κοινή δομή: στη πρώτη κατάτμιση τα tuples σε ομάδες, στην συνέχεια αντιστοιχίση τα ίδια γενικευμένα ψευτοαναγνωριστικά (QID) στα tuples της ίδιας ομάδας. Η ομάδα των tuples με τα ίδια QID λέγεται κατηγορία ισοδυναμίας. Μια τέτοια προσέγγιση, την οποία αποκαλούμε *ομοιογενή γενίκευση.*

Υπαρχει μια ακόμη μέθοδος που ονομάζεται μη-ομοιογενής γενίκευση. Η διαφορά της σε σχέση με την ομοιογενή είναι ότι μπορεί να επιτύχει κάποιος υψηλότερη χρηστικότητα των δεδομένων απο τα δημοσιευμένα δεδομένα.

Στον Πίνακα 3, ο (b) ειναι τα δημοσιευμένα δεδομένα που προκύπτουν απο τον (a) χρησιμοποιώντας ομοιογενής γενίκευση, και ειναι ξεκάθαρο ότι (t_1', t_2', t_5') ειναι τάξεις ισοδυναμίας και (t_3', t_4') ειναι άλλες. Σε κλάση ισοδυναμίας, όλες οι γενικευμένες QID τιμές ειναι οι ίδιες. Ωστόσο, σε μη ομοιογενή γενικιευμένο πίνακα (c), t_1', t_2' και t_5' έχουν διαφορετικές γενικευμένες QID τιμές. Ενώ και οι δύο πίνακες (b) (c) έχουν 2-ανωνυμία, ο (c) προσφέρει υψηλότερη χρηστικότητα απο την στιγμή που τα γενικευμένα QID εύρη στον (c) είναι μικρότερα ή ίσοδύναμα με τα αντίστοιχα του πίνακα (b). Αυτό υποδεικνύει ότι αν κάποιος χρησιμοποιεί μη-ομοιογενής γενίκευση μπορεί να επιτύχει υψηλότερη χρηστικότητα των δεδομένων απο τα δημοσιευμένα δεδομένα.

Πίνακας 3

Παράδειγμα k-ανωνυμίας (k=2) απο ομοιογενή σε μη-ομοιογενή γενίκευση

(a) Αρχικά Δεδομένα			
Tuple ID	**Γένος**	**Ηλικία**	**Τ.Κ**
t_1	Θηλυκό	17	21103
t_2	Αρσενικό	29	21110
t_3	Αρσενικό	27	21210
t_4	Αρσενικό	15	21102
t_5	Θηλυκό	22	21109

(b) Ανταλλαγή δεδομένων που δημιουργούνται με ομοιογενή γενίκευση			
Tuple ID	**Γένος**	**Ηλικία**	**Τ.Κ**
t_1’	Θ ή Α	17-29	2110*
t_2’	Θ ή Α	17-29	211*
t_3’	Αρσενικό	15-27	212*
t_4’	Αρσενικό	15-27	211*
t_5’	Θ ή Α	17-29	211*

(c) Κοινή χρήση δεδομένων που δημιουργούνται από μη-ομοιόμορφη γενίκευση			
Tuple ID	**Γένος**	**Ηλικία**	**Τ.Κ**
t_1’	Θηλυκό	17-29	2110*
t_2’	Αρσενικό	17-29	211*
t_3’	Αρσενικό	15-27	212*
t_4’	Αρσενικό	15-27	211*
t_5’	Θ ή Α	17-29	211*

Κεφάλαιο 5. Επιθέσεις στην k-ανωνυμία

Η έκδοση πολλων διαφορετικών συνολων δεδομένων που βασίζεται στην ίδια όμάδα κάτοχων δεδομένων δημιουργεί περισσότερους παράγοντες επίθεσης. 3 τέτοιες επιθέσεις αναφέρονται απο κάτω. Όταν δημοσιοποιούνται αντίστοιχα των ακόλουθων δεδομένα καποιες απο τις ακόλουθες τακτικές μπορούν να αποτρέψουν τέτοιες επιθέσεις.

i) Επίθεση μη ομαδοποιημένης στοίχισης

Όταν δύο πίνακες δημοσιοποιούνται μπορεί να χρησιμοποιηθούν για να συνδέσουν σύνολα δεδομένων, εαν βασίζονται στον ιδιο αρχικό πίνακα και η θέση των tuples ειναι η ίδια οπως ηταν στον προηγούμενο. Οπως το μοντέλο της k-ανωνυμίας χρησιμοποιεί το ίδιο “σχεσιακό” μοντέλο, θεωρητικά δεν θα υπάρχει καποια προδιεγεγραμμένη σειρά των tuples. Σχεσεις ειναι μια ομάδα απο tuples και στις ομάδες δεν υπάρχει καμία σειρά. Όταν δημοσιοποιούνται αληθινα δεδομένα, υπάρχει μεγάλη πιθανότητα οι tuples να ταξινομούνται βάση κάποιου γνωρίσματος.

Ο σκοπός της γενίκευσης ειναι να ταξινομεί τα δεδομένα κατα αύξουσα ή φθίνουσα σειρα βάση ενός ή περισσοτέρων γνωρισμάτων των ψευδοαναγνωριστικών.

	Quasi – Identifier - Ψευδοαναγνωριστικά			**Ευαίσθητα Δεδομένα**
	Ηλικία	**Γένος**	**Τ.Κ**	**Έσοδα**
1	27	*	*****	40,000
2	40	*	*****	40,000
3	35	*	*****	300,000
4	40	*	*****	1,000,000
5	60	*	*****	20,000
6	46	*	*****	25,000
7	60	*	*****	100,000
8	48	*	*****	1,500,000
9	27	*	*****	25,000
10	34	*	*****	30,000
11	48	*	*****	30,000
12	27	*	*****	60,000
13	46	*	*****	70,000

Πίνακας 4:Γενικευμένος Πίνακας G2

Ο G2 στον 4ο πίνακα βασίζεται στον P(Πίνακας 1a). Η σειρά ειναι η ίδια με τον G1(Πίνακας 1b). Ο G1 ικανοποιεί την k-ανωνυμία με k=4, ο G2 με k=2. Εαν και οι δύο δημοσιοποιηθούν, ο επιτιθέμενος θα μπορέσει να συνδέσει τα tuples βασισμένα στην θέση τους. Ετσι θα μπορέσει να μάθει περισσότερα για την ηλικία και το φύλο της κάθε tuple, αρκετή γνώση για αναγνώριση ενος ανθρώπου.

Αυτη η επίθεση μπορεί να αντιμετωπιστεί εύκολα με την τυχαιοποίηση της σειράς των tuples, όταν απελευθερώνουμε τα δεδομένα.

ii) Επίθεση συμπληρωματικής δημοσιοποίησης

Έαν ένας πίνακας που περιλαμβάνει ενα υποσύνολο προηγούμενου δημοσιοποιημένου πίνακα δημοσιοποιηθει εκ νέου, μια τέτοιου είδους επίθεση μπορεί να ειναι εφικτή. Γνωρίσματα τα οποια δεν ειναι μερη των ψευδοαναγνωριστικών του πρώτου πίνακα μπορούν να χρησιμοποιηθούν για να διασυνδεθούν με τους νέους πίνακες.

	Quasi – Identifier - Ψευδοαναγνωριστικά			**Ευαίσθητα Δεδομένα**
	Ηλικία	**Γένος**	**Τ.Κ**	**Έσοδα**
1	46	*	*****	25,000
2	27	*	*****	40,000
3	48	*	*****	30,000
4	40	*	*****	40,000
5	40	*	*****	1,000,000
6	27	*	*****	60,000
7	46	*	*****	70,000
8	60	*	*****	20,000
9	60	*	*****	100,000
10	27	*	*****	25,000
11	34	*	*****	30,000
12	35	*	*****	300,000
13	48	*	*****	1,500,000

Πίνακας 5: Γενικευμένος Πινακας G3

Ο G3 (Πίνακας 5) ειναι μια ανώνυμη μορφή του P. Έαν δημοσιοποιηθεί μετα τον G1, κάποια tuples μπορούν να συνδεθούν, ακόμα και αν οι θέσεις τους ειναι τυχαίες. Π.χ υπάρχει μονο ένα άτομο με το εισόδημα του 1.500.000. Αυτό ειναι αρκετό για να αντιστοιχηθεί με τα αντιστοιχα tuples και να αποκτήσουμε πληροφορία για τον Qp. Η συνδέουσα τιμή για αυτό το tuple είναι [48, Αρσενικό, 7****, 1,500,000]. Για όλες τις tuples με μοναδικό συνδυασμό των τιμών για {Q G1 ∪Έσοδα}

Για να αποτρέψουμε μια τετοια επίθεση, όλα τα γνωρίσματα της πρωτης έκδοσης/δημοσιοποιήσης θα πρέπει να χειριστούν σαν ψευδοαναγνωριστικά στον δεύτερο πίνακα. Σε αυτή την περίπτωση ο πίνακας δεν πρέπει να δημοσιοποιηθει με αυτό τον τρόπο καθότι δεν καλύπτει k-ανωνυμία.

Επίσης αν χρησιμοποιηθεί ο G1 σαν βάση επίσης δεν θα έχουμε συντελεστή επίθεσης, καθότι αν έχουμε το ίδιο ανώνυμο σχεδιάγραμμα, δεν μπορεί να ληφθεί επιπρόσθετη πληροφορία.

iv) Χρονική Επίθεση:

Καθώς τα δεδομένα αυξάνονται και μαζεύονται τακτικά, τα σύνολα τους ενδέχεται να αυξηθούν χρονικά. Αλλαγές ή αφαιρέσεις των tuples ειναι εξίσου πιθανες. Γι αυτό ακριβώς τον λόγο, δημοσιοποιούνται συχνά επιπρόσθετα δεδομένα. Αυτές οι δημοσιοποιήσεις μπορεί να ειναι ευπαθής στην σύνδεση των πινάκων.

π.χ Ας υποθέσουμε οτι μια ερευνα μάζεψε δεδομένα όπως ο P, και δημοσιοποίησε G1. Αργότερα επιπρόσθετοι tuples μαζεύτηκαν και ενημέρωσα τον P, οπότε ο P_{t1} έγινε P U [51, Θηλυκό, 83581, 28,000], [51, Αρσενικό, 81019, 32,000]}. Βάση του P_{t1} ενας πίνακας γενίκευσης G_{t1} ως G3 {[51, *, *****, 28,000], [51, *, *****, 32,000]} εκδίδεται. Όπως φάνηκε και στο παράδειγμα 5 οι πίνακες G1 & G3 μπορούν να διασυνδεθούν.

Για να αποφύγουμε τέτοιες επιθέσεις, μπορούμε να προβούμε ως εξής:

G1 ∪ (P t1 − P). Δηλαδή G1 ∪ {[*, Θηλυκό,83***, 28,000], [≥45, Αρσενικό, 81***, 32,000]}

Τα προαναφερόμενα ήδη επιθέσεων, όπως παρατηρήσαμε, αντιμετωπίζονται χρησιμοποιώντας αλγοριθμους είτε/και επιπρόσθετες προσωπικές ενέργεις - μεθόδους προσέγγισης χωρίς να αποτελούν πανάκεια - καθότι υπάρχουν νέες επιθέσεις και νέες αντιμετωπίσεις που μπορούν να προσφέρουν ισχυρότερη ανωνυμία λειτουργώντας εναλλακτικά είτε συμπληρωματικά. Παρακάτω θα εξεταστούν αναλυτικότερα οι δύο βασικότερες: η l- ποικιλομορφία και η t – εγγύτητα.

Κεφάλαιο 6. Σύνθετες επιθέσεις στην k-ανωνυμία και τρόποι επίλυσης

6.1 Εισαγωγή

Η προστασία που η k-ανωνυμία προσφέρει είναι απλή και εύκολη στην κατανόηση. Εάν ένας πίνακας ικανοποιεί k-ανωνυμία για κάποια τιμή k, τότε ο οποιοσδήποτε γνωρίζει μόνο τις τιμές του ψευδο-αναγνωριστικού ενός συγκεκριμένου ατόμου δεν μπορεί να προσδιορίσει – αναγνωρίσει την καταγραφή που αρμόζει σε αυτό το συγκεκριμένο άτομο με σιγουριά μεγαλύτερη του 1/k.

Ενώ η k-ανωνυμία προστατεύει την αποκάλυψη ταυτότητας, δεν προσφέρει επαρκής προστασία έναντι της αποκάλυψης του γνωρίσματος. Αυτό αναγνωρίστηκε απο αρκετούς συγγραφείς. Δυο είναι οι επιθέσεις που βρέθηκαν: Η επίθεση της ομοιογένειας και η επίθεση με πρότερη γνώση Παρακάτω θα εξετάσουμε 2 παραδείγματα για την καλύτερη αντίληψη των ευπαθειών.

6.2. Επίθεση της Ομοιογένειας

Η επίθεση της Ομοιογένειας αξιοποιεί την περίπτωση όπου όλες οι τιμές που μπορεί να πάρει μια ευαίσθητη τιμή μέσα σε ένα σύνολο k εγγραφών, είναι ίδιες.

Για παράδειγμα, ας υποθέσουμε ότι έχουμε μια ομάδα k διαφορετικών αρχείων και όλα μοιράζονται ένα συγκεκριμένο ψευδοαναγνωριστικό. Ένας επιτιθέμενος δεν μπορεί να εντοπίσει το άτομο βάσει των ψευδοαναγνωριστικών. Αλλά αν ενδιαφέρεται για τα ευαίσθητα γνωρίσματα και όλες οι ομάδες έχουν την ίδια τιμή τότε τα δεδομένα έχουν διαρρεύσει. Παραθέτουμε και ένα παράδειγμα για την καλύτερη κατανόηση:

Το σενάριο μας περιλαμβάνει την *Αλίκη* και τον *Πέτρο* ως “θύτης” και “θύμα”. Η Αλίκη - βάση του ρόλου της - , παρακολουθεί την προσωπική ζωή του Πέτρου με απώτερο στόχο να γνωρίσει τις αδυναμίες του. Ανακαλύπτει οτι ο Πέτρος επισκέπτεται τακτικά το νοσοκομείο. Πρέπει να μάθει αν υποφέρει απο κάποια αρρώστια. Εισέρχεται στην προσοχή της ένας 4ης – Ανωνυμίας πίνακας ενός δημοσιευμένου εγγράφου των ασθενών του νοσοκομείου που επισκέπτεται ο Πέτρος. (Πίνακας 7). Έτσι, γνωρίζει ότι σε κάποια απο τις εγγραφές αντιστοιχούν τα δεδομένα του Πέτρου. Η Αλίκη γνωρίζει οτι ο Πέτρος ειναι κάπου στα 30 με 35, Αμερικανός και τον Τ.Κ (13053) του. Έτσι λοιπόν γνωρίζει οτι οι τελευταίες εγγραφές (9-12) ανταποκρίνονται στα γνωστά της δεδομένα. Βγαίνει λοιπόν το πόρισμα οτι ο Πέτρος έχει Καρκίνο.[16]

1η Παρατήρηση Η k-ανωνυμία μπορεί να δημιουργήσει ομάδες που μπορούν να διαρρεύσουν πληροφορίες χάριν της ελλείψεως διαφορετικότητας στα ευαίσθητα γνωρίσματα.
Τέτοιες περιπτώσεις δεν ειναι ασυνήθιστες, και πόσο μάλλον όταν εξετάζονται δεδομένα 60.000 διαφορετικών tuples, οπου τα ευαίσθητα γνωρίσματα θα μπορούν να πάρουν μονο 3 διακριτές τιμές. Οπότε, προτείνεται επιπρόσθετα να εφαρμόζεται “διαφορετικότητα” σε όλα τα tuples που μοιράζονται τις ίδιες τιμές των ψευδο-αναγνωριστικων τους, να έχουν διαφορετικές τιμές στα ευαίσθητα γνωρίσματα.

	Μη - Ευαίσθητα			**Ευαίσθητα**
	Τ.Κ.	**Ηλικία**	**Εθνικότητα**	**Κατάσταση**
1	13053	28	Ρωσική	Καρδιακή Ασθένεια
2	13068	29	Αμερικάνικη	Καρδιακή Ασθένεια
3	13068	21	Γαλική	Ίωση
4	13053	23	Αμερικάνικη	Ίωση
5	14853	50	Ινδική	Καρκίνος
6	14853	55	Ρωσική	Καρδιακή Ασθένεια
7	14850	47	Αμερικάνικη	Ίωση
8	14850	49	Αμερικάνικη	Ίωση
9	13053	31	Αμερικάνικη	Καρκίνος
10	13053	37	Ινδική	Καρκίνος
11	13068	36	Ιαπωνική	Καρκίνος
12	13068	35	Αμερικάνικη	Καρκίνος

Πίνακας 6. Μικροδεδομένα Ασθενών

	Μη - Ευαίσθητα			**Ευαίσθητα**
	Τ.Κ.	**Ηλικία**	**Εθνικότητα**	**Κατάσταση**
1	130**	<30	*	Καρδιακή Ασθένεια
2	130**	<30	*	Καρδιακή Ασθένεια
3	130**	<30	*	Ίωση
4	130**	<30	*	Ίωση
5	1485*	≥40	*	Καρκίνος
6	1485*	≥40	*	Καρδιακή Ασθένεια
7	1485*	≥40	*	Ίωση
8	1485*	≥40	*	Ίωση
9	130**	3*	*	Καρκίνος
10	130**	3*	*	Καρκίνος
11	130**	3*	*	Καρκίνος
12	130**	3*	*	Καρκίνος

Πίνακας 7. 4-ανώνυμιας μικροδεδομένα ασθενών

6.3. Επίθεση με πρότερη γνώση

Αυτή η επίθεση αξιοποιεί μια συσχέτιση μεταξύ ενός ή περισσοτέρων γνωρισμάτων ψευδοαναγνωριστικών με το ευαίσθητο γνώρισμα, με σκοπό την μείωση των πιθανών τιμών του εύρους του ευαίσθητο γνώρισματος [15]

Σε συνέχεια του προηγούμενου σεναρίου μας, η Αλίκη μετα την επιτυχή επίθεση ομοιογένειας που εφάρμοσε στον (Πίνακα 7), παίρνει θάρρος και δοκιμάζει τις δυνατότητες της στην εύρεση νέου στόχου. Το θύμα της τώρα η Μαρία. Τα γνωστά γι αυτή στοιχεία είναι οτι η Μαρία είναι γειτόνισσα της, άρα έχει το ίδιο Τ.Κ. 13053 ειναι 28 χρονών και νοσηλεύεται τελευταία στο ίδιο νοσοκομείο. Η Μαρία είναι Γαλλίδα. Παράλληλα, η Αλίκη γνωρίζει οτι οι Γάλλοι έχουν τις χαμηλότερες εμφανίσεις εμφραγμάτων παγκόσμια και κατ’ επέκταση τις λιγότερες πιθανότητες να εμφανίσουν καρδιαγγειακά νοσήματα. Άρα συμπεραίνει ότι η Μαρία έχει μια απλή ίωση.

2η Παρατήρηση Η k-ανωνυμία δεν μπορεί να προστατεύσει απο επιθέσεις που βασίζονται στην πρότερη γνώση.

6.4 l – Ποικιλομορφία (l-diversity)

Ένας νέος ορισμός προστασίας της ιδιωτικότητας που ήρθε να επιλύσει το πρόβλημα της αποκάλυψης ευαίσθητων γνωρισμάτων που συνάδει με την k-ανωνυμία είναι η l-ποικιλομορφία, και όπως προτάσει και το όνομα, διασφαλίζει την διαφορετικότητα – ποικιλομορφία των τιμών των ευαίσθητων γνωρισμάτων σε κάθε κλάση ισοδυναμίας, διατηρώντας το ελάχιστο μέγεθος του k-συνόλου.

Η βασικός τρόπος λειτουργίας της απαιτεί κάθε ομάδα ψευδοαναγνωριστικών, να έχει τουλάχιστον l - “καλά εκπροσωπούμενες” διαφορετικές τιμές, οι οποίες θα χρησιμοποιηθούν για να κάνουν μια tuple, ασαφή. Για να προσδιορίσουμε του πόσο καλά οι τιμές των γνωρισμάτων θα εκπροσωπευτούν, μπορούμε να εφαρμόσουμε τα ακόλουθα πιθανά μοντέλα.

i) **Διακριτή l - ποικιλομορφία**. Η απλούστερη κατανόηση του όρου “καλά εκπροσωπούμενοι” θα ήταν η εξασφάλιση ότι υπάρχουν το λιγότερο l διακριτές τιμές για τα ευαίσθητα γνωρίσματα σε κάθε μια κλάση ισοδυναμίας. Η διακριτή l-ποικιλομορφία δεν εμποδίζει πιθανολογικές επιθέσεις συμπερασμάτων. Μια κλάση ισοδυναμίας μπορεί να εμφανίζει μια τιμή πιο συχνά σε σχέση με άλλες, επιτρέποντας στον επιτιθέμενο να καταλήξει στο συμπέρασμα ότι μια οντότητα ,σε αυτή την κλάση ισοδυναμίας, είναι πολύ πιθανό να έχει ίδια τιμή. Αυτό οδήγησε στην ανάπτυξη των 2 ακόλουθων ισχυρότερων εννοιών της l - ποικιλομορφίας.

ii)**Εντροπία l - ποικιλομορφίας**. Η εντροπία *E* μιας κλάσης ισοδυναμίας ορίζεται ως:

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

Όπου *S είναι* το πεδίο ορισμού του ευαίσθητου γνωρίσματος, και *p(E, s)* είναι το κλάσμα των εγγραφών της εντροπίας *E* με *s* την είσοδο των γνωρισμάτων.

Ένας πίνακας λέμε ότι έχει εντροπία *l-ποικιλομορφίας εάν* για κάθε κλάση ισοδυναμίας E, Entropy(E) ≥ log l. Δηλαδή, η εντροπία ολόκληρου του πίνακα πρέπει να είναι το λιγότερο log(l). Μερικές φορές αυτό μπορεί να είναι πολύ περιοριστικό, καθώς η εντροπία ολόκληρου του πίνακα μπορεί να είναι μικρή εάν λίγες τιμές είναι πολύ συνηθισμένες. Αυτό οδηγεί στη παρακάτω, λιγότερο περιοριστική έννοιά της l-ποικιλομορφίας.

iii)**Επαναλαμβανόμενη (c, l) – ποικιλομορφία.** Η επαναλαμβανόμενη ποικιλομορφία εξασφαλίζει ότι η πιο συχνή τιμή δεν θα εμφανίζεται πολύ συχνά, και οι λιγότερο συχνές τιμές δεν θα εμφανίζονται πολύ σπάνια. Εάν *m* είναι ο αριθμός των τιμών σε μια κλάση ισοδυναμίας, και $r_{i,}$ με 1≤ i ≤ m είναι ο αριθμός των φορών όπου ο $i^{η}$ πιο συχνή τιμή εμφανίζεται σε μια κλάση ισοδυναμίας *E. Τότε E* λέγεται οτι έχει επαναλαμβανόμενη (c, l) - ποικιλομορφία εάν $r_1 < c(r_l + r_{l+1} + ... + r_m)$.
Ο πίνακας αναφέρεται οτι έχει επαναλαμβανόμενη (c, l) - ποικιλομορφία εάν όλες οι ισοδύναμες τάξεις της έχουν επαναλαμβανόμενη(c, l) – ποικιλομορφία.

6.4.1 Περιορισμοί l - ποικιλομορφίας

Ενώ η αρχή - κανόνας της l - ποικιλομορφίας αντιπροσωπεύει ένα σημαντικό βήμα στην βελτίωση της k-ανωνυμίας, ως προς την προστασία της αποκάλυψης γνωρισμάτων, έχει αρκετές ελλείψεις οπου και θα αναφέρουμε.

Η εφαρμογή της l - ποικιλομορφίας μπορεί να ειναι δύσκολη και αχρείαστη.

Παράδειγμα 1
Έαν υποθέσουμε οτι τα αρχικά δεδομένα έχουν μονο ένα ευαίσθητο γνώρισμα: π.χ το αποτέλεσμα ενός συγκεκριμένου ιού. Θα παίρνει τότε μονο 2 τιμές: θετικό και αρνητικό. Ακόμη, αν υποθέσουμε οτι υπάρχουν 10000 εγγραφές, με το 99% να ειναι αρνητικές, και μονο το 1% θετικές. Τότε οι 2 τιμές έχουν διαφορετικό βαθμό ευαισθησίας. Ο ένας δεν θα πείραζε να ειναι γνωστό ότι βγήκε αρνητικός, καθότι ειναι ίδιος με το υπόλοιπο 99% του πληθυσμού, αλλά ο άλλος δεν θα ήθελε να ειναι γνωστό ότι βγήκε θετικός. Σε αυτή την περίπτωση η 2- διαφορετικότητα ειναι αχρείαστη σε μια κλάση ισοδυναμίας που περιλαμβάνει εγγραφές που είναι μονο αρνητικές. Για να μπορέσουμε να έχουμε πίνακα διακριτής 2 – διαφορετικότητας πρέπει να ειναι το πολύ 10000 x 1% = 100 ισοδύναμες τάξεις και η απώλεια της πληροφορίας να ειναι μεγάλη. Επίσης, άξιο παρατήρησης ειναι οτι εξαιτίας της πολύ μικρής εντροπίας των ευαισθητων γνωρισμάτων του πίνακα, έαν κάποιος χρησιμοποιήσει εντροπία l - ποικιλομορφίας, το l θα πρέπει να εχει μικρή τιμή.

Η l - ποικιλομορφία ειναι ανεπαρκής στην αποτροπή της αποκάλυψης των γνωρισμάτων.

Ακολουθούν 2 επιθέσεις στην l - ποικιλομορφία.

i) Επίθεση αλλοίωσης (Skewness Attack): Όταν η συνολική κατανομή ειναι στρεβλή, η εφαρμογή της l – ποικιλομορφίας δεν πρόκειται να αποτρέψει την αποκάλυψη των γνωρισμάτων.

Κρίνοντας απο το παραπάνω Παράδειγμα (1) αν υποθέσουμε ότι μια κλάση ισοδυναμίας έχει ίδιο αριθμό αρνητικών όσο και θετικών εγγραφών. Ικανοποιεί διακριτή 2- ποικιλομορφία, εντροπία 2-ποικιλομορφίας, και οποιαδήποτε επαναλαμβανόμενη (c,2) – ποικιλομορφία μπορεί να επιβληθεί. Ωστοσο, αυτό αποτελεί σοβαρή απειλή για την ιδιωτικότητα, γιατι ο οποιοσδηποτε μέσα στην τάξη μπορεί να θεωρηθεί οτι έχει 50% πιθανότητα να ειναι θετικός, συγκρινόμενος με το 1% του συνολικού πληθυσμού.

ii) Επίθεση ομοιότητας (Similarity Attack): Όταν οι τιμές των ευαίσθητων γνωρίσματων στην κλάση ισοδυναμίας ειναι διακριτές αλλα σημασιολογικά παρόμοια, ο επιτιθέμενος μπορεί να μάθει σημαντικές πληροφορίες. Εξατάζουμε το παρακάτω παράδειγμα:

Παράδειγμα 2 Ο πίνακας 8.1 ειναι ο αρχικός πίνακας, και ο Πίνακας 8.2 δείχνει την ανώνυμη εκδοχή του (με εφαρμογή διακριτής και εντροπία 3 – ποικιλομορφίας). Υπάρχουν 2 ευαίσθητα γνωρίσματα:
Ο Μισθός και η Αρρώστια. Εάν υποθέσουμε ότι η καταγραφή του Πέτρου βρίσκεται σε μια απο τις 3 πρώτες εγγραφές, τότε σίγουρα γνωρίζουμε οτι ο μισθός του ειναι μεταξύ 3K – 5K και μπορούμε να υποθέσουμε οτι είναι σχετικά χαμηλός. Αυτή η επίθεση ισχύει όχι μονο για αριθμητικά γνωρίσματα αλλα και γνωρίσματα κατηγορίας όπως η “Αρρώστια”. Γνωρίζοντας λοιπόν οτι οι εγγραφές του Πέτρου βρίσκονται στην πρώτη τάξη ισοδυναμίας συμπεραίνουμε ότι ο Πέτρος έχει προβλήματα με το στομάχι του, γιατι όλες αυτές οι καταγραφές αφορούν αρρώστιες του στομάχου.

Αυτή η ροή ευαίσθητων πληροφοριών συμβαίνει παρόλο που εφαρμόζεται l-ποικιλομορφία, η οποία εγγυάται ποικιλομορφία στις ευαίσθητες τιμές της κάθε ομάδας, καθότι δεν λαμβάνει υπόψη τη σημασιολογική εγγύτητα αυτών των τιμών.

	T.K	Ηλικία	Μισθός	Αρρώστια
1	47677	29	3.000	Γαστρικό έλκος
2	47602	22	4.000	Γαστρίτιδα
3	47678	27	5.000	Καρκίνος στομάχου
4	47905	43	6.000	Γαστρίτιδα
5	47909	52	11.000	Γρίπη
6	47906	47	8.000	Βρογχίτιδα
7	47605	30	7.000	Βρογχίτιδα
8	47673	36	9.000	Πνευμονία
9	47607	32	10.000	Καρκίνος στομάχου

Πίνακας 8.1: Πίνακας αρχικών στοιχείων μισθού/αρρώστιας

	Τ.Κ	Ηλικία	Μισθός	Αρρώστια
1 2 3	476** 476** 476**	2* 2* 2*	3.000 4.000 5.000	Γαστρικό έλκος Γαστρίτιδα Καρκίνος στομάχου
4 5 6	4790* 4790* 4790*	≥40 ≥40 ≥40	6.000 11.000 8.000	Γαστρίτιδα Γρίπη Βρογχίτιδα
7 8 9	476** 476** 476**	3* 3* 3*	7.000 9.000 10.000	Βρογχίτιδα Πνευμονία Καρκίνος στομάχου

Πίνακας 8.2 Πίνακας 3 – ποικιλομορφίας

Για την αντιμετώπιση τέτοιων επιθέσεων, ορίστηκε η t- εγγύτητα, μια συμπληρωματική έννοια ιδιωτικότητας που αναπαριστά το γενικότερο γνωστικό πεδίο του επιτιθέμενου πάνω στην κατανομή των τιμών του ευαίσθητου γνωρίσματος.

6.5 t- εγγύτητα

"Μια τάξη ισοδυναμίας λέγεται οτι έχει t- εγγύτητα εάν, η απόσταση μεταξύ της κατανομής ενός ευαίσθητου γνωρίσματος σε αυτή την τάξη και της κατανομή του γνωρίσματος σε ολόκληρο τον πίνακα δεν υπερβαίνει ενα t όριο. Τότε λέγεται ότι ένας πίνακας έχει t-εγγύτητα, όταν όλες οι ισοδύναμες τάξεις έχουν t-εγγύτητα".[17]

Η βασική ιδέα της t-εγγύτητας απαιτεί ότι, η κατανομή κάθε ευαίσθητου γνωρίσματος σε κάθε ομάδα θα πρέπει να είναι παρόμοια με εκείνη του συνολικού πίνακα. Επιπλέον, για την μέτρηση της προαναφερόμενης μέτρησης της απόστασης, οι συγγραφείς εισήγαγαν έναν νέο όρο – μέτρο μέτρησης της απόστασης, την EMD (Earth Mover Distance). Η σταθερά t χρησιμοποείται ως όριο για την ικανοποίηση της t- εγγύτητας.

Η ιδιωτικότητα μετράται απο το κέρδος των πληροφοριών που θα αποκομίσει ένας παρατηρητής. Πριν δεί το δημοσιοποημένο πίνακα, ο παρατηρητής έχει μονο κάποια σχετική ίδεα σχετικα με τα ευαίσθητα γνωρίσματα – πληροφορίες ενός ατόμου. Απο την στιγμή που δεί τον δημοσιοποιημένο πίνακα, αποκτάει μια διαφορετική εικόνα - άποψη. Το κέρδος της πληροφορίας μπορεί να αναπαρασταθεί ως η διαφορά μεταξύ της προηγούμενης ιδέας και της μεταγενέστερης ιδέας.

Για την καλύτερη κατανόηση του συγκεκριμένου κανόνα αλλα και την έκφραση του με τύπο, ακολουθεί παράδειγμα.

Παράδειγμα:

Υποθέτουμε πρότερη εικόνα - ιδέα παρατηρητή B_0 για τα ευαίσθητα γνωρίσματα καποιου ιδιώτη. Ύστερα απο κάποια υποθετική ενέργεια, δίνεται στον παρατηρητή μια πλήρη γενικευμένη έκδοση ενός πίνακα δεδομένων όπου όλα τα γνωρίσματα στα ψευδο-αναγνωριστικά έχουν αφαιρεθεί (ή έχουν γενικευθεί στις πιο γενικευμένες τιμές τους). Η ιδέα του παρατηρητή επηρεάζεται απο την κατανομή Q της τιμής του ευαίσθητου γνωρίσματος σε όλο τον πίνακα και αλλάζει σε B_1. Τέλος, δίνεται στον παρατηρητή ο δημοσιευμένος πίνακας. Γνωρίζοντας τις τιμές των ψευδο-αναγνωριστικών, ο παρατηρητής δύναται να αναγνωρίσει την τάξη ισοδυναμίας, οπου βρίσκεται η καταγραφή του ιδιώτη, και μαθαίνει την κατανομή P των τιμών των ευαίσθητων γνωρισμάτων στην τάξη. Η άποψη του παρατηρητή αλλάζει σε B_2.

Η σκοπός της l – ποικιλομορφίας ειναι ο περιορισμός της μείωσης της διαφοράς μεταξύ B_0 και B_2. . Μειώνοντας τη διαφορά μεταξύ B_1 και B_2 καταλήγουμε στο συμπέρασμα ότι Q η κατανομή των ευαίσθητων γνωρισμάτων στον συνολικό πληθυσμό του πίνακα, ειναι η δημόσια πληροφορία. Δεν περιορίζεται το κέρδος της πληροφορίας των παρατηρητών ως σύνολο, αλλα περιορίζεται ο βαθμός στον οποίο ο παρατηρητής μπορεί να μάθει περισσότερα πράγματα για συγκεκριμένα άτομα.[17]

Για να δικαιολογηθεί η παραδοχή οτι η Q θα πρέπει να διαχειριστεί σαν δημόσια πληροφορία, παρατηρείται ότι με τις γενικεύσεις το μέγιστο που κάποιος μπορεί να καταφέρει ειναι να γενικεύσει όλες τις τιμές των ψευδο-αναγνωριστικών στην πιο γενική τιμή τους. Έτσι με την δημοσίευση μιας έκδοσης των δεδομένων θα δημοσιευθεί παράλληλα και μια κατανομή Q. Η μεγάλη διαφορα του B_0 στο B_1, σημαίνει οτι ο πίνακας περιέχει περισσότερα στοιχεία και νέα πληροφορία ως σύνολο. Όσο μεγαλύτερη είναι η διαφορά του B_0 στο B_1 τόσο πιο πολύτιμα ειναι τα δεδομένα. Δεδομένου ότι η αύξηση της πληροφορίας μεταξύ B_0 και B_1 αφορά ολόκληρο το πληθυσμό, δεν περιορίζουμε αυτό το κέρδος.

Περιορίζουμε το κέρδος μεταξύ του B_1 B_2, περιορίζοντας την απόσταση μεταξύ του P και Q. Έτσι αν P = Q τότε $B_1 = B_2$.. Αν P και Q ειναι κοντά, τότε ανάλογα B_1 και B_2 θα ειναι το ίδιο κοντά, ακόμα και αν B0 μπορεί να ειναι διαφορετικό απο B1 και B2.[17]

Απαιτώντας όμως τα P και Q να ειναι κοντά θα περιόριζε το σύνολο της οφέλιμης δημοσιευμένης πληροφορίας, καθότι περιορίζει την συσχέτιση μεταξύ των ψευδο-αναγνωριστικών και των ευαίσθητων γνωρισμάτων Όμως αυτό ακριβώς ειναι το ζητούμενο. Εάν λάβει καθαρή εικόνα σχετικά με αυτή την συσχέτιση, τότε επέρχεται και η αποκάλυψη των γνωρισμάτων. Η παράμετρος *t* στην t- εγγύτητα επιτρέπει την ανταλλαγή μεταξύ χρηστικότητας και ιδιωτικότητας.

Το επόμενο βήμα ειναι η μέτρηση της απόστασης μεταξύ δυο πιθανολογικών κατανομών. Υπάρχουν αρκετοί τρόποι για να οριστεί η απόσταση μεταξύ τους. Βάση των δύο κατανομών
$P = (p_1, p_2, \ldots, p_m)$, $Q = (q_1, q_2, \ldots, q_m)$ δύο γνωστά μέτρα απόστασης είναι τα ακόλουθα:

i) Η μεταβολική απόσταση (variational distance) ορίζεται ως:

$$\mathsf{D}[\mathbf{P}, \mathbf{Q}] = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|.$$

και

ii) Η Kullback-Leibler (KL) απόσταση ορίζεται ως:

$$\mathsf{D}[\mathbf{P}, \mathbf{Q}] = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i} = H(\mathbf{P}) - H(\mathbf{P}, \mathbf{Q})$$

όπου

$H(\mathbf{P}) = \sum_{i=1}^{m} p_i \log p_i$ είναι η εντροπία του P και

$H(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{m} p_i \log q_i$ ειναι η cross-εντροπία του P και Q.

Αυτά τα μέτρα απόστασης δεν αντικατοπτρίζουν τη σημασιολογική απόσταση μεταξύ των τιμών

Αν και ορίζονται, η μεταβολική απόσταση και η απόσταση Kullback-Leibler, ο σκοπός της μέτρηση της απόστασης έχει σκοπό την αναπαράσταση της διαφοροποίησης των τιμών του ευαίσθητου γνωρίσματος με τέτοιο τρόπο ώστε μικρότερη απόσταση να ερμηνεύεται ως: λιγότερο διαφορετικές τιμές και συνεπώς μεγαλύτερος κόπος εξόρυξης της προσωπικής πληροφορίας. Την καλύτερη αναπαράσταση της απόστασης, επιτυγχάνει η μετρική Earth Mover's Distance (που αρχικώς προαναφέραμε), η οποία ορίζεται από το πρόβλημα μετακίνησης. Βασίζεται στο ελάχιστο απαιτούμενο έργο για την μετατροπή της μίας κατανομής στην άλλη με την μαζική ανακατανομή ανάμεσά τους.

Έτσι αν P = (p_1, p_2, …, p_m), Q = (q_1, q_2, …, q_m) με d_{ij} τη σταθερή απόσταση μεταξύ του στοιχείου i του P και στοιχείου j του του Q. Θέλουμε να βρούμε μια ροή F = [f_{ij}] όπου f_{ij} η ροή της μάζας απο το p_i στο q_j έτσι ώστε να ελαχιστοποιηθεί το συνολικό έργο το οποίο είναι ίσο με την μετρική EMD και δίνεται από τη σχέση:

$$WORK(\mathbf{P}, \mathbf{Q}, F) = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} f_{ij}$$

με τους ακόλουθους περιορισμούς:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq m \qquad (c1)$$

$$p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i \qquad 1 \leq i \leq m \quad (c2)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1 \qquad (c3)$$

Αυτοι οι περιορισμοι εγγυώνται ότι ο P μετατρέπεται σε Q απο την ροή της μάζας F. Απο την στιγμή που το πρόβλημα μεταφοράς λυθεί ο EMD ορίζεται ως συνολικό έργο

$$\mathsf{D}[\mathbf{P}, \mathbf{Q}] = WORK(\mathbf{P}, \mathbf{Q}, F) = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} f_{ij}$$

Βάση τη t- εγγύτητα και τους πίνακες που την ικανοποιούν όπως και τη χρήση της EMD προκύπτουν οι δύο κάτωθι ιδιότητες:

Ιδιότητα γενίκευσης: *Έστω T ένας πίνακας δεδομένων και A, B δύο πίνακες-γενικεύσεις του T τέτοιες ώστε ο A να είναι πιο γενικευμένος από τον B. Αν ο T ικανοποιεί την t-εγγύτητα με χρήση του B, τότε ο T ικανοποιεί την t-εγγύτητα και με χρήση του πίνακα A.*

Ιδιότητα υποσυνόλου: *Έστω T ένας πίνακας δεδομένων και C το σύνολο των γνωρισμάτων που συμμετέχουν στα δεδομένα του T. Αν ο T ικανοποιεί την t- εγγύτητα ως προς το σύνολο C τότε ικανοποιεί την t-εγγύτητα ως προς κάθε υποσύνολο γνωρισμάτων $D \subset C$.*

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

Παρόλό το σημαντικό πλεονέκτημα χρήσης του εν λόγω αλγόριθμου - αρχής, δημιουργούνται κάποια ζητήματα:

i)Δεν είναι εύκολη η προστασία της ιδιωτικότητας σύμφωνα με πολλά διαφορετικά επίπεδα ασφάλειας

ii)Η εισαχθείσα, μέτρηση απόστασης (EMD) εκλείπει της ευελιξίας της διαχείρισης αριθμητικών γνωρισμάτων

iii) Ενδέχεται πιθανότητα “θυσίας” της χρηστικότητας των δημοσιευμένων δεδομένων διότι είναι πολύ δύσκολο ένας κανόνας να επιτρέπει όλες τις διανομές γνωρισμάτων να ειναι μεταξύ τους παρόμοιες.

6.6 δ – Παρουσία (δ-Presence)

Η δ-Παρουσία αποσκοπεί στην προστασία του συνόλου των δεδομένων απο την αποκάλυψη της ιδιότητας του μέλους. Απαιτεί την ρητή διαμόρφωση του συνόλου δεδομένων που έχει αποκαλυφθεί σε υποσυνόλο μεγαλύτερου συνόλου δεδομένων που αντιπροσωπεύει τις γνώσεις του επιτιθέμενου. Το κριτήριο επιτάσει, περιορισμούς στις πιθανότητες με τις οποίες μπορεί να προσδιοριστεί, εάν ή όχι ένα άτομο απο το ολικό σύνολοδο δεδομένων περιλαμβάνεται στο υποσύνολο της έρευνας. Τα ανώτερα όρια για αυτές τις πιθανότητες υπολογίζονται με βάση τα μεγέθη ισοδύναμων κλάσεων.

6.7 Ανωνυμοποίηση βάση κινδύνου

Η ανωνυμοποίηση βάση του κινύνδου μπορεί να χρησιμοποιηθεί για την αυτόματη μετατροπή των συνόλων δεδομένων για την διασφάλιση ότι οι εκτιμώμενοι κίνδυνοι επαναπροσφιορισμού θα υποχωρήσουν – πέσουν κάτω ενός συγκεκριμένου ορίου

Κεφάλαιο 7. Εργαλεία Ανωνυμοποίησης

Για την εφαρμογή ανωνυμοποίησης σε ένα σύνολο δεδομένων υπάρχουν διαθέσιμα πολλά εργαλεία. Χωρίζονται ανάλογα με το πως διατίθενται – δωρεάν, επι πληρωμή, ανοιχτού κώδικα, κλειστού κώδικα - και κατα πόσο ολοκληρομένα είναι όσον αφορά την ποικοιλία των αλγόριθμων ανωνυμοποίησης που υποστηρίζουν όπως των μεθόδων που χρησιμοποιούν.

Θα ασχοληθούμε αποκλειστικά με το ARX, το οποίο θεωρείται ενα απο τα πληρέστερα όσον αφορά τα προαναφερθέντα και πρόκειται για ανοιχτού κώδικα λογισμικό. Ακολουθεί μια συνοπτική παρουσίαση αντίστοιχων εργαλειών.

7.1 UTD

Το UTD Anonymization Toolbox είναι γραμμένο σε JAVA και είναι ανοιχτού κώδικα. Υποστηρίζει 3 διαφορετικά μοντέλα αυθεντικοποίησης (k-anonymity, l-diversity και t-closeness). Το εργαλείο εφαρμόζει τον Incognito και τον Datafly (αλγόριθμος ευρετικής αναζήτησης). Όταν χρησιμοποιείται ο l-ποικιλομορφία ή ο t-εγγύτητα, ο μετασχηματισμό δεδομένων περιόρίζεται σε γενίκευση πλήρους τομέα, διότι το εργαλείο δεν είναι δυνατό να διαχειριστεί μη μονοτονικά προβλήματα ιδιωτικότητας. Αυτό μπορεί να οδηγήσει σε δεδομένα κακής ποιότητας. Επίσης υποστηρίζει ακόμα δύο μοντέλα γενίκευσης: πολυδιάσταση ολική γενίκευση με χρήση του Mondrian αλγόριθμου και του ανατομισμού (μορφή “τεμαχισμού”...δεν αναφέρεται στην παρούσα εργασία). Το εργαλείο χρησιμοποιεί SQLite συστήμα υποστήριξης βάσης δεδομένων. Δεν έχει γραφικό περιβάλλον και σε σχέση με το ARX δεν περιλαμβάνει μεθόδους για ανάλυση κινδύνου ή ανωνυμοποίηση βάση του ρίσκου.

7.2 CAT

Το εργαλείο ανωνυμοποίησης του πανεπιστημίου του Cornell (CAT) είναι γραμμένο σε C++. Εφαρμόζει l-ποικιλομορφία και t-εγγύτητα. Σαν μοντέλο γενίκευσης χρησιμοποιεί Ολική Γενίκευση (γενίκευση πλήρους τομέα). Υποστηρίζει μόνο μη αυτόματη απόκρυψη tuple. Για την είσοδο των δεδομένων απαιτεί αυστηρά καθορισμένη μορφή και εφαρμόζει τον αλγόριθμο Incognito. Τέλος παρέχει λίγες μεθόδους αυτόματης αξιολόγησης δεδομένων και των κινδύνων γνωστοποίησης.

7.3 **TIAMAT**

Το TIAMAT ειναι κλειστού κώδικα λογισμικό και είναι γραμμένο σε JAVA. Υποστηρίζει μόνο το μοντέλο της k-ανωνυμίας. Εφαρμόζει τον Mondrian και τον *k*-Member, (αλγόριθμος ομαδοποίησης). Το μοντέλο γενίκευσης είναι πολυδιάστατης ολικής επαναγενίκευσης με γενίκευση. Το εργαλείο περιλαμβάνει απλό γραφικό πρόγραμμα επεξεργασίας των γενικευμένων ιεραρχείων. Για την αυτόματη μέτρηση της χρησιμότητας των δεδομένων το TIAMAT χρησιμοποιεί την μέθοδο GCP και Classification Metric[19] Ο στόχος αυτού του εργαλείου είναι να συγκρίνει τον αλγόριθμο του Mondrian με την προσέγγιση του k-Member και εξ αυτού εφαρμόζει δίαφορες οπτικοποιήσεις που ομαδοποιούν τους χρόνους εκτέλεσης τους, για καλύτερη χρηστικότητα των δεδομένων

7.4 SECRETA

Το SECRETA (Ελληνικής ομάδας ερευνητών) ειναι κλειστού κώδικα λογισμικό και ειναι γραμμένο σε C++. Αναπτύχθηκε για να την σύγκριση των μεθόδων ανωνυμοποίησης σχεσιακών και συναλλακτικών δεδομένων. Για τα δεδομένα συσχέτισης εφαρμόζει αλγόριθμο ομαδοποίησης και Incognito. Τα μοντέλα γενίκευσης για τα δεδομένα συσχέτισης περιλαμβάνουν πλήρη γενίκευση, γενίκευση υποδέντρου και τοπική γενίκευση. Ενσωματώνει εννέα γνωστούς αλγόριθμους ανωνυμοποίησης καθώς και 3 μεθόδους δέσμευσης για τον συνδυασμό των προαναφερόμενων αλγθορίθμων κάτω απο κοινό πλαίσιο. Επίσης έχει γραφικό περιβάλλον οπτικής αναπαράστασης των χαρακτηριστικών των δεδομένων (όπως το ARX & CAT) με μεθόδους εμπειρικών στατιστικών. Το εργαλείο εφαρμόζει μια προχωρημένη μέθοδο μέτρησης της χρηστικότητας των δεδομένων, ορίζοντας τον φόρτο των αναζητήσεων και προσδιορίζοντας το μέσο σχετικό σφάλμα των αποτελεσμάτων της αναζήτησης το βοηθητικό πρόγραμμα δεδομένων, καθορίζοντας φόρτου εργασίας ερωτήματος και προσδιορίζοντας το μέσο σχετικό σφάλμα (average relative error) των αποτελεσμάτων του ερωτήματος.

Δεδομένου ότι ο στόχος του εργαλείου είναι να συγκρίνει διαφορετικές στρατηγικές ανωνυμοποίησης, χρησιμποίει διάφορες οπτικοποιήσεις χρόνων εκτέλεσης και χρηστικότητας δεδομένων.

7.5 ARX

Το ARX είναι ένα ολοκληρωμένο λογισμικό ανοιχτού κώδικα για την ανωνυμοποίηση των ευαισθητων προσωπικών δεδομένων. Υποστηρίζει μια ευρεία ποικιλία (1) μοντέλων ιδιωτικότητας και κινδύνου, (2) μεθόδους μετασχηματισμού δεδομένων και (3) μεθόδους για την ανάλυση της χρηστικότητας των δεδομένων εξόδου.

Τα βασικά πλεονεκτήματα του ARX είναι[21] :

- **Η ανάλυση του κινδύνου**: Το ARX εφαρμόζει πολλαπλές μεθόδους εκτίμησης του κινδύνου του επαναπροσδιορισμού, συμπεριλαμβάνοντας 2 διαφορετικά μοντέλα υπερπληθυσμού.

- **Ανωνυμοποίηση βάση κινδύνου**: Το ARX παρέχει κριτήρια απορρήτου που χρησιμοποιούν τα ενσωματωμένα μοντέλα κινδύνου. Μπορεί αυτόματα να μετασχηματίσει σύνολα δεδομένων προς την εξασφάλιση ότι οι κίνδυνοι είναι κάτω του ορίου που ορίζει ο χρήστης.

- **Συντακτικά κριτήρια απορρήτου**: Το APX υλοποιεί μια ποικιλία απο πολλαπλά συντακτικά κριτήρια απορρήτου , συμπεριλαμβανομένης της k-ανωνυμίας, 3 παραλλαγών της l-diversity, 2 παραλλαγές του t-closeness και δ-presence (δεν αναφέρεται στην παρούσα). Επιπλέον το ARX υποσηρίζει αυθαίρετους συνδυασμούς κριτηρίων προστασίας της ιδιωτικότητας (συμπεριλαμβανομένων των μεθόδων που βασίζονται στον κίνδυνο)

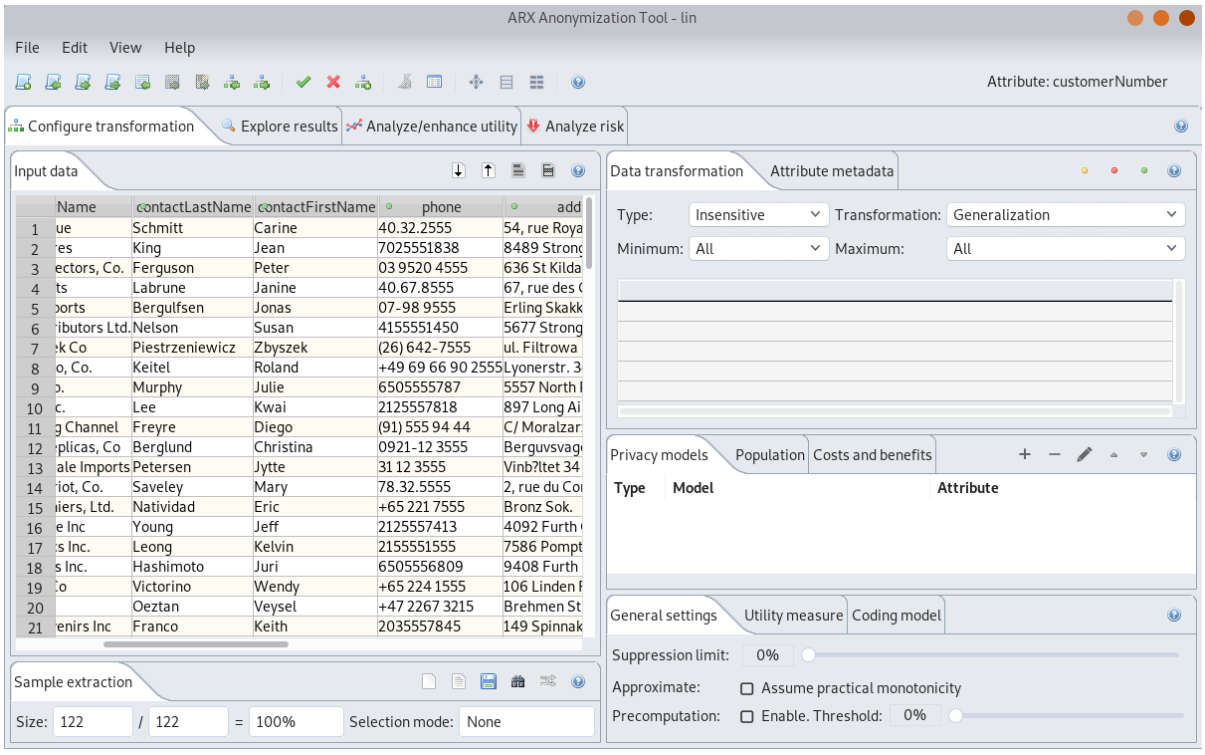
- **Αξιολόγηση χρηστικότητας**: Το ARX εφαρμόζει δίαφορες μεθόδους αξιολόγης της χρησιμότητας των δεδομένων, τόσο αυτόματα όσο και κατα προσωπική παραμετροποίηση. Όσον αφορά τον πρώτο (αυτόματα) οι μέθοδοι είναι πλήρως παραμετροποίησιμοι.

- **Μοντέλο “διαισθητικού” μετασχηματισμού**: Το ARX χρησιμοποιεί την ολική γενίκευση μέσω της πλήρους γενίκευσης τομέα – domain σε συνδυασμό με την τοπική γενίκευση μέσω της απόκρυψης tuple. Αυτο το σχήμα γενίκευσης έχει προταθεί και στον τομέα της βιοϊατρικής. Επιπλέον, υποστηρίζει την top – και bottom γενίκευση καθώς και της μικροσυσσωμάτωσης.
- **Βέλτιστη λειτουργία**: Το ARX υλοποιεί μια ολική βέλτιστη στρατηγική αναζήτησης όπου αυτόματα θα ταξινομεί το χώρο λύσης και θα καθορίζει τον βέλτιστο μετασχηματισμό σύμφωνα με το καθορισμένο μέτρο χρησιμότητας
- **Επεκτασιμότητα**: Το ARX έχει σχεδιαστεί απο την αρχή στο να είναι ιδιαίτερα επεκτάσιμο, και ειναι ικανό να διαχειριστεί πολυ μεγάλα σύνολα δεδομένων (εκατομμυρίων εισαγωγών) σε οικιακά μηχανήματα.

- **Συμβατότητα**: Το ARX υποστηρίζει δεδομένα εισόδου απο πηγές format, όπως .csv αρχεία, excel φύλλα, και σχεσιακών βάσεων δεδομένων συστήματα (MS SQL Server, MySQL, PostgreSQL)

- **Πλήρες γραφικό περιβάλλον**: Το ARX παρέχει ενα ολοκληρωμένο γραφικό περιβάλλον με χαρακτηριστικά όπως οδηγούς δημιουργίας ιεραρχιών γενίκευσης και διάφορες οπτικοποιήσεις των εκτιμήσεων κινδύνου, λύσεων βασιζόμενων στον χώρο και της χρηστικότητας των δεδομένων

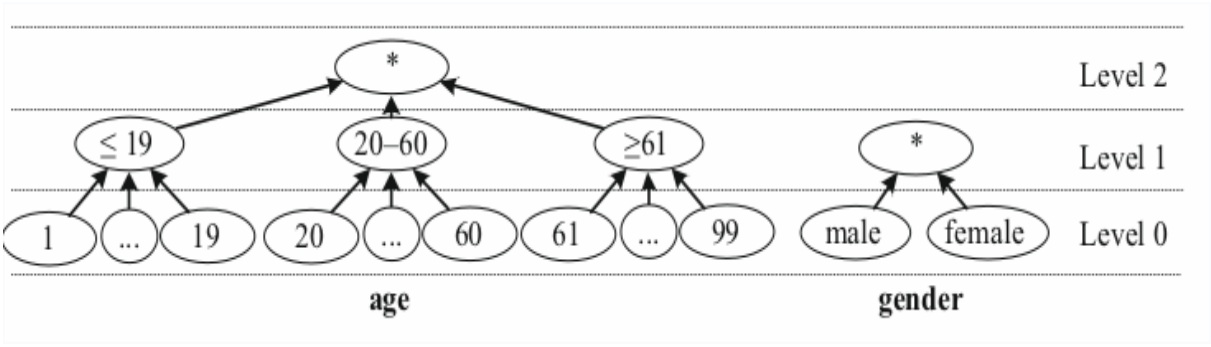
- **Προσεχτικά σχεδιασμένο API**: Το ARX προσφέρει επίσης ένα προγραμματιστικό περιβάλλον (API) που επιτρέπει πρόσβαση σε όλες τις λειτουργίες του. Τόσο οι μέθοδοι όσο και το γραφικό περιβάλλον είναι προσεκτικά σχεδιασμένα.



Εικόνα 1: Το γραφικό περιβάλλον του ARX σε Linux

- **Cross-platform**: Το ARX έχει υλοποιηθεί σε Java και ειναι διάθεσιμο σε όλα τα κοινά λειτουργικά συστήματα.

- **Ανοιχτού Κώδικα**: Το ARX ειναι ανοιχτού κώδικα λογισμικό, παροτρύνοντας τις κριτικές απο την κοινότητα και επιτρέποντας στους χρήστες να το προσαρμόζουν στις δικές του απαιτήσεις



Σχήμα 2: Ιεραρχείες γενίκευσης για τα χαρακτηριστικά ηλικία και γένος

Στην ενότητα **1.1** αναλύσαμε τις τρείς απειλές της ιδιωτικότητας

Για την αντιμετώπιση των απειλών κατά της ιδιωτικότητας με το ARX, τα δεδομένα μετασχηματίζονται με ιεραρχίες γενίκευσης. Το εργαλείο υποστηρίζει ιεραρχίες τόσο για κατηγορηματικά όσο και για συνεχή γνωρίσματα.

Παραδείγματα φαίνονται στο σχήμα 2. Εδώ, οι τιμές του χαρακτηριστικού ηλικία μεταβάλλονται σε “ηλικιακές ομάδες”, ενώ οι τιμές της ιδιότητας του φύλου μπορούν μόνο να κρυφτούν.

Οι ιεραρχίες γενίκευσης είναι κατάλληλες για κατηγορηματικά γνωρίσματα. Μπορούν επίσης να χρησιμοποιηθούν για συνεχή χαρακτηριστικά, εκτελώντας επί τόπου κατηγοριοποίηση. Στο ARX, αυτή η κατηγοριοποίηση υλοποιείται με την απεικόνιση στρατηγικών γενίκευσης με ένα λειτουργικό τρόπο, π.χ., ως ένα σύνολο διαστημάτων.

Για την αύξηση της χρηστικότητας των δεδομένων, η γενίκευση των γνωρισμάτων συνδυάζεται με την απόκρυψη των εγγραφών των αρχείων. Αυτό σημαίνει ότι οι γραμμές που “παραβιάζουν” τα κριτήρια ιδιωτικότητας διαγράφονται αυτόματα απο το σύνολο των δεδομένων. Το σύνολο των διεγεγραμμένων καταγραφών φυλάσεται κάτω απο ένα καθορισμένο όριο, το οποίο ονομάζεται όριο απόκρυψης (suppresion limit). Ως αποτέλεσμα αυτού απαιτείται λιγότερη γενίκευση για την εξασφάλιση της ικανοποιήσης του μοντέλου ιδιωτικότητας απο τις υπόλοιπες εγγραφές.[21]

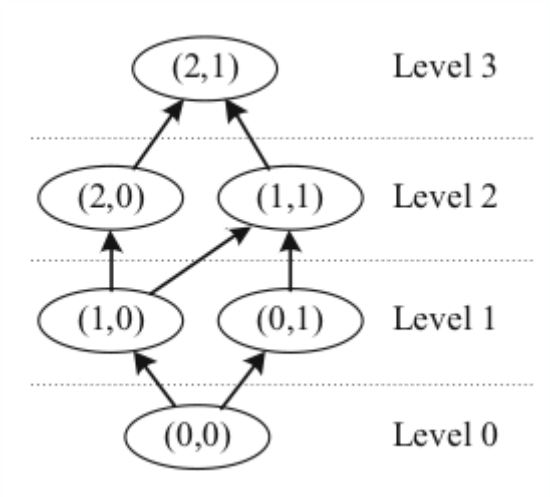
Το ARX εφαρμόζει πολυ-διάσταση ολική γενίκευση με πλήρους τομέα γενίκευση και τοπική γενίκευση με άποκρυψη tuble.

Αυτός ο συνδυασμός των μεθόδων εχει βρεθεί να ταιρίαζει τέλεια στο τομέα της βιοϊατρικής, καθότι είναι: i) διαισθητική , ii) Παράγει σύνολα δεδομένων που είναι κατάλληλα αναλύσεων απο επιδημιολόγους και iii) μπορούν να διαμορφώνονται απο μη ειδικούς. Η παραμετροποίηση (και ως εκ τούτου η ισορροπία της ιδιωτικότητας και της χρηστικότητας) μπορεί να πραγματοποιηθεί μεταβάλλοντας τις ιεραρχείες γενίκευσης ή απλά επιλέγοντας ένα κατάλληλο μετασχηματισμό απο το σύνολο των λύσεων (solution space). Για συνεχείς μεταβλητές, το ARX υποστηρίζει την μικροσυσσωμάτωση. Αυτό σημαίνει ότι οι τιμές του quasi-identifier μέσα σε μια ομάδα μετρασχηματίζονται εφαρμόζοντας αθροιστικές λειτουργίες, όπως το αριθμητικό ή το γεωμετρικό μέσο. Σε έναν ανώνυμο πίνακα, όλη η ομάδα των τιμών αντικαθιστάται απο το

αποτέλεσμα της συναθροιστικής συνάρτησης. Σε αντίθεση με την γενίκευση ή την απόκρυψη, αυτή η μέθοδος μπορεί να χρησιμοποιηθεί για τη διατήρηση των τύπων των δεδομένων και κλίμακα μέτρησης μιας αριθμητικής μεταβλητής.[21]

Όταν χρησιμοποιούμε ολική γενίκευση με γενίκευση πλήρους τομέα, ο χώρος αναζήτησης μπορεί να διαμορφωθεί ως επαναλαμβανόμενη δομή γενίκευσης, η οποία ειναι ένα σύνολο διατεταγμένων πιθανών συνδυασμών των επιπέδων γενίκευσης για κάθε γνώρισμα. Οι επαναμβανόμενες δομές μπορούν να απεικονιστούν ώς διαγράμματα Hasse, που στην περίπτωση του πλαισίου μας σημαίνει, την σχεδίαση της μεταβατικής μείωσης ενός συνόλογου απο μεασχηματισμούς σε ένα άκυκλο γράφημα, όπου ο κάθε κόμβος θα είναι συνδεδεμένος με όλους τους άμεσους διαδόχους και προκάτοχους του. Ο κάθε κόμβος αντιπροσωπεύει ένα μετασχηματισμό και ορίζει τα επίπεδα γενίκευσης όλων των quasi-identifiers. Ένα βέλος υποδηλώνει ότι ένας μετασχηματισμός είναι μια άμεση γενίκευση ενός πιο εξειδικευμένου μετασχηματισμού, που μπορεί να δημιουργεί αυξάνοντας τα επίπεδα γενίκευσης που ορίζει ο προκάτοχος του.

Ένα παράδειγμα που χρησιμοποιεί τις ιεραρχίες απο το σχήμα 2, απεικονίζεται στο σχήμα 3.



Σχήμα 3: Παράδειγμα χώρου αναζήτησης

7.5.1 Η Ανάλυση κινδύνου και η Ανωνυμοποίηση βάση κινδύνου με την χρήση του ARX

Στο ARX, μπορούν να χρησιμοποιηθούν μοντέλα υπερπληθυσμού για την εκτίμηση της μοναδικότητας του πληθυσμού, δηλαδή το κλάσμα των καταχωρήσεων στο σύνολο των δεδομένων που ειναι μοναδικά στο σύνολο του πληθυσμού.

Αυτές οι στατιστικές μεθόδοι υπολογίζουν χαρακτηριστικά του συνολικού πληθυσμού με κατανομές πιθανοτήτων που παραμετροποιούνται με τα χαρακτηριστικά του δείγματος.
Επίσης, οι εκτιμήσεις κινδύνου μπορούν να χρησιμοποιηθούν για την ανωνυμποίηση βάση κινδύνου. Το στατικό μοντέλο k-ανωνυμίας ορίζει ένα ανώτερο όριο στον επαναπροσδιορισμό του κινδύνου, που υπολογίζεται βάση δείγματος συγχνοτήτων. Το ARX υποστηρίζει πολλαπλές παραλλαγές αυτού του κριτηρίου που χρησιμοποιούν τα προαναφερόμενα μοντέλα κινδύνου επιβεβαιώνοντας ότι οι εκτιμήσεις κινδύνου υπολείπονται ενός συγκεκριμένου ορίου.

7.5.2 Αξιολόγηση χρησιμότητας

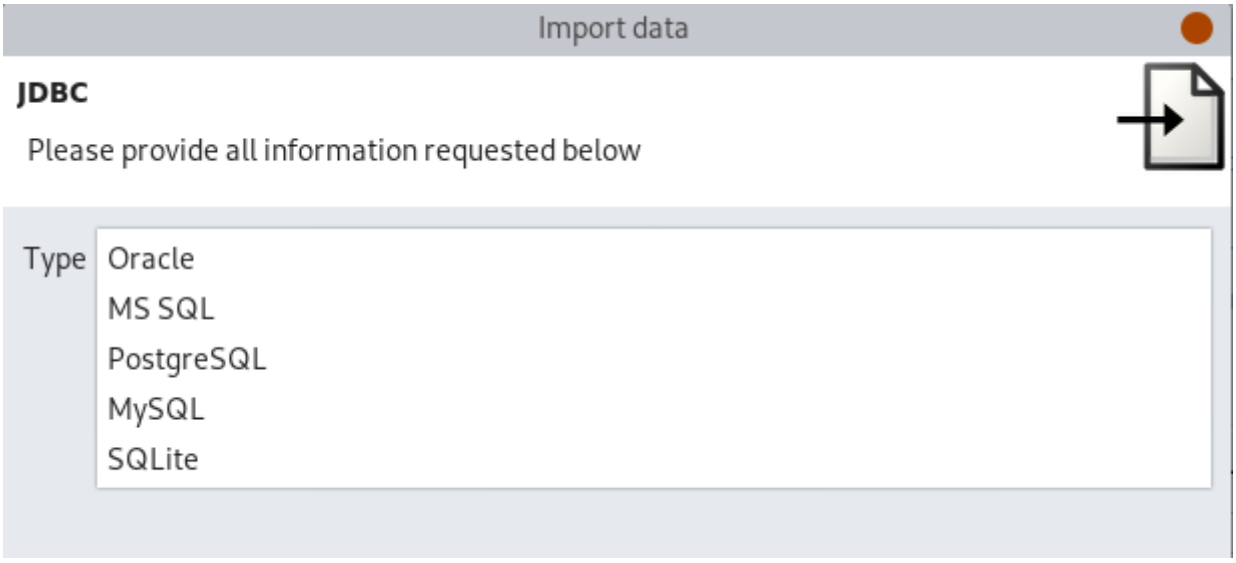
Για την αυτόματη μέτρηση της χρησιμότητας των δεδομένων, το ARX υποστηρίζει μεθόδους που βασίζονται σε κλάσεις ισοδυναμίας , που ονομάζονται μετρήσεις μονοδιάστατης χρηστικότητας μέσα στο εργαλείο. Παραδείγματα περιλαμβάνουν την *Διακριτότητα και το Μέγεθος μέσης ισοδύναμης κλάσης*. Και τα 2 μέτρα υπολογίζουν το περιοχόμενο της πληροφορίας βάσει του μεγέθους των τάξερων ισοδυναμίας , ενώ η ευκρίνεια συνεπάγεται επίσης της ποινής για τις κρυμμένες tuples.

Επιπρόσθετα, το ARX υποστηρίζει μεθόδους που βασίζονται σε τιμές πραγματικών δεδομένων.
Καθώς αυτές οι μέθοδοι υπολογίζουν ανεξάρτητες τιμές για κάθε γνώρισμα, οι οποιες στην συνέχεια μεταγλωττίζονται σε μια ολική τιμή και αποκαλούνται πολυ-διάστατες μετρήσεις χρηστικότητας. Επίσης το ARX υποστηρίζει αρκετές συναθροιστικές λειτουργίες για την σύνταξη μέτρων ολικής χρηστικότητας, συμπεριλαμβανομένου του αθροίσματος, του αριθμητικού και του γεωμετρικού μέσου. [21]

7.5.3 Επιπρόσθετα Χαρακτηριστικά

Το ARX είναι συμβατό με ένα ευρύ φάσμα εργαλείων επεξεργασίας δεδομένων. Στην παρούσα υποστηρίζει διεπαφές για την εισαγωγή δεδομένων για αρχεία CSV, MS Excel φύλλα, συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS), οπως MS SQLServer, PostgreSQL and MySQL.

Το συντακτικό των CSV αρχείων αναγνωρίζεται αυτόματα. Σε σημασιολογικό επίπεδο, το ARX υποστηρίζει διαφορετικούς τύπους δεδομένων και κλίμακες μέτρησης. Τα format των τύπων δεδομένων εντοπίζονται εξίσου αυτόματα. Επίσης υποστηρίζει εκκαθάριση δεδομένων κατα την διάρκεια της εισόδου, εννοώντας ότι τα μη-έγκυρα δεδομένα αντικαθίστονται με μια τιμή που αντιπροσωπεύει τα στοιχεία που λείπουν. Παράλληλα χειρίζεται με ασφάλεια τις ελλείπουσες τιμές, διασφαλίζοντας ότι οι διαφορετικές τιμές που λείπουν, αντιστοιχούν μεταξύ τους.



Εικόνα 2: Επιλογές βάσεων δεδομένων

7.5.4 Χρήση του ARX

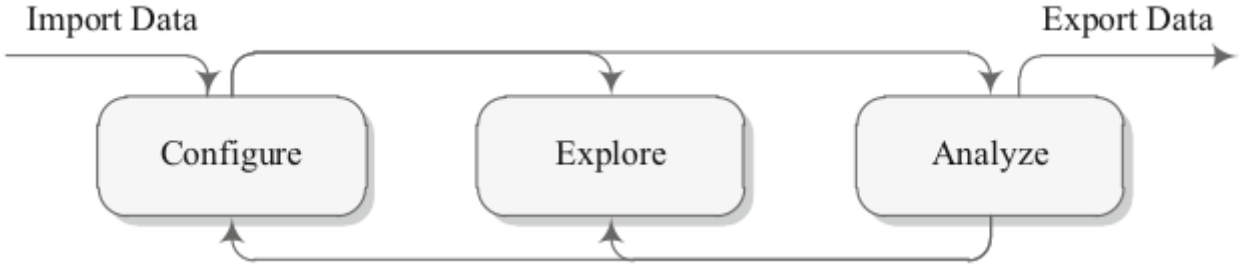
Στην κάτωθι ενότητα θα περιγραφεί αρχικώς η ροή εργασιών που υποστηρίζεται απο το γραφικό περιβάλλον χρήστη του ARX. Εν συνεχεία θα παρουσιαστεί αναλυτικότερα το γραφικό περιβάλλον και θα περιγραφούν αναλυτικότερα τα προαναφερόμενα

7.5.4.1 Η διαδικασία της ανωνυμοποίησης

Η κεντρική πρόκληση στην ανωνυμοποίηση δεδομένων είναι η επίτευξη ισορροπίας μεταξύ της χρηστικότητας των δεδομένων και της ιδιωτικότητας. Στο ARX, οι μέθοδοι που μοντελοποιούν τις διαφορετικές πτυχές αυτής της διαδικασίας συνδυάζονται σε μια ροή εργασίας πολλαπλών σταδίων που επιτρέπει στους χρήστες να προσαρμόζουν με παραμέτρους τις παραμέτρους, μέχρις ότου το αποτέλεσμα να ανταποκρίνεται στις προσδοκίες τους. Όπως απεικονίζεται στο Σχήμα 3, τα βασικά βήματα αποτελούνται από

i) Τη παραμετροποίηση του μοντέλου ιδιωτικότητας και μετασχηματισμού
ii) Εξερεύνηση του χώρου της λύσης
iii) Ανάλυση εισερχόμενων και εξερχόμενων δεδομένων. Στη φάση διαμόρφωσης, τα δεδομένα εισόδου φορτώνονται και σχολιάζονται, δημιουργούνται ή εισάγονται ιεραρχίες γενίκευσης και καθορίζονται όλες οι υπόλοιπες παράμετροι, όπως τα κριτήρια ιδιωτικότητας.

Όταν ο χώρος λύσης χαρακτηρίζεται από την εκτέλεση του ARX αλγόριθμου ανωνυμοποίησης, η φάση εξερεύνησης υποστηρίζει την αναζήτηση μετασχηματισμών δεδομένων που προστατεύουν την ιδιωτικότητα και πληρούν τις απαιτήσεις του χρήστη. Για να εγκριθεί η καταλληλότητα, η φάση της ανάλυσης επιτρέπει τη σύγκριση μετασχηματισμένων συνόλων δεδομένων στο αρχικό σύνολο δεδομένων εισόδου με μεθόδους περιγραφικών στατιστικών. Επιπλέον, μπορούν να πραγματοποιηθούν αναλύσεις κινδύνου για δεδομένα εισόδου όπως και μετασχηματισμένες παραστάσεις τους. Βάση αυτων των αναλύσεων, μπορούν να εξεταστούν και περαιτέρω πιθανές λύσεις και να αξιολογηθούν ή η διαμόρφωση της διαδικασίας ανωνυμίας μπορεί να μεταβληθεί.



Διάγραμμα 1. Η διαδκασία της ανωνυμοποίησης, εφαρμοσμένη στο γραφικό περιβάλλον του ARX

Τα 3 βήματα της διαδικασίας της ανωνυμοποίησης αντιστοιχίζονται σε τέσσερις προοπτικές στο γραφικό περιβάλλον του ARX:

Παραμετροποίηση:

Σε αυτή την προοπτική, *πρώτον*, ένα σύνολο δεδομένων μπορεί να εισαχθεί στο εργαλείο και να επισημανθεί , π.χ., χαρακτηρισμός γνωρίσματων.

Δεύτερον, με απευθείας εισαγωγή ή χρήση του οδηγού, οι ιεραρχίες γενίκευσης για τα ψευδοαναγνωριστικά ή ευαίσθητα γνωρίσματα μπορούν να δημιουργηθούν ημι-αυτόματα

Τρίτον: Μπορούν να προσδιοριστούν τα κριτήρια ιδιωτικότητας, η μέθοδος μέτρησης χρηστικότητας δεδομένων και άλλες παράμετροι, όπως οι ιδιότητες του μοντέλου μετασχηματισμού.

Εξερεύνηση:

Ως αποτέλεσμα της διαδικασίας της ανωνυμοποίησης, δημιουργείται ένας χώρος λύσεων και χαρακτηρίζεται βάση των δεδομένων παραμέτρων. Αυτή η προοπτική επιτρέπει στους χρήστες την περιήγηση των διαθέσιμων μετασχηματισμένων δεδομένων, να οργανώνουν και να τα φιλτράρουν ανάλογα με τις ανάγκες τους, και να επιλέγουν επιπλέον μετασχηματισμούς για ανάλυση.

Αξιολόγηση χρησιμότητας:

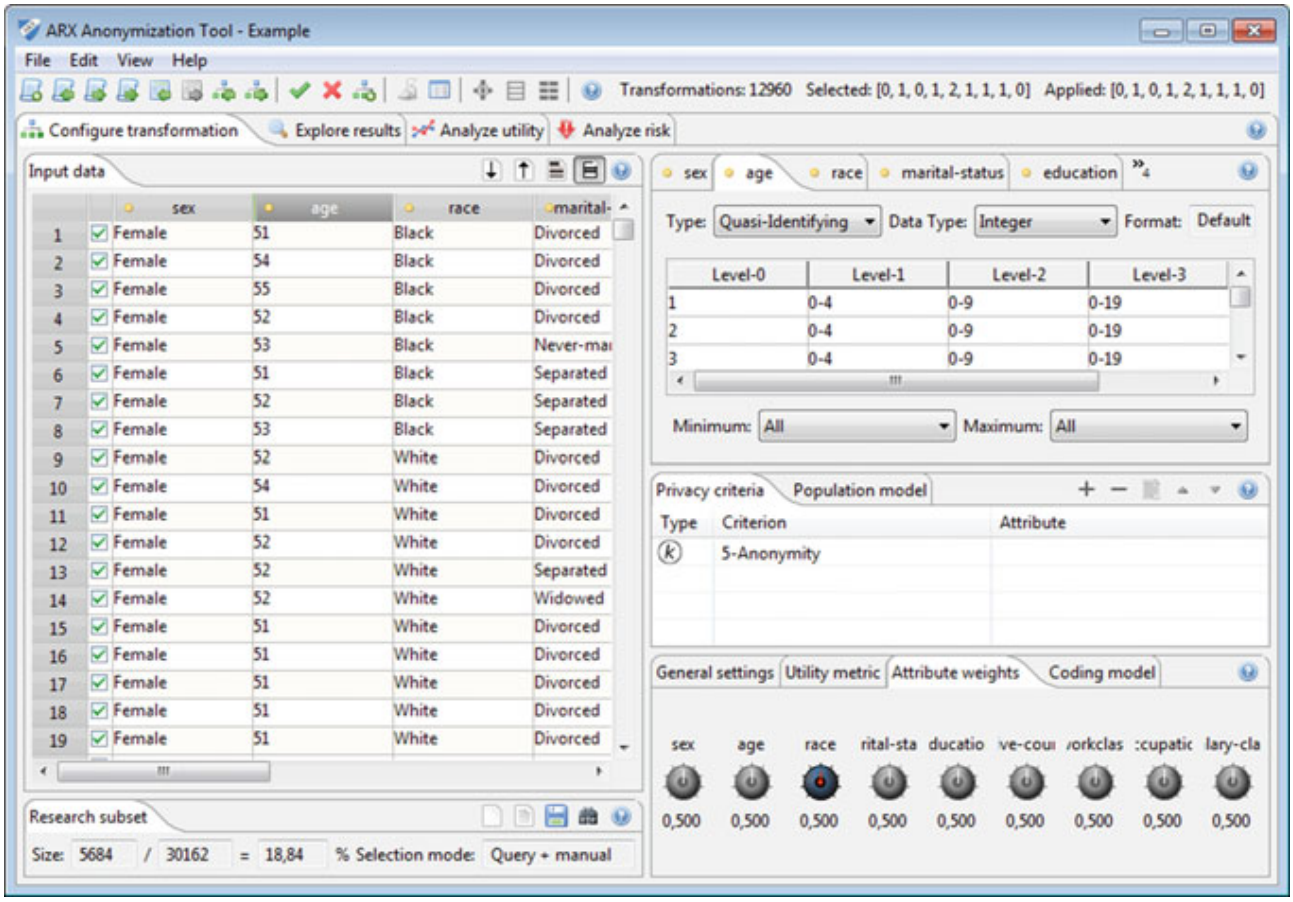
Η συγκεκριμένη προοπτική επιτρέπει τη σύγκριση μετασχηματισμών του συνόλου δεδομένων εισόδου με τα αρχικά δεδομένα, για να αξιολόγησει της καταλληλότητας ενός συγκεκριμένου μετασχηματισμού για το δεδομένο σενάριο χρήσης. Για το σκοπό αυτό, ενσωματώνει διάφορες γραφικές παραστάσεις των αποτελεσμάτων των στατιστικών αναλύσεων ενώ επιτρέπει την σύγκριση των πεδίων 1-1.

Ανάλυση Κινδύνου:

Σε αυτή την προοπτική, μπορεί να αναλυθεί η κατανομή των μεγεθών τάξης, οι κίνδυνοι που σχετίζονται με τα μεμονομένα ψευδοαναγνωριστικά, όπως επίσης εκτιμήσεις κινδύνου βάση δειγμάτων και πληθυσμού. Εμφανίζονται επίσης λεπτομέρειες σχετικά με εκτιμώμενους κινδυνους επαναπροσδιορισμού που προέρχονται απο διαφορειτκά μοντέλα.

7.5.4.2 Παραμετροποιώντας την διαδικασία ανωνυμοποίησης

Σε αυτήν την προοπτική, που φαίνεται και στην εικόνα 3, μπορεί να “σχολιαστεί” (χαρακτηριστεί) ένα σύνολο δεδομένων και να παραμετρποιηθεί η διαδικασία της ανωνυμοποίησης. Με την έννοια του σχόλιου, εννούμε τον χαρακτηρισμό ενός γνώρισματος ως ψευδογνώρισμα ή ευαίσθητο, ο ορισμός ιεραρχείων γενίκευσης, ο ορισμός υποσυνόλου έρευνας και η εισαγωγή πληροφοριων του συνολικού πληθυσμού. Η παραμετροποίηση σημαίνει, τον προσδιορισμό του μοντέλου ιδιωτικότητας και διαμόρφωση της διαδικασίας μετασχηματισμού. Στη αριστερή πλευρά της εικόνας



Εικόνα 3: Η παραμετροποίηση του ARX

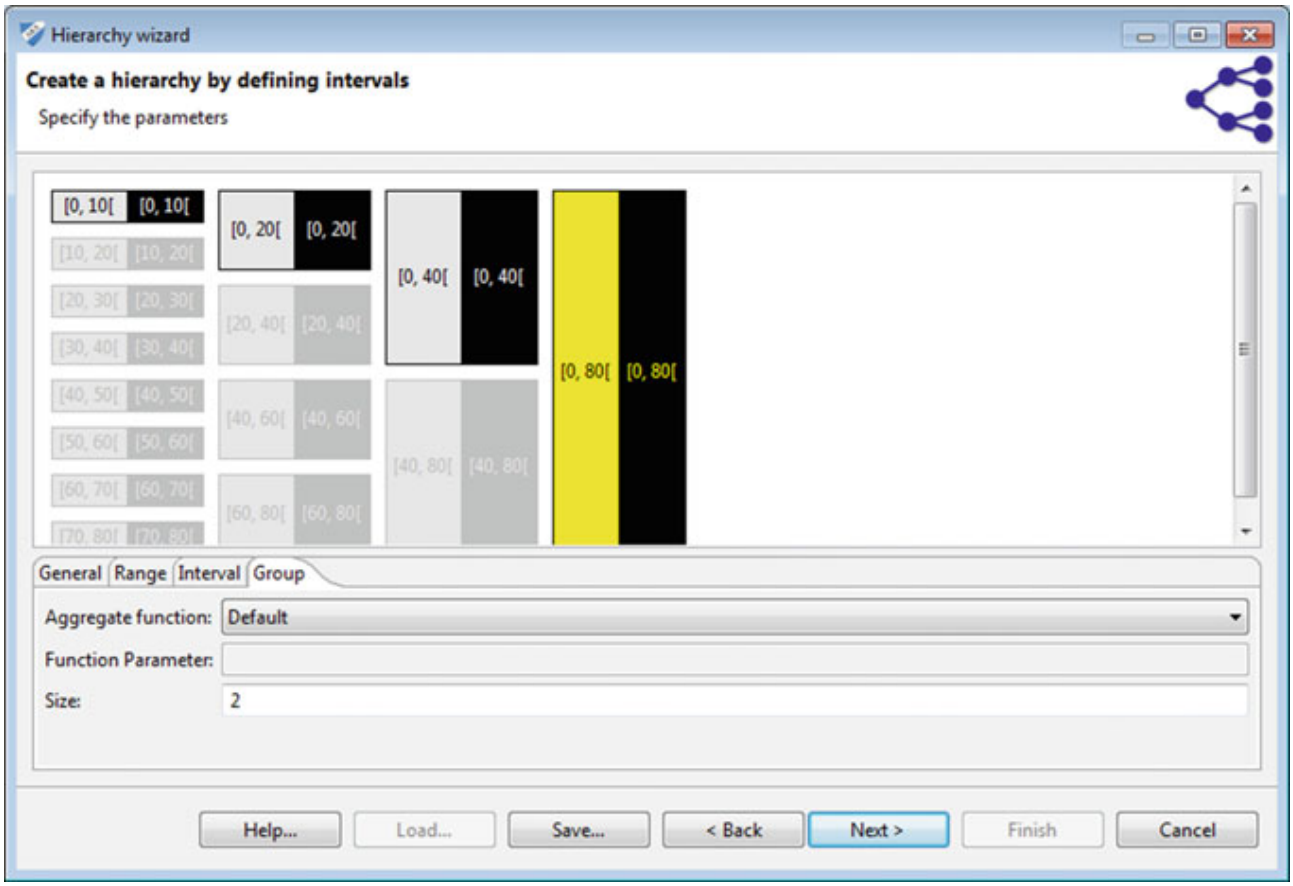
Μια διεπαφή αναζήτησης, η οποία βρίσκεται στην κάτω αριστερή γωνία μπορεί να χρησιμοποιηθεί για την επιλογή ενός υποσυνόλου αναζήτησης, το οποίο ειναι ουσιαστικά ένα υποσύνολο αρχείων δεδομένων τα οποία θα συμπεριληφθούν στο τελικό ανώνυμο σύνολο δεδομένων. Αυτο ο μηχανισμός μπορεί να χρησιμοποιηθεί για διάφορετικούς σκοπούς.

Πρώτον, μπορεί να χρησιμοποιηθεί για την επιβολή δ-παρουσίας στο υποσύνολο, για να προστατέψει το υποσύνολο απο την αποκάλυψη της ιδιότητας του μέλους, εμποδίζοντας τους επιτιθέμενους που γνωρίζουν τη συνολική ομάδα δεδομένων να καθορίσουν εάν ένα συγκεκριμένο περιέχεται ή όχι στο υποσύνολο.

Δεύτερον, μπορεί να χρησιμοποιηθεί για τη χειροκίνητη κατάργηση των καταχωρήσεων των δεδομένων. Η περιοχή στην επάνω δεξιά γωνία περιγράφει όλα τα χαρακτηριστικά και δίνει την επιλογή των “χαρακτηρισμών”, καθορίζοντας τύπους δεδομένων και τύπους γνωρισμάτων. Στην ίδια περιοχή εμφανίζονται επίσης οι ιεραρχίες γενίκευσης που σχετίζονται με τα γνωρίσματα.

Το ARX διακρίνει τέσσερις διαφορετικούς τύπους γνωρίσματων:

1. Χαρακτηριστικά γνωρίσματα, θα αφαιρεθούν από το σύνολο δεδομένων. Τέτοιου τύπου είναι π.χ. τα ονόματα ή οι αριθμοί κοινωνικής ασφάλισης.
2. Τα ψευδοαναγνωρίστικά θα μετασχηματιστούν μέσω της γενίκευσης, απόκρυψης ή μικροσυσσωμάτωσης. Τυπικά παραδείγματα είναι το φύλο, η ημερομηνία γέννησης και οι ταχυδρομικοί κώδικες.
3. Τα ευαίσθητα γνωρίσματα θα διατηρηθούν χωρίς τροποποίηση, αλλά ενδέχεται να υπόκεινται σε περαιτέρω περιορισμούς, όπως t-εγγύτητα t ή l- ποικιλομορφία. Χαρακτηριστικά παραδείγματα είναι οι διαγνώσεις.
4. Τα μη ευαίσθητα γνωρίσματα θα παραμείνουν αμετάβλητα.



Εικόνα 4: Οδηγός δημιουργίας της ιεραρχείας γενίκευσης με διαστήματα

Για τον καθορισμό των ιεραρχιών γενίκευσης, το εργαλείο προσφέρει έναν οδηγό, ο οποίος υποστηρίζει τη δημιουργία ιεραρχιών για γνωρίσματα με διαφορετικές κλίμακες μέτρησης. Ο οδηγός για τη δημιουργία ιεραρχιών με διαστήματα εμφανίζεται στην εικόνα 4.

Αρχικά, μπορεί να οριστεί μια ακολουθία διαστημάτων. Στο επόμενο βήμα, μπορούν να οριστούν τα επόμενα επίπεδα της ιεραρχίας, τα οποία αποτελούνται από ομάδες διαστημάτων από το προηγούμενο επίπεδο. Όπως μπορεί να φανεί, κάθε ακολουθία διαστημάτων ή ομάδων επαναλαμβάνεται αυτόματα για να καλύψει το πλήρες εύρος του γνωρίσματος. Κάθε στοιχείο συσχετίζεται με μια συνθετική συνάρτηση. Αυτές οι συναρτήσεις ενσωματώνουν μεθόδους για την δημιουργία ετικετών για τα διαστήματα, ομάδες και τιμές που πρέπει να μεταφραστούν σε μια κοινή γενικευμένη τιμή.

Από τη προοπτική της παραμετροποίησης, οι ιεραρχίες οπτικοποιούνται με την μορφή πίνακων, όπου κάθε σειρά τους περιέχει τον κανόνα γενίκευσης για μια μοναδική τιμή γνωρίσματος.

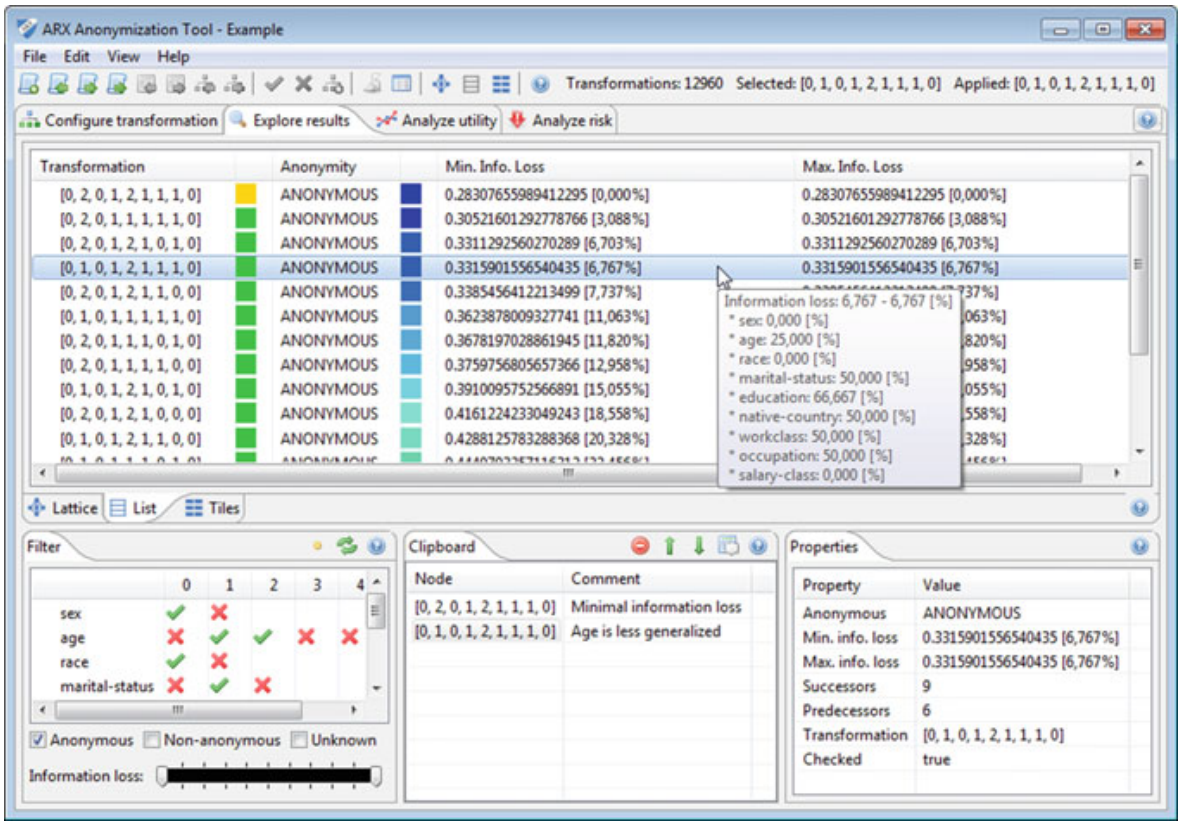
Αυτή η “διαισθητική” αναπαράσταση επιτρέπει τη συμβατότητα με τρίτες εφαρμογές όπως τα προγράμματα υπολογιστικών φύλλων (excel). Οι ιεραρχίες μπορούν επίσης να ρυθμιστούν με έναν ενσωματωμένο επεξεργαστή, ο οποίος είναι διαθέσιμος κάνοντας δεξί κλικ στην πινακοποιημένη αναπαράστασή τους.

Η άποψη στο κέντρο της δεξιάς πλευράς της προοπτικής υποστηρίζει την προδιαγραφή του μοντέλου ιδιωτικότητας. Για το σκοπό αυτό, εμφανίζει μια λίστα με τα κριτήρια ιδιωτικότητας που είναι ενεργοποιημένα και υποστηρίζονται σαν επιλογές η προσθήκη, η αφαίρεση και διαμόρφωση των κριτηρίων μέσω των αντίστοιχων πλαισίων. Στη δεύτερη καρτέλα, μπορούν να οριστούν τα χαρακτηριστικά του πληθυσμού που απαιτούνται για αναλύσεις κινδύνου και ανωνυμοποίηση βάσει κινδύνου.

Επιλογές για τη παρεμετροποίηση της διαδικασίας μετασχηματισμού ενσωματώνονται στις προβολές στο κάτω μέρος της δεξιάς πλευράς της προοπτικής. Στις γενικές ρυθμίσεις, μπορεί να καθοριστεί το όριο κατάργησης και να προσαρμοστούν περαιτέρω παράμετροι που σχετίζονται με την απόδοση. Στη δεύτερη καρτέλα, μπορούν να επιλεγούν και να διαμορφωθούν μέτρα χρηστικότητας. Οι άλλες δύο καρτέλες επιτρέπουν στους χρήστες να παραμετροποιήσουν το μοντέλο μετασχηματισμού με τη σταθμίζοντας γνωρίσματα και με την ιεράρχηση διαφορετικών τύπων γενίκευσης δεδομένων.

7.5.4.3 Διαχείριση του χώρου των αποτελεσμάτων

Όταν ο χώρος αναζήτησης έχει χαρακτηριστεί με βάση τις συγκεκριμένες παραμέτρους, η προοπτική εξερεύνησης επιτρέπει στους χρήστες να αναζητούν πιθανές λύσεις, να τις οργανώνουν και να τις φιλτράρουν σύμφωνα με τις απαιτήσεις τους. Σκοπός της προοπτικής (εικόνα 5) είναι η επιλογή ενός συνόλου ενδιαφερόντων μετασχηματισμών για περαιτέρω αξιολόγηση.



Εικόνα 5: Η προοπτική εξερεύνησης του ARX

Στο κέντρο αυτής της οπτικής εμφανίζεται ένα υποσύνολο του χώρου λύσης. Στο παράδειγμα μας, οι μετασχηματισμοί παρουσιάζονται ως λίστα, ο οποίος ταξινομείται βάση χρηστικότητας δεδομένων. Κάθε μετασχηματισμός εκπροσωπείται από τα επίπεδα γενίκευσης που καθορίζει για τα ψευδοαναγνωριστικά στην είσοδο του σύνολο δεδομένων. Οι μετασχηματισμοί χαρακτηρίζονται από τέσσερα διαφορετικά χρώματα:

το πράσινο υποδεικνύει ότι ένας μετασχηματισμός οδηγεί σε ένα ανώνυμο σύνολο δεδομένων.

Το κόκκινο υποδεικνύει ότι ο μετασχηματισμός δεν έχει ως αποτέλεσμα ένα ανώνυμο σύνολο δεδομένων.

Το κίτρινο υποδεικνύει ότι ο μετασχηματισμός είναι ο βέλτιστος σε πλήρες επίπεδο, και το γκρι δείχνει ότι η ιδιότητα ανωνυμίας ενός μετασχηματισμού είναι άγνωστη ως αποτέλεσμα του κλάδου.

Αν ένας τέτοιος μετασχηματισμός εφαρμοστεί σε σύνολο των δεδομένων, η πραγματική ιδιότητά της ανωνυμίας θα υπολογιστεί στο παρασκήνιο και η κατάσταση του χώρου αναζήτησης θα ενημερωθεί.

Το κάτω αριστερό σημείο επιτρέπει το φιλτράρισμα του εμφανιζόμενου υποσυνόλου του χώρου λύσης, π.χ., περιορίζοντας τους μετασχηματισμούς σε ορισμένα επίπεδα γενίκευσης ή καθορίζοντας κατώτατα όρια στην χρηστικότητα των δεδομένων. Οι μετασχηματισμοί μπορούν επίσης να προστεθούν στο πρόχειρο, το οποίο βρίσκεται στο κάτω-κεντρικό τμήμα της οπτικής μας. Εδώ, οι μετασχηματισμοί μπορούν να σχολιαστούν και να οργανωθούν. Η κάτω δεξιά περιοχή εμφανίζει βασικές πληροφορίες για τον τρέχων επιλεγμένο μετασχηματισμό. Κάνοντας δεξί κλικ σε ένα μετασχηματισμό εμφανίζεται ένα μενού περιεχομένου, το οποίο επιτρέπει την εφαρμογή του στο σύνολο δεδομένων. Το αποτέλεσμα μπορεί στη συνέχεια να εξαχθεί ή να αναλυθεί.

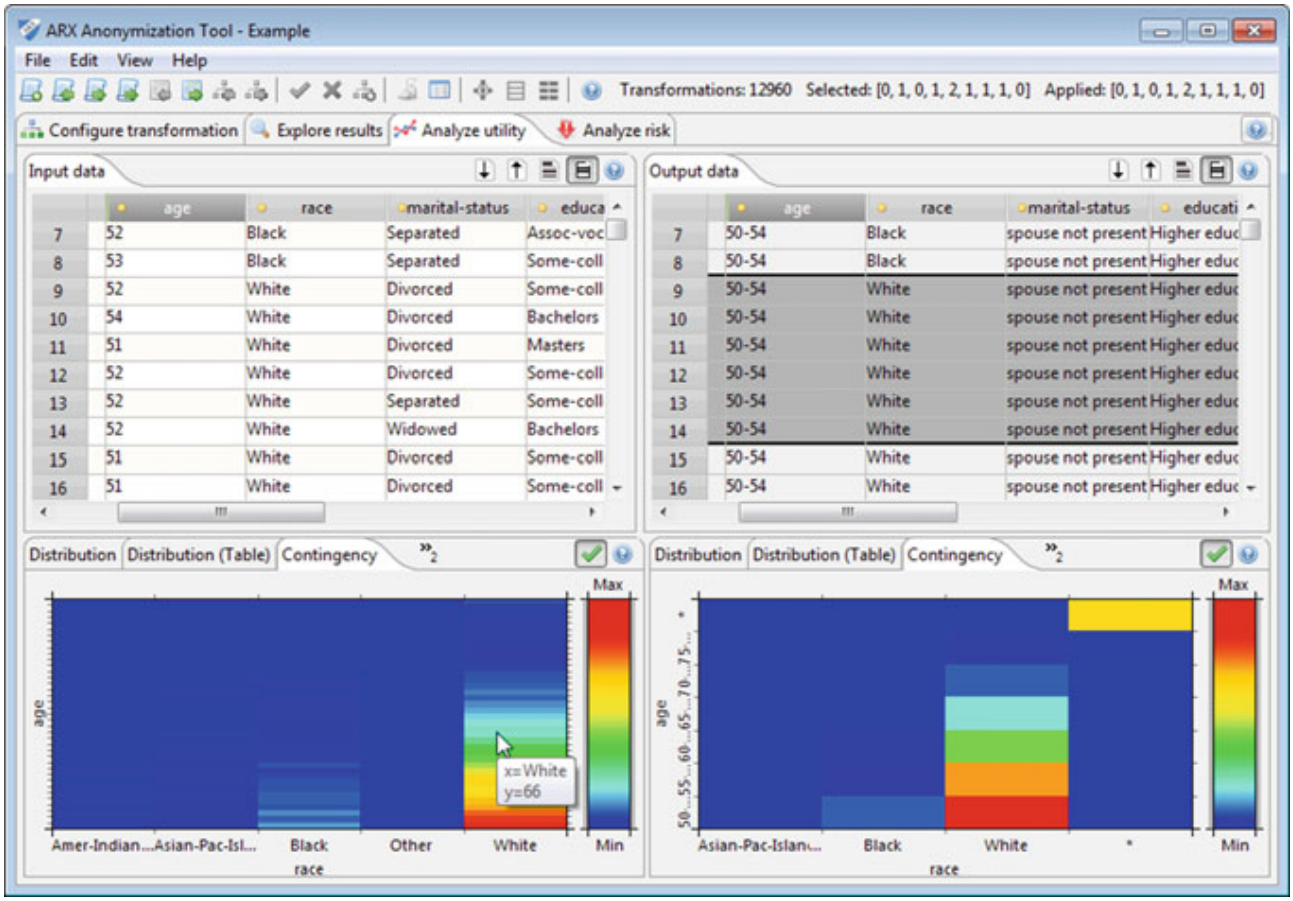
Δύο επιπλέον οπτικοποιήσεις του χώρου λύσεων παρέχονται. Πρώτον, μπορεί να απεικονιστεί ως διάγραμμα Hasse. Εδώ, χρησιμοποιούνται τα ίδια χρώματα όπως στην προβολή λίστας για την κωδικοποίηση της ιδιότητας ανωνυμίας. Η προβολή είναι διαδραστική και υποστηρίζει ζουμ. Δεύτερον, ο χώρος των αποτελεσμάτων μπορεί επίσης να απεικονιστεί ως σύνολο “πλακιδίων”, τα οποία ταξινομούνται και χρωματίζονται με βάση την χρηστικότητα των δεδομένων. Σε αντίθεση με τις άλλες δύο παραστάσεις, αυτή η οπτική είναι ικανή να εμφανίζει ταυτόχρονα μεγάλο αριθμό μετασχηματισμών.

7.5.4.4 Αξιολόγηση Χρηστικότητας Δεδομένων

Αυτή η προοπτική επιτρέπει στους χρήστες να συγκρίνουν τις παραμετροποιημένες απεικονίσεις δεδομένων με το αρχικό σύνολο δεδομένων. Ένα παράδειγμα φαίνεται στη εικόνα 6. Εμφανίζει το σύνολο δεδομένων εισόδου στην αριστερή πλευρά και το σύνολο δεδομένων εξόδου στη δεξιά πλευρά. Και οι δύο πίνακες είναι συγχρονίζονται κατά την κύλιση, υποστηρίζοντας συγκρίσεις μεταξύ των πεδίων. Τα δεδομένα μπορούν να ταξινομηθούν κατα ένα μόνο χαρακτηριστικό ή κατα όλα τα ψευδοαναγνωριστικά, τα οποία επίσης θα επισημάνουν τις οι προκύπτουσες κλάσεις ισοδυναμίας, εάν εφαρμοστούν στην έξοδο του σύνολο δεδομένων.

Για τη σύγκριση δύο απεικονίσεων των δεδομένων, η προβολή εμφανίζει τυπικές περιγραφικές στατιστικές, τόσο σε μορφή πίνακα όσο και σε μορφή γραφικών. Περιλαμβάνουν μεθόδους μοναδικής και ποικίλης ποικιλίας, όπως εμπειρικές κατανομές, κεντρική τάση και

διασπορά, καθώς και πίνακες διασταύρωσης. Μαζί με τις στατιστικά μετρήσεις, αυτή η οπτική εμφανίζει επίσης πληροφορίες αντανακλώντας τη διαδικασία ανωνυμοποίησης.



Εικόνα 6 Αξιολόγηση Χρηστικότητας Δεδομένων στο ARX

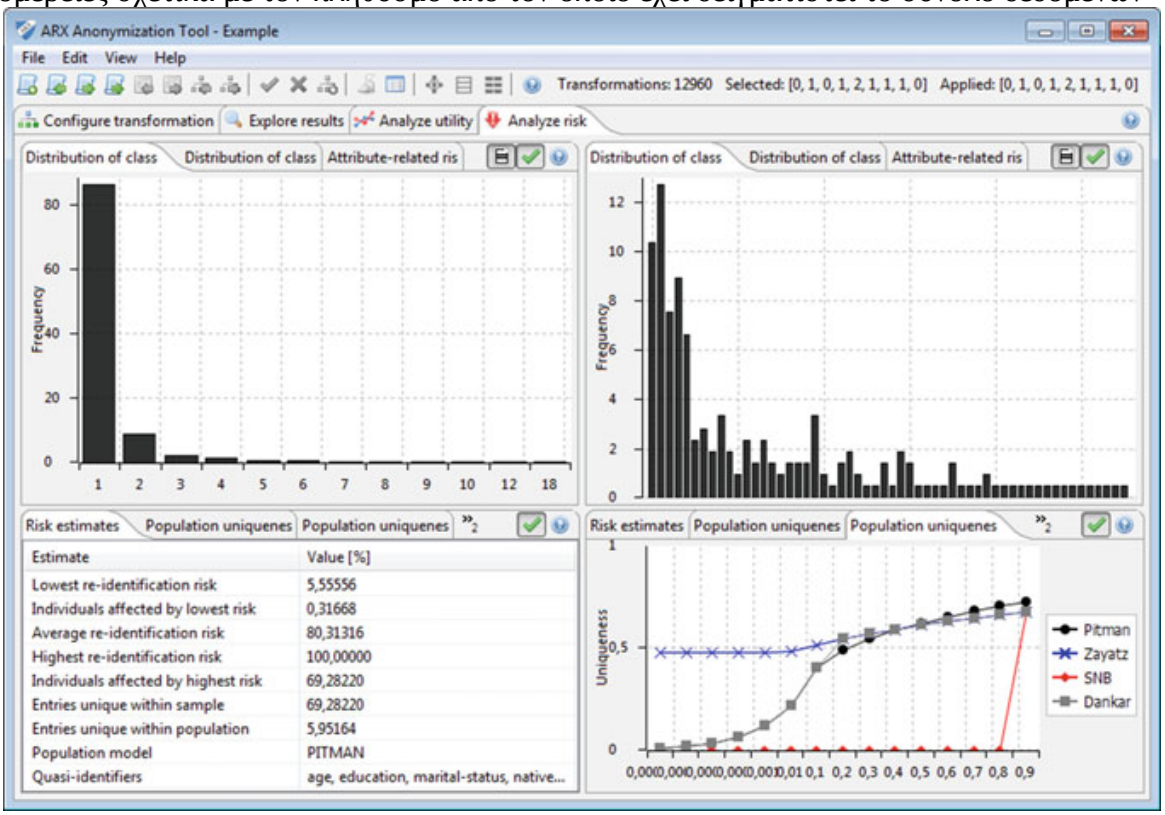
7.5.4.5 Ανάλυση κινδύνων επαναπροσδιορισμού

Με αυτή την προοπτική, οι κίνδυνοι επαναπροσδιορισμού ενός συνόλου δεδομένων εισόδου ή εξόδου, μπορούν να αναλυθούν με διάφορα μέτρα βασισμένα στο δείγμα ή στο πληθυσμό. Επιπλέον, αυτή η προοπτική παρέχει μια σύνοψη της κατανομής των μεγεθών των κλάσεων ισοδυναμίας στο σύνολο δεδομένων και υποστηρίζει την ανάλυση των κινδύνων που σχετίζονται με ομάδες γνωρισμάτων για την εύρεση ψευδοαναγνωριστικών.

Στο εικόνα 7, εμφανίζεται η κατανομή των μεγεθών τάξης τόσο είσοδου όσο και έξοδου. Η άποψη στην κάτω δεξιά γωνία δείχνει μια σύγκριση της εκτιμώμενης μοναδικότητας του πληθυσμού από τρία μοντέλα κινδύνου για διαφορετικά κλάσματα δειγμάτων. Παρουσιάζει επίσης τα αποτελέσματα του κανόνα απόφασης που προτείνεται και επικυρώνεται για τα κλινικά σύνολα δεδομένων από τους Dankar et al., τα οποία επιλέγουν ένα διαφορετικό μοντέλο κινδύνου για τα διαφορετικά κλάσματα του δείγματος [20]

Κατα τον υπολογισμό εκτιμήσεων κινδύνου, το ARX πρέπει να επιλύει αριθμητική συστήματα εξισώσεων. Ο υπολογιστής που χρησιμοποιείται από το ARX μπορεί να ρυθμιστεί στο πλαίσιο των ρυθμίσεων. Εδώ μπορείτε να ορίσετε επιλογές όπως ο συνολικός αριθμός επαναλήψεων, ο μέγιστος αριθμός επαναλήψεων ανά δοκιμή και η απαιτούμενη ακρίβεια. Αυτό μπορεί να επηρεάσει την ακρίβεια των αποτελεσμάτων και των χρόνων εκτέλεσης. Επιπλέον, στην κάτω αριστερή γωνία, η προβολή εμφανίζει βασικές εκτιμήσεις κινδύνου που υπολογίζονται είτε από το ίδιο το δείγμα είτε από ένα μοντέλο κινδύνου.

Αυτή η ανάλυση χρησιμοποιεί εκτιμήσεις της μοναδικότητας είτε βασισμένη σε ένα δείγμα είτε σε ένα στατιστικό μοντέλο. Όταν ένα σύνολο γνωρισμάτων έχει επιλεγεί για ανάλυση, το ARX θα καθορίσει τον μέσο όρο κινδύνου επαναπροσδιορισμού που σχετίζεται με το σύνολο ισχύος αυτών των γνωρισμάτων. Αυτή η πληροφορία μπορεί να φανεί χρήσιμη στην απόφαση του ποια γνωρίσματα πρέπει να γενικευτούν. Επιπλέον, μπορούν να ρυθμιστούν σε ξεχωριστό πλαίσιο οι λεπτομέρειες σχετικά με τον πληθυσμό από τον οποίο έχει δειγματιστεί το σύνολο δεδομένων



Εικόνα 7: Η ανάλυση κινδύνου του ARX

7.6 Τελικές παρατηρήσεις

Μέσα απο αυτή την συνοπτική παρουσίαση των εργαλείων ανωνυμοποίησης, κρίνοντας όμως κυρίως απο το ARX που δοκιμάσαμε, παρατηρούμε ότι η πρόοδος που υπήρξε τα τελευταία χρόνια πάνω σε ένα όχι τόσο γνωστό θέμα στο ευρύ κοινό - αυτό της ανωνυμοποίησης – ειναι αξιοσημείωτη. Τόσο απο πλευράς πληρότητας όσο ποικιλίας, δεδομένου ότι η ανωνυμοποίηση των δεδομένων παραμένει ένα πολύπλοκο ζήτημα που θα πρέπει να εξετάζεται αλλά και εκτελείται απο ειδικούς. Δυστυχώς δεν υπάρχει ενιαίο μέτρο που να είναι ικανό να προστατέψει τις βάσεις δεδομένων απο όλους τους πιθανούς κινδύνους, ειδικά αν ειναι αρκετά ευέλικτο για να υποστηρίξει πολλά σενάρια χρήσης.

Σαν εργαλείο το ARX θεωρείται πλήρες και με συνεχόμενη εξέλιξη, καθότι έχει ενεργή ομάδα ανάπτυξης και ειναι ανοιχτού κώδικα. Με κάθε νέα έκδοση γίνονται νέες προσθήκες για να υποστηρίχτουν νέες λειτουργίες.

Κεφάλαιο 8. Επίλογος

Με την παρούσα διπλωματική εργασία έγινε μια προσπάθεια παρουσίασης του προβλήματος της εφαρμογής ανωνυμίας – απο την απλή ομαδοποίηση μέχρι τις πιο σύνθετες γενικεύσεις - σαν μια οργανική συνολική κατάσταση εξέλιξης και βελτίωσης του τρόπου αντιμετώπισης του μέσα απο την σκοπιά των νέων κανόνων, αλγόριθμων και εφαρμογών που την απαρτίζουν.

Δώσαμε μεγαλύτερη βάση στην ανάλυση της k-ανωνυμίας, τις αδυναμίες της, τις επιθέσεις που τις εκμεταλλεύονται και τους τρόπους επίλυσης μέσω νέων αλγορίθμων ανωνυμοποίησης. Βάση αυτού αναλύουμε περισσότερο τους δύο πιο δυνατούς και καταλήγουμε στην παρουσίαση έτοιμων εφαρμογών, ακαδημαικού χαρακτήρα, που μπορούν να τους εφαρμόσουν. Επικεντρωνόμαστε στο ARX, και αναλύουμε τον τρόπο ανάλυσης του προβλήματος της ανωνυμίας μέσω των ανάλυσης που γίνεται απο την φύση του κώδικα του.

Τέλος, να αναφέρουμε, ότι το συνεχόμενο ενδιαφέρον των ερευνητικών ομάδων γι αυτή – ανωνυμία- συντελεί στην εξέλιξη της και στους νέους τρόπους εφαρμογής της. Εκμεταλλευόμενη την εποχής της συνειδητοποίησης του τρόπου λειτουργίας του διαδικτύου, χρησιμοποίειται περισσότερο και απο περισσότερους.

Δείγμα εφαρμογής της μπορούμε να δούμε στις υλοποιήσεις που γίνονται, εκτός των εργαλείων που προαναφέραμε. Τέτοιο παράδειγμα αποτελλεί το https://haveibeenpwned.com/ όπου ο δημιουργός του συγκερκιμένου εφαρμόζει k-ανωνυμία, των queries προς την βάση σε σύγκριση με της εισόδου του επισκεπτόμενου.

Κύριο κίνητρο για την συγγραφή της ήταν η παρουσίαση αυτής και του τρόπου που μπορεί να γίνει επιτεύξιμη. Ενώ παραμένει η ελπίδα χρήσης της όλο περισσότερο απο τις Ελληνικές Αρχές, Υπηρεσίες, Οργανισμούς και Εταιρείες στο πλαίσιο ικανοποίησης του GDPR.

Βιβλιογραφία

1] Raymond Choo, Man Ho Au. Mobile Security and Privacy

2] Aris Gkoulalas-Divanis, Grigorios Loukides. Medical Data Privacy Handbook && Li, T., Li, N., Zhang, J., Molloy, I.: Slicing: a new approach for privacy preserving data publishing.

3] Sweeney L. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzz. Knowl. Based Syst. 2002;10(5):557–570.

4] Iwuchukwu T., Naughton J. K-anonymization as spatial indexing: toward scalable and incremental anonymization. In: Koch C., ed. Proceedings of the 33rd International Conference on Very Large Data Bases, September 23–27, 2007, University of Vienna, Austria; ACM; 2007:746–757.

5] Ghinita G., Karras P., Kalnis P., Mamoulis N. *Fast data anonymization with low information loss.* In: Koch C., ed. Proceedings of the 33rd International Conference on Very Large Data Bases, 23-27 September 2007, University of Vienna, Austria; ACM; 2007:758–769.

6] LeFevre K., DeWitt D., Ramakrishnan R. Workload-aware anonymization techniques for large-scale datasets. *ACM Trans. Database Syst.* 2008;33(3):1–47.

7]LeFevre K., DeWitt D.J., Ramakrishnan R. *Incognito: efficient full-domain k-anonymity.* In: Özcan F., ed. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 14–16, 2005, Baltimore, MD; ACM; 2005:49–60.

8]El Emam K., Dankar F.K., Issa R., Jonker E., Amyot D., Cogo E., Corriveau J.P., Walker M., Chowdhury S., Vaillancourt R., Roffey T., Bottomley J. Research paper: a globally optimal k-anonymity method for the de-identification of health data. *JAMIA.* 2009;16(5):670–682.

9] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, Ada Wai-Chee Fu. Utility-based anonymization using local recoding

10] A. Gionis and T. Tassa, “k-Anonymization with minimal loss of information,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 2, 2009, pp. 206–219.

11] LeFevre K., DeWitt D., Ramakrishnan. Mondrian Multidimensional K-Anonymity

12] Tang, Qingming, et al. "Improving Strict Partition for Privacy Preserving Data Publishing." Networking and Distributed Computing (ICNDC), 2010 First International Conference on. IEEE, 2010.

13]Divya Sadhwani, Dr. Sanjay Silakari, Mr. Uday Chourasia. Preserving Privacy during Big Data Publishing using K-Anonymity Model – A Survey

14] Janosch Maier . Anonymity: Formalization of Privacy – k-anonymity
Wong W.K., Mamoulis N., Cheung D.W.L. *Non-homogeneous generalization in privacy preserving data publishing.* In: Elmagarmid A., Agrawal D., eds. Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2010), June 6–10, Indianapolis, IN; ACM; 2010:747–758.

15]Charu C. Aggarwal, Philip S. Yu . Privacy-Preserving Data Mining

16]Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam . ℓ -Diversity: Privacy Beyond k -Anonymity

17]Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian (2007). "t-Closeness: Privacy beyond k-anonymity and l-diversity

18]Yu Zong, Guandong Xu . Applied Data Mining by Zhenglu Yang

19] Iyengar, V.: Transforming data to satisfy privacy constraints. In: Proceedings of International Conference on Knowledge Discovery and Data Mining, pp. 279–288 (2002)

20] Dankar, F., Emam, K.E., Neisa, A., Roffey, T.: Estimating the re-identification risk of clinical data sets. BMC Med. Inform. Decis. Mak. 12(1), 66 (2012)

21] Fabian Prasser, Florian Kohlmayer, Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool

Aggarwal G., Panigrahy R., Feder T., Thomas D., Kenthapadi K., Khuller S., Zhu A. Achieving anonymity via clustering. *ACM Trans. Algor.* 2010;6(3):1–19.

Li N., Qardaji W.H., Su D. Provably private data anonymization: or, k-anonymity meets differential privacy. *CoRR.* 2011 Abs/1101.2604.

A Framework for Efficient Data Anonymization under Privacy and Accuracy Constraints
GABRIEL GHINITA, PANAGIOTIS KARRAS, and PANOS KALNIS National University of Singapore and NIKOS MAMOULIS University of Hong Kong, p9 (2009)

Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Samarati, P.: Microdata protection. In: Yu, T., Jajodia, S. (eds.) Secure Data Management in Decentralized Systems. Advances in Information Security, vol. 33, pp. 291–321. Springer, Berlin (2007) – μικροσυσσωμάτωση

Nergiz, M.E., Atzori, M., Clifton, C.: Hiding the presence of individuals from shared data-bases. In: Proceedings of International Conference on Management of Data,(2007)