



ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

Μεταπτυχιακή εξειδίκευση στα Πληροφοριακά Συστήματα

# Μελέτη και Αξιολόγηση Τεχνικών Ιδιωτικότητας στην Ανάλυση Δεδομένων

Γεράσιμος Βαρδακαστάνης

Επιβλέπων Α': Δημήτριος Καραπιπέρης

Επιβλέπων Β': Βασίλειος Ζορκάδης

26 Σεπτεμβρίου 2018





©ΕΑΠ, 2018

Η παρούσα διατριβή, η οποία εκπονήθηκε στα πλαίσια της ΘΕ ΠΛΣΔΕ, και τα λοιπά αποτελέσματα της αντίστοιχης Διπλωματικής Εργασίας (ΔΕ) αποτελούν συνιδιοκτησία του ΕΑΠ και του φοιτητή, ο καθένας από τους οποίους έχει το δικαίωμα ανεξάρτητης χρήσης και αναπαραγωγής (στο σύνολο ή τμηματικά) για διδακτικούς και ερευνητικούς σκοπούς, σε κάθε περίπτωση αναφέροντας τον τίτλο και το συγγραφέα και το ΕΑΠ όπου εκπονήθηκε η ΔΕ καθώς και τον επιβλέποντα και την επιτροπή κρίσης.

---

## Ευχαριστίες

Η παρούσα διπλωματική εργασία συντάχθηκε στα πλαίσια του μεταπτυχιακού προγράμματος σπουδών του τμήματος «Μεταπτυχιακή Εξειδίκευση στα Πληροφοριακά Συστήματα» του Ελληνικού Ανοικτού Πανεπιστημίου.

Θα ήθελα να εκφράσω τις βαθύτατες ευχαριστίες μου στον επιβλέποντα καθηγητή κ. Δημήτριο Καραπιπέρη, για τη συνεργασία, την καθοδήγηση, τις χρήσιμες συμβουλές και την ενθάρρυνσή του κατά τη διάρκεια της συγγραφής της εργασίας αυτής.

Στη συνέχεια θα ήθελα να ευχαριστήσω τους καθηγητές μου, κ. Βασίλειο Βερύκιο και κ. Ιωάννη Μοσχολιό γιατί με βοήθησαν να αποκτήσω τις απαραίτητες γνώσεις προκειμένου να ολοκληρώσω τις σπουδές μου, αλλά και να διευρύνω τους πνευματικούς μου ορίζοντες.

Τέλος ευχαριστώ τους γονείς μου Μαρία και Διονύση για την υποστήριξη τους όλα αυτά τα χρόνια και τη σύντροφό μου Μαρία που με υπομονή και κατανόηση πρόσφερε την απαραίτητη ηθική συμπαράσταση για την ολοκλήρωση της μεταπτυχιακής μου εργασίας.



## Περίληψη

Ζούμε στην ψηφιακή εποχή, μια εποχή που όλο και περισσότερα δεδομένα συλλέγονται, αποθηκεύονται και επεξεργάζονται. Το ζήτημα της ανάλυσης μεγάλων ποσοτήτων δεδομένων, διατηρώντας παράλληλα την ιδιωτικότητα, είναι ένα από τα πλέον επίκαιρα θέματα του παγκόσμιου κοινωνικού διαλόγου, απασχολώντας μια ευρεία γκάμα επιστημόνων. Κατά τη διάρκεια της σύντομης ψηφιακής ιστορίας, έγιναν πολλές αποτυχημένες προσπάθειες, δείχνοντας ότι η συλλογιστική για την προστασία της ιδιωτικότητας των δεδομένων είναι γεμάτη παγίδες. Αυτό προκάλεσε αυξημένο ενδιαφέρον για έναν μαθηματικά αξιόπιστο ορισμό της ιδιωτικότητας.

Η παρούσα εργασία ασχολείται με την διασφάλιση της ιδιωτικότητας σε συλλογές προσωπικών δεδομένων. Αρχικά αναλύουμε τις σύγχρονες μεθόδους γενίκευσης, συγκρίνουμε τις επιδόσεις τους και τονίζουμε τις αδυναμίες τους, υπογραμμίζοντας ότι είναι αδύνατη η απόλυτη πρόληψη αποκαλύψεων. Στη συνέχεια παρουσιάζουμε την κυριότερη μέθοδο τυχαιοποίησης, την Διαφορική Ιδιωτικότητα, η οποία αντιμετωπίζει όλες τις επί του παρόντος γνωστές επιθέσεις, έχει πολλές πρακτικές υλοποιήσεις και γνωρίζει πολλές επεκτάσεις που την καθιστούν εφαρμόσιμη σε ευρύ φάσμα καταστάσεων.

Στο τελευταίο κομμάτι της εργασίας, αναπτύσσουμε αλγόριθμους τυχαιοποίησης της Διαφορικής Ιδιωτικότητας σε γλώσσα προγραμματισμού Python, και αναλύουμε τις επιδόσεις τους.





---

## Abstract

We live in the digital age, a time when more and more data is collected, stored and processed. The question of analyzing large amounts of data, while preserving privacy, is one of the most up-to-date issues of the global social dialogue, employing a wide range of scientists. In the course of history, many failed attempts have been made, indicating that the reasoning behind data privacy is full of traps. This has prompted increased interest in a mathematically robust definition of privacy.

This paper is concerned with ensuring privacy in personal data sets. Initially, we analyze modern generalization methods, compare their performance and emphasize their weaknesses, pointing that absolute disclosure prevention is impossible. Afterwards, we present the main method of randomization, Differential Privacy, which addresses all the currently known attacks, it has many practical implementations and knows many extensions that make it applicable to a wide range of situations.

In the last part of the thesis, we develop randomization algorithms of Differential Privacy in Python programming language and we analyze their performance.



# Περιεχόμενα

Ευχαριστίες

Περίληψη

Λίστα Σχημάτων

<b>1</b>	<b>Εισαγωγή</b>	<b>1</b>
1.1	Γενικά	1
1.1.1	Ανάγκη για προστασία των δεδομένων	1
1.1.2	Παροχή ασφάλειας	3
1.1.3	Ροή πληροφοριών	4
1.1.4	Διατήρηση ποιότητας	5
1.2	Σκοπός	5
1.3	Δομή Εργασίας	6
1.4	Related Work	7
1.4.1	Blockchain	7
1.4.2	Cloud Privacy	7
<b>2</b>	<b>Ανεπάρκεια Προστασίας των Δεδομένων</b>	<b>9</b>
2.1	Παραδείγματα αποκάλυψης πληροφορίας	9
2.2	Ανεπαρκείς προσεγγίσεις προστασίας	13
<b>3</b>	<b>Μηχανισμοί γενίκευσης και ομαδοποίησης</b>	<b>17</b>
3.1	Θεμελίωση	17
3.2	k-ανωνυμία	18
3.2.1	Επίτευξη k-anonymity	21
3.2.2	Επιθέσεις κατά της k-ανωνυμίας	21
3.3	l-Διαφορετικότητα	23
3.3.1	Επίτευξη l-διαφορετικότητας	24
3.3.2	Επιθέσεις κατά της l-διαφορετικότητας	24
3.4	t-Εγγύτητα	26
3.5	Επίθεση Τομής	26
3.6	Αλγόριθμοι Γενίκευσης	29
3.6.1	Mondrian	29
3.6.2	Συσταδοποίηση (Clustering)	30
3.6.3	k - OPTIMIZE	30
<b>4</b>	<b>Διαφορική Ιδιωτικότητα</b>	<b>33</b>

4.1	Η έννοια της διαδραστικότητας . . . . .	33
4.2	Θεμελίωση . . . . .	34
4.2.1	Θεώρημα Σύνθεσης . . . . .	37
4.3	Προσθήκη θορύβου και Μηχανισμοί . . . . .	39
4.3.1	Μηχανισμός Laplace . . . . .	39
4.3.2	Εκθετικός Μηχανισμός . . . . .	41
4.3.3	Μηχανισμός Gauss . . . . .	42
4.4	Κατασκευή σύνθετων Μηχανισμών . . . . .	43
<b>5</b>	<b>Εφαρμογή Αλγορίθμων Διαφορικής Ιδιωτικότητας</b>	<b>45</b>
5.1	Υλοποίηση . . . . .	45
5.2	Εφαρμογή και αξιολόγηση . . . . .	48
5.2.1	Δημοφιλή ονόματα . . . . .	48
5.2.2	Μέση τιμή . . . . .	50
5.2.3	Πλήθος Τρίτεχνων . . . . .	52
<b>6</b>	<b>Συμπεράσματα - Μελλοντική Έρευνα</b>	<b>55</b>
6.1	Σύνοψη και συμπεράσματα . . . . .	55
6.2	Νέες τεχνικές - συνδυασμοί . . . . .	56
	 <b>Παράρτημα Κώδικα</b>	 <b>58</b>

# Κατάλογος Σχημάτων

1.1	Στάδια ελέγχου ροής πληροφοριών . . . . .	4
2.1	Χρήστες AOL από το 2002 . . . . .	10
2.2	Συνδιασμός κοινών γνωρισμάτων . . . . .	11
2.3	Πληθος χρηστών Netflix, προς πλήθος βαθμολογιών . . . . .	12
3.1	Ένα 4-ανώνυμο σύνολο δεδομένων . . . . .	20
3.2	Ένα 3-διαφορετικό σύνολο δεδομένων . . . . .	23
3.3	Μοντελοποίηση του αλγορίθμου k-optimize . . . . .	31
4.1	Η έννοια της διαδραστικότητας . . . . .	33
4.2	Λειτουργία μηχανισμού ΔΙ . . . . .	37
4.3	Συνάρτηση πυκνότητας πιθανότητας κατανομής Laplace, με παράμετρο κλίμα- κας $b = 1$ . . . . .	40
5.1	Πίνακας αναπληρωτών εκπαιδευτικών ΠΕ19, μετά από ανωνυμοποίηση. . . . .	47
5.2	Δημοφιλέστερα ονόματα - χωρίς θόρυβο (πάνω) και με θόρυβο ( $\epsilon=1$ ) (κάτω) . . . . .	49
5.3	Ιστόγραμμα συχνοτήτων των βαθμών . . . . .	51
5.4	Αποκάλυψη τιμής ευαίσθητου γνωρίσματος . . . . .	52
6.1	Πλεονεκτήματα και μειονεκτήματα τεχνικών ανωνυμοποίησης . . . . .	56



# Κεφάλαιο 1

## Εισαγωγή

### 1.1 Γενικά

Στην σύγχρονη κοινωνία παρουσιάζεται εκθετική αύξηση της συλλογής δεδομένων, καθώς η υπολογιστική τεχνολογία, η δικτυακή συνδεσιμότητα και ο ψηφιακός χώρος αποθήκευσης γίνονται συνεχώς πιο προσιτά. Κάθε δευτερόλεπτο, εκατομμύρια καταχωρήσεις από οργανισμούς, εταιρίες και μέσα κοινωνικής δικτύωσης, έχουν οδηγήσει σε εκτίναξη την παραγωγή δεδομένων σε 3 πεντάκις εκατομμύρια bytes ημερησίως. Χαρακτηριστικό είναι ότι μόνο η google στεγάζει περισσότερα από 10 exabytes δεδομένων.

Για κάθε άνθρωπο, από την στιγμή της γέννησής του, υπάρχει τουλάχιστον μια καταχώρηση σε μια συλλογή δεδομένων. Σε διάφορους τομείς, από ιδιωτικούς και κοινωνικούς φορείς όπως τράπεζες, νοσοκομεία, τηλεφωνικούς παρόχους, ακόμα και επιχειρήσεις, συλλέγονται και αποθηκεύονται αναλυτικά προσωπικές πληροφορίες των πελατών τους. Είναι γεγονός ότι όσο περισσότερες πληροφορίες υπάρχουν για κάποιον στις βάσεις δεδομένων εταιριών και οργανισμών, τόσο καλύτερες υπηρεσίες παρέχονται στο άτομο αυτό, αλλά και αντίστροφα, επωφελούνται οι αναλυτές των βάσεων δουλεύοντας με καλύτερο/μεγαλύτερο δείγμα: Στατιστικά δεδομένα υγείας για ένα άτομο που συλλέγονται είτε από τους άμεσα εμπλεκόμενους φορείς, είτε έμμεσα από μικροσυσσκευές (smartwatches, smartphones, κλπ) μπορούν να βοηθήσουν στην πρόληψη μιας σοβαρής ασθένειας, μικροδεδομένα μετακινήσης από τους κατοίκους μιας πόλης συμβάλουν στην βελτίωση της κυκλοφορίας, ενώ ανάλυση σε βάσεις δεδομένων εγκληματιών οδηγεί σε πιθανή εξιχνίαση κάποιας υπόθεσης.

#### 1.1.1 Ανάγκη για προστασία των δεδομένων

Η αποκάλυψη προσωπικών στοιχείων και προτιμήσεων είναι, θα τολμούσαμε να πούμε, αναγκαία για την ομαλότερη και αποτελεσματικότερη λειτουργία εφαρμογών και υπηρεσιών. Η συνεχής αύξηση όμως της συλλογής δεδομένων οδηγεί σε υπερβολική συγκέντρωση ατομικών πληροφοριών. Χαρακτηριστικό είναι ότι περισσότερες από 15.000 ειδικευμένες βάσεις

περιέχουν δυο δισεκατομμύρια ονόματα καταναλωτών μαζί με μεγάλο όγκο προσωπικών πληροφοριών, ενώ ο μέσος καταναλωτής βρίσκεται καταχωρημένος τουλάχιστον σε 25 διαφημιστικές λίστες. Πολλές από αυτές τις λίστες, αφού οργανωθούν σύμφωνα με γνωρίσματα όπως εισόδημα, ηλικία, πολιτικές και θρησκευτικές απόψεις, αγοράζονται και πωλούνται καθημερινά. Η κατάχρηση συνεπώς των βάσεων δεδομένων είναι αναπόφευκτη.

Οι κάτοχοι των συνόλων δεδομένων συνήθως δίνουν εγγύηση ότι οι καταχωρήσεις τους είναι ασφαλείς, όμως πολλές είναι οι περιπτώσεις που χρησιμοποιήθηκαν βάσεις δεδομένων για παραβίαση της ιδιωτικότητας κάποιου ατόμου. Αξίζει να αναφέρουμε τους λόγους:

- Εξαπάτηση
- Στρατηγικές ενέργειες
- Σφάλματα

Προκύπτει λοιπόν η ανάγκη για αυστηρότερη προστασία της ιδιωτικότητας των πληροφοριών.

Εντυπωσιακό είναι το αποτέλεσμα μιας πρόσφατης δημοσκόπησης (YouGov, 2017) που αναφέρει ότι το 1/5 των κατοίκων του Ηνωμένου Βασιλείου ανησυχούν για την πώληση των προσωπικών τους δεδομένων σε άλλες εταιρίες, ενώ το 12% ανησυχεί ότι τα ήδη καταχωρημένα δεδομένα τους μπορεί να υποκλαπούν. Αν αναλογιστούμε ότι τα ποσοστά αυτά αντιστοιχούν σε εκατομμύρια πολιτών, συμπεραίνουμε ότι η απαίτηση για ασφάλεια στις συλλογές δεδομένων, τόσο από την μεριά των διαχειριστών όσο και των υποκειμένων, είναι ολοένα και μεγαλύτερη.

Αυτό, δεν αφήνει ασυγκίνητες ούτε τις κυβερνήσεις ανά τον κόσμο, τοποθετώντας την προστασία των δεδομένων ψηλά στην ατζέντα τους. Η Ευρωπαϊκή Ένωση ψήφισε πρόσφατα το GDPR - General Data Protection Regulation -με ημερομηνία εφαρμογής τον Μάη του 2018- έναν κανονισμό πρώτον για την προστασία των φυσικών προσώπων έναντι της επεξεργασίας των προσωπικών δεδομένων τους και δεύτερον για την ελεύθερη διακίνηση αυτών.

Η προστασία της ιδιωτικότητας είναι αδιαμφισβήτητα μια ανάγκη της εποχής, καθώς η συλλογή δεδομένων αυξάνεται με γεωμετρική πρόοδο. Το μεγάλο ζήτημα είναι η παροχή των απαραίτητων πληροφοριών και η ορθή χρήση τους απο τρίτους, αποφεύγοντας ταυτόχρονα οποιαδήποτε παραβίαση. Με το πέρασμα των χρόνων έχουν διατυπωθεί ποικίλες μέθοδοι και τεχνικές προστασίας της ιδιωτικότητας των ευαίσθητων δεδομένων με σκοπό την διατήρηση της ισορροπίας ανάμεσα στην εκμετάλλευση των προσωπικών δεδομένων και τον σεβασμό προς τα άτομα.



### 1.1.2 Παροχή ασφάλειας

Ο έλεγχος πρόσβασης (access control) στην εκάστοτε βάση, χρησιμοποιώντας κλασικές μεθόδους κρυπτογράφησης, αποτελεί την καθολική τεχνική προστασίας των δεδομένων. Αυτό δυστυχώς δεν εφαρμόζεται πάντα, αφού συνήθως τα σύνολα δεδομένων προς ανάλυση είναι δημοσίως διαθέσιμα. Η μέθοδος αυτή είναι χρήσιμη μόνο αν εφαρμοστεί ορθόδοξα, ώστε να αποτρέπει την πρόσβαση στα δεδομένα σε οποιονδήποτε επιτιθέμενο αλλά εμφανίζοντας όλη τη βάση στον καλόβουλο αναλυτή. Η μελέτη της μεθόδου αυτής ωστόσο ξεφεύγει από τα όρια της εργασίας μας.

Η παραδοσιακή προσέγγιση του προβλήματος της προστασίας των εγγραφών είναι η «ανωνυμοποίηση» του συνόλου δεδομένων αποκρύπτοντας ή κρυπτογραφώντας χαρακτηριστικά τα οποία θα μπορούσαν να αποκαλύψουν την ταυτότητα ενός ατόμου, όπως όνομα, διεύθυνση, ημερομηνία γέννησης. Ως αποτέλεσμα, προβάλλεται μια «εξυγιασμένη» (sanitized) βάση δεδομένων. Παρ' όλ' αυτά, αποδεικνύεται ότι η απλή απόκρυψη πληροφορίας δεν είναι αρκετή, διότι σε συνδιασμό με κατάλληλη βοηθητική γνώση είναι δυνατόν να αποκαλυφθεί η ταυτότητα ενός υποκειμένου.

Κατά καιρούς, έχουν γίνει προσπάθειες για την μεγιστοποίηση της ασφάλειας βελτιώνοντας τις τεχνικές ανωνυμοποίησης. Οι περισσότερες τεχνικές χρησιμοποιούν καποιά μορφή μετασχηματισμού των δεδομένων, με στόχο την γενίκευση/ομαδοποίηση των πληροφοριών ώστε να διασφαλιστεί η ιδιωτικότητα. Παρακάτω θα αναφέρουμε τις κυριότερες από αυτές, τις οποίες θα παρουσιάσουμε αναλυτικά στα επόμενα κεφάλαια.

- **k-anonymity:** Το μοντέλο αυτό μειώνει την διακριτικότητα των πληροφοριών με χρήση της τεχνικής της γενίκευσης, εξασφαλίζοντας ότι ο επιτιθέμενος δεν καταφέρνει να προσδιορίσει μοναδικά κάποια εγγραφή σε μια βάση δεδομένων αφού θα υπάρχουν τουλάχιστον  $k - 1$  εγγραφές με ίδια ακριβώς χαρακτηριστικά.
- **l-diversity, t-closeness:** Τα μοντέλα αυτά σχεδιάστηκαν για να διαχειριστούν τις αδυναμίες της μεθόδου της k-anonymity, η οποία είναι ανεπαρκής στο να προστατέψει από την αποκάλυψη τιμών γνωρισμάτων των εγγραφών. Η l-diversity το αποτρέπει αυτο ομαδοποιώντας τις καταχωρήσεις σε κλάσεις με  $l$  τουλάχιστον διαφορετικές τιμές απο κάθε χαρακτηριστικό. Η t-closeness πάει ένα βήμα παραπέρα, ορίζοντας συγκεκριμένο κατώφλι για την διαφορά της κατανομής ενός χαρακτηριστικού σε μια κλάση απο την κατανομή σε όλη την βάση.

Ένας άλλος τρόπος προσέγγισης του προβλήματος είναι η τυχαιοποίηση, στην οποία δεν δημιουργείται μια νέα εξυγιασμένη βάση όπως στις παραπάνω μεθόδους, αλλά παρέχεται ιδιωτικότητα μέσω της προσθήκης θορύβου στα δεδομένα. Ο κύριος εκφραστής της διαδραστικής μεθόδου είναι η Διαφορική Ιδιωτικότητα.

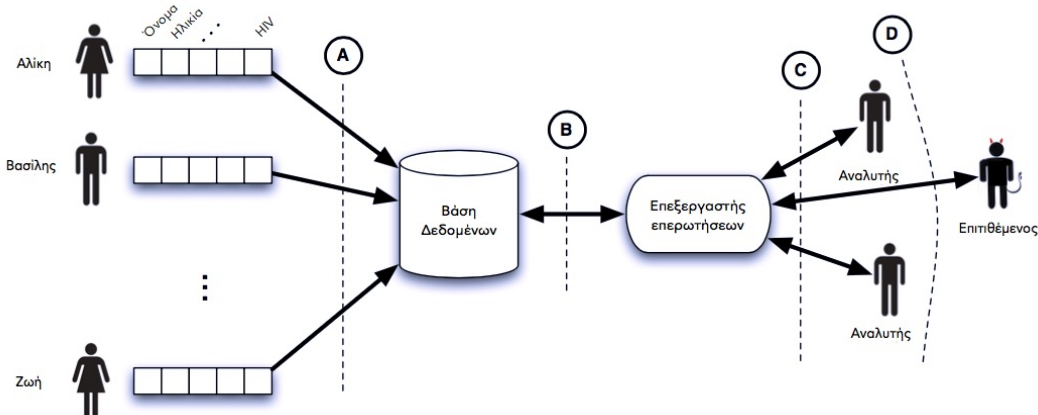
Η τεχνική αυτή δεν στηρίζεται στον διαχωρισμό της βάσης δεδομένων σε υποσύνολα όπως οι προαναφερθείσες, αλλά έχει ως σκοπό τον περιορισμό της βαρύτητας των δεδομένων στα αποτελέσματα που εξορύσσονται από την βάση.

Όπως είδαμε, η ιδιωτικότητα σε μια βάση μπορεί να επιτευχθεί με διάφορους τρόπους. Παρακάτω ταξινομούμε τα είδη προστασίας που μπορούν να εφαρμοστούν σε διάφορα επίπεδα της ροής πληροφοριών κατά την ανάλυση δεδομένων.

### 1.1.3 Ροή πληροφοριών

Η παρακάτω προσέγγιση δείχνει την ροή πληροφορίας ξεκινώντας από τις προσωπικές πληροφορίες ενός ατόμου και καταλήγοντας στον αναλυτή. Τα δεδομένα συλλέγονται σε μια βάση που την ελέγχει ένας διαχειριστής. Το επιθυμητό είναι οι αναλυτές να μπορούν να πραγματοποιούν επερωτήσεις και υπολογισμούς σε αυτήν ώστε να συλλέξουν στατιστικά στοιχεία, ενώ ένας επιτιθέμενος να μην μπορεί να εξάγει ευαίσθητα δεδομένα για συγκεκριμένα άτομα.

Για να προστατευτεί η ιδιωτικότητα των ατόμων αυτών πρέπει να υπάρχει έλεγχος της ροής πληροφοριών κατά την διαδρομή τους από το ένα άκρο στο άλλο [Hay, 2010]



ΣΧΗΜΑ 1.1: Στάδια ελέγχου ροής πληροφοριών

A. Διαταραχή εισόδου: Τα ίδια τα άτομα μπορούν να αλλοιώσουν τα δεδομένα που δίνουν στη βάση, προσθέτοντας θόρυβο. Με αυτόν τον τρόπο ο επιτιθέμενος δεν θα είναι ποτέ σίγουρος ότι οι ευαίσθητες πληροφορίες για κάποιο άτομο είναι σωστές, ενώ τα στατιστικά που συγκεντρώνουν οι αναλυτές δεν υφίστανται αξιόλογη μεταβολή.

B. Μετασχηματισμός δεδομένων: Η αρχική βάση δεδομένων αναδομείται σε μια νέα εξυγιασμένη βάση στην οποία οι ευαίσθητες πληροφορίες για συγκεκριμένα άτομα δεν είναι πια διακριτές. Σε αυτό το επίπεδο εφαρμόζεται το μοντέλο της k-anonymity.

C. Διαταραχή απάντησης επερωτήσεων: Οι αναλυτές μπορούν να επεξεργαστούν τα δεδομένα αλλά για να διασφαλιστεί η ιδιωτικότητα έχει προστεθεί θόρυβος. Εδώ εφαρμόζεται το μοντέλο της Διαφορικής Ιδιωτικότητας.

D. Έλεγχος Πρόσβασης: Η βάση δεν είναι δημοσίως διαθέσιμη και μόνο οι έμπιστοι ερευνητές έχουν πρόσβαση. Έτσι αποτρέπεται η διαρροή ευαίσθητων δεδομένων σε κακόβουλα άτομα.

#### 1.1.4 Διατήρηση ποιότητας

Όπως είναι προφανές, η απόκρυψη και η γενίκευση των δεδομένων έχει ως αποτέλεσμα την αλλοίωση της πληροφορίας και την απώλεια της αποτελεσματικότητας. Είναι γεγονός ότι το ποσό ανωνυμίας που έχει μια βάση δεδομένων είναι αντιστρόφως ανάλογο με την ποιότητα:

Η δημοσίευση της βάσης δεδομένων στο ακέραιο παρέχει την καλύτερη ποιότητα, ενώ η πλήρης απόκρυψη την καλύτερη ιδιωτικότητα. [Sweeney, 2001]



Κανένα από τα δυο άκρα δεν είναι επιθυμητό να προσεγγίζεται από ένα σύνολο δεδομένων. Αφενός διότι στο ένα δεν θα παρέχεται καμία ασφάλεια στις εγγραφές και αφετέρου διότι στο άλλο τα δεδομένα θα είναι τόσο διαταραγμένα, που η ανάλυση θα είναι αδύνατη. Στο πλαίσιο της εργασίας αυτής θα συγκρίνουμε στο δυνατόν τις τεχνικές ιδιωτικότητας ως προς την ισορροπία που διατηρούν μεταξύ της προστασίας των εγγραφών και της τελικής ποιότητας των δεδομένων που παρέχονται στον αναλυτή.

## 1.2 Σκοπός

Η παρούσα διπλωματική εργασία έχει τους εξής στόχους:

- Παρουσίαση τεχνικών και μεθοδολογιών που χρησιμοποιούνται για την ενίσχυση της ιδιωτικότητας στις βάσεις δεδομένων.
- Επισκόπηση και κριτική αξιολόγηση από την σκοπιά της ασφάλειας και ιδιωτικότητας των υπό εξέταση προτάσεων/μεθοδολογιών.
- Ανάδειξη κρίσιμων και ανοικτών ζητημάτων, καθώς και κατάδειξη προοπτικών για περαιτέρω έρευνα.

### 1.3 Δομή Εργασίας

Η εργασία είναι δομημένη ως εξής:

Στο δεύτερο κεφάλαιο παρουσιάζονται τα βασικά στοιχεία του προβλήματος. Γίνεται παρουσίαση και ανάλυση της μεθόδου  $k$ -ανωνυμία ( $k$ -anonymity). Δίνονται παραδείγματα και εξετάζονται οι αδυναμίες που την χαρακτηρίζουν. Στη συνέχεια, παρουσιάζονται δυο τεχνικές βελτιστοποίησης της μεθόδου της  $k$ -ανωνυμίας, η  $l$ -διαφορετικότητα ( $l$ -diversity) και η  $t$ -εγγύτητα ( $t$ -closeness). Αναλύεται η δυνατότητα και το κόστος εφαρμογής τους και εξετάζονται τα μειονεκτήματα χρήσης τους.

Στο τρίτο κεφάλαιο γίνεται θεμελίωση και ανάλυση του μοντέλου της Διαφορικής Ιδιωτικότητας. Δίνονται παραδείγματα, αναλύεται η επίδοσή των μηχανισμών και εξετάζεται η χρησιμότητά τους.

Στο τέταρτο κεφάλαιο περιγράφονται αλγόριθμοι ανωνυμοποίησης που χρησιμοποιούν κυρίως τις τεχνικές γενίκευσης από το δεύτερο κεφάλαιο.

Στο πέμπτο κεφάλαιο αναπτύσσονται αλγοριθμικές εφαρμογές που υλοποιούν τα μοντέλα της Διαφορικής Ιδιωτικότητας.

Στο έκτο κεφάλαιο συνοψίζονται τα αποτελέσματα της διπλωματικής εργασίας, αναφέρονται πιθανά ανοιχτά ζητήματα καθώς επίσης και η προοπτική για περαιτέρω έρευνα.

## 1.4 Related Work

### 1.4.1 Blockchain

Η τεχνολογία Blockchain είναι ένα από τα δημοφιλέστερα ζητήματα των τελευταίων ετών και έχει ήδη αρχίσει να αλλάζει τον τρόπο ζωής της σύγχρονης κοινωνίας εξ' αιτίας της επιρροής της σε επιχειρήσεις και βιομηχανίες. Παρόλο που η τεχνολογία αυτή είναι ευρέως γνωστό ότι παρέχει αξιόπιστες υπηρεσίες, τα θέματα ασφάλειας και οι προκλήσεις πίσω από αυτή την καινοτόμο μεθόδο είναι κάτι που δεν πρέπει να μας εφησυχάζει.

Το μοντέλο blockchain δεν εκφράζει μια απλή τεχνική, αλλά έναν συνδυασμό κρυπτογραφικών, μαθηματικών, αλγορίθμικών και οικονομικών μεθόδων, πάνω σε peer-to-peer δίκτυα, που στοχεύουν στην επίλυση παραδοσιακών κατανεμημένων προβλημάτων συγχρονισμού σε βάσεις δεδομένων. Τα βασικά στοιχεία που χαρακτηρίζουν το μοντέλο είναι η αποκέντρωση, λόγω της μη χρήσης κεντρικών εξυπηρετητών, η διαφάνεια, το open source λογισμικό, η αδυναμία μεταβολής των εγγραφών και η ανωνυμία.

Κάθε εγγραφή που δημιουργείται από έναν κόμβο σε ένα δίκτυο του μοντέλου blockchain, αφού ελεγχθεί για την εγκυρότητα της, τοποθετείται σε ένα block μαζί με άλλες εγγραφές. Κάθε block επισφραγίζεται με ένα «Proof of Work», υπολογισμένο από έναν από τους κόμβους του δικτύου μέσω μιας συνάρτησης κατακερματισμού η οποία συνδυάζει στοιχεία του τρέχοντος block αλλά και του προηγούμενου. Στη συνέχεια συνδέεται στην «αλυσίδα». Με αυτόν τον τρόπο καθίσταται αδύνατη η παραβίαση κάποιου block, εκτός αν οι επιτιθέμενοι κατέχουν την πλειοψηφία της υπολογιστικής ισχύς των κόμβων του δικτύου.

Πέραν αυτής της υποθετικής κατάστασης, γνωστή και ως «51% attack», που θα έβαζε σε κίνδυνο την αλυσίδα[Bastiaan, 2015], πρόσφατες έρευνες έχουν αναδείξει αρκετά θέματα ασφάλειας στα blockchain. Ένα από τα σημαντικότερα είναι το Hard/Soft Fork, που περιγράφει την εμφάνιση προβλημάτων από την μη ταυτόχρονη αναβάθμιση λογισμικού των κόμβων.

Δεν υπάρχει αμφιβολία ότι το μοντέλο blockchain είναι ένα κρίσιμο θέμα της εποχής. Όσο το χρησιμοποιούμε και επωφελούμαστε των δυνατοτήτων του, τόσο πρέπει να επιφυλασσόμαστε και να εξετάζουμε πιθανά θέματα ασφάλειας που θα προκύπτουν.

### 1.4.2 Cloud Privacy

Οι υπηρεσίες νέφους έχουν μπει για τα καλά στη ζωή μας, αφού οι απαιτήσεις χώρου, χρόνου και ταχύτητας συνεχώς αυξάνονται. Πολλές εταιρίες Πληροφορικής και οργανισμοί προσφέρουν αλλά και χρησιμοποιούν εφαρμογές και υπηρεσίες νέφους, μια τεχνολογία η οποία κατα κύριο λόγο χρησιμοποιεί υπολογιστές και εξυπηρετητές ενός δικτύου για την μεταφορά, την επεξεργασία και την αποθήκευση των δεδομένων.

Όπως σε κάθε νέα τεχνολογία, έτσι και στο Cloud Computing παρουσιάζονται αρκετά θέματα ασφαλείας, ένα από τα σημαντικότερα είναι η προστασία των δεδομένων. Οι οργανισμοί δεν πρόκειται να μεταφέρουν τα δεδομένα τους σε απομακρυσμένους εξυπηρετητές αν δεν λάβουν εγγύηση για προστασία των δεδομένων από τους παρόχους υπηρεσιών νέφους.

Πολλές τεχνικές προστασίας έχουν διατυπωθεί για την προστασία των δεδομένων, αλλά υπάρχουν ακόμα ανοιχτά ζητήματα. Οι πιο δημοφιλείς μέθοδοι περιλαμβάνουν χρήση SSL κρυπτογράφησης, συστήματα ανίχνευσης εισβολών και έλεγχο πρόσβασης βασισμένο σε Multi Tenancy.

## Κεφάλαιο 2

# Ανεπάρκεια Προστασίας των Δεδομένων

Οι εργασίες για την προστασία των προσωπικών δεδομένων έχουν ξεκινήσει εδώ και αρκετές δεκαετίες. Δυστυχώς, παρά τον σχετικά ορθό σχεδιασμό τους και το μεγάλο μέγεθος των συνόλων δεδομένων που αυτές εφαρμόζονται, η αποκάλυψη και η ταυτοποίηση εγγραφών είναι όπως θα δούμε αναπόφευκτη.

Στο κεφάλαιο αυτό αναφέρουμε αρχικά περιπτώσεις διαρροής πληροφοριών κατά την επεξεργασία συνόλων δεδομένων. Στη συνέχεια, περιγράφουμε τη λειτουργία βασικών τεχνικών προστασίας και τονίζουμε τα σημεία στα οποία αυτές αποτυγχάνουν.

### 2.1 Παραδείγματα αποκάλυψης πληροφορίας

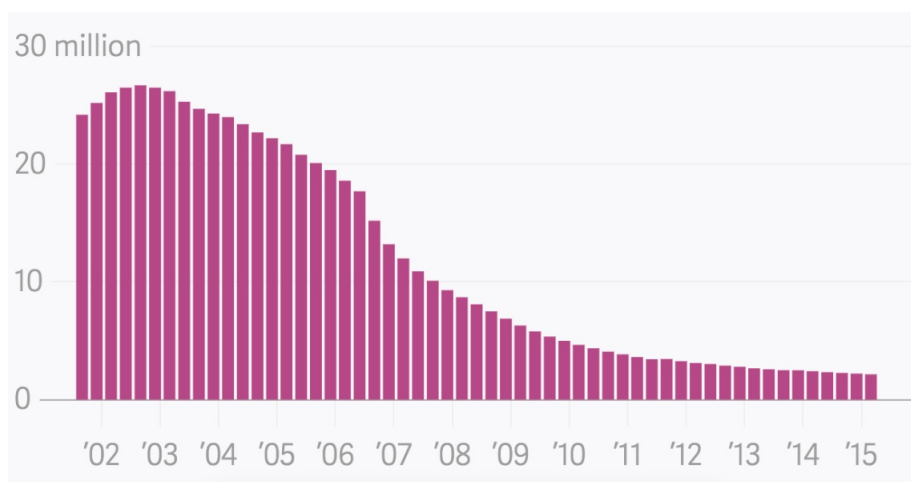
Κατά καιρούς έχουν προταθεί αρκετοί, διαφορετικοί μεταξύ τους, ορισμοί για την έννοια της διαρροής και αποκάλυψης πληροφοριών κατά την επεξεργασία δεδομένων. Η αποκάλυψη σχετίζεται με την απόδοση σημαντικών πληροφοριών σε έναν ερωτώμενο, είτε πρόκειται για ένα άτομο είτε για έναν οργανισμό. Διακρίνουμε τρία είδη αποκάλυψης: όταν το άτομο μπορεί να ταυτοποιηθεί από το σύνολο δεδομένων (αποκαλυψη ταυτότητας), όταν διαρρεύσουν ευαίσθητες πληροφορίες για κάποιο άτομο (αποκάλυψη γνωρίσματος), και όταν τα δημοσιευμένα δεδομένα καθιστούν δυνατό τον ακριβέστερο προσδιορισμό της τιμής κάποιου γνωρίσματος μιας εγγραφής από ό, τι θα ήταν εφικτό (επαγωγική αποκάλυψη).

Παρακάτω κάνουμε μια συλλογή παραδειγμάτων αποκάλυψης και επιθέσεων στα οποία βλέπουμε πως από ένα σύνολο δεδομένων απο το οποίο απουσιάζουν ή αποκρύπτονται προσωπικά στοιχεία, ο επιτιθέμενος καταφέρνει να εξάγει την απαραίτητη πληροφορία ώστε να μπορεί να ταυτοποιηθεί τουλάχιστον ένα άτομο, ή η τιμή ενός ευαίσθητου γνωρίματος μιας εγγραφής, εντός του συνόλου. Τις περισσότερες φορές αυτό γίνεται με τη χρήση βοηθητικών πληροφοριών. Αυτό μπορεί να είναι οποιαδήποτε πρόσθετη πληροφορία που έχει

πρόσβαση ο επιτιθέμενος, συμπεριλαμβανομένων παλαιότερων εκδόσεων της αρχικής βάσης δεδομένων.

- **Η διαρροή δεδομένων αναζήτησης της AOL**

Τον Αύγουστο του 2006, η AOL Research δημοσιεύει σε μια από τις ιστοσελίδες της έναν συμπιεσμένο φάκελο ο οποίος περιείχε 20 εκατομμύρια επερωτήσεις από 650.000 χρήστες. Σε μια προσπάθεια διατήρησης της ιδιωτικότητας, αντικατέστησαν τα ονόματα χρηστών με έναν αριθμό που δημιουργήθηκε τυχαία, ενώ επίσης επεξεργάστηκαν προσεκτικά αποκαλυπτικές επερωτήσεις (παράδειγμα, όταν τα άτομα αναζητούν το δικό τους όνομα ή αριθμό ασφαλείας). Ωστόσο, μπορούσε να εξαχθεί το πλήρες ιστορικό αναζήτησης ενός ατόμου. Αυτό με τη σειρά του, επέτρεπε σε οποιονδήποτε να εντοπίζει τα άτομα με βάση το ιστορικό αναζήτησης και, συνεπώς, να παραβιάσει την ιδιωτικότητά τους. Συγκεκριμένα, ερευνητές της New York Times κατάφεραν να εντοπίσουν ένα άτομο από τα δημοσιευμένα και ανώνυμα αρχεία αναζήτησης, διασταυρώνοντάς τα με λίστες τηλεφωνικών καταλόγων. Η AOL αναγνώρισε το λάθος της και αφαίρεσε τα δεδομένα την αμέσως επόμενη ημέρα. Ωστόσο, η ζημιά είχε ήδη γίνει. Τα δεδομένα αναδιανεμήθηκαν από άλλους και μπορούν ακόμη και σήμερα να μεταφορτωθούν από mirror sites.



ΣΧΗΜΑ 2.1: Χρήστες AOL από το 2002

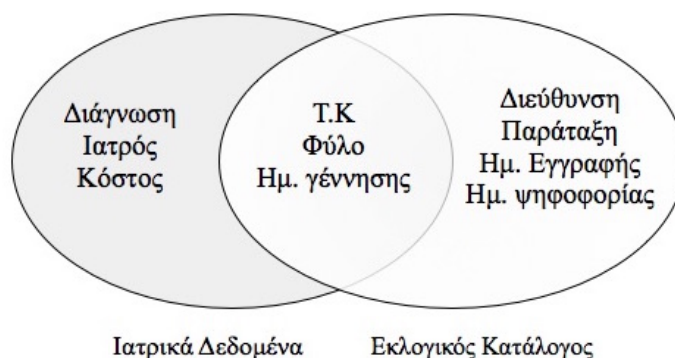
Εντούτοις, δεν είναι σωστό να θεωρούμε ότι η προστασία της ιδιωτικότητας ενός συνόλου δεδομένων είναι δύσκολο έργο, επηρεασμένοι από το παραπάνω γεγονός. Αυτό συνέβει επειδή δεν δόθηκε ουσιαστική προσοχή στην ανωνυμοποίηση των αποτελεσμάτων αναζήτησης, πράγμα που συμπεραίνουμε από την άμεση απομάχρυνση από την εταιρία του υπευθύνου της κοινοποίησης των δεδομένων. Επιπλέον, ο προϊστάμενος της AOL Research απολύεται ενώ ο CTO παραιτήθηκε. Τελικά το γεγονός οδήγησε στην άμεση απομάχρυνση πολλών χρηστών, με αποτέλεσμα την επιτάχυνση της ήδη φθίνουσας πορείας της εταιρίας.



- **Διασταύρωση ιατρικών δεδομένων**

Μιά από τις δημοφιλέστερες αναδείξεις παραβίασης είναι η ταυτοποίηση του κυβερνήτη της Μασαχουσέτης σε λίστα ιατρικών δεδομένων.

Είναι αποδεδειγμένο ότι το 63% του πληθισμού των Ηνωμένων Πολιτειών μπορεί να ταυτοποιηθεί αν είναι γνωστά μόνο: ο ταχυδρομικός κώδικας, το φύλο και η ημερομηνία γέννησης ενός ατόμου. Σύμφωνα με ένα άρθρο [Sweeney, 2001], αν ασκηθεί επίθεση συνδεσιμότητας σε δύο βάσεις δεδομένων που μοιράζονται αυτή την τριάδα γνωρισμάτων, τότε μπορεί να ταυτοποιηθεί μια τουλάχιστον εγγραφή. Συγκεκριμένα, στο πείραμα χρησιμοποιήθηκε ως πρώτη βάση ο κατάλογος ψηφοφόρων του Cambridge και ως δεύτερη η βάση δεδομένων υγείας του Massachusetts Group Insurance Commission (GIC). Τόσο ο κατάλογος ψηφοφόρων όσο και η βάση δεδομένων υγείας κάνουν χρήση των γνωρισμάτων { T.K, φύλο, ημερομηνία γέννησης }, και αφού συνδυάστηκαν δυο εγγραφές, μια σε κάθε βάση, ταυτοποιήθηκαν τα στοιχεία του πρώην κυβερνήτη William Weld.

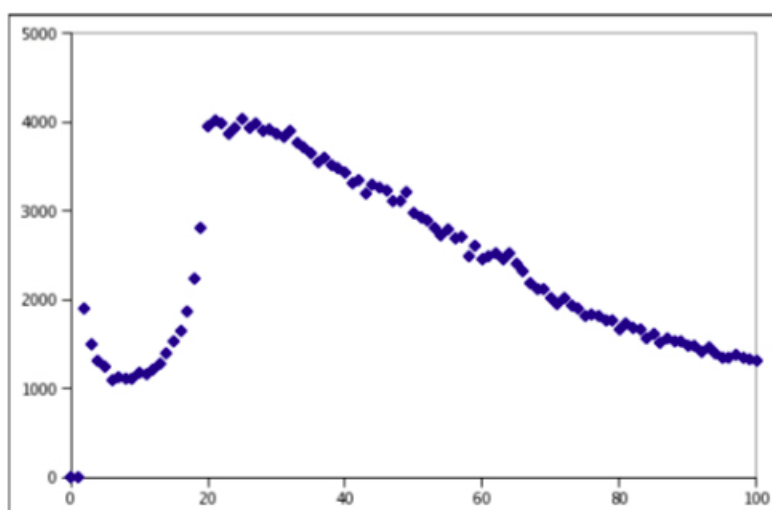


ΣΧΗΜΑ 2.2: Συνδιασμός κοινών γνωρισμάτων

Αυτό που προκάλεσε ανησυχία δεν είναι το ότι κατάφερε να ταυτοποιηθεί ένα διάσημο πρόσωπο συνδυάζοντας τα στοιχεία δυο βάσεων, αλλά ότι τέτοιου τύπου βάσεις, κυρίως με ιατρικά δεδομένα, είχαν ήδη διαμοιραστεί σε εταιρίες και ερευνητές πιστεύοντας λανθασμένα ότι τα δεδομένα είναι ανωνυμοποιημένα. Παρόλο που τα σύνολα δεδομένων δεν περιείχαν προσωπικά αναγνωριστικά, όπως το όνομα ή τον αριθμό κοινωνικής ασφάλισης ενός ασθενούς, τα άτομα μπορούσαν ακόμα να αναγνωριστούν συνδυάζοντας δύο βάσεις δεδομένων.

- Το βραβείο του 1 εκατομμυρίου της Netflix

Η Netflix είναι μια εταιρία που παρέχει συνδρομητικές υπηρεσίες τηλεόρασης σε μορφή on demand. Για να βελτιώσει τον αλγόριθμο εμφάνισης προτεινόμενων ταινιών, η εταιρία δημοσίευσε αρχικά ένα σύνολο δεδομένων που περιείχε περισσότερες από 100 εκατομμύρια βαθμολογίες ταινιών από 480 χιλιάδες χρήστες σε σχεδόν 18 χιλιάδες τίτλους, και στη συνέχεια προσέφεραν ένα βραβείο αξίας 1 εκατομμυρίου δολαρίων σε εκείνους που θα μπορούσαν να βελτιώσουν το τρέχων σύστημα κατά τουλάχιστον 10%. Σε μια προσπάθεια διασφάλισης της ιδιωτικότητας των χρηστών τους, αφαιρούν δεδομένα προσωπικού χαρακτήρα. Το μόνο που απομένουν είναι οι βαθμολογίες (βαθμολογία μεταξύ 1 και 5), η ημερομηνία που δόθηκε η βαθμολογία, το ανώνυμο αναγνωριστικό του χρήστη που έδωσε την αξιολόγηση και τέλος ο τίτλος της ταινίας που αξιολογήθηκε. Η Netflix δήλωσε επίσης ότι το σύνολο που δημοσιεύθηκε αποτελούσε λιγότερο από το ένα δέκατο του πλήρους συνόλου δεδομένων και ότι τα δεδομένα είναι διαταραγμένα. Επίσης δήλωσε, ότι αυτό το ποσοστό επιλέχτηκε τυχαία. Στην πραγματικότητα, αν παρατηρήσουμε το πλήθος βαθμολογιών για κάθε χρήστη βλέπουμε ότι ελάχιστοι έχουν βαθμολογήσει κάτω από 20 ταινίες [Narayanan and Shmatikov, 2008].



ΣΧΗΜΑ 2.3: Πλήθος χρηστών Netflix, προς πλήθος βαθμολογιών

Επιπλέον, τέθηκε υπό αμφισβίτηση η υποτιθέμενη διάταραξη των δεδομένων, όταν ταυτοποιήθηκαν δυο συνδρομητές εντός των δεδομένων. Ο ένας από αυτούς είχε τροποποιημένες 1 από τις 306 αξιολογήσεις ενώ ο άλλος είχε 5 από 229, άρα το ποσοστό θορύβου ήταν ελάχιστο. Εδώ βέβαια, είναι σημαντικό να ληφθεί υπόψη ότι το σύνολο δεδομένων δημοσιεύτηκε με σκοπό την ανάπτυξη βελτιωμένων αλγορίθμων.

Επομένως αν υπήρχε υπερβολική διαταραχή, θα είχε μειωθεί σημαντικά η χρησιμότητα του συνόλου.

Το άρθρο του Narayanan αποδεικνύει ότι απαιτείται ελάχιστη συνδεσιμότητα με εξωτερικά στοιχεία ώστε να ταυτοποιηθεί ένας συνδρομητής. Η σύνδεση στην προκειμένη περίπτωση γίνεται με εγγραφές από την βάση δεδομένων του IMDb. Μέσα από μόνο 50 εγγραφές, ταυτοποιήθηκαν δυο χρήστες.

Είναι εύλογο να αναρωτηθεί κάποιος για το αν αξίζει να ανησυχούμε για την ιδιωτικότητα δεδομένων τέτοιου τύπου. Αν όμως αναλογιστούμε το πόσο αποκαλυπτικό για τις πολιτικές πεποιθήσεις ενός ατόμου μπορεί να είναι η προτίμηση ταινιών όπως Fahrenheit 9/11 και Death by China, συμπεραίνουμε ότι η παροχή προστασίας της ιδιωτικότητας είναι αναγκαία.

## 2.2 Ανεπαρκείς προσεγγίσεις προστασίας

Ως κριτήρια της αποτελεσματικής ανωνυμοποίησης θεωρούνται τα παρακάτω:

- Μη δυνατότητα εντοπισμού φυσικού προσώπου
- Μη συνδεσιμότητα μεταξύ καταχωρήσεων που αντιστοιχούν σε ένα πρόσωπο
- Μη δυνατότητα εξαγωγής συμπερασμάτων σχετικά με ένα φυσικό πρόσωπο.

Έχουν παρουσιαστεί κατά καιρούς αρκετές προτάσεις για τη διαφύλαξη της ιδιωτικότητας κατά την ανάλυση δεδομένων. Ωστόσο, οι περισσότερες από αυτές δεν ικανοποιούν τις παραπάνω εγγυήσεις απορρήτου. Αξίζει όμως να αναφερθούν αυτές οι προσπάθειες, διότι οφείλουμε να τονίσουμε ότι η διατήρηση της ιδιωτικότητας μπορεί να είναι ένα δύσκολο έργο με απροσδόκητα προβλήματα.

### • Επερωτήσεις μεγάλων συνόλων

Μια κοινή ιδέα είναι να απαγορευτεί η εκτέλεση επερωτήσεων σχετικά με ένα συγκεκριμένο άτομο ή ένα μικρό σύνολο ατόμων. Είναι εύκολο να καταλάβουμε ότι αυτή η στρατηγική είναι ανεπαρκής όταν γνωρίζουμε ότι ένα συγκεκριμένο άτομο  $X$  είναι στη βάση δεδομένων. Αυτό που μπορούμε να κάνουμε είναι να εφαρμόσουμε δυο «μεγάλες» επερωτήσεις. Η πρώτη, για παράδειγμα, μπορεί να είναι «Πόσα άτομα στη βάση δεδομένων έχουν διαβήτη;» και η δεύτερη «Πόσα άτομα στη βάση δεδομένων, που δεν ονομάζονται  $X$ , έχουν διαβήτη;». Από τα αποτελέσματα των δυο επερωτήσεων μπορούμε να συμπεράνουμε την κατάσταση του ατόμου  $X$ . Παρόλο που και οι δυο επερωτήσεις θεωρούνται μεγάλες, παραβιάστηκε η ιδιωτικότητα μιας εγγραφής.

- **Έλεγχος Επερωτήσεων**

Μια άλλη στρατηγική είναι η αξιολόγηση κάθε επερώτησης στη βάση δεδομένων στο πλαίσιο του ιστορικού των επερωτήσεων, ώστε να προσδιοριστεί εάν μια απάντηση θα ήταν αποκαλυπτική. Εάν η επερώτηση θεωρηθεί ότι παραβιάζει την ιδιωτικότητα, απορρίπτεται και δεν επιστρέφεται απάντηση. Αυτή η μέθοδος μπορεί να χρησιμοποιηθεί για την ανίχνευση της επερώτησης στο προηγούμενο παράδειγμα με το αν ένα άτομο X έχει ή όχι διαβήτη. Παρόλο που η τεχνική αυτή μπορεί να φαίνεται πολλά υποσχόμενη, η παρακολούθηση όλων των επερωτήσεων είναι υπολογιστικά ανέφικτη [Kleinberg and Papadimitriou, 1998]. Επιπλέον, η ίδια η άρνηση απάντησης σε μία επερώτηση μπορεί να είναι αποκαλυπτική. Για αυτούς τους λόγους, η παραπάνω προσέγγιση είναι πρακτικά μη εφαρμόσιμη.

- **Δειγματοληψία**

Η δειγματοληψία μερικές φορές θεωρείται ως πιθανή λύση στο πρόβλημα. Εδώ ένας τυχαίος αριθμός εγγραφών στη βάση δεδομένων θα επιλεγεί τυχαία και οι πληροφορίες τους θα επιστραφούν προς ανάλυση. Ιδιότητες και στατιστικά στοιχεία μπορούν στη συνέχεια να υπολογιστούν στο νέο. Αν γίνει επιλογή ενός αρκετά μεγάλου υποδείγματος τα αποτελέσματα είναι αντιπροσωπευτικά της βάσης ως σύνολο.

Λαμβάνοντας υπόψη ότι το υποσύνολο είναι συνήθως μικρό σε σύγκριση με το μέγεθος της βάσης δεδομένων, είναι απίθανο να επιλεγεί μια συγκεκριμένη εγγραφή. Ωστόσο, κάθε φορά που επιλέγεται η ιδιωτικότητα της παραβιάζεται. Αυτό δεν είναι αποδεκτό, δεδομένου ότι θέλουμε να προστατεύσουμε την ιδιωτικότητα κάθε ατόμου στη βάση δεδομένων. Οπότε και πάλι αυτή η προσέγγιση δεν είναι εφαρμόσιμη σε πραγματικές καταστάσεις.

- **Τυχαία Απάντηση (Randomized Response)**

Η προσέγγιση αυτή αποτελεί μια ισχυρή τεχνική παροχής ιδιωτικότητας. Εδώ τα δεδομένα τυχαιοποιούνται πριν αποθηκευτούν και τα στατιστικά στοιχεία μπορούν να υπολογιστούν από τα τυχαίοποιημένα δεδομένα. Η μέθοδος της τυχαίας απάντησης λειτουργεί με την ρίψη ενός νομίσματος, και με βάση το αποτέλεσμα η επερώτηση απαντάται αληθώς ή τυχαίως. Η ιδιωτικότητα είναι εγγυημένη λόγω της αβεβαιότητας ως προς τον τρόπο ερμηνείας μιας δεδομένης τιμής.

Αυτή η μέθοδος δημιουργήθηκε για καταστάσεις όπου τα άτομα δεν εμπιστεύονται τον υπεύθυνο/διαχειριστή δεδομένων. Η πιθανότητα επιστροφής τυχαίων δεδομένων επιλέγεται έτσι ώστε κατά τον υπολογισμό των στατιστικών στοιχείων για όλα τα δεδομένα που επιστρέφονται, το σφάλμα να παραμένει εντός ενός αποδεκτού διαστήματος. Ενώ η τυχαία απάντηση παρέχει προστασία της ιδιωτικότητας, το μειονέκτημα είναι ότι μπορεί να εισάγει παραπάνω θόρυβο στα δεδομένα και έτσι η προσέγγιση αυτή καθίσταται αδύνατη για σύνθετα δεδομένα.

- **Τυχαίο Αποτέλεσμα (Randomized Output)**

Κατά την μέθοδο αυτή προστίθεται θόρυβος στην έξοδο. Αυτό διαφέρει από την προηγούμενη τεχνική επειδή εκεί τα δεδομένα είναι τυχαιοποιημένα, ενώ σε αυτή την περίπτωση ο διαχειριστής έχει πρόσβαση στα αρχικά δεδομένα. Σε αυτή την τεχνική αυτός είναι που θα προσθέσει τυχαίο θόρυβο στο αποτέλεσμα των επερωτήσεων.

Σημειώνουμε ότι αν γίνει αόριστα η παραπάνω προσέγγιση θα αποτύχει. Για παράδειγμα, ας υποθέσουμε ότι ο θόρυβος έχει μέση τιμή μηδέν και ότι χρησιμοποιείται ανεξάρτητη τυχαία συχνότητα για τη δημιουργία κάθε απάντησης. Σε αυτήν την περίπτωση, αν η ίδια επερώτηση δωθεί αρκετές φορές οι απαντήσεις μπορούν να αθροίζονται, να υπολογισθεί ο μέσος όρος και η αληθινή απάντηση θα αποκαλυφθεί τελικά.

Παρατηρούμε λοιπόν ότι οι παραπάνω προτάσεις προστασίας των δεδομένων αδυνατούν τελικά να διαφυλάξουν την ιδιωτικότητα των εγγραφών ενός συνόλου. Πάνω στις αδυναμίες αυτές στηρίζονται, όπως θα δούμε στη συνέχεια, οι μέθοδοι γενίκευσης, οι οποίες προσδίδουν σε ικανοποιητικό βαθμό τις επιθυμητές εγγυήσεις απορρήτου.



## Κεφάλαιο 3

# Μηχανισμοί γενίκευσης και ομαδοποίησης

Οι μέθοδοι γενίκευσης αποτελούν την πιο ευρέως διαδεδομένη τεχνική ανωνυμοποίησης. Η συγκεκριμένη προσέγγιση συνίσταται στην ομαδοποίηση των εγγραφών ή την αλλοίωση των γνωρισμάτων στα οποία ανφέρονται τα δεδομένα, μέσω της τροποποίησης της αντίστοιχης κλίμακας μεγέθους. Παρόλο που αυτή η μέθοδος μπορεί να αποβεί αποτελεσματική για την πρόληψη του εντοπισμού ενός προσώπου, δεν εξασφαλίζει αποτελεσματική ανωνυμοποίηση σε όλες τις περιπτώσεις. Για αυτό το λόγο έχουν παρουσιαστεί ειδικές και τεχνολογικά εξελιγμένες ποσοτικές προσεγγίσεις για την πρόληψη του συνδιασμού συνόλων δεδομένων, προς εξαγωγή συμπερασμάτων για μεμονομένα φυσικά πρόσωπα.

Το κεφάλαιο αυτό ξεκινά με την παρουσίαση του μοντέλου της  $k$ -ανωνυμίας. Στη συνέχεια περιγράφονται οι τεχνικές της  $l$ -διαφορετικότητας και  $t$ -εγγύτητας που επεκτείνουν της δυνατότητες της  $k$ -ανωνυμίας. Τέλος γίνεται παρουσίαση σύγχρονων αλγορίθμων ανωνυμοποίησης που υλοποιούν αυτές τις μεθόδους.

### 3.1 Θεμελίωση

Όπως είδαμε στις προηγούμενες παραγράφους, μια από τις αποδοτικότερες λύσεις για την προστασία της ιδιωτικότητας των δεδομένων, αλλά ταυτόχρονα διατήρηση της χρησιμότητάς τους, είναι η ανωνυμοποίηση. Παρακάτω δίνουμε μερικούς ορισμούς και προτάσεις που θα χρησιμοποιήσουμε κατά την διάρκεια της διαδικασίας αυτής.

**Ορισμός 3.1.** -Περι συνόλων δεδομένων-

- Ένα **σύνολο δεδομένων**(dataset) είναι μια πεπερασμένη συλλογή στοιχείων.

- Κάθε στοιχείο του, ή εγγραφή, είναι μια διατεταγμένη λίστα τιμών και αντιστοιχεί σε ένα πρόσωπο, στο οποίο αναφέρονται τα δεδομένα τιμών για κάθε **γνώρισμα**(attribute).

Το σύνολο δεδομένων το μοντελοποιούμε σαν μια πλειάδα στοιχείων-γραμμών  $D = (x_1, x_2, \dots, x_n)$ .

- Η ανωνυμοποιημένη προβολή του  $D$  συμβολίζεται ως  $R$ .
- Ορίζουμε ως  $A = A_1, A_2, \dots, A_r$  μια συλλογή από  $r$  **γνωρίσματα**.
- Συμβολίζουμε ως  $t$  κάθε πλειάδα στο  $R$ .
- Κάθε  $t[A_i]$  με  $i \in [1, r]$  εκφράζει την τιμή του γνωρίσματος  $A_i$  στο  $R$  για την  $t$ .
- Συμβολίζουμε  $t[A] = (t[A_1], t[A_2], \dots, t[A_r])$ .

Θεωρητικά, ως ανωνυμοποίηση ενός συνόλου δεδομένων  $D$  θεωρείται ένα σύνολο γενικεύσεων των γνωρισμάτων του, αποτελούμενο από μια ακριβώς γενίκευση ανά γνώρισμα του  $D$ . Αυτές οι γενικεύσεις μετασχηματίζουν το  $D$  σε ένα νέο σύνολο δεδομένων  $D'$ . Επιπλέον, η διαδικασία της γενίκευσης είναι αυτή που δίνει το όνομά της στην οικογένεια μεθόδων που αναλύουμε στη συνέχεια.

## 3.2 k-ανωνυμία

Η μέθοδος της  $k$ -ανωνυμίας<sup>1</sup> είναι μια μη διαδραστική<sup>2</sup> τεχνική ιδιωτικότητας. Αρχικά, η  $k$ -ανωνυμία θεωρήθηκε ως πιθανή λύση στο πρόβλημα της ιδιωτικότητας, κυρίως λόγω της εννοιολογικής απλότητας και των αποτελεσματικών αλγορίθμων που εγγυώνται την εφαρμογή της. Δυστυχώς προέκυψε άμεσα ότι έχει ελλείψεις και δεν είναι σε θέση να προστατεύσει την ιδιωτικότητα όλων των ατόμων. Θα αναλύσουμε την μέθοδο της  $k$ -ανωνυμίας και στη συνέχεια, αφού δούμε τα κενά ασφαλείας, θα παρουσιάσουμε βελτιώσεις.

Η ιδέα πίσω από την  $k$ -ανωνυμία είναι ότι όταν δίνονται πληροφορίες για ένα άτομο από μια εξωτερική πηγή, όπως το όνομά του, η ημερομηνία γέννησής, ο ταχυδρομικός κώδικας, το φύλο κλπ., θα πρέπει να είναι αδύνατο να βρεθεί, με ακρίβεια, στο ανωνυμοποιημένο σύνολο-πίνακα η γραμμή-πλειάδα που αντιστοιχεί στο άτομο αυτό [Sweeney, 2002]. Διαισθητικά, η διαδικασία πρέπει να έχει ως αποτέλεσμα οποιοσδήποτε χρήστης της βάσης να μην είναι σε θέση να αποκτήσει πληροφορίες σχετικά με ένα συγκεκριμένο άτομο, αφού δεν μπορεί να εντοπίσει τις πληροφορίες του στη βάση δεδομένων. Αυτό μπορεί να επιτευχθεί με την κατάργηση ορισμένων γνωρισμάτων σε συνδυασμό με την τροποποίηση ορισμένων τιμών πριν την απελευθέρωση της βάσης δεδομένων.

<sup>1</sup>k-anonymity

<sup>2</sup>Με τον όρο μη διαδραστική τεχνική εννοούμε ότι, μετά την εφαρμογή της, δημοσιοποιείται μια εξυγιασμένη προβολή της βάσης



Το πρώτο βήμα κατά τη δημιουργία της ανώνυμης βάσης δεδομένων  $R$ , είναι η κατάργηση γνωρισμάτων που προσδιορίζουν σαφώς τα άτομα. Είναι αυτά που συνήθως χρησιμοποιούμε ως πρωτεύοντα κλειδιά<sup>3</sup>: Γνωρίζοντας την τιμή ενός γνωρίσματος κάποιου ατόμου μας διευκολύνει να βρούμε, με μεγάλη πιθανότητα, την πλειάδα που αντιστοιχεί σε αυτό το άτομο, παραδείγματος χάρη η διεύθυνση, ο αριθμός δελτίου ταυτότητας, το όνομα, το τηλέφωνο, κλπ. Ωστόσο, υπάρχουν γνωρίσματα τα οποία, όταν λαμβάνονται μαζί, είναι δυνατό να ταυτοποιηθεί ένα άτομο. Μια τέτοια συλλογή γνωρισμάτων ονομάζεται *quasi-identifier*. Σε γενικές γραμμές, ένα *quasi-identifier* περιέχει γνωρίσματα τα οποία είναι πιθανόν να υπάρχουν και σε άλλες βάσεις.

Ο ακριβής ορισμός ενός *quasi-identifier*<sup>4</sup> βασίζεται στον διαχωρισμό μεταξύ ευαίσθητων και μη ευαίσθητων γνωρισμάτων. Η τιμή ενός ευαίσθητου γνωρίσματος πρέπει να παραμείνει μυστική, πράγμα που σημαίνει ότι σε μια ανωνυμοποιημένη βάση δεδομένων θα πρέπει να είναι αδύνατο να συνδεθεί μια ευαίσθητη τιμή (συγκεκριμένα, μια πλειάδα που περιέχει μια ευαίσθητη τιμή) σε ένα συγκεκριμένο άτομο. Με αυτόν τον τρόπο είναι αδύνατο να μάθει κανείς την ευαίσθητη τιμή ενός ατόμου από την ανωνυμοποιημένη βάση δεδομένων και έτσι διατηρείται η ιδιωτικότητα. Όλα τα γνωρίσματα εκτός των ευαίσθητων θα ονομάζονται μη ευαίσθητα. Θεωρείται επίσης ότι τα επιλεγμένα ευαίσθητα χαρακτηριστικά δεν εμφανίζονται σε άλλες βάσεις δεδομένων. Επομένως, όταν είναι γνωστή η τιμή ενός ευαίσθητου γνωρίσματος, δεν είναι δυνατόν να βρεθεί το άτομο που αντιστοιχεί σε αυτήν την τιμή, πράγμα που σημαίνει επίσης ότι ένα ευαίσθητο γνώρισμα δεν περιλαμβάνεται ποτέ σε ένα *quasi-identifier*. Καταλήγουμε στο συμπέρασμα ότι ένα *quasi-identifier* αποτελείται μόνο από μη ευαίσθητα γνωρίσματα.

### Ορισμός 3.2. (*quasi-identifier*)

Ένα σύνολο από μη ευαίσθητα γνωρίσματα  $Q = Q_1, Q_2, \dots, Q_r$  καλείται *quasi-identifier* για μια βάση δεδομένων  $D$  αν υπάρχει  $t \in D$  ώστε η πρόταση

$$t = t' \vee t[Q] \neq t'[Q]$$

για κάθε  $t' \in D$  να είναι ψευδής.

Με άλλα λόγια, υπάρχει τουλάχιστον μια εγγραφή στην αρχική βάση δεδομένων που μπορεί να προσδιοριστεί μοναδικά χρησιμοποιώντας μόνο τιμές γνωρισμάτων του συνόλου  $Q$ . Δηλώνουμε το σύνολο όλων των *quasi-identifier* με το σύμβολο  $QI$ .

Είναι προφανές ότι αν ένα γνώρισμα μπορεί να οριστεί ως πρωτεύον κλειδί, τότε αποτελεί ένα *quasi-Identifier*, όπως επίσης συλλογές γνωρισμάτων όπως {ημερομηνία γέννησης, Τ.Κ., φύλο}.

<sup>3</sup>primary keys

<sup>4</sup>Σε διάφορα ελληνικά άρθρα έχει χρησιμοποιηθεί ο όρος «ψευδοανγνωριστικό σύνολο», που είναι εμφανώς μακριά από την αγγλική έννοια

Στην συνέχεια θα αναφερθούμε σε ομάδες ατόμων τα οποία θα έχουν ίδιες τιμές για ένα σύνολο από quasi-identifiers.

### Ορισμός 3.3. (Κλάση Ισοδυναμίας)

Μια κλάση ισοδυναμίας για έναν πίνακα  $R$  σε σχέση με τα γνωρίσματα στο  $A$  είναι ένα σύνολο πλειάδων  $E = \{t_1, t_2, \dots, t_i\} \in R$  για τα οποία  $t_1[A] = t_2[A] = \dots = t_i[A]$ .

Δηλαδή, η προβολή κάθε πλειάδας επί των γνωρισμάτων στο  $A$  είναι ακριβώς ίδια.

TK	Ηλικία	Περιοχή	Ασθένια
172**	<30	*	Καρδιακό Νόσημα
172**	<30	*	Καρκίνος
172**	<30	*	Καρκίνος
172**	<30	*	Καρδιακό Νόσημα
163**	>40	*	Καρκίνος
163**	>40	*	Ηπατίτιδα
163**	>40	*	Καρδιακό Νόσημα
163**	>40	*	Καρδιακό Νόσημα
172**	>40	*	Καρκίνος
172**	>40	*	Καρκίνος
172**	>40	*	Καρκίνος
172**	>40	*	Καρκίνος

ΣΧΗΜΑ 3.1: Ένα 4-ανώνυμο σύνολο δεδομένων

Στον παραπάνω πίνακα έχουμε 4 γνωρίσματα από τα οποία το ένα είναι ευαίσθητο (Ασθένια). Τα στοιχεία είναι ανωνυμοποιημένα. Επεμβαίνοντας στην Περιοχή και στα δυο τελευταία ψηφία του TK παρατηρούμε ότι δημιουργούνται τρεις ισοδύναμες κλάσεις, οι οποίες περιέχουν 4 εγγραφές οι κάθε μία.

Η  $k$ -ανωνυμία λειτουργεί εξασφαλίζοντας ότι υπάρχουν τουλάχιστον  $k - 1$  άλλα άτομα (δηλ. γραμμές στη βάση δεδομένων) που έχουν τις ίδιες τιμές για όλα τα quasi-identifiers. Αυτό θα πρέπει να διασφαλίζει ότι αν ψάχνουμε για ένα συγκεκριμένο άτομο, θα έχουμε πάντα τουλάχιστον  $k$  αποτελέσματα και έτσι δεν μπορούμε να προσδιορίσουμε την συγκεκριμένη πλειάδα.

### Ορισμός 3.4. ( $k$ -ανωνυμία)

Μια προβολή  $R$  της βάσης  $D$  θα είναι  $k$ -ανώνυμη αν για κάθε πλειάδα  $t \in R$ , και για κάθε quasi-identifier  $A \in IQ$ , υπάρχουν τουλάχιστον  $k - 1$  άλλες πλειάδες  $t_1, \dots, t_{k-1} \in R$  ώστε  $t_1[A] = t_2[A] = \dots = t_{k-1}[A]$ .

Μια τεχνική για την επίτευξη  $k$ -ανωνυμίας είναι η γενίκευση γνωρισμάτων, η οποία αντικαθιστά τις τιμές των quasi-identifiers με τιμές που είναι λιγότερο συγκεκριμένες αλλά σημασιολογικά ορθές. Ως αποτέλεσμα, περισσότερες εγγραφές θα έχουν το ίδιο σύνολο τιμών

γνωρισμάτων. Παρατηρούμε ότι ο παραπάνω πίνακας είναι 4-ανώνυμος, ενώ έχουμε quasi-identifier το σύνολο γνωρισμάτων {TK, Ηλικία, Περιοχή}. Ο ορισμός της  $k$ -ανωνυμίας προϋποθέτει ότι ο υπεύθυνος της βάσης είναι σε θέση να αναγνωρίσει με ακρίβεια τα quasi-identifiers. Ωστόσο, το έργο αυτό είναι πολύ δύσκολο στην εφαρμογή και είναι εύκολο να γίνει λάθος. Ειδικά ο διαχωρισμός ευαίσθητων και μη ευαίσθητων γνωρισμάτων μπορεί να είναι προβληματικός. Αυτό είναι σαφώς ένα από τα μειονεκτήματα της  $k$ -ανωνυμίας: Η υπόθεση ότι κάποιος είναι με ακρίβεια ικανός να βρει όλα τα quasi-identifiers μοιάζει να είναι υπερβολικά απαιτητική.

### 3.2.1 Επίτευξη $k$ -anonymity

Υπάρχουν δυο κοινώς χρησιμοποιούμενες μέθοδοι για την επίτευξη της  $k$ -ανωνυμίας:

- Η πρώτη είναι η τεχνική γενίκευσης - generalization, όπου μια τιμή για ένα γνώρισμα μετατρέπεται σε μια γενικότερη και πιθανώς «αφηρημένη» τιμή. Παραδείγματος χάρη η τιμή του γνωρίσματος TK, όπου λείπουν τα δυο τελευταία ψηφία. Η απώλεια πληροφορίας είναι αναγκαστικά το τίμημα που πρέπει να πληρώσουμε ώστε να επιτευχθεί ιδιωτικότητα.
- Η δεύτερη μέθοδος είναι αυτή της απόκρυψης - suppression. Όπως ορίζεται από την ίδια τη λέξη, εδώ η πληροφορία αποκρύπτεται εντελώς. Παραδείγματος χάρη η τιμή του γνωρίσματος Περιοχή. Η μέθοδος της απόκρυψης μπορεί να θεωρηθεί και σαν εφαρμογή της γενίκευσης, στον μέγιστο δυνατό βαθμό για κάποιο γνώρισμα.

Μπορούμε τώρα να διατυπώσουμε το πρόβλημα της επίτευξης  $k$ -ανωνυμίας ελαχιστοποιώντας ταυτόχρονα τον αριθμό των γενικεύσεων ή αποκρύψεων που εφαρμόζονται. Αυτό θα είχε ως αποτέλεσμα τον βέλτιστο ανώνυμο πίνακα, όπου το βέλτιστο σημαίνει ότι η πληροφορία έχει παραμορφωθεί στο ελάχιστο, και έτσι μπορεί να θεωρηθεί ως η πιο χρήσιμη ανώνυμοποιημένη βάση δεδομένων. Ακόμα κι αν περιορίζουμε το πρόβλημα μόνο στην καταστολή των τιμών, μπορούμε να αποδείξουμε ότι το πρόβλημα βελτιστοποίησης είναι NP-Hard [Meyerson and Williams, 2004]. Επομένως στην πράξη εφαρμόζονται μόνο αλγόριθμοι προσέγγισης. Παρακάτω θα δείξουμε ότι η  $k$ -ανωνυμία αποτυγχάνει να προστατεύσει από πολλαπλές, πιθανώς ανεξάρτητες, προβολές.

### 3.2.2 Επιθέσεις κατά της $k$ -ανωνυμίας

Ενώ το μοντέλο της  $k$ -ανωνυμίας αποτελεί θεμελιώδη έννοια πάνω στην προστασία των δεδομένων, δεν εξασφαλίζει απόλυτα την ιδιωτικότητα σε συγκεκριμένες επιθέσεις. Όπως έχει αποδειχθεί, η  $k$ -ανωνυμία δεν εγγυάται πλήρως την μη αποκάλυψη της τιμής ευαίσθητων γνωρισμάτων των εγγράφων. Σε πολλές περιπτώσεις στατιστικών ερευνών απαιτείται

η δημοσίευση των τιμών ενός ευαίσθητου γνώρισματος ως έχουν, οι οποίες προφανώς δεν επηρεάζονται από την εφαρμογή  $k$ -ανωνυμοποίησης. Θεωρείται ότι εφόσον δεν μπορεί να ταυτοποιηθεί ένα άτομο με μια πλειάδα, δεν μπορεί να προκύψει συμπέρασμα για την τιμή που αυτό λαμβάνει για ένα ευαίσθητο γνώρισμα. Μπορεί, ωστόσο, κάποιος να συμπεράνει την τιμή του ευαίσθητου γνώρισματος μιας ή περισσότερων ομάδων πλειάδων. Αν επιπλέον υπάρχει η δυνατότητα συνδυασμού κάποιων τιμών του quasi-identifier που γνωρίζει ο επιτιθέμενος με κάποιες από αυτές που εμφανίζονται, θα μπορούσε να συμπεράνει την κλάση ισοδυναμίας που ανήκει η εγγραφή και πιθανότατα πληροφορίες σχετικά με το ευαίσθητο γνώρισμα.

Οι ακόλουθες δύο κατηγορίες επιθέσεων αναλύθηκαν κατά την παρουσίαση της  $l$ -διαφορετικότητας [Machanavajjhala et al., 2006].

- **Επίθεση ομοιογένειας**

Αποδεικνύεται ότι εάν δεν υπάρχει διαφοροποίηση στην τιμή των ευαίσθητων γνωρισμάτων, παραβιάζεται η ιδιωτικότητα της εγγραφής. Εάν ο αριθμός των πλειάδων υπερβαίνει κατά πολύ τις πιθανές τιμές των ευαίσθητων χαρακτηριστικών, αυτό μπορεί να είναι μια κοινή κατάσταση. Παρατηρώντας τον προηγούμενο πίνακα βλέπουμε ότι αν ο επιτιθέμενος γνωρίζει ότι το άτομο που αναζητά βρίσκεται στη βάση δεδομένων, ενώ ταυτόχρονα ξέρει ότι είναι πάνω από 40 ετών και τα τρία πρώτα ψηφία του T.K. είναι 172, συμπεραίνει με βεβαιότητα ότι το άτομο πάσχει από καρκίνο. Ενώ η ταυτότητα του ατόμου προστατεύεται (ο επιτιθέμενος δεν γνωρίζει ποιά πλειάδα αντιστοιχεί στο άτομο), προκύπτει αποκάλυψη τιμής του ευαίσθητου γνώρισματος.

- **Επίθεση με πρότερη γνώση**

Ο επιτιθέμενος μπορεί επίσης να έχει γνώση σχετικά με τη κατανομή των ευαίσθητων τιμών. Στο προηγούμενο παράδειγμα, αν ο επιτιθέμενος γνωρίζει ότι το άτομο που αναζητά βρίσκεται στη βάση, ενώ ταυτόχρονα ξέρει ότι είναι κάτω από 30 ετών, τότε η πλειάδα που αντιστοιχεί βρίσκεται στην πρώτη κλάση ισοδυναμίας. Αν επιπλέον γνωρίζει ότι είναι αθλητικός τύπος και προσέχει τη διατροφή του, εύκολα συμπεραίνει ότι είναι απίθανο να πάσχει από καρδιακό νόσημα, άρα το άτομο έχει, με μεγάλη πιθανότητα, καρκίνο.

Μπορούμε να αποτρέψουμε την επίθεση ομοιογένειας διασφαλίζοντας ότι υπάρχει αρκετή ποικιλομορφία στις ευαίσθητες τιμές. Η προστασία από πρότερη γνώση είναι σχεδόν αδύνατη. Μπορούμε, ωστόσο, να δυσκολέψουμε τον επιτιθέμενο: Εάν υπάρχει μεγαλύτερη ποικιλία στις τιμές των ευαίσθητων γνωρισμάτων, θα χρειαστεί περισσότερη γνώση για την ανάκτηση της ακριβούς τιμής.

### 3.3 $l$ -Διαφορετικότητα

Το μοντέλο της  $l$ -διαφορετικότητας επεκτείνει την τεχνική της  $k$ -ανωνυμίας, έτσι ώστε να διασφαλιστεί ότι είναι αδύνατη η εφαρμογή επιθέσεων εξαγωγής συμπερασμάτων, εξασφαλίζοντας ότι κάθε γνώρισμα θα έχει τουλάχιστον  $l$  διαφορετικές τιμές για κάθε κλάση ισοδυναμίας.

**Ορισμός 3.5.** (Εντροπία  $l$ -διαφορετικότητας)

Για μια κλάση ισοδυναμίας  $E$ , έστω το  $S$  το πεδίο τιμών των ευαίσθητων γνωρισμάτων και το  $Pr[E, s]$  ο λόγος των εγγραφών στο  $E$  που έχουν ευαίσθητη τιμή  $s$ , τότε το  $E$  είναι  $l$ -διαφορετικό αν:

$$-\sum_{s \in S} Pr[E, s] \log(Pr[E, s]) \geq \log(l)$$

**Ορισμός 3.6.** Ένα σύνολο δεδομένων είναι  $l$ -διαφορετικό αν όλες οι ισοδύναμες κλάσεις είναι  $l$ -διαφορετικές.

Μιά απλούστερη εκδοχή του ορισμού είναι ότι κάθε κλάση ισοδυναμίας θα πρέπει να έχει τουλάχιστον  $l$  διαφορετικές τιμές για το ευαίσθητο γνώρισμα (Διακριτή  $l$ -διαφορετικότητα).

TK	Ηλικία	Περιοχή	Ασθένια
172**	<30	*	Καρδιακό Νόσημα
172**	<30	*	Καρκίνος
172**	<30	*	Ηπατίτιδα
172**	<30	*	Καρδιακό Νόσημα
163**	>40	*	Καρκίνος
163**	>40	*	Ηπατίτιδα
163**	>40	*	Καρδιακό Νόσημα
163**	>40	*	Καρδιακό Νόσημα
172**	>40	*	Καρκίνος
172**	>40	*	Καρκίνος
172**	>40	*	Ηπατίτιδα
172**	>40	*	Καρδιακό Νόσημα

ΣΧΗΜΑ 3.2: Ένα 3-διαφορετικό σύνολο δεδομένων

Η σχέση στον ορισμό είναι σχεδόν η ίδια με την εντροπία του Shannon, όπου οι πιθανότητες δίδονται τώρα ως κλάσματα συχνοτήτων των ευαίσθητων γνωρισμάτων. Όπως επισημάνθηκε στον παραπάνω ορισμό, για να υπάρχει  $l$ -διαφορετικότητα για κάθε κλάση ισοδυναμίας, η εντροπία ολόκληρου του πίνακα πρέπει να είναι τουλάχιστον  $\log(l)$ . Κάποιες φορές αυτό μπορεί να είναι υπερβολικά περιοριστικό, καθώς η εντροπία ολόκληρου του πίνακα θα είναι αρκετά μικρή εάν κάποιες τιμές είναι πολύ συχνές. Αυτό οδηγεί στην ακόλουθη λιγότερο συντηρητική έννοια της  $l$ -διαφορετικότητας.

**Ορισμός 3.7.** (Αναδρομική  $(c, l)$ -διαφορετικότητα)

Έστω  $m$  ο αριθμός των πιθανών τιμών ευαίσθητου γνωρίσματος σε μια κλάση ισοδυναμίας

και  $r_i$ , με  $i \in [1, m]$ , το πόσες φορές η  $i$ -οστή συχνότερη τιμή εμφανίζεται στην κλάση  $E$ . Τότε η  $E$  θα έχει  $(c, l)$ -διαφορετικότητα αν:

$$r_1 \leq c(r_l + r_{l+1} + \dots + r_m)$$

**Ορισμός 3.8.** Ένα σύνολο δεδομένων είναι  $(c, l)$ -διαφορετικό αν όλες οι ισοδύναμες κλάσεις είναι  $l$ -διαφορετικές.

Η αναδρομική  $(c, l)$ -διαφορετικότητα εξασφαλίζει ότι η πιο συχνή τιμή δεν εμφανίζεται πολύ συχνά και οι λιγότερο συχνές τιμές δεν εμφανίζονται πολύ σπάνια.

Γενικά, συμπεραίνουμε ότι η μέθοδος της  $l$ -διαφορετικότητας αντιμετωπίζει τις επιθέσεις ομοιογένειας, ενώ ταυτόχρονα δυσκολεύει τους επιτιθέμενους με πρότερη γνώση. Όσο υψηλότερη είναι η τιμή του  $l$ , τόσο περισσότερη γνώση απαιτείται για να αποκαλυφθεί η τιμή ενός ευαίσθητου χαρακτηριστικού ενός ατόμου.

### 3.3.1 Επίτευξη $l$ -διαφορετικότητας

Η εισαγωγή της τεχνικής αυτής αποτελεί μια σημαντική βελτίωση της  $k$ -ανωνυμίας. Ωστόσο, παρατηρείται ιδιόταιρη δυσκολία στην εφαρμογή της.

Ας θεωρήσουμε μια βάση δεδομένων μεγέθους  $n = 100000$  με ένα ευαίσθητο γνώρισμα δύο και μόνο πιθανών τιμών, το να έχει ή όχι τον ιό *HIV*. Έστω ότι το 99% του δείγματος έχει την τιμή OXI (είναι αρνητικοί στον ιό). Αν επιθυμούμε οποιαδήποτε τιμή προσέγγισης  $l$ -διαφορετικότητας, κάθε κλάση ισοδυναμίας θα πρέπει να έχει και τις δυο τιμές. Αυτό μεταφράζεται στην δημιουργία 1000 κλάσεων, που θα έχει ως αποτέλεσμα μεγάλη απώλεια πληροφορίας κατά την εφαρμογή γενίκευσης. Σημειώνουμε επίσης ότι επειδή η εντροπία του ευαίσθητου γνωρίσματος στον πίνακα είναι πολύ μικρή, η  $l$ -διαφορετικότητα μπορεί να επιτευχθεί μόνο εάν διαλέξουμε ένα αρκετά μικρό  $l$ , καθώς για μεγάλη τιμή θα ήταν στην πραγματικότητα αδύνατο να εξασφαλίσει διαφορετικότητα. Παρόλο που αυτό δεν αποτελεί ελάττωμα, το αναφέρουμε για να δείξουμε τη δυνητική δυσκολία να επιτύχουμε την  $l$ -διαφορετικότητα.

### 3.3.2 Επιθέσεις κατά της $l$ -διαφορετικότητας

Παρακάτω αναφέρουμε δυο είδη επιθέσεων κατά της μεθόδου αυτής.

- **Ασύμμετρη Επίθεση**

Όταν η συνολική κατανομή είναι ασύμμετρη, η εφαρμογή  $l$ -διαφορετικότητας δεν εξασφαλίζει την μη αποκάλυψη τιμής ενός γνωρίσματος. Στο προηγούμενο παράδειγμα όπου το 99% ενός πληθυσμού έχει την τιμή ευαίσθητου γνωρίσματος OXI, η αρχή της

*l*-διαφορετικότητας επιτρέπει να υπάρχει μια κλάση ισοδυναμίας με ίσο αριθμό θετικών/αρνητικών στον *l*ό ατόμων. Παρατηρούμε ότι η ιδιωτικότητα κάθε ατόμου που ανήκει σε αυτή την κλάση χάνεται επειδή θεωρείται ότι έχει 50% πιθανότητα να είναι θετικός στον *l*ό, αντί της πραγματικής 1%.

Ας θεωρήσουμε τώρα μια κλάση ισοδυναμίας που έχει 98 θετικές εγγραφές και μόνο 2 αρνητικές. Αυτή η κλάση είναι 2-διαφορετική και έχει μεγαλύτερη εντροπία από το σύνολο του πίνακα, ικανοποιώντας έτσι κάθε *l*-διαφορετικότητα εντροπίας που κάποιος μπορεί να εφαρμόσει. Ωστόσο, ένα τυχαίο άτομο στην κλάση θα θεωρείται 98% θετικός, αντί για 1%. Στην πραγματικότητα, αυτή η κλάση ισοδυναμίας έχει ακριβώς την ίδια διαφορετικότητα με μια τάξη που έχει 2 θετικές και 98 αρνητικές εγγραφές, παρόλο που οι δύο κλάσεις παρουσιάζουν πολύ διαφορετικά επίπεδα κινδύνων ιδιωτικότητας.

- **Επίθεση ομοιότητας**

Ένα άλλο πιθανό πρόβλημα παρουσιάζεται όταν οι τιμές ευαίσθητων χαρακτηριστικών είναι διακριτές αλλά σημασιολογικά παρόμοιες. Για παράδειγμα, μία κλάση ισοδυναμίας μπορεί να περιέχει διάφορους τύπους καρκίνων. Σε αυτή την περίπτωση γνωρίζουμε ότι όλοι στην ομάδα έχουν καρκίνο. Ένα άλλο παράδειγμα είναι ένα γνώρισμα που περιγράφει τον μισθό ενός συνόλου εργαζομένων. Κάθε άτομο σε μια κλάση μπορεί να έχει μια μοναδική τιμή, αλλά το συνολικό διάστημα μπορεί ακόμα να είναι μικρό. Ως εκ τούτου θα μπορούσαμε να μαντέψουμε με ακρίβεια το μισθό κάθε ατόμου λαμβάνοντας τον μέσο όρο στην κλάση.



### 3.4 $t$ -Εγγύτητα

Μια βελτίωση της  $l$ -διαφορετικότητας που προσπαθεί να λύσει τα προβλήματα που παρουσιάσαμε ονομάζεται  $t$ -εγγύτητα. Εδώ προσπαθούμε να διασφαλίσουμε ότι η κατανομή ενός ευαίσθητου γνωρίσματος στην κλάση είναι ίδια με την κατανομή του σε ολόκληρο τον πληθυσμό. Αυτό έχει ως αποτέλεσμα τον ακόλουθο ορισμό.

**Ορισμός 3.9.** Μια κλάση ισοδυναμίας  $E$  θα έχει  $t$ -εγγύτητα εάν η απόσταση μεταξύ της κατανομής ενός ευαίσθητου γνωρίσματος στην κλάση αυτή και της κατανομής του γνωρίσματος σε ολόκληρο τον πίνακα δεν υπερβαίνει ένα κατώφλι  $t$ .

**Ορισμός 3.10.** Ένα σύνολο δεδομένων θα έχει  $t$ -εγγύτητα, αν όλες οι κλάσεις ισοδυναμίας του έχουν  $t$ -εγγύτητα.

Το ακριβές μέτρο απόστασης που χρησιμοποιείται δεν έχει μεγάλη σημασία στο πλαίσιο της εργασίας αυτής. Παρόλο που η εγγύτητα είναι μια σαφής βελτίωση των προηγούμενων τεχνικών προστασίας της ιδιωτικότητας, θα δούμε στη συνέχεια ότι εξακολουθεί να είναι ευάλωτη σε μια επίθεση συνδεσιμότητας. Αυτό συμβαίνει κυρίως διότι η  $t$ -εγγύτητα - όπως επίσης και οι προηγούμενες μέθοδοι - επικεντρώνονται σε στατικά δεδομένα που μένουν αμετάβλητα. Επομένως, περιορίζονται σε μια και μόνο δημοσίευση και δεν υποστηρίζουν την αναδημοσίευση μιας νέας προβολής της βάσης δεδομένων.

Αυτό προκαλεί προβλήματα επειδή μπορεί να υπάρχει μια βάση δεδομένων, από διαφορετική εταιρία/οργανισμό, όπου είναι αποθηκευμένες σχεδόν οι ίδιες πληροφορίες. Αν αυτός ο οργανισμός κοινοποιήσει μια ανωνυμοποιημένη βάση δεδομένων, η παραδοχή μιας «μοναδικής» δημοσίευσης παύει πλέον να ισχύει. Στην πραγματικότητα δηλαδή είναι αδύνατον να αποτρέψουμε πολλαπλές δημοσιεύσεις της ίδιας βάσης δεδομένων, και επομένως μηχανισμός ιδιωτικότητας που προϋποθέτει μοναδική δημοσίευση, θεωρείται πρακτικά επισφαλής.

### 3.5 Επίθεση Τομής

Μια επίθεση που δεν μπορεί να αντιμετωπίσει καμία από τις προαναφερθείσες τεχνικές, και ουσιαστικά καμία εκ των μεθόδων γενίκευσης, είναι η επίθεση τομής. Είναι μια ιδιαίτερη περίπτωση επίθεσης σύνθεσης, όπου συνδυάζονται δύο ή περισσότερες κοινοποιήσεις της βάσης δεδομένων ή επικαλυπτόμενες βάσεις, δημοσιευμένες από διαφορετικές οντότητες [Ganta et al., 2009]. Η επίθεση τομής εξαρτάται από μια σημαντική ιδιότητα που μπορεί να διαθέτει ένας μηχανισμός προστασίας ιδιωτικότητας, εν ονόματι εντοπισμός (locatability).

**Ορισμός 3.11.** (Εντοπισμός)

Έστω  $Q$  το σύνολο των quasi-identifiers τιμών μιας εγγραφής στην αρχική βάση δεδομένων  $D$ . Ένας μηχανισμός ανωνυμοποίησης  $M$ , που παράγει μια ανωνυμοποιημένη προβολή  $R$



δεδομένης της βάσης  $D$ , ικανοποιεί την ιδιότητα του εντοπισμού αν κάποιος μπορεί να ταυτοποιήσει ένα σύνολο πλειάδων  $\{t_1, t_2, \dots, t_k\}$  του  $R$  που να αντιστοιχούν στο  $Q$ .

Εν ολίγοις, δηλώνεται ότι, δεδομένων των τιμών των quasi-identifiers ενός ατόμου, μπορούμε να βρούμε την κλάση ισοδυναμίας που ανήκει.

Όπως αναφέραμε και στην παράγραφο 3.2, οι περισσότεροι αλγόριθμοι ανωνυμοποίησης παράγουν μια ανώνυμη βάση, δεδομένης την αρχικής. Η πλειοψηφία αυτών των μηχανισμών ικανοποιεί την ιδιότητα εντοπισμού, πράγμα που δεν ισχύει αναγκαστικά για όλους. Για αυτούς που δεν ικανοποιούν αυστηρά την ιδιότητα, τα πειράματα αποκαλύπτουν ότι απλές επιθέσεις συνδυαστικού τύπου μπορούν ακόμα να εντοπίσουν την κλάση ισοδυναμίας ενός ατόμου με ικανοποιητικού βαθμού πιθανότητα. Η επίθεση τομής προϋποθέτει ότι ο μηχανισμός που χρησιμοποιείται για τη δημιουργία της ανώνυμης βάσης ικανοποιεί την ιδιότητα εντοπισμού.

Παρακάτω παρουσιάζουμε την αλγόριθμο της επίθεσης τομής:

**Data:**  $R_1, R_2, \dots, R_n$  Οι  $n$  ανώνυμες προβολές  
 $P$  ένα σύνολο ατόμων, κοινό στις  $n$  δημοσιεύσεις  
**for**  $i$  **in**  $P$  **do**  
    **for**  $j=1$  **to**  $n$  **do**  
         $e_{ij} \leftarrow \text{GetEqClass}(R_j, i)$   
         $s_{ij} \leftarrow \text{SensitiveValueSet}(e_{ij})$   
    **end**  
     $S_i \leftarrow s_{i1} \cap s_{i2} \cap \dots \cap s_{in}$   
**end**  
**return**  $S_1, S_2, \dots, S_{|P|}$

**Algorithm 1:** Intersection Attack

Έστω  $R_1, R_2, \dots, R_n$  οι  $n$  ανώνυμες προβολές μιας βάσης  $D$ . Έστω  $P$  ένα επικαλυπτόμενο υποσύνολο ατόμων, των οποίων γνωρίζουμε την τιμή των quasi-identifiers, που εμφανίζεται σε όλες τις δημοσιεύσεις. Η συνάρτηση  $\text{GetEqClass}$  επιστρέφει την κλάση ισοδυναμίας που ανήκει το άτομο, βασιζόμενη στις τιμές των quasi-identifiers του. Εφόσον υποθέσαμε ότι ο μηχανισμός προστασίας ιδιωτικότητας που παράγει τις ανωνυμοποιημένες προβολές  $R_j$  ικανοποιεί την ιδιότητα εντοπισμού, αυτή η συνάρτηση υπάρχει σε κάθε περίπτωση. Η συνάρτηση  $\text{SensitiveValueSet}$  επιστρέφει το σύνολο των (ξεχωριστών) ευαίσθητων τιμών για τα άτομα σε μια δεδομένη κλάση ισοδυναμίας.

Αυτό που κάνει ο βρόχος είναι, από κάθε ανώνυμη προβολή να εξάγει το πιθανό σύνολο τιμών για το ευαίσθητο χαρακτηριστικό. Στη συνέχεια παίρνουμε την τομή όλων αυτών των συνόλων. Αν καταλήξουμε σε μία μόνο τιμή, μόλις αποκαλύψαμε το ευαίσθητο χαρακτηριστικό.

Έστω για παράδειγμα, ότι θέλουμε να εξάγουμε συμπέρασμα για την ασθένεια ενός ατόμου  $a$ , για το οποίο γνωρίζουμε ότι βρίσκεται στην ανωνυμοποιημένη βάση  $R$  και στην επίσης ανωνυμοποιημένη  $B$ . Έχουμε φτάσει στο συμπέρασμα με βάση την επεξεργασία της  $R$ , ότι ο  $a$  έχει Καρκίνο ή Διαβήτη, ενώ από την  $B$ , ότι πάσχει από Καρδιακό νόσημα είτε από Καρκίνο. Προκύπτει λοιπόν το συμπέρασμα:

$$S_a = \{\text{Καρκίνος, Διαβήτης}\} \cap \{\text{Καρδιά, Καρκίνος}\} = \{\text{Καρκίνος}\}$$

Προφανώς η ιδιωτικότητα του ατόμου  $a$  έχει παραβιαστεί. Αυτή είναι μια «τέλεια παραβίαση» επειδή μπορούμε να συμπεράνουμε την ακριβή ευαίσθητη τιμή του ατόμου. Με άλλα λόγια ο επιτιθέμενος μαθαίνει την τιμή του ευαίσθητου γνωρίσματος ενός ατόμου με βεβαιότητα 100%. Μια άλλη μορφή παραβίασης είναι η «μερική παραβίαση», η οποία μπορεί να συμβεί όταν ο επιτιθέμενος είναι σε θέση να συμπυκνώσει τις πιθανές ευαίσθητες τιμές σε λίγες μόνο, οι οποίες θα μπορούσαν να αποκαλύψουν πολλές πληροφορίες.

Αν πάρουμε πάλι ως παράδειγμα την αποκάλυψη της ασθένειας του ατόμου  $a$  και προκύψει μετά την τομή των αποτελεσμάτων:

$$S_a = \{\text{Υπέρταση, Ανεύρυσμα, Φλεβοθρόμβωση}\}$$

προκύπτει το συμπέρασμα ότι το άτομο αυτό πάσχει από μια καρδιακή νόσο. Η πιθανότητα εύρεσης της νόσου σε αυτή την περίπτωση είναι 33%.

Η επίθεση τομής εφαρμόστηκε σε δεδομένα ανωνυμοποιημένα με την μέθοδο της  $k$ -ανωνυμίας,  $l$ -διαφορετικότητας και  $t$ -εγγύτητας. Το συμπέρασμα ήταν ότι και οι τρεις μηχανισμοί αποτυγχάνουν να προστατεύσουν την ιδιωτικότητα όλων των ατόμων. Η  $l$ -διαφορετικότητα και η  $t$ -εγγύτητα αποδίδουν καλύτερα από την  $k$ -ανωνυμία, ωστόσο παρατηρούνται ακόμη περιπτώσεις παραβίασης. Κάτι που προκύπτει επιπλέον, είναι ότι για την  $l$ -διαφορετικότητα και  $t$ -εγγύτητα πρέπει να δημιουργηθούν μεγάλες κλάσεις ισοδυναμίας, με αποτέλεσμα μεγάλη απώλεια πληροφορίας.

Το συμπέρασμα λοιπόν είναι ότι όλες οι μέθοδοι γενίκευσης παρέχουν ικανοποιητική ιδιωτικότητα, προστατεύοντάς τα δεδομένα είτε από αποκάλυψη ταυτότητας, είτε από αποκάλυψη τιμής ευαίσθητου γνωρίσματος. Σε ιδιέταιρες καταστάσεις όμως, καμία από τις μεθόδους αυτές δεν μπορεί να αντιμετωπίσει την επίθεση τομής.

## 3.6 Αλγόριθμοι Γενίκευσης

Οι περισσότεροι μηχανισμοί ανωνυμοποίησης που παρουσιάσαμε στα προηγούμενα κεφάλαια, έχουν προσεγγιστεί κατά καιρούς με αλγορίθμους σε διάφορες γλώσσες προγραμματισμού, και έχουν αξιολογηθεί σε έρευνες. Στην παράγραφο αυτή επιδιώκουμε να συγκεντρώσουμε τις βέλτιστες και ταχύτερες προσεγγίσεις.

Τα τελευταία χρόνια έχουν δημοσιευτεί δεκάδες άρθρα που περιγράφουν και αναλύουν αλγορίθμους ανωνυμοποίησης συνόλων δεδομένων. Στόχος των αλγορίθμων αυτών είναι η δημιουργία μιας τροποποιημένης έκδοσης του συνόλου δεδομένων, έτσι ώστε η ιδιωτικότητα των εγγραφών να προστατεύεται επαρκώς, ενώ ταυτόχρονα, να διατηρείται η χρηστικότητα των δημοσιευμένων δεδομένων.

### 3.6.1 Mondrian

Ο Mondrian είναι ένας από τους κλασικότερους αλγορίθμους γενίκευσης και παρουσιάστηκε αρχικά ως εφαρμογή του μοντέλου της  $k$ -ανωνυμίας. Ως είσοδο δέχεται το σύνολο των αρχικών δεδομένων και επιστρέφει την βέλτιστη γενίκευση<sup>5</sup> του. Θεωρείται ένας από τους πιο σύγχρονους αλγορίθμους ανωνυμοποίησης και είναι ελκυστικός τόσο λόγω των λύσεων που παρέχει όσο και του χαμηλού χρόνου εκτέλεσης.

Η γενική ιδέα πίσω από τον αλγόριθμο είναι ένας top-down διαχωρισμός των δεδομένων. Όλα τα αρχεία ανήκουν αρχικά στην ίδια κλάση ισοδυναμίας και αναδρομικά επιλέγεται μια διάσταση για να χωριστούν οι κλάσεις ισοδυναμίας, μέχρις ότου δεν υπάρχει καμία διάσταση στην οποία μπορεί να χωριστεί η κλάση για να παράγει έγκυρες  $k$ -ανώνυμες συστάδες.

Ο αλγόριθμος ακολουθεί την εξής διαδικασία:

- Ορίζονται οι περιοχές που καλύπτουν το χώρο των πεδίων τιμών του quasi-identifier
- Επιλέγεται η διάσταση με την οποία θα γίνει ο διαχωρισμός των κλάσεων. Υπάρχουν πολλοί τρόποι επιλογής, συνήθως όμως επιλέγεται η διάσταση με το μεγαλύτερο εύρος τιμών .
- Εφαρμόζεται ο διαχωρισμός κατά την παραπάνω επιλεγμένη διάσταση, βάσει της μέσης τιμής του αντίστοιχου γνωρίσματος, έτσι ώστε οι τιμές που είναι μικρότερες ή ίσες από αυτήν να βρίσκονται στην αριστερή κλάση και οι υπόλοιπες να βρίσκονται στην δεξιά κλάση ισοδυναμίας.
- Η διαδικασία επαναλαμβάνεται για κάθε μία από τις δύο προκύπτουσες κλάσεις ισοδυναμίας αναδρομικά μέχρι να μην υπάρχει άλλη επιτρεπόμενη πολυδιάστατη τομή για διαχωρισμό σε καμία διάσταση.

<sup>5</sup>Ως βέλτιστη γενίκευση ορίζεται το ποσό γενίκευσης κατά το οποίο παρέχεται η μέγιστη δυνατή ιδιωτικότητα, εξασφαλίζοντας την ελάχιστη απώλεια πληροφορίας

- Ως έξοδος, προκύπτει ο βέλτιστος διαχωρισμός και συνεπώς η κατάλληλη πολυδιάστατη γενίκευση που θα χρησιμοποιηθεί για να ανακωδικοποιηθούν τα δεδομένα.

Σε ψευδογλώσσα:

**Συνάρτηση** *partition\_anon*(*P*)

**Data:** Διαμέριση *P*

**if** είναι αδύνατη περεταίρω πολυδιάστατη τομή της *P* **then**

**return** *P*

**else**

$dim \leftarrow choosedimension(P)$

$lhs \leftarrow \{t \in \text{διαμεριση} : t.dim = false\}$

$rhs \leftarrow \{t \in \text{διαμεριση} : t.dim = true\}$

**return** *partition\_anon*(*lhs*)  $\cup$  *partition\_anon*(*rhs*)

**end**

**Algorithm 2:** Mondrian

### 3.6.2 Συσταδοποίηση (Clustering)

Πέραν της χρήσης του *Mondrian* για την ανωνυμοποίηση των δεδομένων, έχουν υιοθετηθεί μοντέλα από τον τομέα της εξόρυξης δεδομένων. Κλασικό παράδειγμα είναι η εφαρμογή αλγορίθμων συσταδοποίησης, η οποία θεωρείται πολλά υποσχόμενη μέθοδος.

Για την ανάπτυξη τέτοιων αλγορίθμων είναι απαραίτητο να ξεπεραστούν τα προβλήματα εγγύτητας με την αναζήτηση πλησιέστερων γειτόνων που χρησιμοποιεί κάθε φορά για να επιλέξει εγγραφές για συσταδοποίηση. Χρησιμοποιείται η εξής διαδικασία: αντί να βρεθεί ο πλησιέστερος γείτονας μεταξύ όλων των εγγραφών, αρκεί να βρεθεί ο πλησιέστερος ανάμεσα σε κάποιο δείγμα ενός σταθερού αριθμού εγγραφών και να το δεχτεί.

### 3.6.3 k - OPTIMIZE

Ο αλγόριθμος αυτός παρουσιάστηκε ως η βέλτιστη-ταχύτερη μέθοδος *k*-ανωνυμοποίησης. Χρησιμοποιεί προτάσεις της Θεωρίας Γραφημάτων, ενώ αποφεύγονται δαπανηροί αλγριθμικοί υπολογισμοί, όπως η ταξινόμηση [Bayardo and Agrawal, 2005].

Ο αλγόριθμος ξεκινά υπολογίζοντας μια τυχαία ανωνυμοποίηση του συνόλου. Στη συνέχεια εισέρχεται σε μια φάση γενίκευσης στην οποία οι τιμές αφαιρούνται διαδοχικά (πάντοτε επιλέγοντας εκείνη που βελτιώνει περισσότερο το χρονικό κόστος) έως ότου αυτό να μην μπορεί πλέον να βελτιωθεί. Στη συνέχεια, μεταβαίνει σε μια φάση «εξειδίκευσης» στην οποία οι τιμές προστίθενται επαναληπτικά (και πάλι πάντα επιλέγοντας εκείνη που παρέχει τη μεγαλύτερη βελτίωση του κόστους) έως ότου δεν είναι δυνατή η βελτίωση. Ο αλγόριθμος εκτελεί επαναληπτικά αυτές τις δύο φάσεις μέχρις ότου καμία φάση δεν είναι ικανή να βελτιώσει το score (υποδηλώνοντας ότι επιτυγχάνεται τοπικό ελάχιστο). Στο

σημείο αυτό, καταγράφεται το κόστος ανωνυμοποίησης και ο αλγόριθμος επαναλαμβάνει ολόκληρη τη διαδικασία. Σημειώνεται ότι ο αλγόριθμος αυτός δεν έχει συνθήκες διακοπής και αντίθετα προσπαθεί συνεχώς να βελτιώσει την καλύτερη λύση που βρέθηκε μέχρι να σταματήσει ο χρήστης.

```

K-OPTIMIZE( $k$ , head set  $H$ , tail set  $T$ , best cost  $c$ )
    ;; This function returns the lowest cost of any
    ;; anonymization within the sub-tree rooted at
    ;;  $H$  that has a cost less than  $c$  (if one exists).
    ;; Otherwise, it returns  $c$ .
     $T \leftarrow \text{PRUNE-USELESS-VALUES}(H, T)$ 
     $c \leftarrow \min(c, \text{COMPUTE-COST}(H))$ 
     $T \leftarrow \text{PRUNE}(H, T, c)$ 
     $T \leftarrow \text{REORDER-TAIL}(H, T)$ 
    while  $T$  is non-empty do
         $v \leftarrow$  the first value in the ordered set  $T$ 
         $H_{\text{new}} \leftarrow H \cup \{v\}$ 
         $T \leftarrow T - \{v\}$       ;; preserve ordering
         $c \leftarrow \text{K-OPTIMIZE}(k, H_{\text{new}}, T, c)$ 
         $T \leftarrow \text{PRUNE}(H, T, c)$ 
    return  $c$ 

PRUNE( $k$ , head set  $H$ , tail set  $T$ , best cost  $c$ )
    ;; This function creates and returns a new
    ;; tail set by removing values from  $T$  that
    ;; cannot lead to anonymizations with cost
    ;; lower than  $c$ 
    if  $\text{COMPUTE-LOWER-BOUND}(k, H, H \cup T) \geq c$ 
        then return  $\emptyset$ 
     $T_{\text{new}} \leftarrow T$ 
    for each  $v$  in  $T$  do
         $H_{\text{new}} \leftarrow H \cup \{v\}$ 
        if  $\text{PRUNE}(H_{\text{new}}, T_{\text{new}} - \{v\}, c) = \emptyset$ 
            then  $T_{\text{new}} \leftarrow T_{\text{new}} - \{v\}$ 
    if  $T_{\text{new}} \neq T$  then return  $\text{PRUNE}(H, T_{\text{new}}, c)$ 
    else return  $T_{\text{new}}$ 

```

ΣΧΗΜΑ 3.3: Μοντελοποίηση του αλγορίθμου k-optimize

Οφείλουμε να παρατηρήσουμε ότι αρκετοί από τους αλγορίθμους αυτούς χρησιμοποιούνται στην πράξη, ενώ διαθέσιμοι σε open-source μορφή υπάρχουν από πολλές πηγές, όπως Anonymization Toolbox από το Πανεπιστήμιο του Ντάλας και το ARX - Powerful Data Anonymization, τα οποία υποστηρίζουν κυρίως τα μοντέλα  $k$ -ανωνυμίας και  $l$ -διαφορετικότητας. Στις βιβλιοθήκες ανωνυμοποίησης συνήθως συναντάμε και αλγορίθμους όπως ο Datafly και ο Incognito, οι οποίοι όμως δεν είναι τόσο αποτελεσματικοί όσο αυτοί που αναφέραμε παραπάνω.



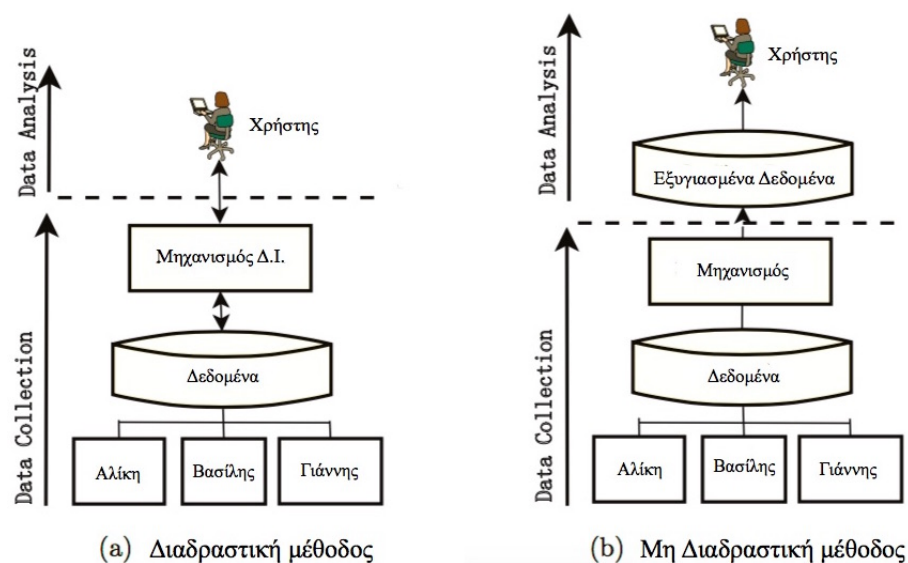
Ένα βασικό σημείο στο οποίο υστερούν οι προαναφερθέντες αλγόριθμοι αποτελεί ο χρόνος εκτέλεσης, όπου κάποιες φορές είναι πιθανό να καθιστά αδύνατη την πρακτική εφαρμογή τους. Πρόσφατες μελέτες διερευνούν τη χρήση συμπληρωματικών μεθόδων ώστε να καταστούν οι μηχανισμοί αυτοί αποδοτικότεροι σε σχέση με την υπολογιστική τους πολυπλοκότητα, κάνοντάς τους φιλικότερους στη χρήση σε πρακτικές εφαρμογές. [\[Mohammadian et al., 2014\]](#)

## Κεφάλαιο 4

# Διαφορική Ιδιωτικότητα

Στο προηγούμενο κεφάλαιο αναλύσαμε τεχνικές στις οποίες ο διαχειριστής του συνόλου δεδομένων δημοσιεύει μια εξυγιανσμένη εκδοχή τους. Ωστόσο, είδαμε ότι η ανωνυμοποίηση του συνόλου δεδομένων αρκετές φορές δεν επαρκεί για να προστατέψει τα δεδομένα από έναν ισχυρό και καλά προετοιμασμένο επιτιθέμενο. Σε μια βάση  $n$  στοιχείων για παράδειγμα, ένας γνώστης συγκεκριμένου γνωρίσματος των  $n - 1$  αντικειμένων, μπορεί εύκολα να συμπεράνει την τιμή του γνωρίσματος του ατόμου που απομένει. Στη συνέχεια θα παρουσιάσουμε και θα αναλύσουμε την διαφορική ιδιωτικότητα, μια διαδραστική μέθοδο η οποία προστατεύει τα δεδομένα, ακόμη και από επιτιθέμενους με πρότερη γνώση.

### 4.1 Η έννοια της διαδραστικότητας



ΣΧΗΜΑ 4.1: Η έννοια της διαδραστικότητας



Μία ελπιδοφόρα προσέγγιση που θα μπορούσε να καλύψει τα κενά των τεχνικών ομαδοποίησης που παρουσιάσαμε, είναι να μεσολαβήσει της πρόσβασης στη βάση δεδομένων μια αξιόπιστη διασύνδεση η οποία θα απαντάει τις επερωτήσεις των αναλυτών.

Έχουν προταθεί κατά καιρούς αρκετές κρυπτογραφικές μέθοδοι, κυρίως μέσω προσθήκης θορύβου, ώστε να διασφαλιστεί η ιδιωτικότητα των μηχανισμών επερωτήσεων και να εξασφαλιστεί η ανώνυμη επικοινωνία Βάσης-Αναλυτή. Σε αυτό το σημείο όμως, είναι εύλογο το ερώτημα:

«Μήπως οι τιμές εξόδου των μηχανισμών αυτών, ήδη αποκαλύπτουν υπερβολικά πολλές πληροφορίες;»

Μία λύση στο πρόβλημα αυτό είναι η τυχαιοποίηση των αποτελεσμάτων του μηχανισμού. Υποθέστε ότι κάποιος θέλει να μάθει αν έχουμε κάποιο συγκεκριμένο χαρακτηριστικό. Σε κάθε τέτοια ερώτηση απαντάμε με τον ακόλουθο τρόπο:

1. Ρίχνουμε ένα νόμισμα
2. Αν έρθει γράμματα, τότε απαντάμε αληθώς.
3. Αν έρθει κορώνα, τότε ρίχνουμε ένα δεύτερο νόμισμα και απαντάμε «Ναι» αν έρθει κορώνα και «Όχι» αν έρθει γράμματα.

Συμπεραίνουμε ότι δεν μπορεί να προκύψει ακριβές συμπέρασμα για το αν έχουμε ή όχι το χαρακτηριστικό. Αυτό το παράδειγμα παρουσιάζει έναν απλό μηχανισμό τυχαιοποίησης των αποτελεσμάτων των επερωτήσεων. Τον φορμαλισμό της σκέψης αυτής έρχεται να εκφράσει η διαφορική ιδιωτικότητα.

## 4.2 Θεμελίωση

Πρωτού δώσουμε τον ορισμό της διαφορικής ιδιωτικότητας θα δούμε δύο ιδιότητες όπου κάθε μέθοδος ιδιωτικότητας οφείλει να κατέχει.

### **Ανθεκτικότητα σε πρότερη γνώση:**

Στο εισαγωγικό παράδειγμα είδαμε ότι υπάρχει πιθανότητα για έναν επιτιθέμενο να έχει γνώση για όλα σχεδόν τα στοιχεία της βάσης. Γενικά οφείλουμε πάντα να λογαριάζουμε οποιαδήποτε πληροφορία μπορεί ήδη να γνωρίζει κάποιος για ένα υποσύνολο δεδομένων. Επειδή είναι πολύ δύσκολο να προσδιοριστεί ποσοτικά, υποθέτουμε ότι ιδανικά ο επιτιθέμενος γνωρίζει τα πάντα, εκτός από τις ατομικές πληροφορίες ενός ατόμου.

### **Ανθεκτικότητα σε πολλαπλές εφαρμογές - Σύνθεση (composition):**

Αν εφαρμόσουμε την μέθοδο αρκετές φορές σε συγγενείς βάσεις, ο επιτιθέμενος δεν επιτρέπεται να μπορεί να συνδυάσει τα αποτελέσματα ώστε να εξακριβώσει την ταυτότητα



κάποιου ατόμου. Εκεί είναι που αποτυγχάνουν οι περισσότερες μη διαδραστικές τεχνικές. Θα αναλύσουμε περισσότερο την ιδιότητα αυτή μετά τον ορισμό της διαφορικής ιδιωτικότητας.

Έστω  $x = (x_1, x_2, \dots, x_n)$  ένα σύνολο δεδομένων, και  $x_i$  μια εγγραφή του.

#### Ορισμός 4.1. (Γειτνίαση)

Δύο σύνολα δεδομένων  $x, x'$  **γειτνιάζουν**, αν για κάθε στοιχείο τους ισχύει  $x_i = x'_i$  για κάθε  $i \in [1, n]$ , εκτός ενός το πολύ στοιχείου.

Δηλαδή δυο γειτονικά σύνολα δεδομένων πρέπει να διαφέρουν το πολύ σε ένα στοιχείο τους. Συμβολίζουμε  $x \sim x'$  [Vadhan, 2017].

Όταν προτάθηκε για πρώτη φορά διαφορική ιδιωτικότητα, η σχέση γειτνίασης καθορίστηκε με έναν ελαφρώς διαφορετικό τρόπο: δύο βάσεις δεδομένων είναι γειτονικές εάν και μόνο εάν μια βάση δεδομένων είναι αποτέλεσμα της προσθήκης / αφαίρεσης ενός χρήστη από την άλλη βάση δεδομένων [Dwork, 2006b].

Το κίνητρο πίσω από τον αρχικό ορισμό είναι να γίνει απόκρυψη της συμμετοχής οποιουδήποτε ατόμου στη βάση δεδομένων. Ένα τυπικό παράδειγμα είναι μια βάση δεδομένων για ασθενείς με συγκεκριμένο τύπο ασθένειας. Ο ορισμός γενικεύει την αρχική έννοια της σχέσης γειτνίασης προκειμένου να χειριστεί βάσεις δεδομένων που αποτελούνται από αριθμητικές τιμές. Στην πραγματικότητα, ο ορισμός αυτός μπορεί να επεκταθεί περαιτέρω για να ενσωματώσει πιο πολύπλοκα αντικείμενα, όπως διανυσματικά μεγέθη.

Τονίζουμε ότι η Διαφορική Ιδιωτικότητα είναι σε θέση να εγγυηθεί ότι το αποτέλεσμα υπολογισμών σε μια βάση δεδομένων δεν μεταβάλλεται πολύ όταν οποιοσδήποτε μεμονωμένος χρήστης στη βάση δεδομένων αλλάζει τις πληροφορίες του. Με άλλα λόγια, η διατήρηση της ιδιωτικότητας είναι ισοδύναμη με την απόκρυψη αλλαγών στη βάση δεδομένων

#### Ορισμός 4.2. (Τυχαιοποιημένη συνάρτηση)

Έστω  $X$  το σύνολο όλων των πεπερασμένων συνόλων δεδομένων και  $B$  το σύνολο όλων των τυχαίων μεταβλητών με εικόνα  $B$ . Ορίζουμε μια τυχαιοποιημένη συνάρτηση  $M$ :

$$M : X \longrightarrow B$$

Στη βιβλιογραφία συναντάμε την έννοια της τυχαιοποιημένης συνάρτησης και ως τυχαιοποιημένου αλγορίθμου ή μηχανισμού.

#### Ορισμός 4.3. (Διαφορική Ιδιωτικότητα)

Δεδομένου  $\epsilon > 0$ , μια τυχαιοποιημένη συνάρτηση  $M$  αποδίδει  $\epsilon$ -διαφορική ιδιωτικότητα, αν για κάθε ζεύγος συνόλων δεδομένων  $x, x'$  με  $x \sim x'$  και κάθε  $S \subseteq R_M$ , όπου  $R_M$  το σύνολο τιμών της  $M$ , ισχύει:

$$P[M(x) \in S] \leq e^\epsilon \cdot P[M(x') \in S]$$

Ως  $\epsilon$  θεωρούμε έναν μικρό, όχι αμελητέο, θετικό αριθμό, συνήθως στο διάστημα  $(0.01, \ln 2)$ . Όσο μικρότερη η τιμή του, τόσο μεγαλύτερη η προστασία των εγγραφών. Είναι προφανές ότι ο ορισμός παύει να είναι χρήσιμος αν  $\epsilon < \frac{1}{n}$ . Επίσης θεωρούμε το  $n$  ως καθολικά γνωστή πληροφορία.

Παρατηρούμε ότι η σχέση μπορεί να γραφεί ισοδύναμα

$$P[M(x') \in S] \leq e^\epsilon \cdot P[M(x) \in S]$$

εξ αιτίας της συμμετρίας που προκύπτει από τον ορισμό της γειτνίασης των βάσεων. Η έννοια της διαφορικής ιδιωτικότητας μας βεβαιώνει ότι ο επιτιθέμενος δεν μπορεί να συμπεράνει από την εικόνα της  $M$ , με μεγάλη πιθανότητα, αν τα δεδομένα από μια και μόνο εγγραφή έχουν μεταβληθεί.

Σε ορισμένες περιπτώσεις είναι χρήσιμο να θεωρίσουμε μια γενίκευση του ορισμού:

**Ορισμός 4.4.** Δεδομένου  $\epsilon, \delta > 0$ , μια τυχαιοποιημένη συνάρτηση  $M$  αποδίδει  $(\epsilon, \delta)$ -διαφορική ιδιωτικότητα, αν για κάθε ζεύγος συνόλων δεδομένων  $x, x'$  με  $x \sim x'$  και κάθε  $S \subseteq R_M$ , όπου  $R_M$  το σύνολο τιμών της  $M$ , ισχύει:

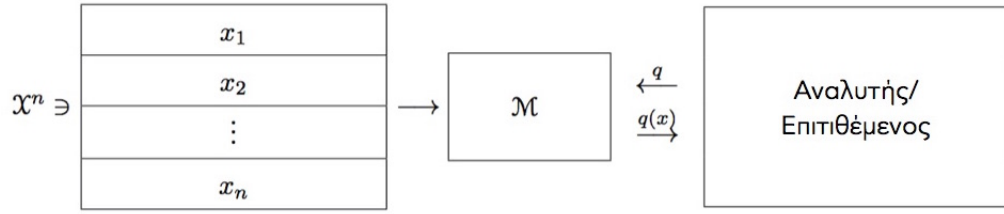
$$P[M(x) \in S] \leq e^\epsilon \cdot P[M(x') \in S] + \delta$$

Προφανώς, όσο μεγαλύτερο είναι το  $\delta > 0$  τόσο πιο εύκολα ένας επιτιθέμενος μπορεί να ξεχωρίσει το ποια βάση είναι η  $x'$  και ποιά η  $x$ . Ο αρχικός ορισμός (με  $\delta = 0$  δηλαδή) είναι ασφαλέστερος. Συνοπτικά, ο όρος  $\delta$  αντιπροσωπεύει την πιθανότητα ότι μερικά άτομα μπορεί να συμβεί να χάσουν περισσότερη ιδιωτικότητα από ό, τι τα υπόλοιπα, ότι το πλασιαστικό φράγμα δεν ισχύει για όλους. Αν το  $\delta$  είναι πολύ μικρό, ο κίνδυνος αυτός είναι πολύ μικρός.

**Counting Query:** Ένας βασικός τύπος επερωτήσεων που θα εξετάσουμε εκτενώς είναι το counting query, το οποίο καθορίζεται από μια δίτιμη απεικόνιση (predicate) στις εγγραφές του συνόλου  $q : X \rightarrow \{0, 1\}$ , και γενικεύεται σε ολόκληρο το σύνολο δεδομένων  $x \in X^n$ , υπολογίζοντας τον λόγο των εγγραφών που ικανοποιούν την απεικόνιση:

$$q(x) = \frac{1}{n} \sum_{i=1}^n q(x_i)$$

Όπως είδαμε και στο παράδειγμα της εισαγωγής, η ιδιωτικότητα δεν πρέπει να θεωρείται τετριμμένη ακόμη και αν δίνονται μόνο counting επερωτήσεις στη βάση, επειδή τα αποτελέσματα μπορούν να συνδυαστούν με συνέπεια την αποκάλυψη πληροφορίας για κάποια εγγραφή.



ΣΧΗΜΑ 4.2: Λειτουργία μηχανισμού ΔΙ

#### 4.2.1 Θεώρημα Σύνθεσης

Στην εισαγωγή του κεφαλαίου αναφέραμε ότι η διαφορική ιδιωτικότητα χαρακτηρίζεται από δύο βασικές ιδιότητες. Αρχικά, εξασφαλίζει ότι τα αποτελέσματα θα είναι ανεπηρέαστα από οποιαδήποτε εξωτερική γνώση. Δεύτερον, η μέθοδος παρουσιάζει ανθεκτικότητα στην μετεπεξεργασία. Όπως είδαμε, το αποτέλεσμα από την εφαρμογή ενός μηχανισμού ιδιωτικότητας κοινοποιείται, και εν συνεχεία αυτό μπορεί είτε να χρησιμοποιηθεί αυθαίρετα από άλλους είτε να χρησιμοποιηθεί ως δείγμα ώστε να εφαρμοστεί ένας νέος μηχανισμός ακόμη και από τον ίδιο χρήστη. Η ανθεκτικότητα στην επαναληπτική εφαρμογή μηχανισμών εγγυάται ότι δεν πρόκειται να χαθεί μέρος της ιδιωτικότητας, παρά την επαναχρησιμοποίηση [Dwork et al., 2010].

**Θεώρημα 4.5.** (Σύνθεση) Έστω μηχανισμός  $M : X \rightarrow B$  που παρέχει  $\epsilon$ -διαφορική ιδιωτικότητα. Τότε για κάθε συνάρτηση  $f$ , η σύνθεση  $f \circ M$  διατηρεί  $\epsilon$ -διαφορική ιδιωτικότητα.

Στη συνέχεια εισάγουμε δυο κανόνες σύνθεσης που χρησιμοποιούνται συχνά για την δημιουργία νέων μηχανισμών ιδιωτικότητας. Ο νόμος της ακολουθιακής σύνθεσης παρουσιάζεται παρακάτω και χρησιμοποιείται συνήθως όταν απαιτείται ο υπολογισμός πολλών πληροφοριών από την ίδια βάση.

**Θεώρημα 4.6.** (Ακολουθιακή σύνθεση - *sequential composition*)

Έστω μηχανισμός  $M_1 : X \rightarrow B$  που παρέχει  $\epsilon_1$ -διαφορική ιδιωτικότητα και μηχανισμός  $M_2 : X \rightarrow B$  που παρέχει  $\epsilon_2$ -διαφορική ιδιωτικότητα. Ορίζουμε νέο μηχανισμό  $M(x) = (M_1(x), M_2(x))$ , ο οποίος θα διατηρεί  $(\epsilon_1 + \epsilon_2)$ -διαφορική ιδιωτικότητα.

Ας υποθέσουμε ότι έχουμε μια βάση δεδομένων με τους μισθούς μιας εταιρίας. Έστω ότι κάποιος θέλει να κοινοποιήσει την μέση τιμή, αλλά και την διακύμανση των μισθών. Θα χρησιμοποιήσει έναν μηχανισμό ιδιωτικότητας για τον μέσο και έναν για την διακύμανση, οι οποίοι στη συνέχεια μπορούν να συνδυαστούν δίνοντας την τάξη της ιδιωτικότητας που παρουσιάζεται στο θεώρημα. Τέλος, παρατηρούμε ότι το θεώρημα συνεπάγεται την εξής

πρόταση: Όσο περισσότερες επερωτήσεις τίθενται στην ίδια βάση, τόσο περισσότερη ιδιωτικότητα χάνεται.

Για συγκεκριμένες εφαρμογές που απαιτείται επαναχρησιμοποίηση, ο μηχανισμός που θα επιθυμούσαμε να σχεδιάσουμε είναι ένα αποτέλεσμα «προσαρμοστικής» σύνθεσης διάφορων μηχανισμών. Παρόμοια εφαρμόζονται και οι αλγορίθμοι μηχανικής μάθησης, όπου το αποτέλεσμα κάθε βήματος εξαρτάται από τα προηγούμενα βήματα. Σε περιπτώσεις λοιπόν που απαιτούνται επαναληπτικοί υπολογισμοί χρησιμοποιούμε αυτόν τον κανόνα για διατήρηση της ιδιωτικότητας:

**Θεώρημα 4.7.** (Προσαρμοστική σύνθεση - *Adaptive composition*)

Έστω μηχανισμός  $M_1 : X \rightarrow B_1$  που παρέχει  $\epsilon_1$ -διαφορική ιδιωτικότητα και μηχανισμός  $M_2 : X \times B_1 \rightarrow B_2$  ώστε  $M_2(\cdot, y_1)$  που παρέχει  $\epsilon_2$ -διαφορική ιδιωτικότητα για κάθε  $y_1 \in B_1$ . Ορίζουμε νέο μηχανισμό  $M(x) = M_2(x, M_1(x))$ , ο οποίος θα παρέχει  $(\epsilon_1 + \epsilon_2)$ -διαφορική ιδιωτικότητα.

Ο κανόνας αυτός γενικεύει την έννοια της μετεπεξεργασίας που παρουσιάστηκε στο θεώρημα 4.5, επειδή οποιαδήποτε συνάρτηση  $f$  που δεν εξαρτάται από την βάση δεδομένων μπορεί να γίνει ένας μηχανισμός 0-διαφορικής ιδιωτικότητας. Επιπλέον αποτελεί γενίκευση και της ακολουθιακής σύνθεσης.

Τα παραπάνω θεωρήματα σύνθεσης ισχύουν και για τον γενικευμένο ορισμό της διαφορικής ιδιωτικότητας. Συγκεκριμένα, μετά την εφαρμογή της σύνθεσης σε μηχανισμούς που παρέχουν  $(\epsilon_1, \delta_1)$ -διαφορική ιδιωτικότητα και  $(\epsilon_2, \delta_2)$ -διαφορική ιδιωτικότητα, παρέχεται  $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$ -διαφορική ιδιωτικότητα. Παρακάτω παρουσιάζουμε ένα ακόμη θεώρημα σύνθεσης.

**Θεώρημα 4.8.** (Ανώτερη σύνθεση - *Advanced composition*)

Για κάθε  $\epsilon, \delta, \delta' \geq 0$ , ο μηχανισμός που δημιουργείται από την προσαρμοστική σύνθεση  $k$  μηχανισμών με  $(\epsilon, \delta)$ -διαφορική ιδιωτικότητα, παρέχει  $(\epsilon', k\delta + \delta')$ -διαφορική ιδιωτικότητα με

$$\epsilon' = \sqrt{2k \log(1/\delta')} \epsilon + k\epsilon(e^\epsilon - 1)$$

Παρατηρούμε ότι όταν το  $\epsilon$  είναι κοντά στο 0, τότε υπερισχύει ο πρώτος προσθετέος.

Τα παραπάνω θεωρήματα σύνθεσης καλύπτουν τόσο την επαναληπτική εφαρμογή μηχανισμών διαφορικής ιδιωτικότητας στην ίδια βάση, όσο και την επαναληπτική εφαρμογή τους σε διαφορετικές βάσεις δεδομένων που όμως μπορεί να περιέχουν πληροφορίες σχετικές με μια συγκεκριμένη εγγραφή.

### 4.3 Προσθήκη θορύβου και Μηχανισμοί

Έστω  $q$  μια επερώτηση τύπου counting query. Προσπαθώντας να προστατέψουμε την ιδιωτικότητα με προσθήκη θορύβου προκύπτει η σχέση:

$$M(x) = q(x) + noise$$

Πρέπει όμως να είμαστε προσεκτικοί στην ποσότητα θορύβου που θα προσθέσουμε. Η παρακάτω πρόταση περιγράφει επακριβώς τις δυο ακραίες καταστάσεις της σχέσης ποιότητας-ιδιωτικότητας:

«Δημοσιεύοντας ένα σύνολο δεδομένων στο ακέραιο παρέχεται η καλύτερη δυνατή ποιότητα, ενώ με την πλήρη απόκρυψη παρέχεται η καλύτερη δυνατή ιδιωτικότητα.»

Δεδομένων συνόλων  $x \sim x'$  μεγέθους  $n$ , παρατηρούμε ότι  $|q(x) - q(x')| \leq 1/n$ . Συμπεραίνουμε ότι θόρυβος μεγέθους  $1/\epsilon n$  θα είναι αρκετός για να κάνει τις συναρτήσεις (μηχανισμούς)  $M(x)$  και  $M(x')$  «ε-πανομοιότυπες» με την έννοια που απαιτείται από τον ορισμό της διαφορικής ιδιωτικότητας. Έτσι, για κάθε αποτέλεσμα  $y$  της επερώτησης  $q$ , πρέπει η κατανομή των απαντήσεων από τις  $x$  και  $x'$  να μοιάζει, κατά έναν παράγοντα της τάξης  $e^\epsilon$ . Αν  $z$  είναι ο θόρυβος που προστείνεται, παρατηρούμε:

$$y = q(x) + z \iff z = y - q(x)$$

και

$$y = q(x') + z' \iff z' = y - q(x')$$

Προκύπτει  $|z - z'| \leq 1/n$ . Βλέπουμε ότι αρκεί η συνάρτηση κατανομής θορύβου να μεταβάλλεται κατά το πολύ  $e^\epsilon$  σε διαστήματα μήκους  $1/n$ . Αυτό οδηγεί σε υιοθέτηση μοντέλων, όπως οι παρακάτω μηχανισμοί.

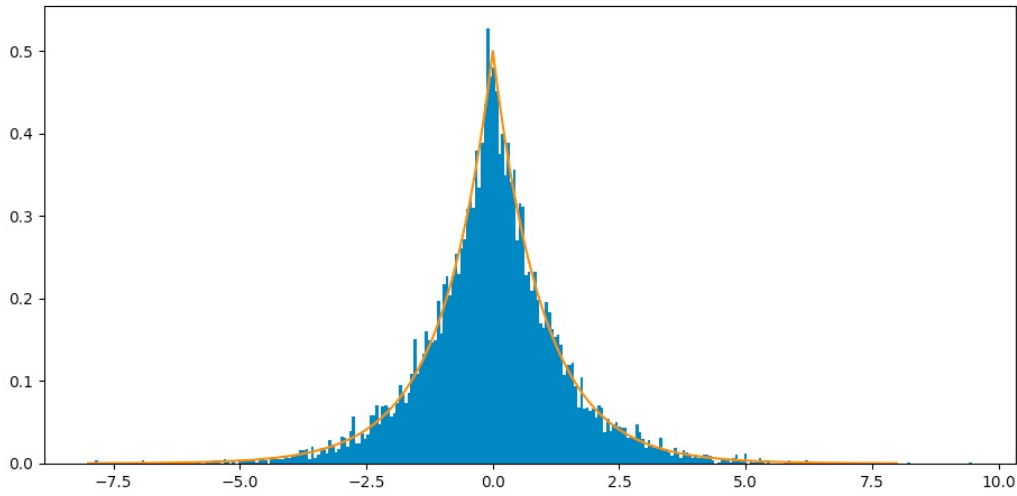
#### 4.3.1 Μηχανισμός Laplace

**Ορισμός 4.9.** (Κατανομή Laplace)

Η κατανομή Laplace με παράμετρο κλίμακας  $b > 0$  (θεωρώντας παράμετρο θέσης 0) ορίζεται ως η κατανομή με συνάρτηση πυκνότητας πιθανότητας

$$Lap(x|b) = \frac{1}{2b} e^{-\frac{|x|}{b}}$$

Σημειώνεται ότι η διακύμανση της κατανομής είναι  $\sigma^2 = 2b^2$  και η μέση τιμή 0.



ΣΧΗΜΑ 4.3: Συνάρτηση πυκνότητας πιθανότητας κατανομής Laplace, με παράμετρο κλίμακας  $b = 1$

#### Ορισμός 4.10. (Ευαισθησία)

Η  $l_1$ -ευαισθησία (sensitivity) μιας συνάρτησης  $f : X \rightarrow \mathbb{R}^k$  είναι:

$$\Delta = \max_{x \sim x'} \|f(x) - f(x')\|$$

Η ευαισθησία μίας επερώτησης  $q$  καταγράφει το μέγεθος με το οποίο τα δεδομένα μιας και μόνο εγγραφής μπορούν να μεταβάλλουν την επερώτηση στην «χειρότερη» περίπτωση, και αντίστοιχα, τον θόρυβο που πρέπει να εισάγουμε στο αποτέλεσμα ώστε να αποκρύψουμε την συμμετοχή μιας εγγραφής στη βάση. Είναι προφανές ότι ένα counting query έχει ευαισθησία  $\Delta = 1$ .

#### Θεώρημα 4.11. (Μηχανισμός Laplace)

Έστω επερώτηση  $q$  με σύνολο τιμών  $\mathbb{R}$ , και έστω  $\Delta$  η  $l_1$ -ευαισθησία της. Τότε ο μηχανισμός

$$M(x) = q(x) + z$$

με  $z \sim \text{Lap}(\Delta|\epsilon)$ , παρέχει  $\epsilon$ -διαφορική ιδιωτικότητα.

Το μέγεθος του θορύβου εξαρτάται από το είδος της επερώτησης και την επιλογή του  $\epsilon$ . Άρα για ένα counting query θέλουμε θόρυβο τάξης  $\sim \text{Lap}(1|\epsilon)$ , ενώ όσο η τιμή του  $\epsilon$  μικραίνει, τόσο το αποτέλεσμα γίνεται περισσότερο ανακριβές.

Θωρείστε την βάση δεδομένων  $x$  με τους μισθούς που αναφέραμε προηγουμένως, και την επερώτηση για τον μέσο μισθό:

$$q(x) = \frac{\sum_{i=1}^n x_i}{n}$$

με  $x_i \in [0, x_{max}]$ . Αν χρησιμοποιήσουμε μια σχέση γειτνίασης τύπου  $|x_i - x'_i| \leq x_{max}$  τότε η ευαισθησία της επερώτησης θα είναι

$$\Delta = \max_{x \sim x'} |q(x) - q(x')| = \frac{1}{n} \max_{x_i, x'_i} |x_i - x'_i| \quad i \in [1, n]$$

Έτσι έχουμε

$$\Delta = \frac{x_{max}}{n}$$

Σύμφωνα με το παραπάνω θεώρημα λοιπόν, ο μηχανισμός

$$M(x) = \frac{\sum_{i=1}^n x_i}{n} + \text{Lap}\left(\frac{x_{max}}{n\epsilon}\right)$$

θα διατηρεί ε-διαφορική ιδιωτικότητα. Παρατηρούμε ότι το μέγεθος του θορύβου που προκύπτει από τον μηχανισμό Laplace είναι αντιστρόφως ανάλογο του πλήθους  $n$  των εγγραφών, πράγμα αναμενόμενο αφού διασθητικά αναμένουμε καλύτερη ιδιωτικότητα αν το μέγεθος της βάσης είναι μεγάλο [Cortés et al., 2016].

### 4.3.2 Εκθετικός Μηχανισμός

Ενώ ο μηχανισμός Laplace παρέχει διαφορική ιδιωτικότητα, δεν αρκεί για να καλύψει όλες τις ανάγκες, αφού εφαρμόζεται κυρίως σε επερωτήσεις που επιστρέφουν αριθμητικά αποτελέσματα. Μια χρήσιμη και αποτελεσματική μέθοδος για μη αριθμητικές επερωτήσεις είναι ο εκθετικός μηχανισμός [Dwork, 2008]. Έδω απαιτείται η χρήση scoring συναρτήσεων όπου αντιστοιχίζουν τα ζεύγη αποτελεσμάτων σε utility scores.

#### Ορισμός 4.12. (Scoring Function)

Έστω  $X$  το σύνολο όλων των πεπερασμένων συνόλων δεδομένων και  $R$  το σύνολο τιμών ενός μηχανισμού  $M$ . Ως scoring συνάρτηση ορίζεται:

$$u = X \times R \longrightarrow \mathbb{R}$$

και αποδίδει έναν βαθμό (score) σε κάθε ζεύγος  $(x, r) \in X \times R$ .

Όσο υψηλότερος είναι ο βαθμός του ζεύγους, τόσο καλύτερο το αποτέλεσμα. Η εφαρμογή που αποζητείται εδώ είναι να δίνεται μια βάση δεδομένων  $x$  και ο μηχανισμός να επιστρέφει το  $r \in R$  που μεγιστοποιεί τον βαθμό  $u(x, r)$ , ενώ παρέχει διαφορική ιδιωτικότητα.

Πριν παρουσιάσουμε τον εκθετικό μηχανισμό θα δώσουμε τον ορισμό της ευαισθησίας μιας scoring συνάρτησης.

**Ορισμός 4.13.** Η ευαισθησία μιας scoring συνάρτησης  $u = X \times R \rightarrow \mathbb{R}$  είναι

$$\Delta_u = \max_{r \in R} (\max_{x \sim x'} |u(x, r) - u(x', r)|)$$

**Ορισμός 4.14.** (Εκθετικός Μηχανισμός)

Ως εκθετικός μηχανισμός ορίζεται μια τυχαιοποιημένη συνάρτηση  $M_E : X \rightarrow R$ , η οποία δεδομένης βάσης  $x \in X$  και παραμέτρου  $\epsilon$ , παράγει ένα στοιχείο  $r \in R$  με πιθανότητα ανάλογη του

$$e^{\frac{\epsilon}{2\Delta_u} u(x, r)}$$

Όπου  $u$  είναι μια scoring συνάρτηση και  $\Delta_u$  η ευαισθησία της. Ο ορισμός συνεπάγεται το γεγονός ότι η πιθανότητα επιστροφής μιας τιμής  $r$  αυξάνεται εκθετικά με την μεγιστοποίηση της τιμής  $u(x, r)$ . Η τιμή  $r$  που μεγιστοποιεί την  $u(x, r)$  έχει την μεγαλύτερη πιθανότητα.

**Θεώρημα 4.15.** Ένας εκθετικός μηχανισμός παρέχει  $\epsilon$ -διαφορική ιδιωτικότητα.

Σε σύγκριση με τον μηχανισμό Laplace, ο εκθετικός μηχανισμός είναι γενικότερος στο ότι δεν περιορίζει το ερώτημα να είναι αριθμητικό. Στην πράξη, ο εκθετικός μηχανισμός χρησιμοποιείται ευρύτερα σε περιπτώσεις όπου το σύνολο τιμών της επερώτησης είναι πεπερασμένο, έτσι ώστε η συνάρτηση πυκνότητας πιθανότητας να μπορεί να υπολογιστεί.

### 4.3.3 Μηχανισμός Gauss

Εκτός από τον Μηχανισμό Laplace, για την εφαρμογή  $(\epsilon, \delta)$ -διαφορικής ιδιωτικότητας σε αριθμητικές επερωτήσεις επιλέγεται και ο παρακάτω. Αρχικά δίνουμε τον ορισμό της  $l_2$  ευαισθησίας.

**Ορισμός 4.16.** Η  $l_2$ -ευαισθησία (sensitivity) μιας συνάρτησης  $f : X \rightarrow \mathbb{R}^k$  είναι:

$$\Delta_2 = \max_{x \sim x'} \|f(x) - f(x')\|_2$$

Υπενθυμίζουμε την λειτουργία της ευκλείδειας νόρμας:  $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$

**Θεώρημα 4.17.** Έστω επερώτηση  $q$  με σύνολο τιμών  $\mathbb{R}$  και  $\Delta_2$  ευαισθησία. Για  $\epsilon > 0$  και  $\delta > 0$  ο μηχανισμός

$$M(x) = q(x) + z$$

με  $z$  ένα τυχαίο διάνυσμα ανεξάρτητων τυχαίων μεταβλητών που ακολουθούν την Κανονική κατανομή με μέση τιμή 0 και διακύμανση  $\sigma^2 = c^2 \Delta_2^2$ , με  $c = \sqrt{2 \log(1.25/\delta)/\epsilon}$ , παρέχει  $(\epsilon, \delta)$ - διαφορική ιδιωτικότητα.



Δεν προκαλεί έκπληξη το γεγονός ότι εμφανίζεται ο όρος  $\delta$ : η κατανομή Laplace είναι ιδανική για πολλαπλασιαστικό φράγμα, αλλά η Κανονική κατανομή όχι. Αν βέβαια το  $\delta$  είναι επαρκώς (π.χ. λογαριθμικά) μικρό, στην πράξη δεν θα βιώσουμε ποτέ αδυναμία εγγύησης της ιδιωτικότητας.

Επίσης, ένα να απο τα πλεονεκτήματα αυτού του μηχανισμού είναι ότι το άθροισμα δυο μηχανισμών Gauss είναι μηχανισμός Gauss, πράγμα που κάνει ευκολότερη την κατανόηση της εφαρμογής του κατά την ανάλυση δεδομένων. Οι δύο μηχανισμοί δίνουν την ίδια απώλεια αθροιστικά κατά τη σύνθεση, οπότε αν και η εγγύηση απορρήτου είναι ασθενέστερη για κάθε μεμονωμένο υπολογισμό, τα αθροιστικά αποτελέσματα μετά από πολλούς υπολογισμούς είναι ανταγωνιστικά. Τελικά προκύπτει ότι η χρήση του μηχανισμού Laplace είναι καθαρότερη, ενώ οι δύο μηχανισμοί συμπεριφέρονται παρόμοια σε περιπτώσεις σύνθεσης [Dwork et al., 2010].

#### 4.4 Κατασκευή σύνθετων Μηχανισμών

Οι μηχανισμοί που παρουσιάστηκαν παραπάνω είναι αρκετά γενικοί και απλοί στην εφαρμογή. Όπως θα δούμε και στη συνέχεια, η μόνη ποσότητα που χρειάζεται να υπολογιστεί για την υλοποίηση αυτών των αλγορίθμων είναι η ευαισθησία. Παρόλο που συνήθως υπολογίζεται εύκολα για απλές επερωτήσεις, μπορεί να είναι δύσκολο να υπολογιστεί για περίπλοκα ερωτήματα (π.χ., η βέλτιστη λύση ενός προβλήματος μη γραμμικής βελτιστοποίησης). Όταν οι επερωτήσεις είναι περίπλοκες, μια κοινή στρατηγική είναι η αποδόμηση της υπο έρευνα επερώτησης έτσι ώστε η ευαισθησία κάθε μέρους της να μπορεί εύκολα να υπολογιστεί. Παραδείγματος χάριν, αν και η ευαισθησία της βέλτιστης λύσης ενός προβλήματος μη γραμμικής βελτιστοποίησης μπορεί να μην είναι εύκολο να υπολογιστεί, η ευαισθησία των ενδιάμεσων αποτελεσμάτων που χρησιμοποιούνται για να υπολογιστεί επαναληπτικά η βέλτιστη λύση είναι συχνά πιο απλό να υπολογιστεί. Έτσι, χρησιμοποιώντας τους κανόνες σύνθεσης, μπορεί κανείς να κατασκευάσει τον επιθυμητό μηχανισμό ιδιωτικότητας.

Τολμούμε να πούμε ότι η Διαφορική Ιδιωτικότητα αποτελεί την «ασφαλέστερη» επιλογή από τους μηχανισμούς προστασίας που αναλύσαμε. Ο κύριος λόγος είναι ότι η εφαρμογή της δεν επηρεάζεται από την γνώση που κατέχει ο επιτιθέμενος. Επιπλέον, λόγω αυτού, δεν υπάρχει και η ανάγκη μοντελοποίησης της πρότερης γνώσης, οδηγώντας σε ελαφρώς ταχύτερους αλγορίθμους.

Στην πραγματικότητα, η Διαφορική Ιδιωτικότητα υπόσχεται να προστατεύει τα άτομα από κάθε πιθανή απειλή που θα μπορούσαν να αντιμετωπίσουν λόγω της ύπαρξης των δεδομένων τους στην ιδιωτική βάση δεδομένων  $x$  και που δεν θα αντιμετώπιζαν εάν τα δεδομένα τους δεν ήταν μέρος της  $x$ . Παρόλο που οι εγγραφές μπορούν πράγματι να απειληθούν μόλις απελευθερωθούν τα αποτελέσματα ενός μηχανισμού  $\Delta I$ , η διαφορική ιδιωτικότητα υπόσχεται ότι η πιθανότητα διαρροής δεν αυξάνεται σημαντικά από την επιλογή συμμετοχής τους.



Κατά κάποιο τρόπο αξιολογείται η απόφαση ενός ατόμου αν θα συμπεριλάβει ή όχι τα δεδομένα του σε μια βάση δεδομένων που θα χρησιμοποιηθεί ένας μηχανισμός ΔΙ. Εξετάζεται η διαφορά δηλαδή μεταξύ της πιθανότητας διαρροής δεδομένου ότι συμμετέχει στη βάση, σε σύγκριση με την πιθανότητα διαρροής δεδομένου ότι δεν συμμετέχει.

## Κεφάλαιο 5

# Εφαρμογή Αλγορίθμων Διαφορικής Ιδιωτικότητας

Στο κεφάλαιο αυτό θα περιγράψουμε τη δημιουργία προγραμμάτων ανωνυμοποίησης, χρησιμοποιώντας μηχανισμούς Διαφορικής Ιδιωτικότητας. Στη συνέχεια κάνουμε χρήση των αλγορίθμων αυτών με επερωτήσεις πάνω σε σύνολα δεδομένων. Τέλος εξάγουμε συμπεράσματα από την εφαρμογή των μηχανισμών.

### 5.1 Υλοποίηση

Η αρχιτεκτονική των μοντέλων που διαχειρίζονται τους μηχανισμούς τυχαιοποίησης είναι εντελώς διαφορετική από αυτήν των μηχανισμών γενίκευσης που αναλύσαμε. Όπως είδαμε στο κεφάλαιο 3, η τεχνική που χρησιμοποιούν οι περισσότεροι αλγόριθμοι είναι να δέχονται ένα σύνολο δεδομένων, και να επιστρέφουν μια ανωνυμοποιημένη εκδοχή του. Σε αυτή τη νέα βάση εργάζονται στη συνέχεια οι αναλυτές. Η χρήση διαδραστικών τεχνικών απαιτεί διαφορετική μοντελοποίηση: Από την μια μεριά, ο αναλυτής θέτει την επερωτήσή του στην βάση δεδομένων, ενώ από την άλλη ο διαχειριστής επιλέγει το μέγεθος ιδιωτικότητας που επιθυμεί<sup>1</sup>. Τα δεδομένα επεξεργάζονται, προστίθεται ο θόρυβος και το αποτέλεσμα επιστρέφει στον αναλυτή.

Στο πρόγραμμα μας, θα κάνουμε χρήση κυρίως του μηχανισμού Laplace, ο οποίος ικανοποιεί το κριτήριο της ΔΙ όπως δείξαμε στο προηγούμενο κεφάλαιο.

Η συνάρτηση πυκνότητας πιθανότητας για μέσο  $\mu = 0$  είναι

$$f(x|0, b) = \frac{1}{2b} e^{-\frac{|x|}{b}}$$

---

<sup>1</sup>Επιλογή της παραμέτρου  $\epsilon$

άρα έχουμε την αθροιστική συνάρτηση κατανομής:

$$F(x) = \int_{-\infty}^x f(u)du = \int_{-\infty}^x \frac{1}{2b} e^{-\frac{|u|}{b}} du = \begin{cases} \frac{1}{2} e^{\frac{x}{b}} & x < 0 \\ 1 - \frac{1}{2} e^{-\frac{x}{b}} & x \geq 0 \end{cases}$$

Η αντίστροφη συνάρτηση είναι:

$$F^{-1}(x) = \begin{cases} b \cdot \ln(2x) & 0 < x < 1/2 \\ -b \cdot \ln(2 - 2x) & 1/2 \leq x \leq 1 \end{cases}$$

Θέτοντας  $u = x - 1/2$  καταλήγουμε σε μια γεννήτρια τυχαίων μεταβλητών:

$$X = -b \cdot \text{sgn}(u) \ln(1 - 2|u|) \quad u \in \left(-\frac{1}{2}, \frac{1}{2}\right]$$

Έτσι, επιλέγοντας τυχαίες μεταβλητές  $u$  από την ομοιόμορφη κατανομή στο διάστημα  $[-0.5, 0.5]$ , η τυχαία μεταβλητή  $X$  θα ανήκει στην κατανομή Laplace με παράμετρο κλίμακας  $b$ .

Κατασκευάζουμε δυο διαφορετικές συναρτήσεις. Η μια ως είσοδο δέχεται παράμετρο  $b$  και πλήθος μεταβλητών που επιθυμούμε, ενώ η δεύτερη την παράμετρο  $\epsilon$  και διάσταση βάσης δεδομένων. Η σχέση που συνδέει το  $\epsilon$  με την παράμετρο  $b$  είναι

$$b = \frac{\Delta f}{\epsilon}$$

με  $\Delta f$  να συμβολίζει την ευαισθησία μιας συνάρτησης  $f : X \rightarrow \mathbb{R}^k$ :

$$\Delta f = \max_{x \sim x'} \|f(x) - f(x')\|$$

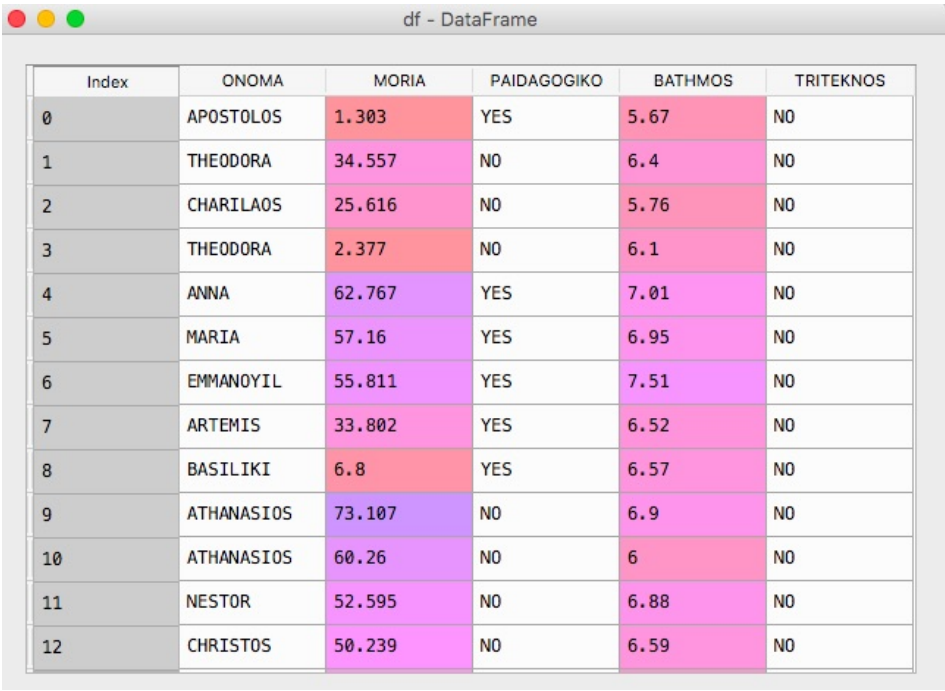
όπου  $x, x'$  δύο γειτνιαζουσες βάσεις δεδομένων<sup>2</sup>.

<sup>2</sup> Σε περίπτωση που απαιτηθεί αυστηρή εφαρμογή του ορισμού της ευαισθησίας: Για την δημιουργία της βάσης  $x'$ , έχουμε αναπτύξει την συνάρτηση «falsedata», η οποία έχει ως είσοδο την αρχική βάση  $x$  και ως έξοδο μια νέα βάση η οποία διαφέρει από την αρχική το πολύ σε μια εγγραφή. Η συνάρτηση επιλέγει τυχαία μια πλειάδα της βάσης, και μεταβάλλει τις τιμές των αντίστοιχων γνωρισμάτων με τυχαίο τρόπο:

- Για αλφαριθμητικά γνωρίσματα, επιλέγει μια τιμή από το πεδίο τιμών του αντίστοιχου γνωρίσματος.
- Για αριθμητικά γνωρίσματα, η συνάρτηση εντοπίζει την ελάχιστη και την μέγιστη τιμή του γνωρίσματος και επιλέγει μια τυχαία τιμή σε αυτό το διάστημα. Έτσι διασφαλίζεται η επιπλέον διατήρηση της ιδιωτικότητας ειδικά αν αυτό το γνώρισμα είναι ευαίσθητο.
- Για τα λογικά γνωρίσματα (στην περίπτωση της δικής μας βάσης, οι τιμές είναι YES ή NO), επιλέγεται τυχαίως μια εκ των δύο τιμών, αλλά με πιθανότητα αντίστοιχη της εμφάνισης στη βάση. Δηλαδή αν το γνώρισμα παίρνει τις τιμές 0 ή 1, και οι άσσοι είναι 9πλάσιοι από τα μηδενικά, θα επιλεγεί η τιμή 1 με πιθανότητα 90%.

Το σύνολο δεδομένων που χρησιμοποιούμε στα πειράματα είναι ένας πίνακας εκπαιδευτικών ΠΕ19 Πληροφορικής του Υπουργείου Παιδείας<sup>3</sup>. Αφού μετατρέψαμε τους χαρακτήρες από ελληνικά σε λατινικά, αφαιρέσαμε περιττά και γνωρίσματα που οδηγούν σε ταυτοποίηση. Καταλήγουμε στην τελική μορφή ενός συνόλου δεδομένων που αποτελείται από 1400 περίπου εγγραφές και έχει τα γνωρίσματα {Όνομα, Μόρια, Παιδαγωγικό, Βαθμός, Τρίτεκνος}. Συγκεκριμένα:

- Όνομα: Το μικρό όνομα του ατόμου
- Μόρια: Τα μόρια που έχει συγκεντρώσει (θεωρείται ευαίσθητο χαρακτηριστικό)
- Παιδαγωγικό: YES Αν διαθέτει παιδαγωγική πιστοποίηση, NO διαφορετικά.
- Βαθμός: Βαθμός πτυχίου (θεωρείται ευαίσθητο χαρακτηριστικό)
- Τρίτεκνος: YES Αν έχει τρία παιδιά τουλάχιστον, NO διαφορετικά.



Index	ONOMA	MORIA	PAIDAGOGIKO	BATHMOS	TRITEKNOS
0	APOSTOLOS	1.303	YES	5.67	NO
1	THEODORA	34.557	NO	6.4	NO
2	CHARILAOS	25.616	NO	5.76	NO
3	THEODORA	2.377	NO	6.1	NO
4	ANNA	62.767	YES	7.01	NO
5	MARIA	57.16	YES	6.95	NO
6	EMMANOYL	55.811	YES	7.51	NO
7	ARTEMIS	33.802	YES	6.52	NO
8	BASILIKI	6.8	YES	6.57	NO
9	ATHANASIOS	73.107	NO	6.9	NO
10	ATHANASIOS	60.26	NO	6	NO
11	NESTOR	52.595	NO	6.88	NO
12	CHRISTOS	50.239	NO	6.59	NO

ΣΧΗΜΑ 5.1: Πίνακας αναπληρωτών εκπαιδευτικών ΠΕ19, μετά από ανωνυμοποίηση.

Η επιλογή της βάσης έγινε λόγω των χαρακτηριστικών της, τα οποία θα μας καλύψουν στους υπολογισμούς που επιθυμούμε να πραγματοποιήσουμε. Επιπλέον, το μέγεθος είναι ιδανικό για να εμφανιστούν τυχόν σφάλματα από κακή επιλογή παραμέτρων.

<sup>3</sup><http://e-aitisi.sch.gr>

## 5.2 Εφαρμογή και αξιολόγηση

Η ανάπτυξη του αλγορίθμου γίνεται σε γλώσσα Python, έκδοσης 2.7.15, σε σύστημα OS X υπολογιστή Mac, με διπύρινο επεξεργαστή συγχρονισμένο στα 2.4 GHz. Κάνουμε χρήση των βιβλιοθηκών *Numpy* και *Pandas*, ενώ η υλοποίηση γίνεται στο περιβάλλον προγραμματισμού Spyder.

### 5.2.1 Δημοφιλή ονόματα

Στο πείραμα αυτό θα εφαρμόσουμε την επερώτηση “ποιο είναι το πιο συνηθισμένο όνομα στη βάση δεδομένων;”.

Υλοποιούμε μια συνάρτηση που δεχεται ως όρισμα την βάση δεδομένων, μετρά το πλήθος του κάθε ονόματος και επιστρέφει αυτό με την μέγιστη τιμή. Υλοποιούμε και μια συνάρτηση τυχαιοποίησης, που δέχεται ως όρισμα την βάση και την παράμετρο  $\epsilon$  και πραγματοποιεί ακριβώς την ίδια διαδικασία, αλλά στο τέλος προσθέτει θόρυβο από την κατανομή Laplace.

Η ευαισθησία της επερώτησης θα είναι  $\Delta_f = 1$ , οπότε η παράμετρος κλίμακας της Laplace είναι  $b = \frac{1}{\epsilon}$ .

Για  $\epsilon = 1$  έχουμε το αποτέλεσμα:

- Χωρίς θόρυβο: ‘GEORGIOS’
- Με θόρυβο: ‘GEORGIOS’

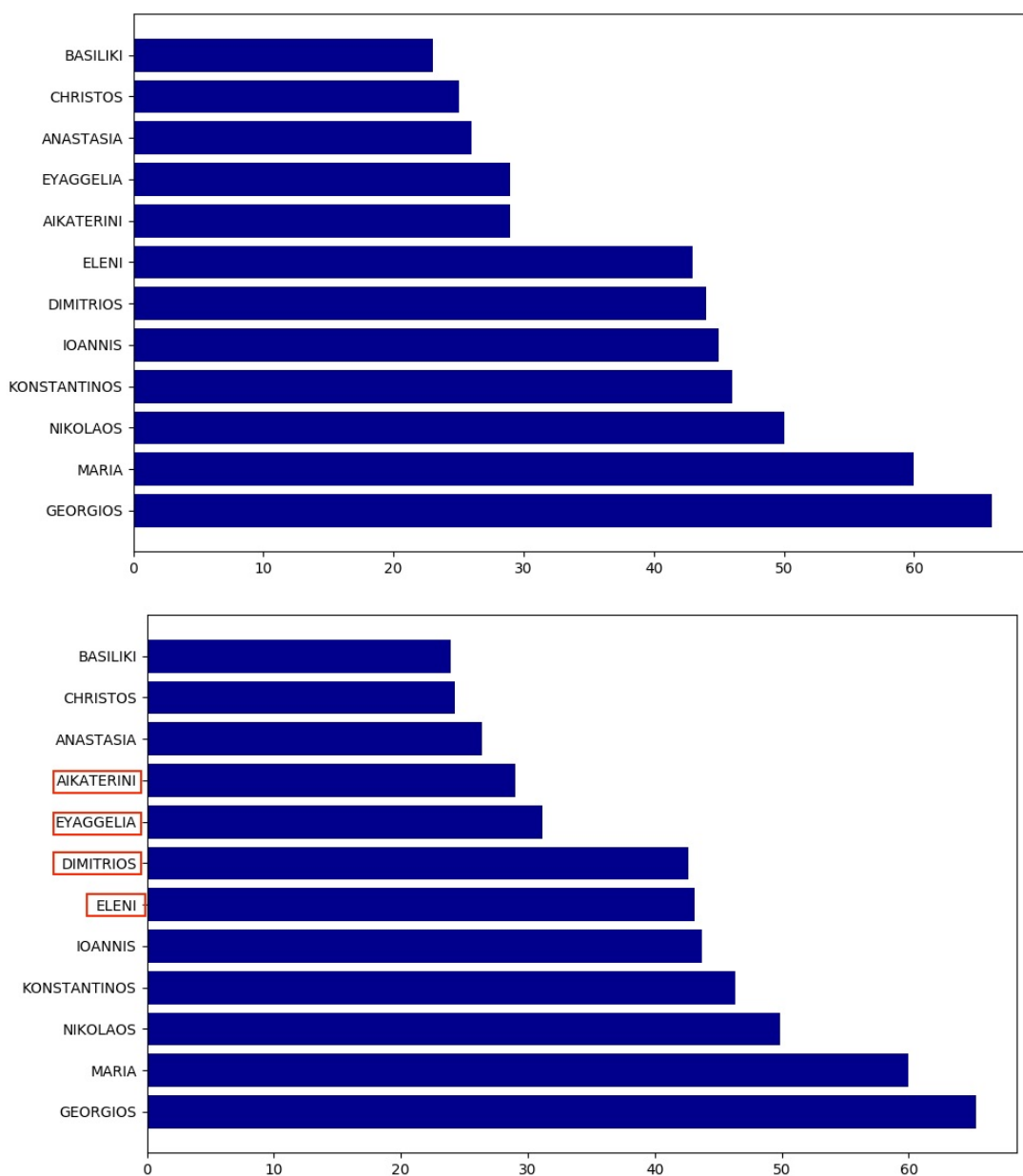
Για  $\epsilon = 0.4$ , μετά από λίγες εκτελέσεις προκύπτει το αποτέλεσμα:

- Χωρίς θόρυβο: ‘GEORGIOS’
- Με θόρυβο: ‘MARIA’

Οπότε επαληθεύουμε την πρόταση ότι όσο μικραίνει η τιμή του  $\epsilon$ , τόσο η απώλεια πληροφορίας μεγιστοποιείται.

Επεκτινύουμε την επερώτηση, αναζητώντας περισσότερα του ενός ονόματα. Η επερώτηση μας είναι τύπου Ιστογράμματος (histogram query), οπότε εξακολουθεί η ευαισθησία να είναι  $\Delta_f = 1$ . Επιλέγουμε λοιπόν να υπολογιστούν τα 12 δημοφιλέστερα ονόματα στη βάση, επιλέγοντας  $\epsilon = 1$ .

Παρατηρούμε ότι η πρώτη τριάδα παραμένει ανεπηρέαστη. Αντίθετα στις μεσαίες τιμές, όπου οι συχνότητες των ονομάτων είναι αρκετά κοντά, παίρνουμε εσφαλμένα αποτελέσματα. Έκτο δημοφιλέστερο όνομα στην βάση προκύπτει το ‘ELENI’, αντί του πραγματικού



ΣΧΗΜΑ 5.2: Δημοφιλέστερα ονόματα - χωρίς θόρυβο (πάνω) και με θόρυβο ( $\epsilon=1$ ) (κάτω)

‘DIMITRIOS’, παρά το ότι στην εκτέλεση της αρχικής επερώτησης η επιλογή του  $\epsilon$  να έχει την τιμή 1 ήταν αποτελεσματική.

Αντιλαμβανόμαστε το μέγεθος του ζητήματος για την ορθή επιλογή της παραμέτρου  $\epsilon$ , το οποίο οφείλει να προβάλλει τα αποτελέσματά των επερωτήσεων σε απόλυτη ισορροπία: διατηρώντας την ιδιωτικότητα των δεδομένων του συνόλου και διασφαλίζοντας την ακρίβεια του αποτελέσματος.

### 5.2.2 Μέση τιμή

Στο πείραμα αυτό θα εφαρμόσουμε διάφορες μορφές της κλασικής επερώτησης για την εύρεση της μέσης τιμής ενός ευαίσθητου, αριθμητικού χαρακτηριστικού.

Ξεκινάμε με την εύρεση της μέσης τιμής του γνωρίσματος 'BATHMOS'. Θα έχουμε λοιπόν:

$$f(x) = \frac{1}{n} \sum_{i=1}^n b_i$$

Οι τιμές των βαθμών κειμένονται στο διάστημα  $[5, b_{max}]$ . Παρατηρούμε ότι

$$|b_i - b'_i| \leq b_{max} - 5$$

, συνεπώς η ευαισθησία μπορεί να υπολογιστεί ως εξής:

$$\Delta_f = \max_{x \sim x'} |q(x) - q(x')| = \frac{1}{n} \max_{x_i, x'_i} |b_i - b'_i| \quad i \in [1, n]$$

Έτσι έχουμε

$$\Delta_f = \frac{b_{max} - 5}{n}$$

Γνωρίζουμε ότι ο μηχανισμός

$$M(x) = \frac{\sum_{i=1}^n b_i}{n} + \text{Lap}\left(\frac{b_{max} - 5}{n\epsilon}\right)$$

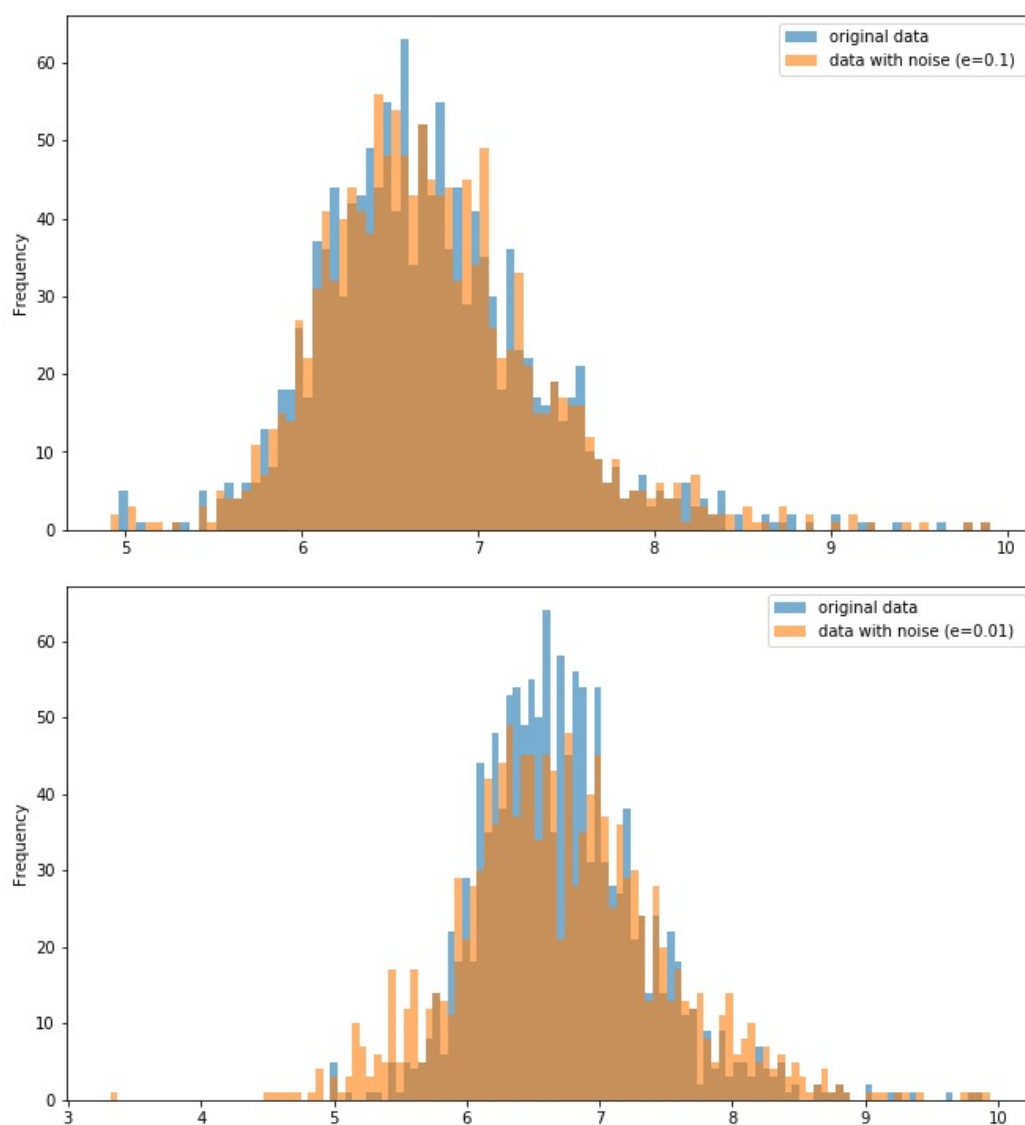
θα διατηρεί  $\epsilon$ -διαφορική ιδιωτικότητα [Dwork and Aaron, 2014]. Εφαρμόζοντας τον τύπο στους υπολογισμούς μας παίρνουμε τα αποτελέσματα για τον μέσο βαθμό

- Αρχικά: 6.74
- Με  $\epsilon = 1$ : 6.74
- Με  $\epsilon = 0.1$ : 6.73
- Με  $\epsilon = 0.01$ : 6.81

Μια άλλη εκδοχή, είναι να προσθέσουμε θόρυβο Laplace σε κάθε έναν από τους βαθμούς και να υπολογιστεί στη συνέχεια η μέση τιμή τους. Λόγω του θεωρήματος της σύνθεσης ικανοποιείται η  $\epsilon$ -διαφορική ιδιωτικότητα. Με αυτόν τον τρόπο μπορούμε να απεικονίσουμε και το ιστόγραμμα συχνοτήτων των βαθμών.

Βλέποντας τα δυο διαγράμματα συμπεραίνουμε ότι για  $\epsilon = 0.01$ , ενώ η μέση τιμή είναι σχεδόν ανεπηρέαστη, θα υπάρχει σημαντική απόκλιση στα αποτελέσματα. Επιπλέον η διασπορά της κατανομής *Laplace* οδηγεί στην εμφάνιση τιμών κάτω από τη βάση του 5, πράγμα αδύνατον.





ΣΧΗΜΑ 5.3: Ιστόγραμμα συχνοτήτων των βαθμών

### 5.2.3 Πλήθος Τρίτεχνων

Για τη συνέχεια των δοκιμών μας θεωρούμε το γνώρισμα 'TRITEKNOS' ως ευαίσθητο, συνεπώς επερωτήσεις για συγκεκριμένες εγγραφές ως προς την τιμή του χαρακτηριστικού αυτού αποκλείονται. Μπορούμε, ωστόσο, να επερωτήσουμε συλλογικά τη βάση και να εξάγουμε συμπεράσματα. Τι γίνεται όμως στην περίπτωση που ένας επιτιθέμενος κατέχει πολύ μεγάλη γνώση για τα άτομα στην βάση;

Η επερώτηση για το πλήθος των εγγραφών με θετική τιμή στο TRITEKNOS επιστρέφει 32. Αυτό σημαίνει ότι η πιθανότητα ένα άτομο στη βάση να είναι θετικός στο γνώρισμα αυτό, είναι περίπου 2,3%. Στη συνέχεια εφαρμόζουμε την επερώτηση «ποιό είναι το πλήθος των τρίτεχνων που έχουν βαθμό ίσο με 5». Η απάντηση είναι 1, ενώ το πλήθος των ατόμων με βαθμό 5 είναι τέσσερα. Η πιθανότητα επιλογής τρίτεχνου, δηλαδή, σε αυτό το υποσύνολο της βάσης είναι 25%, δεκαπλάσια και πλέον από αυτήν επί του αρχικού συνόλου.

Υποθέτουμε τώρα ότι ο επιτιθέμενος γνωρίζει καλά τα περισσότερα άτομα στη βάση. Συγκεκριμένα ξέρει ότι ο Δημήτρης, η Φαίη και ο Ανδρέας έχουν βαθμό ίσο με 5 ακριβώς, ενώ είναι σίγουρος ότι κανείς τους δεν είναι τρίτεχνος. Σύμφωνα λοιπόν με τα παραπάνω δεδομένα, το τέταρτο άτομο του υποσυνόλου θα έχει θετική τιμή στο γνώρισμα. Προκύπτει λοιπόν αποκάλυψη του ευαίσθητου γνωρίσματος μιας εγγραφής.

ΟΝΟΜΑ	MORIA	PAIDAGOGIKO	BATHMOS	TRITEKNOS
DIMITRIOS	102.048	YES	5	NO
FAII	34.014	NO	5	NO
SOTIRIA	4.161	NO	5	YES
ANDREAS	0.953	NO	5	NO

ΣΧΗΜΑ 5.4: Αποκάλυψη τιμής ευαίσθητου γνωρίσματος

Εφαρμόζοντας, ωστόσο, ελάχιστο θόρυβο με την μέθοδο της Διαφορικής Ιδιωτικότητας, βλέπουμε ότι η πιθανότητα αποκάλυψη της τιμής του γνωρίσματος 'TRITEKNOS' για μια εγγραφή είναι μηδενική.

Αν εισάγουμε θόρυβο στο γνώρισμα BATHMOS όπως στην προηγούμενη παράγραφο, τότε θα είναι αδύνατο να εντοπιστούν τα 4 αυτά άτομα με βαθμό ίσο με 5. Βλέπουμε λοιπόν πως η διαφορική ιδιωτικότητα προστατεύει τα δεδομένα από επιτιθέμενους με γνώση.

Οι παραπάνω εφαρμογές δείχνουν ότι η εισαγωγή θορύβου στα δεδομένα και γενικά η τυχαιοποίηση είναι, θα τολμούσαμε να πούμε, η ασφαλέστερη επιλογή για την διατήρηση της ιδιωτικότητάς τους. Το μεγαλύτερο μειονέκτημα των μεθόδων αυτών είναι η επιλογή του ιδανικού μέτρου θορύβου ώστε να μην προκύψει διαστρέβλωση της πληροφορίας, ενώ ταυτόχρονα να είναι εγγυημένη η προστασία των δεδομένων. Συγκεκριμένα, στα παραπάνω

παραδείγματα παρατηρήσαμε ότι η επιλογή της παραμέτρου  $\epsilon$  είναι μια αρκετά επίπονη διαδικασία. Σχετικά με το θέμα αυτό έχουν γραφτεί πολλά άρθρα και έχουν προταθεί δεκάδες μέθοδοι οι οποίες αναζητούν τη «χρυσή τομή», χωρίς να προκύπτει πάντα το αναμενόμενο αποτέλεσμα από την εφαρμογή τους [Murtagh and Vadhan, 2016].



## Κεφάλαιο 6

# Συμπεράσματα - Μελλοντική Έρευνα

Παρακάτω συνοψίζουμε τις μεθόδους που αναλύσαμε κατά τη διάρκεια της εργασίας αυτής και εξάγουμε συμπεράσματα. Στη συνέχεια αναφέρουμε επεκτάσεις, μελλοντικά έργα και νέα μοντέλα προστασίας που παρουσιάστηκαν πρόσφατα. Τέλος παρουσιάζουμε ορισμένα ανοικτά ζητήματα πάνω στην προστασία της ιδιωτικότητας των δεδομένων που θα μας απασχολήσουν στο άμεσο μέλλον.

### 6.1 Σύνοψη και συμπεράσματα

Στην εργασία αυτή ασχοληθήκαμε με την μελέτη των μεθόδων προστασίας της ιδιωτικότητας των δεδομένων. Μιλήσαμε για την επικινδυνότητα αποκάλυψης πληροφορίας που προκύπτει από την εφαρμογή απλών τεχνικών ανωνυμοποίησης, και την ανάγκη εισαγωγής πολυπλοκότερων μεθόδων. Αναλύσαμε τις κυριότερες τεχνικές γενίκευσης, τονίζοντας τόσο τα πλεονεκτήματα, αλλά και τα μειονεκτήματα της κάθε μιας. Παρουσιάσαμε την επίθεση τομής η οποία διαπερνά, θεωρητικά, κάθε μέθοδο γενίκευσης. Στη συνέχεια αναλύσαμε σε βάθος την μέθοδο της διαφορικής ιδιωτικότητας και τους μηχανισμούς τυχαιοποίησης που την συνοδεύουν. Τέλος, αναπτύξαμε κώδικα σε γλώσσα προγραμματισμού Python που εφαρμόζει τον μηχανισμό Laplace και εκτελέσαμε παραδείγματα τυχαιοποίησης σε ένα σύνολο δεδομένων.

Όπως είδαμε, ένα σύνολο δεδομένων στο οποίο έχουν εφαρμοστεί μηχανισμοί ανωνυμοποίησης μπορεί να εξακολουθεί να παρουσιάζει κινδύνους αποκάλυψης για τις εγγραφές στις οποίες αναφέρονται τα δεδομένα. Ακόμα και σε περίπτωση μη ανάκτησης της ταυτότητας ενός ατόμου, ενδέχεται να είναι εφικτή η αποκάλυψη στοιχείων σχετικά με το συγκεκριμένο άτομο με τη βοήθεια συνήθως άλλων πηγών πληροφοριών. Οφείλουμε λοιπόν να υπογραμμίσουμε ότι καμία από τις τεχνικές που περιγράφονται στην παρούσα εργασία δεν πληροί

με βεβαιότητα τα τρία κριτήρια της αποτελεσματικής ανωνυμοποίησης<sup>1</sup>. Ωστόσο, ορισμένοι απο τους κινδύνους αυτούς ενδέχεται να μπορούν να αντιμετωπιστούν εξ' ολοκλήρου από μια συγκεκριμένη μέθοδο, δεδομένου ότι έχουν γίνει οι απαραίτητοι χειρισμοί κατά την ανάπτυξη της. Ο παρακάτω πίνακας απεικονίζει τις δυνατότητες των μηχανισμών ανωνυμοποίησης που αναλύσαμε.

Μηχανισμός	Αποκάλυψη ταυτότητας	Αποκάλυψη συγκεκριμένου γνωρίσματος	Κίνδυνος Συνδεσιμότητας
k-ανωνυμία	Όχι	Ναι	Ναι
l-διαφορετικότητα	Όχι	Ίσως Όχι	Ναι
t-εγγύτητα	Όχι	Όχι	Ναι
Διαφορική Ιδιωτικότητα	Όχι	Όχι	Ίσως Όχι

ΣΧΗΜΑ 6.1: Πλεονεκτήματα και μειονεκτήματα τεχνικών ανωνυμοποίησης

Συνεπώς, καλό είναι να εφαρμόζεται ο κάθε μηχανισμός ιδιωτικότητας σε αντίστοιχες περιπτώσεις ανωνυμοποίησης ανάλογα με τις απαιτήσεις του υπεύθυνου της βάσης δεδομένων ή ακόμη καλύτερα, η ταυτόχρονη χρήση διαφορετικών μηχανισμών στο ίδιο σύνολο δεδομένων.

## 6.2 Νέες τεχνικές - συνδυασμοί

Εχουν προταθεί τον τελευταίο καιρό νέες μέθοδοι ανωνυμοποίησης δεδομένων αλλά και συνδυασμοί τεχνικών γενίκευσης-τυχαιοποίησης οι οποίες υπόσχονται ακόμη καλύτερα αποτελέσματα.

Πιο συγκεκριμένα, οι συνεχείς απαιτήσεις για προστασία των όλο και μεγαλύτερων συνόλων δεδομένων, οδηγούν σε έρευνες οι οποίες επιδιώκουν να συνδυάσουν τις μεθόδους ανωνυμοποίησης που αναφέραμε, με σκοπό ασφαλέστερα και ταχύτερα αποτελέσματα. Ένας συνδιασμός που αξίζει να ανφέρουμε είναι η χρήση τεχνικών γενίκευσης και διαφορικής ιδιωτικότητας [Domingo-Ferrer and Soria-Comas, 2015]. Στην εργασία αυτή αποδεικνύεται ότι η k-ανωνυμία για την προστασία των quasi-identifiers, σε συνδιασμό με μηχανισμό ε-διαφορικής ιδιωτικότητας για τα ευαίσθητα γνωρίσματα αποδίδει στοχαστική t-εγγύτητα<sup>2</sup>, με  $t = t(k, \epsilon)$  συνάρτηση των k και  $\epsilon$ .

Πέραν αυτού, πρόσφατες έρευνες αναδεικνύουν νέα μοντέλα προστασίας της ιδιωτικότητας των δεδομένων, τόσο διαδραστικά όσο και μη διαδραστικά. Τελευταία, παρουσιάστηκε η τεχνική permutation paradigm για να περιγράψει οποιαδήποτε μέθοδο κάλυψης δεδομένων,

<sup>1</sup>Παράγραφος 2.2

<sup>2</sup>επέκταση της t-εγγύτητας

κυρίως microdata, ως «μεταλλαγή», ανοίγοντας το δρόμο για την πραγματοποίηση ουσιαστικών αναλυτικών συγκρίσεων των μεθόδων [Domingo-Ferrer and Muralidhar, 2016]. Η ιδιωτικότητα που εξασφαλίζεται με αυτή τη μέθοδο μπορεί να επαληθεύεται από κάθε άτομο που παρέχει τα δεδομένα του στη βάση και επίσης, στο επίπεδο συνόλου δεδομένων, από τον διαχειριστή [Ruiz, 2018]. Ακόμη, η τεχνική αυτή μοντελοποιεί την μέγιστη γνώση του επιτιθέμενου, ενώ προσπαθεί να προσεγγίσει την κατάσταση πλήρους διαφάνειας για τον χρήστη δεδομένων ως προς την ανωνυμοποίηση<sup>3</sup>, δηλαδή εφαρμογή της υπόθεσης του Kerkhoff<sup>4</sup>.

Η προστασία της ιδιωτικότητας των δεδομένων απαιτεί, όπως είδαμε, πολλούς πόρους και μεγάλη ακρίβεια. Είναι προφανές ότι είναι απαραίτητη η συνεχής αναζήτηση ταχύτερων αλγορίθμων ανωνυμοποίησης καθώς και ασφαλέστερων τεχνικών, οι οποίες ταυτόχρονα θα αποφέρουν μηδενική απώλεια πληροφορίας.

---

<sup>3</sup>Μόνο η τυχαιοποίηση που χρησιμοποιήθηκε πρέπει να μένει κρυφή

<sup>4</sup>Ένα κρυπτογραφικό σύστημα πρέπει να σχεδιάζεται για να είναι ασφαλές, ακόμη και αν όλες οι λεπτομέρειες του, εκτός από το κλειδί, είναι δημοσίως γνωστές.

# Παράρτημα Κώδικα

---

```

import numpy as np
import random as rd
import pandas as pd
import matplotlib.pyplot as plt
import dp_algos as a
from collections import Counter

#Read data set
def readdata():
    #add=raw_input("Give file name:")
    add='PE19CSV2'
    path='/Users/mikevard/Desktop/PYTHON-DP/'+add+'.csv'
    df = pd.read_csv(path, sep=';')
    return df

#Generates a data set that differs in one element
#from original data set
def falsedata(df):

    data = df
    n = len(data)
    #random name
    names = data["ONOMA"]
    i = rd.randint(0, n)
    rdname = names.loc[i]
    #random moria
    maxm = data["MORIA"].max()
    minm = data["MORIA"].min()
    rdmoria = "%.3f" % rd.uniform(minm, maxm)
    #random vathmos
    maxb = data["BATHMOS"].max()
    minb = data["BATHMOS"].min()
    rdbathmos = "%.3f" % rd.uniform(minb, maxb)

    #random tritekno, with probability
    cbool = Counter(data["TRITEKNOS"])
    p1 = cbool['YES']/float(n)
    p2 = cbool['NO']/float(n)
    rdtrit = np.random.choice(["YES", "NO"], size=None, p=[p1, p2])

    #random paidagogiko, with probability
    cbool = Counter(data["PAIDAGOGIKO"])
    p1 = cbool['YES']/float(n)
    p2 = cbool['NO']/float(n)
    rdpaid = np.random.choice(["YES", "NO"], size=None, p=[p1, p2])

    i = rd.randint(0, n)
    data.loc[i] = [rdname, rdmoria, rdpaid, rdbathmos, rdtrit ]
    return data
  
```



```
def add_noise(data,e):
    n=len(data)
    #lapN = Laplace(2, n)
    count = Counter(data)
    lapN = a.Laplace(1/float(e), n)

    #print "Most common names, initially:", count.most_common(3)
    test = count.most_common(12)
    x=[]
    y=[]
    for i in range(12):
        x.append(test[i][0])
        y.append(test[i][1])
    print "intially:"
    plt.barh(x,y, color='darkblue')
    plt.show()
    i=0
    for item in count:
        #print i, " ", item, "\n"
        count[item]+=lapN[i]
        i+=1

    #print "Most common names, finally:", count.most_common(3)

    test = count.most_common(12)
    x=[]
    y=[]
    for i in range(12):
        x.append(test[i][0])
        y.append(test[i][1])
    print "fianally:"
    plt.barh(x,y, color='darkblue')
    plt.show()

def add_noisem(data,e):
    n=len(data)
    #lapN = Laplace(2, n)
    count = Counter(data)
    lapN = a.Laplacem(count, n)
    #print "Initial: ", count
    print "Most common names, initially:", count.most_common(1)
    i = 0
    for item in count:
        #print i, " ", item, "\n"
        count[item]+=lapN[i]
        i+=1

    print "Most common names, finally:", count.most_common(1)
```

```
df = readdata()
data = df["ONOMA"]
add_noise(data,1)
```

---

```
import math
import random
import numpy as np
import pandas as pd
```

```
#Creates list with Laplace noise (uses builtin function)
```

```
#Parameters: scale and size
```

```
def Laplace(scale,n):
    m=0
    sample = np.random.laplace(m, scale, n)
    listN = []
    for x in range(1500):
        listN.append(sample[x])
    return listN
```

```
#Creates list with Laplace noise
```

```
#Parameters: a dict and the epsilon
```

```
def Laplacem(dic,epsilon):
```

```
    #sensitivity
```

```
    values=[]
```

```
    for x in dic:
```

```
        values.append(dic[x])
```

```
    maxi = max(values)
```

```
    mini = min(values)
```

```
#Generating random variables according to the Laplace distribution
```

```
    sigma = (maxi - mini)/float(epsilon)
```

```
    x = np.random.random_sample((2000,))
```

```
    l=[]
```

```
    for i in range(len(dic)):
```

```
        if x[i]<1/2:
```

```
            noise = (sigma/math.sqrt(2))*math.log(2*x[i])
```

```
        else:
```

```
            noise = -(sigma/math.sqrt(2))*math.log(2*(1-x[i]))
```

```
        l.append(noise)
```

```
    return len(l)
```

```
#Creates list with Laplace noise
```

```
#Parameters: a list and the epsilon
```

```
def Laplace1(dic,epsilon):
```

```
    #sensitivity
```

```
    maxi = dic.max()
```

---

```

mini = dic.min()

#Generating random variables according to the Laplace distribution
sigma = (maxi - mini)/(float(epsilon)*len(dic))
x = np.random.random_sample((2000,))
l=[]
for i in range(len(dic)):
    if x[i]<1/2:
        noise = (sigma/math.sqrt(2))*math.log(2*x[i])
    else:
        noise = -(sigma/math.sqrt(2))*math.log(2*(1-x[i]))
    l.append(noise)
return l

```

---



---

```

from dp_algos import Laplace1

#Returns the average BATHMOS of poeople
def av_query(data):
    df = data
    #df = data[data.TRITEKNOS=="YES"]
    df = df["BATHMOS"]
    mesos = df.mean()
    return mesos

#WITH Differential Privacy
def dpav_query (data, data2, epsilon):
    df2 = data2
    #df = data2[data2.TRITEKNOS=="YES"]
    df2 = df2["BATHMOS"]
    df = data["BATHMOS"]
    #generate Laplace_noise
    l=Laplace1(df, epsilon)
    #print l
    m=sum(l)/len(l)
    mesos = df2.mean() + m
    return mesos

```

---



# Bibliography

- [Bastiaan, 2015] Bastiaan, M. (2015). Preventing the 51%-attack: a stochastic analysis of two phase proof of work in bitcoin. In *Available at <http://refereat.cs.utwente.nl/conference/22/paper/7473/preventingthe-51-attack-stochastic-analysis-of-two-phase-proof-of-work-in-bitcoin.pdf>*.
- [Bayardo and Agrawal, 2005] Bayardo, R. J. and Agrawal, R. (2005). Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on Data Engineering, ICDE '05*, pages 217–228.
- [Bun et al., 2015] Bun, M., Nissim, K., Stemmer, U., and Vadhan, S. P. (2015). Differentially private release and learning of threshold functions. *CoRR*, abs/1504.07553.
- [Cortés et al., 2016] Cortés, J., Dullerud, G. E., Han, S., Ny, J. L., and Pappas, G. J. (2016). Differential privacy in control and network systems. *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 4252–4272.
- [Domingo-Ferrer and Muralidhar, 2016] Domingo-Ferrer, J. and Muralidhar, K. (2016). New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *Information Sciences*, 337:11–24.
- [Domingo-Ferrer and Soria-Comas, 2015] Domingo-Ferrer, J. and Soria-Comas, J. (2015). From t-closeness to differential privacy and vice versa in data anonymization. *CoRR*, abs/1512.05110.
- [Domingo-Ferrer and Soria-Comas, 2018] Domingo-Ferrer, J. and Soria-Comas, J. (2018). *Connecting Randomized Response, Post-Randomization, Differential Privacy and t-Closeness via Deniability and Permutation*.
- [Dwork, 2006a] Dwork, C. (2006a). Differential privacy. *Automata, Languages and Programming. ICALP*, 4052(10).
- [Dwork, 2008] Dwork, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer.

- [Dwork and Aaron, 2014] Dwork, C. and Aaron, R. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9:211–407.
- [Dwork et al., 2010] Dwork, C., Rothblum, G. N., and Vadhan, S. (2010). Boosting and differential privacy. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, Washington, DC, USA. IEEE Computer Society.
- [Dwork, 2006b] Dwork, Cynthia, M. F. N. K. S. A. (2006b). Calibrating noise to sensitivity in private data analysis. In Halevi, S. and Rabin, T., editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.
- [Ganta et al., 2009] Ganta, S. R., Kasiviswanathan, S. P., and Smith, A. (2009). Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 265–273. ACM.
- [Ganta et al., 2008] Ganta, S. R., Kasiviswanathan, S. P., and Smith, A. D. (2008). Composition attacks and auxiliary information in data privacy. *CoRR*, abs/0803.0032.
- [Geng et al., 2015] Geng, Q., Kairouz, P., Oh, S., and Viswanath, P. (2015). The staircase mechanism in differential privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9:1–1.
- [Geng and Viswanath, 2016] Geng, Q. and Viswanath, P. (2016). Optimal noise adding mechanisms for approximate differential privacy. *IEEE Transactions on Information Theory*, 62:952–969.
- [Ghinita et al., 2007] Ghinita, G., Karras, P., Kalnis, P., and Mamoulis, N. (2007). Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases*, pages 758–769. VLDB Endowment.
- [Hay, 2010] Hay, M. (2010). *Enabling Accurate Analysis of Private Network Data*. PhD thesis, University of Massachusetts.
- [Kleinberg and Papadimitriou, 1998] Kleinberg, J. and Papadimitriou, C. (1998). Segmentation problems. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 473–482. ACM.
- [Machanavajjhala et al., 2006] Machanavajjhala, A., Gehrke, J., and Kifer, D. (2006).  $\epsilon$ -diversity: Privacy beyond  $k$ -anonymity. In *null*, page 24. IEEE.

- [McGlinchey and Mason, 2017] McGlinchey, A. and Mason, O. (2017). Differential privacy and the  $l_1$  sensitivity of positive linear observers. *IFAC-PapersOnLine*, 50(1):3111 – 3116. 20th IFAC World Congress.
- [Meyerson and Williams, 2004] Meyerson, A. and Williams, R. (2004). On the complexity of optimal  $k$ -anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM.
- [Mohammadian et al., 2014] Mohammadian, E., Noferesti, M., and Jalili, R. (2014). Fast: fast anonymization of big data streams. In *Proceedings of the 2014 International Conference on Big Data Science and Computing*, page 23. ACM.
- [Murtagh and Vadhan, 2016] Murtagh, J. and Vadhan, S. (2016). The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pages 157–175. Springer.
- [Narayanan and Shmatikov, 2008] Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, SP '08, pages 111–125.
- [Ninghui, 2007] Ninghui, Li, T. L. S. V. (2007).  $l$ -diversity: Privacy beyond  $k$ -anonymity. *IEEE 23rd International Conference on Data Engineering*, pages 106–115.
- [Podgursky, 2011] Podgursky, B. (2011). *Practical  $K$ -anonymity on Large Datasets*. PhD thesis, Vanderbilt University.
- [Qu et al., 2017] Qu, Y., Yu, S., Gao, L., and Niu, J. (2017). Big data set privacy preserving through sensitive attribute-based grouping. *IEEE ICC 2017 Communication and Information Systems Security Symposium*.
- [Ruiz, 2018] Ruiz, N. (2018). On some consequences of the permutation paradigm for data anonymization: Centrality of permutation matrices, universal measures of disclosure risk and information loss, evaluation by dominance. *Information Sciences*, 430:620–633.
- [Simi et al., 2017] Simi, M. M. S., Nayaki, M. K. S., and Elayidom, D. M. S. (2017). An extensive study on data anonymization algorithms based on  $k$ -anonymity. *IOP Conference Series: Materials Science and Engineering*, 225(1):012279.
- [Sweeney, 2001] Sweeney, L. (2001). *Computational Disclosure Control: A Primer on Data Privacy Protection*. PhD thesis, Massachusetts Institute of Technology. AAI0803469.

- [Sweeney, 2002] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557 – 570.
- [Vadhan, 2017] Vadhan, S. (2017). The complexity of differential privacy. *Tutorials on the Foundations of Cryptography. Information Security and Cryptography*, pages 347–450. Publisher: Springer, Cham.