10. Νέες Τάσεις

Σύνοψη: *Στο παρόν κεφάλαιο θα αναφερθούμε στις νέες τάσεις των συστημάτων συστάσεων. Πιο συγκεκριμένα, θα αναλύσουμε τις συστάσεις σε* ***group (σύνολα)*** *ανθρώπων. Θα μελετήσουμε επίσης διαφορετικές μεθόδους συγκερασμού των προτιμήσεων πολλών χρηστών (π.χ.* ***Μέθοδος μέγιστης ικανοποίησης χρήστη***, ***Μέθοδος ελάχιστης δυσαρέσκειας χρήστη*** *κ.λπ.). Ακόμη, θα κάνουμε εκτεταμένη αναφορά στα φλέγοντα θέματα* ***ηθικής***, ***δικαιοσύνης***, ***λογοδοσίας*** *και* ***λογοκρισίας*** *που έχουν προκύψει και αποτελούν μια σύγχρονη τάση στα συστήματα συστάσεων. Τέλος, θα εμβαθύνουμε σε τρεις διαφορετικές πολιτικές προστασίας των προσωπικών δεδομένων (****αρχιτεκτονικές τοπολογίες συστημάτων συστάσεων****,* ***αλγοριθμικές τεχνικές****, και* ***κανονιστικό νομικό πλαίσιο****). Εξειδικεύοντας εδώ, μια νέα τάση στα συστήματα συστάσεων είναι τα δεδομένα με τις προτιμήσεις του χρήστη να μην πηγαίνουν στον server της εταιρείας για λόγους προστασίας της ιδιωτικότητάς του. Στην κατεύθυνση αυτή, θα παρουσιάσουμε αλγορίθμους που υλοποιούνται είτε μ' έναν* ***κατανεμημένο τρόπο (federated learning)*** *είτε μ' έναν* ***αποκεντροποιημένο (decentralized matrix factorization with differential privacy)*** *τρόπο εκπαίδευσης των μοντέλων πρόβλεψης.*

Προαπαιτούμενη γνώση: *Ο αναγνώστης καλό θα ήταν να έχει διαβάσει προηγουμένως το τέταρτο κεφάλαιο.*

10.1 Συστήματα Συστάσεων για Ομάδες Ατόμων

Όλες οι μέθοδοι που παρουσιάσαμε μέχρι στιγμής στα προηγούμενα εννέα κεφάλαια του παρόντος βιβλίου χρησιμοποιούνται για τη *σύσταση μιας λίστας προτεινόμενων στοιχείων* σ' **ένα μόνο άτομο**. Σ' αυτήν την ενότητα θα αναφερθούμε στο πώς οι παραπάνω μέθοδοι μπορούν να επεκταθούν και σε **μια ομάδα** που αποτελείται από **δύο ή περισσότερα άτομα**. Σ' αυτό το σημείο θα πρέπει να εστιάσουμε στις παρακάτω δύο στρατηγικές για τη δημιουργία της τελικής λίστας προτεινόμενων στοιχείων για τα μέλη μιας ομάδας:

- **Συγκερασμός των βαθμολογιών της ομάδας**: Η πρώτη στρατηγική αξιοποιεί τις βαθμολογίες των μελών μιας ομάδας για τα υπό εξέταση στοιχεία και βάσει αυτών δημιουργεί την τελική *διατεταγμένη λίστα προτεινόμενων στοιχείων.*

- **Συγκερασμός των λιστών κατάταξης**: Η δεύτερη στρατηγική δημιουργεί αρχικά μια λίστα προτεινόμενων στοιχείων για το κάθε μέλος της ομάδας ξεχωριστά, και στη συνέχεια προσπαθεί να τις συνενώσει σωρευτικά σε μια τελική λίστα με προτεινόμενα στοιχεία που αφορά σ' όλα τα μέλη της ομάδας.

10.1.1 Συγκερασμός των Βαθμολογιών της Ομάδας

Τα *συστήματα συστάσεων για ομάδες ατόμων* αφορούν σε περιπτώσεις όπως, για παράδειγμα, τη σύσταση ταινίας κατάλληλης για μια οικογένεια ή τη σύσταση ενός εστιατορίου για να συναντηθεί μια παρέα φίλων. Εν συνεχεία, θα εξετάσουμε διαφορετικές μεθόδους *συγκερασμού των βαθμολογιών των ατόμων μιας ομάδας* με τη χρήση ενός παραδείγματος:

Παράδειγμα 10.1 Έστω, λοιπόν, ότι ο Πίνακας 10.1 περιέχει τις βαθμολογίες 7 χρηστών σε 4 ταινίες στη βαθμολογική κλίμακα τιμών από 1 έως 5. Θεωρήστε ότι έχουμε μια ομάδα τριών ατόμων βάσει *φύλου* (*gender*) $G = \{,,\}$ που εμφανίζονται στον Πίνακα 10.1. Θεωρήστε, ακόμη, ότι έχουμε ένα σύνολο τεσσάρων ταινιών M = { *Titanic, Top Gun, Spider-Man, Gladiator*} που επίσης εμφανίζονται στον Πίνακα 10.1. Από τις ταινίες του συνόλου M, θέλουμε να προτείνουμε κάποιες ταινίες στην ομάδα ατόμων G, ώστε η λίστα των προτεινόμενων ταινιών να ικανοποιεί όσο το δυνατόν περισσότερο τα μέλη της ομάδας των αντρών.

Ταινίες ⇒ Χρήστες ⇓	Titanic	Top Gun	Spider-Man	Gladiator
Πέτρος	1	4	1	1
Νίκη	4	3	1	2
Θάνος	2	2	4	2
Ειρήνη	1	1	3	5
Κώστας	4	4	4	5
Στέλλα	1	4	1	1
Φωτεινή	3	4	4	1

Πίνακας 10.1: Πίνακας βαθμολογίας Χρηστών-Ταινιών.

Προκειμένου, συνεπώς, να συγκεράσουμε τις βαθμολογίες των μελών μιας ομάδας, μπορούμε να επιλέξουμε διάφορες στρατηγικές και μεθόδους (π.χ. *Μέθοδος Ελάχιστης Δυσαρέσκειας Χρήστη, Μέθοδος Μέγιστης Ικανοποίησης Χρήστη* κ.λπ.). Βάσει της επιλεγόμενης μεθόδου, οι ταινίες που θα

επιτύχουν τις υψηλότερες βαθμολογίες θα είναι και αυτές που θα συστήνονται στον χρήστη. Ακολούθως, θα περιγράψουμε τις *διαφορετικές μεθόδους και στρατηγικές συγκερασμού των βαθμολογιών* των χρηστών.

Μέθοδος Μέσης Ικανοποίησης Χρήστη

Με αυτήν τη μέθοδο υποθέτουμε ότι οι ταινίες που ικανοποιούν κατά μέσο όρο κάθε μέλος της ομάδας είναι οι καλύτερες που μπορούν να προταθούν σ' αυτήν. Συνεπώς, η συγκεντρωτική βαθμολογία R_m της ομάδας για την ταινία m με βάση τη *στρατηγική μεγιστοποίησης της μέσης ικανοποίησης χρήστη* δίνεται από την παρακάτω εξίσωση:

$$R_m = \frac{1}{n} \sum_{u \in G} r_{u,m}$$

όπου $r_{u,m}$ η βαθμολογία ενός χρήστη $u \in G$ για την ταινία $m \in M$, R_m η συγκεντρωτική βαθμολογία της ομάδας για την ταινία m, και n ο αριθμός των μελών της ομάδας. Άρα, για καθεμία από τις τέσσερεις ταινίες του Παραδείγματος 10.1 υπολογίζουμε τη βαθμολογία R_m της ως ακολούθως:

$$R_{\text{Titanic}} = \frac{1+2+4}{3} = 2.33 \qquad R_{\text{Top Gun}} = \frac{4+2+4}{3} = 3.33$$

$$R_{\text{Spider-Man}} = \frac{1+4+4}{3} = 3 \qquad R_{\text{Gladiator}} = \frac{1+2+5}{3} = 2.66$$

Τέλος, θα δημιουργήσουμε μια *διατεταγμένη λίστα προτεινόμενων ταινιών* για τα μέλη της ομάδας των αντρών που βρίσκονται στον Πίνακα 10.1, κατατάσσοντας τις ταινίες σε φθίνουσα διάταξη βάσει της υψηλότερης βαθμολογίας R_m που συγκεντρώνει η κάθε ταινία m. Η κορυφαία ταινία που θα τους προταθεί είναι η *Top Gun*, επειδή συγκέντρωσε τον μεγαλύτερο μέσο όρο βαθμολογίας, ίσο με 3.33 βαθμούς. Παρατηρήστε, βέβαια, το γεγονός ότι η ταινία αυτή δεν αρέσει καθόλου στον Πέτρο, που είναι ένα από τα τρία μέλη της ομάδας των αντρών, και που θα αναγκαστεί τελικά να τη δει. Γι' αυτόν τον λόγο θα εξετάσουμε παρακάτω και εναλλακτικούς τρόπους *συγκερασμού των βαθμολογιών των μελών της ομάδας* που ίσως να είναι περισσότερο δίκαιοι.

Μέθοδος Μέγιστης Ικανοποίησης Χρήστη

Η στρατηγική της *μέγιστης ικανοποίησης χρήστη* θεωρεί τη μέγιστη βαθμολογία που πήρε ένα στοιχείο στην ομάδα ως τη βαθμολογία της ομάδας και διατυπώνεται ως εξής:

$$R_m = \max_{u \in G} \; r_{u,m}$$

Η στρατηγική αυτή θέτει πολύ ψηλά τον πήχη με το να θεωρεί ότι, εφόσον υπάρχει έστω μία υψηλή αξιολόγηση γι' ένα στοιχείο, τότε υπάρχει πιθανότητα να αρέσει και στην ομάδα συνολικά. Συνεπώς, επί των δεδομένων του Παραδείγματος 10.1 και με δεδομένο ότι προκρίνουμε τις ταινίες που

έχουν λάβει τουλάχιστον μία πολύ μεγάλη βαθμολογία R_m, αυτή η στρατηγική εγγυάται ότι οι ταινίες που προτείνονται στην ομάδα θα αρέσουν πολύ **τουλάχιστον σ' ένα** μέλος της ομάδας. Συνεπώς, υπολογίζουμε τη βαθμολογία R_m για καθεμία από τις τέσσερεις ταινίες του Παραδείγματος 10.1 ως ακολούθως:

$$R_{\text{Titanic}} = \max\{1,2,4\} = 4 \qquad R_{\text{Top Gun}} = \max\{4,2,4\} = 4$$
$$R_{\text{Spider-Man}} = \max\{1,4,4\} = 4 \qquad R_{\text{Gladiator}} = \max\{1,2,5\} = 5$$

Κατόπιν, βάσει των δεδομένων του Παραδείγματος 10.1 και της *μεθόδου μέγιστης ικανοποίησης χρήστη* θα προταθεί η ταινία *Gladiator*, επειδή είναι η μόνη που βαθμολογήθηκε με 5 βαθμούς από τον χρήστη (user) Κώστα. Παρατηρούμε παρόλα αυτά στον Πίνακα 10.1 ότι δύο μέλη της ομάδας (Πέτρος και Θάνος) έχουν βαθμολογήσει την ταινία *Gladiator* με 1 και 2 βαθμούς αντίστοιχα. Αυτό σημαίνει ότι, παρόλο που δεν τους αρέσει, είναι "υποχρεωμένοι" να την παρακολουθήσουν.

Μέθοδος Ελάχιστης Δυσαρέσκειας Χρήστη

Η στρατηγική της *ελάχιστης δυσαρέσκειας χρήστη (least misery)* εγγυάται ότι σε κανένα άτομο/χρήστη δεν προτείνεται ένα στοιχείο που δεν του αρέσει καθόλου. Ακολουθώντας λοιπόν τη μέθοδο αυτή η συγκεντρωτική βαθμολογία R_m μιας ταινίας από τα μέλη μιας ομάδας δίνεται με την παρακάτω εξίσωση:

$$R_m = \min_{u \in G} \; r_{u,m}$$

Συνεπώς, για τα δεδομένα του Παραδείγματος 10.1 υπολογίζουμε τη βαθμολογία R_m για καθεμία από τις τέσσερεις ταινίες του ως ακολούθως:

$$R_{\text{Titanic}} = \min\{1,2,4\} = 1 \qquad R_{\text{Top Gun}} = \min\{4,2,4\} = 2$$
$$R_{\text{Spider-Man}} = \min\{1,4,4\} = 1 \qquad R_{\text{Gladiator}} = \min\{1,2,5\} = 1$$

Έπειτα, για τα δεδομένα του Παραδείγματος 10.1 και βάσει της *μεθόδου της ελάχιστης δυσαρέσκειας χρήστη* θα προταθεί η ταινία *Top Gun*, επειδή είναι εκείνη που εισέπραξε την ελάχιστη δυνατή *δυσαρέσκεια*. Αυτό σημαίνει ότι ένα μέλος της ομάδας τη βαθμολόγησε με δύο (2), περισσότερο δηλαδή από τις υπόλοιπες ταινίες που είχαν βαθμολογηθεί από κάποιο μέλος της ομάδας με χειρότερη βαθμολογία (δηλαδή τη μονάδα).

10.1.2 Συγκερασμός Λιστών Κατάταξης

Το πρόβλημα του *συγκερασμού διαφορετικών λιστών κατάταξης (ranking lists aggregation)* προκύπτει όταν διαθέτουμε μια διαφορετική *διατεταγμένη λίστα με προτεινόμενα στοιχεία* για κάθε μέλος

μιας ομάδας, και πρέπει να τις αθροίσουμε σωρευτικά για να κατασκευάσουμε μια νέα *τελική διατεταγμένη λίστα* που να αφορά σ' όλα τα μέλης της ομάδας και να είναι αποτέλεσμα μιας δίκαιης "συναίνεσης". Στο πλαίσιο λοιπόν αυτό των συστημάτων συστάσεων, με το πρόβλημα του *συγκερασμού των διαφορετικών λιστών κατάταξης* επιδιώκεται ο μετριασμός της τυχόν "μεροληψίας" που ενυπάρχει κατά τη φάση της σωρευτικής άθροισης των λιστών κατάταξης. Για τη φάση αυτήν υπάρχουν διάφορες προσεγγίσεις, και μεταξύ των σημαντικότερων εξ αυτών είναι και οι εξής:

- **Καταμέτρηση κατά Borda (Borda Count)**: Καταμετρά πόντους για το κάθε στοιχείο βάσει της θέσης του στις λίστες κατάταξης.
- **Βέλτιστη συνάθροιση κατάταξης κατά Kemeny (Kemeny optimal rank aggregation)**: Επιλέγεται ως τελική διατεταγμένη λίστα εκείνη που έχει τη μικρότερη μέση απόσταση από τις λίστες κατάταξης των μελών της ομάδας.

Εν συνεχεία, θα εξετάσουμε δύο διαφορετικές μεθόδους *συγκερασμού των λιστών κατάταξης των ατόμων μιας ομάδας* με τη χρήση του ακόλουθου παραδείγματος:

Παράδειγμα 10.2 Δίνεται, λοιπόν, ο Πίνακας 10.2 που περιέχει την αξιολόγηση τριών μελών μιας ομάδας για τρεις ταινίες. Καθένα από τα τρία μέλη μιας ομάδας φίλων έχει δηλώσει μια λίστα με τη σειρά προτίμησής του για τις τρεις ταινίες (βλ. τον Πίνακα 10.2). Κάθε μέλος της ομάδας κατατάσσει ξεχωριστά τις υποψήφιες ταινίες στις τρεις θέσεις κατάταξης $\{A, B, \Gamma\}$. Καλείται δε η ομάδα να καταλήξει σε μια τελική κατάταξη αυτών των ταινιών ως αποτέλεσμα μιας δίκαιης "συναίνεσης" μεταξύ των μελών της.

Μέλη Ομάδας ⇒ Ταινίες ⇓	Μέλος 1	Μέλος 2	Μέλος 3
Titanic	A	B	B
Top Gun	B	Γ	A
Gladiator	Γ	A	Γ

Πίνακας 10.2: Συγκριτική κατάταξη των ταινιών από τα μέλη μιας ομάδας.

Καταμέτρηση κατά Borda (Borda count)

Η *καταμέτρηση κατά Borda* προσμετρά πόντους για το κάθε στοιχείο βάσει της θέσης κατάταξης που αυτό έχει στη λίστα προτίμησης του κάθε μέλους της ομάδας. Έτσι, το τελευταίο στοιχείο στη λίστα προτίμησης ενός μέλους της ομάδας παίρνει μηδέν (0) πόντους, το προτελευταίο στοιχείο παίρνει 1 πόντο και ούτω καθεξής. Στην περίπτωσή μας, βάσει της *καταμέτρησης κατά Borda* η ταινία *Titanic* λαμβάνει συνολικά 4 πόντους, οι οποίοι αναλύονται ως ακολούθως:

- 2 βαθμούς, επειδή το μέλος 1 της ομάδας κατέταξε το στοιχείο (δηλαδή την ταινία) πρώτο,
- 1 βαθμό, επειδή το μέλος 2 της ομάδας κατέταξε το στοιχείο δεύτερο, και
- 1 βαθμό, επειδή το μέλος 3 της ομάδας κατέταξε το στοιχείο δεύτερο.

Με αντίστοιχο τρόπο υπολογισμού, η ταινία *Top Gun* συγκεντρώνει 3 πόντους, ενώ 2 η ταινία *Gladiator*. Συνεπώς, η τελική κατάταξη των ταινιών είναι {*Titanic, Top Gun, Gladiator*}.

Βέλτιστη Σωρευτική 'Αθροιση Λιστών Κατάταξης κατά Kemeny

Η συγκεκριμένη μέθοδος απαιτεί την ύπαρξη μιας *συνάρτησης μέτρησης της απόστασης μεταξύ δύο λιστών κατάταξης* (π.χ. η *συνάρτηση απόστασης Kendall* τ). Δοθέντων των λιστών κατάταξης που θέλουμε να *συγκεράσουμε* και μιας συνάρτησης μέτρησης της *απόστασης* μεταξύ τους, ο στόχος είναι να υπολογιστεί μια νέα διατεταγμένη λίστα η οποία θα ελαχιστοποιεί τη *μέση απόσταση* από τις λίστες κατάταξης των μελών της ομάδας.

Η πιο γνωστή *συνάρτηση μέτρησης της απόστασης μεταξύ δύο λιστών κατάταξης* είναι η *Kendall* τ. Πιο συγκεκριμένα, οποιαδήποτε λίστα κατάταξης μπορεί να αναπαρασταθεί με μια ακολουθία από *συγκρίσεις* μεταξύ των στοιχείων της *ανά δύο* (*pairwise comparisons*).

Για δύο λίστες κατάταξης ρ και σ, λοιπόν, η σύγκριση για το ζευγάρι στοιχείων τους (i, j) βρίσκεται σε *συμφωνία (concordant)* εάν η σειρά κατατάξης των στοιχείων $\{\rho_i, \sigma_i\}$ και $\left\{\rho_j, \sigma_j\right\}$ στις δύο λίστες είναι ίδια:

$$\rho_i > \rho_j, \quad \sigma_i > \sigma_j \quad \text{ή} \quad \rho_i < \rho_j, \quad \sigma_i < \sigma_j.$$

Αντίθετα, για δύο λίστες κατάταξης ρ και σ η σύγκριση για το ζευγάρι στοιχείων τους (i, j) βρίσκεται σε *ασυμφωνία (disconcordant)* εάν η σειρά κατάταξης των στοιχείων $\{\rho_i, \sigma_i\}$ και $\left\{\rho_j, \sigma_j\right\}$ στις δύο λίστες είναι αντίθετη:

$$\rho_i > \sigma_j, \quad \rho_i < \sigma_j \quad \text{ή} \quad \rho_i < \sigma_j, \quad \rho_i > \sigma_j$$

Τονίζουμε ότι, όταν $\rho_i = \sigma_j$, τότε το ζευγάρι στοιχείων δεν χαρακτηρίζεται ούτε ως *σύμφωνο* ούτε ως *ασύμφωνο*. Στον Πίνακα 10.2, εάν συγκρίνουμε τα στοιχεία Α και Β για τις λίστες του μέλους 1 {Α, Β, Γ} και του μέλους 2 {Β, Γ, Α} θα δούμε ότι βρίσκονται σε *ασυμφωνία*, διότι στη λίστα κατάταξης του μέλους 1 το Α>Β, ενώ στη λίστα κατάταξης του μέλους 2 το Β>Α.

Έστω δε ότι το c συμβολίζει τον συνολικό αριθμό των *σύμφωνων* ζευγαριών στοιχείων και το d τον συνολικό αριθμό των *ασύμφωνων* ζευγαριών στοιχείων. Η *Kendall* τ υπολογίζει τη διαφορά μεταξύ των δύο κανονικοποιημένη με βάση τον συνολικό αριθμό ζευγαριών στοιχείων $\begin{pmatrix} n \\ 2 \end{pmatrix}$. Υπογραμμίζουμε ότι γι' έναν αριθμό n στοιχείων ο συνολικός αριθμός των ζευγαριών των στοιχείων ανά δύο που μπορούν να παραχθούν είναι $\frac{n!}{2!(n-2)!}$. Δηλαδή, ισχύει $\begin{pmatrix} n \\ 2 \end{pmatrix} = \frac{n!}{2!(n-2)!}$. Στο τρέχον

Παράδειγμα, 10.2, η λίστα κατάταξης του μέλους 1 είναι {Α, Β, Γ} και συνεπώς τα ζευγάρια που μπορούν να συγκριθούν είναι 3: το Α με το Β, το Α με το Γ, και το Β με το Γ. Η *συνάρτηση απόστασης Kendall* τ για δύο λίστες κατάταξης $\rho, \sigma \in S_n$ ορίζεται στην παρακάτω Εξίσωση:

$$\tau(\rho, \sigma) = \frac{c - d}{\binom{n}{2}}.$$

Στον Πίνακα 10.2, εάν συγκρίνουμε τις λίστες κατάταξης του μέλους 1 {Α, Β, Γ} και του μέλους 2 {Β, Γ, Α} βάσει του παραπάνω μαθηματικού τύπου, τότε *Kendall* τ= -0.3333. Στον δε Πίνακα 10.3 εμφανίζονται οι *αποστάσεις/ομοιότητες* μεταξύ όλων (τριών δηλαδή) των λιστών κατάταξης που μπορούν να παραχθούν από τα τρία στοιχεία {Α, Β, Γ} του Παραδείγματος 10.2.

	Α Β Γ	Α Γ Β	Β Α Γ	Β Γ Α	Γ Α Β	Γ Β Α
Α Β Γ	1	0.333	0.333	−0.333	−0.333	−1
Α Γ Β	0.333	1	−0.333	−1	0.333	−0.333
Β Α Γ	0.333	−0.333	1	0.333	−1	−0.333
Β Γ Α	−0.333	−1	0.333	1	−0.333	0.333
Γ Α Β	−0.333	0.333	−1	−0.333	1	0.333
Γ Β Α	−1	−0.333	−0.333	0.333	0.333	1

Πίνακας 10.3: Πίνακας *αποστάσης/ομοιότητας* μεταξύ των λιστών κατάταξης βάσει του μέτρου *Kendall* τ.

Τονίζουμε, εδώ, ότι για έναν αριθμό n στοιχείων το σύνολο S_n όλων των διαφορετικών λιστών κατάταξης που μπορούν να παραχθούν είναι $\boldsymbol{n!}$. Στον πίνακα του Παραδείγματος 10.2 εμφανίζονται μόνο οι 3 από τις 6 συνολικά λίστες κατάταξης που μπορούν να παραχθούν από τη μετάθεση των 3 ταινιών. Δοθέντων ενός υποσυνόλου λιστών κατάταξης $R \subseteq S_n$ των μελών μιας ομάδας και της *συνάρτησης απόστασης Kendall* τ, η *βέλτιστη σωρευτική άθροιση λιστών κατάταξης κατά Kemeny* υπολογίζει μια νέα διατεταγμένη λίστα ρ^* που ελαχιστοποιεί τη *μέση απόσταση* από το σύνολο των λιστών κατάταξης του υποσυνόλου R ως ακολούθως:

$$\rho^* = \underset{\rho \in S_n}{\operatorname{argmin}} \frac{1}{|R|} \sum_{\sigma \in R} \tau(\rho, \sigma)$$

Τονίζουμε ότι η *βέλτιστη σωρευτική άθροιση λιστών κατάταξης κατά Kemeny* θεωρείται ο πιο αξιόπιστος τρόπος για τον *συγκερασμό των λιστών κατάταξης*, εξαιτίας του ότι ικανοποιεί πολλά αξιώματα της *Θεωρίας της Κοινωνικής Επιλογής* (*Social Choice and Arrow's impossibility theorem*), όπως, ενδεικτικά, την ύπαρξη απεριόριστων επιλογών για κάθε άτομο και τη μη εξάρτηση των προτίμησεων του ατόμου από αυτές των άλλων.

Εάν λοιπόν εφαρμόσουμε τον παραπάνω μαθηματικό τύπο στον πίνακα του Παραδείγματος 10.2, τότε η βέλτιστη *λίστα κατάταξης κατά Kemeny* είναι η {Β Α Γ}, δεδομένου ότι έχει τη μεγαλύτερη *μέση ομοιότητα* με τις τρεις λίστες κατάταξης των μελών της ομάδας όπως φαίνεται στον παρακάτω πίνακα:

Λίστα Κατάταξης	Μέση Ομοιότητα = 1-Απόσταση
Β Α Γ	0.220
Β Γ Α	0
Α Β Γ	0
Γ Α Β	−0.553
Γ Β Α	−0.333
Α Γ Β	−0.333

Πιο συγκεκριμένα, για τη λίστα κατάταξης {Β Α Γ} υπολογίζουμε την *ομοιότητά* της με τις λίστες κατάταξης των μελών της ομάδας ως ακολούθως:

$$\rho^*\{\text{Β Α Γ}\} = \frac{\tau(\{\text{Β Α Γ}\},\{\text{Α Β Γ}\}) + \tau(\{\text{Β Α Γ}\},\{\text{Β Γ Α}\})}{3} = \frac{0.333 + 0.333}{3} = 0.22 \tag{10.1}$$

Συνεπώς, βάσει της *βέλτιστης σωρευτικής άθροισης λιστών κατάταξης κατά Kemeny* η τελική λίστα κατάταξης των ταινιών του παραδείγματός μας είναι {*Top Gun, Titanic, Gladiator*}. Να σημειώσουμε εδώ ότι για τα τρία μόνο υποψήφια στοιχεία του Παραδείγματος 10.2 παράγονται 6 διαφορετικές λίστες κατάταξης και απαιτούνται τουλάχιστον 15 συγκρίσεις ζευγαριών στοιχείων μεταξύ τους. Γίνεται επομένως σαφές ότι η προαναφερθείσα *κατά Kemeny κατάταξη*, όπως έχει αποδειχθεί, είναι ένα *NP-hard* (*Non deterministic Polynomial time*) πρόβλημα (δηλαδή δεν λύνεται σε *πολυωνυμικό χρόνο*), το οποίο όμως μπορεί να λυθεί με *γραμμικό προγραμματισμό*, εφόσον προηγουμένως το εκφράσουμε ως ένα *minimum weighted feedback arc* πρόβλημα.

10.2 Ηθική και Συστήματα Συστάσεων

Η ηθική, γενικώς, είναι ένα πλαίσιο κριτηρίων που θέτει η κοινωνία στα μέλη της για να κρίνουν ποιες ανθρώπινες πράξεις είναι αποδεκτές και ορθές και ποιες ανάρμοστες και λανθασμένες. Εξειδικεύοντας τον παραπάνω ορισμό στα δικά μας, η ηθική στα συστήματα συστάσεων σχετίζεται με τον τρόπο που: (i) αναπτύσσουμε "δίκαιους" αλγορίθμους, (ii) εκπαιδεύουμε "αμερόληπτα" μοντέλα πρόβλεψης για την παραγωγή της λίστας προτεινόμενων στοιχείων, (iii) αξιολογούμε το περιεχόμενο που συνεισφέρουν οι χρήστες των μέσων κοινωνικής δικτύωσης (π.χ. προκλητικό ή παραπλανητικό περιεχόμενο).

Αναγνωρίζοντας ότι οι εκάστοτε κοινωνικές προκαταλήψεις μπορεί να είναι εγγενείς στα δεδομένα εκπαίδευσης, μια σύσταση θεωρείται δίκαιη όταν δεν τις ενισχύει. Άρα, θα πρέπει να υπάρχει

*δικαιοσύνη (**fairness**)* ως προς τον τρόπο που εκπαιδεύουμε τα μοντέλα των αλγορίθμων μας βάσει των δεδομένων εκπαίδευσης.

Επιπροσθέτως, οι επιχειρήσεις και λοιποί οργανισμοί πρέπει να θεωρούνται υπεύθυνοι για τις αποφάσεις που λαμβάνονται από τους αλγορίθμους που χρησιμοποιούν, και να είναι σε θέση να *λογοδοτούν* με μετρήσιμο τρόπο (***accountability***) αλλά και να τεκμηριώνουν επαρκώς την αυτοματοποιημένη λήψη μιας απόφασης.

Τέλος, πολλοί χρήστες χρησιμοποιούν καταχρηστικά τα εργαλεία των μέσων κοινωνικής δικτύωσης, ενδεικτικά, για να *παρεμβαίνουν στις εθνικές εκλογές*, να *διαδίδουν ψευδείς ειδήσεις (fake news)*, να *παραπληροφορούν* ή ακόμα και να *υποκινούν τρίτους σε βίαιες πράξεις*. Αποτέλεσμα όλων αυτών είναι η αύξηση του *δυσφημιστικού λόγου (defamation speech)*, η *κοινωνική πόλωση (social polarization)* και ο *εξτρεμισμός (extremism)* στα μέσα κοινωνικής δικτύωσης.

Συνάγουμε το συμπέρασμα, λοιπόν, ότι τα συστήματα συστάσεων θα πρέπει να είναι ικανά να αξιολογούν το προσφερόμενο από τους χρήστες περιεχόμενο μ' έναν δίκαιο και αμερόληπτο τρόπο, ώστε να μην κατηγορούνται για παράνομη **λογοκρισία** (**censorship**) του περιεχομένου που αναρτούν οι χρήστες, αλλά και να μη συμβάλλουν, από την άλλη, στη διασπορά *ψευδών ειδήσεων* και *δυσφημιστικού λόγου*.

Στις ακόλουθες ενότητες λοιπόν θα αναλύσουμε τα θέματα που αφορούν στη *δικαιοσύνη*, τη *λογοδοσία* και τη *λογοκρισία περιεχομένου* στο πλαίσιο των συστημάτων συστάσεων.

10.2.1 Δικαιοσύνη (Fairness)

Στα συστήματα συστάσεων με τον όρο *δικαιοσύνη* αναφερόμαστε στην εξάλειψη κάθε είδους *διακρίσεων ή μεροληψίας* που μπορεί να αποτυπώνονται στη λίστα προτεινόμενων στοιχείων έναντι *προστατευόμενων ομάδων* (π.χ. *άτομα με αναπηρία*). Γενικά, υπάρχουν δύο βασικές προσεγγίσεις επ' αυτού [Pitoura et al., 2021]:

- Η *ατομική δικαιοσύνη*, η οποία βασίζεται στην παραδοχή ότι παρόμοιες "οντότητες" πρέπει να αντιμετωπίζονται με παρόμοιο τρόπο.
- Η *ομαδική δικαιοσύνη*, η οποία ομαδοποιεί τις "οντότητες" καταρχάς με βάση ένα ή περισσότερα προστατευόμενα χαρακτηριστικά (π.χ. τη *σωματική* και *νοητική αναπηρία* κ.λπ.), και στη συνέχεια επιδιώκει να αντιμετωπίζονται όλες οι ομάδες με παρόμοιο τρόπο.

Στα συστήματα συστάσεων, η *ατομική δικαιοσύνη* επιτυγχάνεται όταν η λίστα προτεινόμενων στοιχείων είναι δομημένη με τέτοιον τρόπο, ώστε η αναλογία των στοιχείων που αφορούν σε *διαφορετικές κατηγορίες ενδιαφέροντος* **(output bias)** να είναι η ίδια με την αναλογία που υπάρχει στις *προτιμήσεις εισόδου* του υπό εξέταση χρήστη **(input bias)**. Σε διαφορετική περίπτωση, θα πρέπει να ακολουθηθεί μια μέθοδος επανακατάταξης της λίστας προτεινόμενων στοιχείων για την επίτευξη *δικαιοσύνης*, έτσι ώστε να συμπεριληφθούν αντιπροσωπευτικά στοιχεία που αφορούν σ' όλες τις

κατηγορίες ενδιαφέροντος του χρήστη. Μ' άλλα λόγια, για έναν υπό εξέταση χρήστη, μια λίστα προτεινόμενων στοιχείων είναι δίκαια δομημένη εάν περιέχει στοιχεία από διάφορες κατηγορίες στοιχείων (π.χ. για ειδησεογραφικά άρθρα η κατηγορία *πολιτική*, η κατηγορία *σπορ* κ.λπ.) με αναλογία που να είναι ίση με την αναλογία των κατηγοριών ενδιαφέροντος που υπάρχει στις προτιμήσεις του χρήστη. Για παράδειγμα, εάν σ' έναν χρήστη έχουν αρέσει 7 ρομαντικές ταινίες και 3 δράσης, μια λίστα συστάσεων θα ήταν σχετικά δικαιότερη εάν περιείχε 70% ρομαντικές ταινίες και 30% δράσης.

Επίσης, ένα γνωστό πρόβλημα *μεροληψίας (bias)* που υπάρχει στα συστήματα συστάσεων αφορά στη **μεροληψία θέσης (position bias)** των στοιχείων που προτείνονται, το οποίο συνίσταται στο ότι *οι άνθρωποι τείνουν να εξετάζουν μόνο τα στοιχεία που εμφανίζονται στις πρώτες θέσεις* μιας διατεταγμένης λίστας στοιχείων. Λόγου χάρη, η προσοχή που λαμβάνουν τα στοιχεία που βρίσκονται στο πάνω μέρος μιας ιστοσελίδας (π.χ. οι τίτλοι μιας διαδικτυακής εφημερίδας) είναι πάντα μεγαλύτερη συγκριτικά με τις ειδήσεις που εμφανίζονται στο κάτω μέρος της. Συνεπώς, σύμφωνα με την αρχή της τήρησης **ομοιότητας ως προς την έξοδο (output similarity)** ενός αλγορίθμου θα πρέπει να δίνονται ίσες δυνατότητες κατάταξης παρόμοιων στοιχείων. Είναι λοιπόν προφανές ότι τα στοιχεία που εμφανίζονται στο πάνω μέρος μιας ιστοσελίδας, και με τα οποία αλληλεπιδρούν οι χρήστες περισσότερο, θα προκρίνονται έναντι άλλων στοιχείων κατά τη διάρκεια εκπαίδευσης ενός μοντέλου πρόβλεψης. Συνεπώς, για να είναι περισσότερο *δίκαιοι* οι αλγόριθμοι, θα πρέπει να εμπεριέχουν κάποιο φίλτρο μείωσης της *μεροληψίας* που ενυπάρχει λόγω της αρχικής προνομιακής τοποθέτησης αυτών των στοιχείων. Στο σημείο αυτό να τονίσουμε ότι στις κοινωνικές επιστήμες υπάρχει μια θεμελιώδης διάκριση μεταξύ **ισότητας (equality)** και **ίσων ευκαιριών (equity)**: η *ισότητα* αναφέρεται στην ίση μεταχείριση των *οντοτήτων*, ενώ οι *ίσες ευκαιρίες* αναφέρονται στην αντιμετώπιση των *οντοτήτων* ανάλογα με τις ανάγκες τους, έτσι ώστε όλοι τελικά να λαμβάνουν το ίδιο αποτέλεσμα (ίσες ευκαιρίες), ακόμη και όταν ορισμένα άτομα βρίσκονται σε μειονεκτική θέση.

Επίσης, σε πολλές περιπτώσεις τα συστήματα συστάσεων αναπαράγουν κοινωνικές προκαταλήψεις, εξαιτίας του ότι τα μοντέλα πρόβλεψης εκπαιδεύονται χρησιμοποιώντας *μεροληπτικά (biased)* δεδομένα. Παραδείγματος χάρη, σε μια μηχανή συστάσεων θέσεων εργασίας μια γυναίκα-χρήστης που ενδιαφέρεται να δουλέψει στον στρατό θα ανέμενε να της προτείνονταν θέσεις σ' αυτόν, αντανακλώντας την πραγματική κατανομή στον γενικό πληθυσμό των ανδρών έναντι των γυναικών (περίπου 50%-50%). Εάν μία μηχανή συστάσεων, λοιπόν, προτείνει έναν δυσανάλογα μικρό αριθμό θέσεων εργασίας στον στρατό για τις γυναίκες-χρήστες, τότε μπορεί να εκληφθεί ως *μεροληπτική*. Ένα τέτοιο *μεροληπτικό* περιβάλλον συστάσεων για θέσεις εργασίας, π.χ. σε ανέργους, δυνητικά επηρεάζει τις αντιλήψεις και τις πεποιθήσεις των χρηστών για διάφορα επαγγέλματα. Ως εκ τούτου, και προκειμένου να εξαλειφθεί η παραπάνω *μεροληψία* και να γίνει ένα σύστημα συστάσεων *δικαιότερο*, θα πρέπει πολλές φορές να υπάρξει μια *αντιστάθμιση* (*trade off*) μεταξύ της αποτελεσματικότητας της λίστας προτεινόμενων στοιχείων και του βαθμού *αμεροληψίας (unbiasedness)* του συστήματος συστάσεων, όπως θα φανεί και στο παράδειγμα που ακολουθεί.

Αντιστάθμιση μεταξύ "Ακρίβειας Συστάσεων" και "Αμεροληψίας Συστάσεων"

Παράδειγμα 10.3 Έστω, λοιπόν, ότι διαχειριζόμαστε ένα σύστημα συστάσεων που προτείνει θέσεις εργασίας σε ανέργους. Έστω, ακόμη, ότι το σύστημα συστάσεων προβλέπει για 200 διαφορετικούς ανθρώπους την πιθανότητα να τους αρέσει ή όχι μια θέση εργασίας στον στρατό. Οι σχετικές με το παραπάνω ζητούμενο προβλέψεις του συστήματος συστάσεων εμφανίζονται στον Πίνακα 10.4.

		Πρόβλεψη	
		ΑΛΗΘΕΣ	**ΨΕΥΔΕΣ**
Γεγονός	**ΑΛΗΘΕΣ**	*60*	*20*
	ΨΕΥΔΕΣ	*40*	*80*

Πίνακας 10.4: *Πίνακας συγκεχυμένων (confusion) αποτελεσμάτων πρόβλεψης* σχετικώς με το εάν μια προτεινόμενη θέση εργασίας στον στρατό θα ενδιέφερε ή όχι έναν πληθυσμό από 200 ανέργους.

Όπως φαίνεται στο πάνω αριστερό τμήμα του Πίνακα 10.4, το σύστημα ορθώς προβλέπει για 60 ανθρώπους ότι θα τους αρέσει η συγκεκριμένη θέση εργασίας (*True Positives*). Από την άλλη μεριά, όπως φαίνεται στο πάνω δεξιό τμήμα του Πίνακα 10.4, από τους ανθρώπους για τους οποίους το σύστημα συστάσεων προέβλεψε ότι δεν θα τους αρέσει η θέση εργασίας στον στρατό, σε 20 τελικά εξ αυτών αρέσει (*False Negatives*). Ακόμη, οι προβλέψεις που καταμετρώνται ως *ψευδώς θετικές (False Positive)* αφορούν σε 40 ανθρώπους, δεδομένου ότι ο αλγόριθμος εσφαλμένα προέβλεψε ότι θα τους αρέσει η θέση εργασίας, ενώ στην πραγματικότητα δεν ισχύει αυτό. Τέλος, 80 άνθρωποι καταμετρώνται ως *True Negatives*, επειδή ορθώς το σύστημα συστάσεων προέβλεψε ότι η θέση εργασίας δεν θα τους αρέσει. Από τον Πίνακα 10.4 εύκολα μπορούμε να υπολογίσουμε ότι η *ακρίβεια (precision)* του συστήματος συστάσεων ισούται με 60% (60/(60+40)=0.6), γεγονός που την καθιστά αρκετά ικανοποιητική.
Ας υποθέσουμε, τώρα, ότι μας αποκαλύπτεται το φύλο των 200 ανθρώπων, ώστε να μπορούμε να τους διαχωρίσουμε σε γυναίκες και άνδρες. Στον παρακάτω Πίνακα 10.5 εμφανίζουμε τις προβλέψεις του συστήματος συστάσεων ανά φύλο:

	Δεν αρέσει η θέση εργασίας στον στρατό	Αρέσει η θέση εργασίας στον στρατό	Σύνολο
Γυναίκα	80	20	100
Άντρας	20	80	100
Σύνολο	100	100	200

Πίνακας 10.5: Πίνακας Ενδεχομένων (Contigency Matrix).

Αναλυτικότερα, το σύστημα συστάσεων για το 80% των γυναικών προβλέπει ότι δεν θα τους αρέσει η θέση εργασίας στον στρατό, ενώ για το 80% των ανδρών προβλέπει ότι θα τους αρέσει. Βάσει του παραπάνω *Πίνακα Ενδεχομένων* κάποιος από τους επιστήμονες που σχεδίασαν το σύστημα συστάσεων θα μπορούσε να ισχυριστεί ότι αυτό *μεροληπτεί*, επειδή παράγει πολύ διαφορετικές προβλέψεις για τα δύο φύλα. Συγκεκριμένα, το σύστημα συστάσεων φαίνεται να αναπαράγει την κοινωνική προκατάληψη έναντι των γυναικών ότι μάλλον δεν είναι πολύ σχετικές με τον στρατό και συνεπώς δεν ενδιαφέρονται να εργαστούν εκεί. Υπάρχει, άρα, μεγάλη πιθανότητα τα δεδομένα εκπαίδευσης του συστήματος συστάσεων να ήταν *μεροληπτικά* και επομένως και οι προβλέψεις του συστήματος συστάσεων να αντανακλούν το γεγονός αυτό.

Επίσης, με μια πρώτη ματιά, το σύστημα συστάσεων φαίνεται να είναι αρκετά ακριβές και αποτελεσματικό - από τον *Πίνακα Συγκεχυμένων Αποτελεσμάτων Πρόβλεψης* 10.4 βλέπουμε ότι το σύστημα συστάσεων πετυχαίνει 60% *ακρίβεια* σε σύνολο 200 ανθρώπων. Έστω, λοιπόν, ότι από τους 60 ανθρώπους για τους οποίους έκανε σωστή πρόβλεψη (*true positives* στον Πίνακα 10.4) οι 56 είναι άντρες και οι 4 γυναίκες. Σ' αυτήν την περίπτωση, τα ποσοστά των *αληθινά θετικών (true positives)* είναι 70% (56/80=0.7) για τους άντρες και 20% (4/20=0.2) για τις γυναίκες, διότι στον παραπάνω Πίνακα μπορούμε να δούμε ότι σε 80 άντρες αρέσει η θέση εργασίας, ενώ ως προς τις γυναίκες, αρέσει μόνο σε 20. Ακόμη, το *ψευδώς θετικό* ποσοστό των γυναικών είναι 80% -16 γυναίκες από τις 20 για τις οποίες προβλέφθηκε ότι θα τους άρεσε μια θέση εργασίας στον στρατό-, και είναι σχεδόν τριπλάσιο από το *ψευδώς θετικό* ποσοστό (*false positives*) των αντρών - 24 άντρες από τους 80.

Από τα παραπάνω ποσοστά, επομένως, γίνεται εύκολα αντιληπτό ότι οι προβλέψεις του συστήματος συστάσεων δεν είναι *αμερόληπτες* και *δίκαιες*. Ο λόγος είναι ότι το ποσοστό των γυναικών στις σωστές προβλέψεις είναι εξαιρετικά δυσανάλογο σε σχέση με το ποσοστό των αντρών και άρα το σύστημα συστάσεων *μεροληπτεί*. Αντιθέτως, βάσει της αρχής των *Ίσων Ευκαιριών (Equal Opportunities)*, το σύστημα συστάσεων θα πρέπει να εγγυάται ότι οι διάφοροι τύποι σφαλμάτων κατηγοριοποίησης (π.χ. *ψευδώς αρνητικά, ψευδώς θετικά*) θα πρέπει να είναι ίσοι ως προς τα δύο φύλα.

Τέλος, κάποιος θα μπορούσε να επικαλεστεί το γεγονός ότι οι γυναίκες αποτελούν το 50% του συνολικού πληθυσμού στην Ελλάδα και οι άνδρες το υπόλοιπο 50%, και συνεπώς θα ήταν δίκαιο να έχουν πρόσβαση με ανάλογη ποσόστωση στις προτεινόμενες θέσεις εργασίας. Στην περίπτωσή μας όμως, δεδομένου ότι το σύστημα συστάσεων για το 80% των γυναικών προβλέπει ότι δεν θα τους αρέσει η θέση εργασίας στον στρατό, ενώ για το 80% των ανδρών προβλέπει το αντίθετο, κάτι τέτοιο δεν συμβαίνει. Ως εκ τούτου, για να αντιμετωπιστεί το παραπάνω ζήτημα, θα πρέπει στις λίστες προτεινόμενων θέσεων εργασίας να συμπεριληφθούν με αντιπροσωπευτικό τρόπο και θέσεις που να αφορούν και στα δύο φύλα, ώστε να υπάρξει μια δίκαιη αναλογία (π.χ. 50%- 50%) σε θέσεις εργασίας στον στρατό. Αυτό όμως με

τη σειρά του μπορεί να αυξήσει κατά πολύ το *ψευδώς θετικό* ποσοστό στις προβλέψεις που αφορούν στις γυναίκες και συνεπώς να οδηγήσει σε περαιτέρω μείωση της *ακρίβειας* του συστήματος συστάσεων.
Συνοψίζοντας, επανερχόμαστε σε όσα προείπαμε στην αρχή της υποενότητας "προλογίζοντας" το παραπάνω παράδειγμα, ότι, δηλαδή, πολλές φορές είναι πιθανό να πρέπει να υπάρξει μια *αντιστάθμιση* μεταξύ της *ακρίβειας (precision)* και του βαθμού *αμεροληψίας (unbiasedness)* του συστήματος συστάσεων προκειμένου οι αλγόριθμοί μας να γίνονται *δικαιότεροι.*

10.2.2 Λογοδοσία (Accountability)

Οι αλγόριθμοι *βαθιάς μάθησης (deep learning)* έχουν κυριαρχήσει πλήρως τα τελευταία χρόνια στα συστήματα συστάσεων, εξαιτίας της μεγάλης αποτελεσματικότητας που επιτυγχάνουν ως προς την *ακρίβεια* της λίστας προτεινόμενων στοιχείων, η οποία παράγεται από τη χρήση ενός *νευρωνικού δικτύου πάρα πολλών στρωμάτων (νευρώνων).* Παρόλα αυτά, τέτοιοι αλγόριθμοι δεν είναι εύκολο να *αιτιολογήσουν* επαρκώς τα προτεινόμενα στοιχεία της λίστας συστάσεων, διότι παράγουν πολύπλοκα μοντέλα πρόβλεψης, τα οποία, εύλογα, δεν είναι εύκολα ερμηνεύσιμα λόγω του ότι εμπεριέχουν πολλά διαφορετικά *στρώματα νευρώνων (neuron layers)* - όπως ήδη τονίσαμε. Γι' αυτόν τον λόγο στη βιβλιογραφία αναφέρονται και ως *αλγόριθμοι μαύρα κουτιά (black box algorithms),* εξαιτίας του ότι η αποτελεσματικότητά τους δεν είναι εύκολα ερμηνεύσιμη.

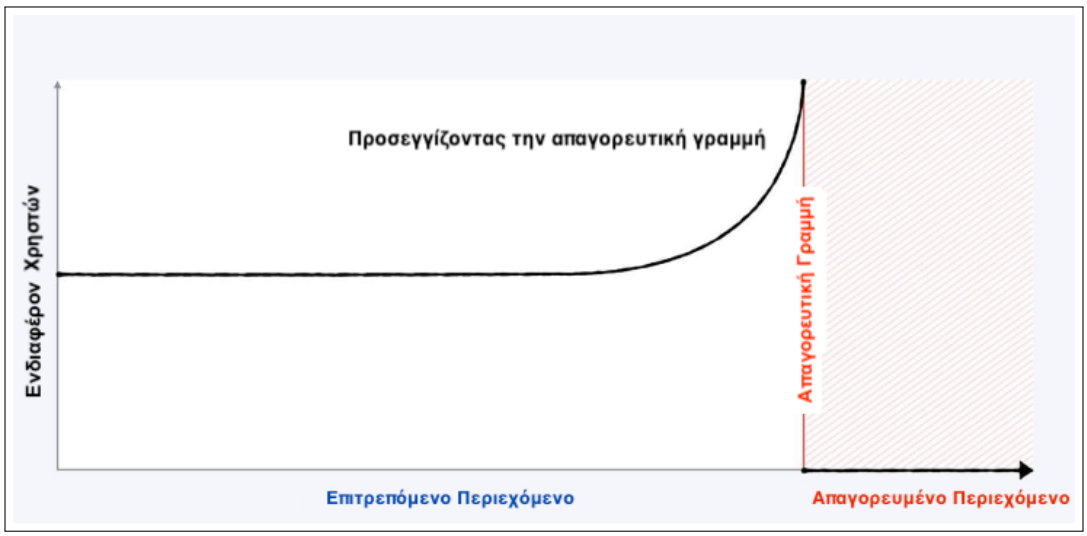
Από την άλλη μεριά όμως, ο τρόπος λειτουργίας των συστημάτων συστάσεων θα πρέπει να είναι *διάφανος* για να αυξηθεί η εμπιστοσύνη των χρηστών σ' αυτά. Λόγου χάρη, μια μικρή τροποποίηση ενός αλγορίθμου συστάσεων μπορεί να έχει ως αποτέλεσμα να προτείνεται συνεχώς ένα προϊόν που είναι σε στοκ και η επιχείρηση απλώς θέλει να το αποσύρει από τις αποθήκες της, ενώ στην πραγματικότητα κανένας δεν θα ενδιαφερόταν να το αγοράσει. Σ' αυτό λοιπόν το σημείο θα πρέπει να υπάρχει η παρέμβαση μιας ρυθμιστικής αρχής που να προστατεύει τον αγοραστή από τέτοιες μη αποδεκτές ενέργειες, ζητώντας για παράδειγμα την επαρκή τεκμηρίωση των λόγων που οδήγησαν στην προώθηση προς επιλογή ενός στοιχείου (εδώ, προϊόντος) από το σύστημα συστάσεων, και οι οποίοι θα πρέπει να είναι πάντα διαθέσιμοι στην περίπτωση που ζητηθούν από τον χρήστη.

Για τον σκοπό αυτό, τα τελευταία χρόνια έχει αναπτυχθεί ένα νέο γνωστικό αντικείμενο της τεχνητής νοημοσύνης που αναφέρεται ως **Explainable AI** *(επεξηγήσιμη τεχνητή νοημοσύνη).* Στόχος του, να αυξηθεί η *επεξηγησιμότητα* των αυτοματοποιημένων αποφάσεων με τρόπο μετρήσιμο και *διάφανο.* Οι δε πάροχοι των συστημάτων συστάσεων θα πρέπει να είναι σε θέση να *εξηγήσουν επαρκώς στον χρήστη* **(the right to be explained)** τους λόγους που οδήγησαν σε μια συγκεκριμένη λίστα προτεινόμενων στοιχείων. Άρα, η αιτιολόγηση των προτεινόμενων στοιχείων θα πρέπει να είναι κατά το δυνατόν *μετρήσιμη και προφανής.*

10.2.3 Λογοκρισία (Censorship)

Οι άνθρωποι που χρησιμοποιούν τα μέσα κοινωνικής δικτύωσης ασχολούνται κατά μέσο όρο περισσότερο με περιεχόμενο που είναι στα όρια του μη επιτρεπτού όπως (i) η *παραπληροφόρηση* και οι *ψεύδεις ειδήσεις (fake news)*, (ii) οι *προκλητικές φωτογραφίες* και (iii) η *(διαδικτυακή) ρητορική μίσους (online hate speech/rhetoric)* ή ο *προσβλητικός/δυσφημιστικός λόγος (offensive/defamation speech).*

Πιο συγκεκριμένα, τα κοινωνικά δίκτυα όπως το *Facebook* ή το *Twitter* συνήθως ορίζουν κανόνες για το επιτρεπτό και μη περιεχόμενο. Στο Σχήμα 10.1 φαίνεται ότι ο αριθμός των χρηστών που ενδιαφέρονται και ασχολούνται με μια υπηρεσία (*user engagement*) αυξάνεται όσο το περιεχόμενο πλησιάζει τα όρια μεταξύ αποδεκτού και μη αποδεκτού. Συνεπώς, και οι χρήστες που συνεισφέρουν περιεχόμενο στα δίκτυα αυτά τείνουν με τη σειρά τους να επιλέγουν υλικό που να είναι πολύ κοντά στα όρια του επιτρεπτού, προκειμένου να προσελκύσουν περισσότερους *ακόλουθους (followers)*, όπως φαίνεται στο Σχήμα 10.1.



Σχήμα 10.1: Καμπύλη αύξησης του ενδιαφέροντος των χρηστών βάσει του προσφερόμενου περιεχομένου.

Οι δε πάροχοι των μέσων κοινωνικής δικτύωσης προσπαθούν να περιορίσουν και να επιβάλλουν κυρώσεις στο μη επιτρεπτό περιεχόμενο, ώστε να έχει μικρότερη διασπορά στην κοινότητα. Επίσης, χρησιμοποιούν αλγορίθμους συστάσεων για την κατάταξη των αναρτήσεων που λαμβάνουν οι χρήστες στις *ροές (news feeds)* τους, ώστε αυτοί να βλέπουν πρωτίστως τις *αναρτήσεις (posts)* που είναι ωφέλιμες, και όχι προκλητικό και παραπλανητικό περιεχόμενο.

Ο έλεγχος, τώρα, της καταλληλότητας του περιεχομένου γίνεται συνήθως από ανθρώπους, γεγονός που συνεπάγεται σημαντικά προβλήματα καθώς δεν μπορεί να αφαιρεθεί το προβληματικό περιεχόμενο προτού να γίνει η σχετική αναφορά από κάποιους χρήστες. Από την άλλη μεριά, η αποτελεσματικότητα του αυτοματοποιημένου εντοπισμού μη επιτρεπτού περιεχομένου έχει αρκετά προβλήματα: π.χ. τα τεχνικά συστήματα λογισμικού που αξιολογούν το περιεχόμενο προσπαθούν

να ακολουθήσουν μια *πολιτική*, αλλά πολλές από τις αποφάσεις τους επιδέχονται εξαιρέσεων. Λόγου χάρη, σε περιπτώσεις *ρητορικής μίσους* έχει σημασία η κατανόηση του **πότε** κάποιος καταδικάζει μια ρατσιστική προσβολή σε αντιδιαστολή με **τη χρήση της** για να επιτεθεί σε άλλους. Από την άλλη όμως -επιστρέφοντας στον αρχικό μας προβληματισμό-, ενώ οι υπολογιστές είναι συνεπείς σε (εξαιρετικά) επαναλαμβανόμενες εργασίες, όπως η *αυτοματική λογοκρισία περιεχομένου*, οι άνθρωποι δεν είναι πάντα το ίδιο συνεπείς στις κρίσεις τους σε ανάλογες περιπτώσεις. Έτσι, προκύπτουν διάφορα ηθικά ζητήματα που πρέπει να αντιμετωπιστούν. Για παράδειγμα:

- Ποια πρέπει να είναι τα όρια έκφρασης των χρηστών;
- Ποιο περιεχόμενο θα πρέπει να διανέμεται και ποιο να αποκλείεται;
- Ποιος θα πρέπει να αποφασίζει για την *πολιτική* έναντι τέτοιων περιπτώσεων (καταλληλότητας περιεχομένου) και να λαμβάνει αποφάσεις επιβολής ποινών: οι άνθρωποι ή αυτοματοποιημένες αλγοριθμικές αποφάσεις; Μήπως και οι δύο;
- Πώς μπορούμε να καταστήσουμε *διαφανείς* αυτές τις αποφάσεις λογοκρισίας; Μπορούμε να παρέχουμε επαρκείς επεξηγήσεις για το τί υποβάλλεται σε τέτοιον έλεγχο και τί όχι;

10.3 Ιδιωτικότητα και Συστήματα Συστάσεων

Ο κίνδυνος να διαρρεύσουν προσωπικά δεδομένα από τα *συγκεντροποιημένα (centralized)* συστήματα συστάσεων είναι υπαρκτός, δεδομένου ότι όλα τα δεδομένα φυλάσσονται στο ίδιο μέρος, γεγονός που τα καθιστά πιο ευάλωτα σε κακόβουλες επιθέσεις. Συνεπώς, οι υπεύθυνοι σχεδίασης ενός συστήματος συστάσεων θα πρέπει να προστατεύουν την ιδιωτικότητα των χρηστών, αλλά ταυτόχρονα να μην υποβαθμίζουν την ποιότητα της λίστας προτεινόμενων στοιχείων. Η πολιτική προστασίας της *ιδιωτικότητας* των χρηστών, λοιπόν, εδράζεται σε τρεις διαφορετικούς τομείς:

- Ο πρώτος αφορά στην *αρχιτεκτονική της τοπολογίας των εξυπηρετητών (data servers)* όπου φυλάσσονται τα δεδομένα των χρηστών, καθώς και στα πρωτόκολλα και πιστοποιητικά που εγγυώνται ότι ο πάροχος συστάσεων προστατεύει τα προσωπικά δεδομένα τους σύμφωνα με την κείμενη νομοθεσία. Τα *κατανεμημένα (distributed)* συστήματα συστάσεων διασφαλίζουν ότι τα δεδομένα δεν αποθηκεύονται σ' ένα και μοναδικό σημείο, όπως συμβαίνει στα *συγκεντροποιημένα (centralized)* συστήματα συστάσεων.
- Ο δεύτερος αφορά στις αλγοριθμικές τεχνικές που χρησιμοποιούνται για την προστασία των προσωπικών δεδομένων των χρηστών. Η βασική ιδέα που τις διέπει είναι ότι, ακόμη και αν τα δεδομένα των χρηστών διαρρεύσουν σε κάποιον κακόβουλο χρήστη, αυτός θα κατέχει μόνο τροποποιημένες ή κωδικοποιημένες πληροφορίες και δεν θα μπορεί να αποκαλύψει τα αρχικά δεδομένα.

- Ο τρίτος αναφέρεται σε νομοθεσία και στρατηγικές (π.χ. *General Data Protection Regulation - GDPR*), οι οποίες μπορούν να επιβληθούν στους παρόχους συστάσεων από κυβερνήσεις χωρών και αυτόνομες ρυθμιστικές αρχές. Για παράδειγμα, υπάρχουν κυβερνητικές πολιτικές που είτε εμποδίζουν τους παρόχους συστάσεων να διατηρούν τα δεδομένα των χρηστών τους για μεγάλα χρονικά διαστήματα είτε τους εμποδίζουν να μοιράζονται ή να ανταλλάσσουν τα δεδομένα με τρίτους. Ένα βασικό πρόβλημα που προκύπτει εδώ είναι ότι η σχετική νομοθεσία μεταξύ των κρατών δεν είναι ενιαία και συνεπώς υπάρχουν χώρες στις οποίες δεν γίνεται επαρκής έλεγχος.

Τονίζουμε στο σημείο αυτό ότι, παρόλο που οι τρεις παραπάνω τομείς φαίνεται να είναι ανεξάρτητοι μεταξύ τους, πολλά συστήματα συστάσεων μπορεί να εφαρμόζουν ανάλογη πολιτική σε περισσότερους τους ενός εκ των τριών προαναφερθέντων, για να προστατεύσουν την *ιδιωτικότητα* των χρηστών τους. Ως εκ τούτου, προτείνεται στους σχεδιαστές συστημάτων συστάσεων να αξιοποιούν συνδυαστικά στρατηγικές και για τους τρεις τομείς όταν σχεδιάζουν τους σχετικούς μηχανισμούς προστασίας της *ιδιωτικότητας*.

Παράδειγμα 10.4 Έστω, λοιπόν, ότι ένα μεγάλο ηλεκτρονικό κατάστημα (*e-shop*) παρέχει προσωποποιημένες συστάσεις σε εκατοντάδες χιλιάδες χρήστες και ότι υλοποιεί ένα *κατανεμημένο* (τα δεδομένα αποθηκεύονται σε διαφορετικά μέρη) *σύστημα διαχείρισης δεδομένων*. Ακόμη, έστω ότι συγχρόνως χρησιμοποιεί αλγοριθμικές τεχνικές για την προστασία των δεδομένων και επιτρέπει την πρόσβαση μόνο σε δεδομένα που είναι κρυπτογραφημένα. Τέλος, έστω ότι το ίδιο κατάστημα, για να αυξήσει την εμπιστοσύνη των χρηστών του, δηλώνει ρητά και υιοθετεί ένα πρωτόκολλο ασφαλείας όπου η συλλογή και χρήση των προσωπικών δεδομένων τους γίνεται με συμμόρφωση στην κείμενη νομοθεσία περί *ιδιωτικότητας*.

Στις παρακάτω ενότητες, λοιπόν, θα περιγράψουμε αναλυτικά τον καθέναν από τους τρεις τομείς της πολιτικής για την προστασία των προσωπικών δεδομένων των χρηστών.

10.3.1 Αρχιτεκτονική Συστημάτων για την Προστασία της Ιδιωτικότητας

Για την αποφυγή της ανάγκης *συγκεντροποιημένης (centralized)* συλλογής δεδομένων και εκπαίδευσης των αλγορίθμων υπάρχουν *αποκεντρωμένες (decentralized)* και *κατανεμημένες (distributed)* αρχιτεκτονικές συστημάτων συστάσεων, οι οποίες και θα αναλυθούν στις επόμενες υποενότητες.

Αποκεντρωμένα Συστήματα Συστάσεων

Οι *αποκεντρωμένες μέθοδοι* επιτρέπουν στους *απομακρυσμένους εξυπηρετητές (servers)* να συνεργάζονται για να "μαθαίνουν" εξατομικευμένα μοντέλα πρόβλεψης χωρίς την ύπαρξη ενός *κεντρικού εξυπηρετητή*. Αναλυτικότερα, η διαδικασία δημιουργίας συστάσεων υλοποιείται σ' ένα σύστημα που αποτελείται από *ομότιμους εξυπηρετητές* **(peer-to-peer system)**, εξαλείφοντας επομένως την

ανάγκη μιας κεντρικής υπηρεσίας. Τα δεδομένα που διαμοιράζονται στο δίκτυο των *ομότιμων εξυπηρετητών* μπορούν να είναι είτε οι αξιολογήσεις/προτιμήσεις των χρηστών είτε κάποιες παράμετροι του μοντέλου πρόβλεψης. Στην πρώτη περίπτωση έχουμε λιγότερη ασφάλεια για τα προσωπικά δεδομένα, ενώ στη δεύτερη, όπου οι εξυπηρετητές διαμοιράζονται μόνο παραμέτρους του μοντέλου εκπαίδευσης, επιτυγχάνεται υψηλότερη ασφάλεια.

Ο Chen και οι λοιποί συν-συγγραφείς [Chen et al., 2018a] πρότειναν έναν *αποκεντρωμένο αλγόριθμο παραγοντοποίησης πίνακα (decentralized matrix factorization)* για συστάσεις *σημείων ενδιαφέροντος (Points of Interest)*. Πιο συγκεκριμένα, δημιούργησαν ένα δίκτυο χρηστών βάσει της γεωγραφικής τους θέσης, ώστε αυτοί να εκπαιδεύουν ένα τοπικό μοντέλο στο κινητό τους μοιραζόμενοι κάποιες *παραμέτρους μάθησης (gradients-ανάδελτα)* με τους (γεωγραφικά) γειτονικούς τους χρήστες για τα σημεία ενδιαφέροντος που επισκέπτονται. Στην ίδια κατεύθυνση, ο Liao και οι συνεργάτες του [Liao et al., 2012] πρότειναν έναν *αποκεντρωμένο αλγόριθμο παραγοντοίησης πίνακα*, για να λύσουν το πρόβλημα της πρόβλεψης των αποστάσεων μεταξύ κόμβων σε δίκτυα. Τονίζουμε, εδώ, ότι και οι δύο παραπάνω εργασίες απαιτούν από τους χρήστες την ανταλλαγή ευαίσθητων πληροφοριών που αφορούν στη γεωγραφική τους θέση και γι' αυτόν τον λόγο η προστασία των δεδομένων αυτών πρέπει να διασφαλίζεται πλήρως από τον εκάστοτε αλγόριθμο συστάσεων. Τέλος, η *PDMFRec* [Duriakova et al., 2019] πρότεινε μια *αποκεντρωμένη παραγοντοποίηση του πίνακα βαθμολογίας χρηστών-στοιχείων*, η οποία επιτρέπει στον χρήστη να ορίσει ποιες πληροφορίες θέλει να μοιραστεί με τους άλλους και ποιες όχι. Ειδικότερα, η *PDMFRec* μεταφέρει τον υπολογισμό του μοντέλου σύστασης στη συσκευή του χρήστη και εξαλείφει την ανάγκη ανταλλαγής ευαίσθητων προσωπικών δεδομένων. Επιπλέον, οι χρήστες μπορούν να επιλέξουν την ποσότητα και τον τύπο των πληροφοριών που θα μοιραστούν, για να επιτύχουν ταυτόχρονα υψηλή προστασία της *ιδιωτικότητας* των προσωπικών δεδομένων τους αλλά και υψηλή *ακρίβεια* συστάσεων.

Μια από τις προκλήσεις στις *αποκεντρωμένες αρχιτεκτονικές υπολογιστών* είναι ότι πολλοί αλγόριθμοι συστάσεων απαιτούν μεγάλη υπολογιστική ισχύ, η οποία δεν προσφέρεται από τις έξυπνες φορητές συσκευές των χρηστών. Αυτός ο περιορισμός δίνει προβάδισμα σε αρχιτεκτονικές προσεγγίσεις που μοιράζουν την εργασία ανάμεσα σε μια δυνατή κεντρική συσκευή εξυπηρετητή και σε μια πιο αδύναμη συσκευή, αυτήν του χρήστη (βλ. επόμενη υποενότητα). Έτσι, οι συστάσεις παράγονται στο επίπεδο του *χρήστη (client)*, αποκαλύπτοντας λιγότερες πληροφορίες στο *κεντρικό σύστημα συστάσεων (server)*.

Κατανεμημένα Συστήματα Συστάσεων

Οι *κατανεμημένες (distributed) μέθοδοι* επιτρέπουν σε απομακρυσμένους κόμβους να μάθουν ένα τοπικό μοντέλο πρόβλεψης, ενώ ο υπολογισμός του γενικού μοντέλου γίνεται στον κεντρικό εξυπηρετητή. Τα τελευταία χρόνια πολλά συστήματα συστάσεων αξιοποιούν συνδυαστικά τις δύο από τις τρεις κατηγορίες πολιτικών προστασίας των προσωπικών δεδομένων (δηλαδή την *αρχιτεκτονική τοπολογίας εξυπηρετητών* και τις *αλγοριθμικές τεχνικές*) που περιγράφτηκαν στην Ενότητα 6.3,

υιοθετώντας την *Ομοσπονδιακή μάθηση* (**Federated Learning - FL**) για την εκπαίδευση των αλγορίθμων συστάσεων ([Chen et al., 2018a] και [Chen et al., 2018b]) (ως "ομοσπονδιακή", εδώ, νοείται η από *πολλαπλούς εξυπηρετητές συγκερασμένη εκπαίδευση ενός μοντέλου πρόβλεψης*). Σύμφωνα με το *FL*, ένα μοντέλο πρόβλεψης εκπαιδεύεται τοπικά στο "έξυπνο" κινητό τηλέφωνο του κάθε χρήστη, ενώ το γενικό μοντέλο πρόβλεψης που αφορά σ' όλους τους χρήστες ενημερώνεται μόνο με την αποστολή κάποιων *παραμέτρων μάθησης (learning parameters)* στον κεντρικό εξυπηρετητή. Μ' άλλα λόγια, το σύστημα συστάσεων φτιάχνει προσωποποιημένα μοντέλα πρόβλεψης στις συσκευές των χρηστών, ενώ ο συγκερασμός των προτιμήσεων των διαφορετικών χρηστών αποστέλλεται στον κεντρικό εξυπηρετητή με τη μορφή παραμέτρων μόνο (και όχι ως βαθμολογίες επί των στοιχείων) για την εκπαίδευση του συνολικού μοντέλου πρόβλεψης.

10.3.2 Αλγοριθμικές Τεχνικές για την Προστασία της Ιδιωτικότητας

Προκειμένου να προστατευθεί η *ιδιωτικότητα* των ευαίσθητων προσωπικών δεδομένων, υπάρχουν αλγόριθμοι οι οποίοι τροποποιούν τα δεδομένα που αφορούν στην ταυτότητα των χρηστών είτε γενικεύοντάς τα (π.χ. μετατρέποντας τη διεύθυνση κατοικίας του χρήστη στο αντίστοιχο επίπεδο της πόλης ή και του νομού διαμονής του) είτε προσθέτοντας "θόρυβο" στις βαθμολογίες των χρηστών επί των στοιχείων/προϊόντων του συστήματος συστάσεων. Η διαδικασία της αφαίρεσης ή γενίκευσης των ευαίσθητων προσωπικών δεδομένων είναι γνωστή ως *ανωνυμοποίηση της ταυτότητας χρήστη (de-identification)*. Η βασική ιδέα στην οποία εδράζεται η πρόσθεση "θορύβου" στα δεδομένα μέσω των τεχνικών *διατάραξης* ή *απόκρυψης* (*perturbation* ή *obfuscation*) είναι ότι στην περίπτωση που ένας κακόβουλος επιτιθέμενος αποκτούσε πρόσβαση στα προφίλ των χρηστών θα τα έβλεπε παραμορφωμένα και όχι ως έχουν πραγματικά.

Επίσης, υπάρχουν οι *αλγόριθμοι κρυπτογράφησης*, οι οποίοι προσφέρουν εργαλεία κωδικοποίησης των δεδομένων για να τα προστατέψουν. Τέλος, υπάρχουν οι αλγόριθμοι που βασίζονται στη *διαφοροποιημένη προστασία ιδιωτικότητας (differential privacy)*, οι οποίοι τροποποιούν κατάλληλα τα αρχικά δεδομένα παρέχοντας ταυτόχρονα μετρήσιμες εγγυήσεις προστασίας της *ιδιωτικότητας* των χρηστών. Στις παρακάτω υποενότητες θα περιγράψουμε λεπτομερώς καθεμία από τις προαναφερθείσες περιπτώσεις.

Τεχνικές Ανωνυμοποίησης

Οι αλγοριθμικές τεχνικές για την προστασία της *ιδιωτικότητας* των ευαίσθητων προσωπικών δεδομένων, από τη μία μεριά προστατεύουν τους χρήστες των συστημάτων συστάσεων μέσω της *ανωνυμοποίησης* των προσωπικών τους δεδομένων, από την άλλη μεριά, όμως, περιορίζουν την ικανότητα των συστημάτων αυτών να κάνουν ακριβέστερες και αποτελεσματικότερες συστάσεις. Για παράδειγμα, η πρόσθεση "θορύβου" στα πραγματικά δεδομένα έχει συνήθως και επιπτώσεις στην *ακρίβεια* των συστάσεων. Αυτό οφείλεται στο γεγονός ότι υπάρχει μια *αντιστάθμιση (trade off)* με-

ταξύ της παραμόρφωσης των δεδομένων για λόγους ασφάλειας και της μειωμένης προβλεπτικής αξίας που έχουν αυτά για την παραγωγή ποιοτικών συστάσεων μετά την αλλοίωσή τους. Αυτά θα τα δούμε αναλυτικότερα στις επόμενες υποενότητες, όπου θα ορίσουμε και θα περιγράψουμε τις έννοιες της *k-ανωνυμίας (k-anonymity)* ενός πίνακα δεδομένων αλλά και του *ποσοστού απώλειας πληροφορίας (information loss ratio)* αυτού μετά την αλλοίωση και παραμόρφωσή του λόγω της προσθήκης «θορύβου».

Παράδειγμα 10.5 Έστω, λοιπόν, οτι ένα νοσοκομείο διαθέτει έναν πίνακα ευαίσθητων προσωπικών δεδομένων με κάποια στοιχεία ασθενών του και θέλει να τον μοιραστεί μ' ένα άλλο νοσοκομείο στα πλαίσια ενός κοινού ερευνητικού προγράμματος για την εξαγωγή συγκεντρωτικών συμπερασμάτων για την υγεία των ασθενών τους (π.χ. ποσοστά πασχόντων ανά ασθένεια). Προκειμένου να προστατευθούν οι ευαίσθητες πληροφορίες των ασθενών, το νοσοκομείο πρέπει να *ανωνυμοποιήσει* τον πίνακα δεδομένων έτσι ώστε ο πίνακας που θα προκύψει να προστατεύει τις ευαίσθητες πληροφορίες των ασθενών.
Στον Πίνακα 10.6 υπάρχει το πεδίο *Ονοματεπώνυμο*, το οποίο μπορεί να αποκαλύψει την ταυτότητα του ασθενούς, και, συνεπώς, το νοσοκομείο θα πρέπει να το αφαιρέσει. Παρόλα αυτά, στον Πίνακα 10.6 τα χαρακτηριστικά *Φύλο, Πόλη* και *Ηλικία* αντιστοιχούν σε προσωπικές πληροφορίες των ασθενών. Τα παραπάνω τρία χαρακτηριστικά, λοιπόν, αποτελούν έναν *quasi-identifier* ή αλλιώς, *ψευδο-προσδιοριστή*, του οποίου ο ακριβής ορισμός δίνεται ακολούθως:

Ορισμός 10.1 *Ψευδο-προσδιοριστής(quasi-identifier)* σ' έναν πίνακα δεδομένων καλείται το σύνολο εκείνων των πεδίων του (π.χ. ηλικία, διεύθυνση κ.λπ.), τα οποία, εφόσον συνδυαστούν κατάλληλα με τα πεδία ενός άλλου δευτερεύοντος και δημόσια διαθέσιμου πίνακα δεδομένων (π.χ. πίνακας ψηφοφόρων ενός εκλογικού κέντρου κοντά στο νοσοκομείο του παραδείγματός μας), μπορούν να *ταυτοποιήσουν* απόλυτα ένα άτομο.

Ονοματεπώνυμο	Φύλο	Πόλη	Ηλικία	Ασθένεια
Πέτρος Πετρίδης	άνδρας	Λάρισα	26	HIV
Γιάννης Γιάννου	άνδρας	Βόλος	29	COVID
Μαρία Μαριανού	γυναίκα	Κιλκίς	36	HIV
Χριστίνα Χριστίδου	γυναίκα	Θεσσαλονίκη	37	FLU
Βασίλης Βασιλειάδης	άνδρας	Καρδίτσα	38	COVID
Γιώργος Γεωργίου	άνδρας	Τρίκαλα	36	HIV

Πίνακας 10.6: Ένας πίνακας με ευαίσθητα ιατρικά δεδομένα.

Γι' αυτόν τον λόγο το νοσοκομείο θα πρέπει να φροντίσει για την περαιτέρω *ανωνυμοποίηση* του πίνακα 10.6 πριν τον μοιραστεί με τρίτους φορείς. Προς τούτο, θα περιγράψουμε κατόπιν μια μέθοδο *ανωνυμοποίησης* που είναι γνωστή ως *γενίκευση* (*generalization*). Πρόκειται, λοιπόν, για μια πράξη που αλλάζει την τιμή ενός χαρακτηριστικού (π.χ. διεύθυνση, ηλικία κ.λπ.) *γενικεύοντάς* την. Εάν η τιμή είναι αριθμητική, τότε μπορεί να αλλάξει σ' ένα εύρος τιμών. Λόγου χάρη, η τιμή 26 για το χαρακτηριστικό *Ηλικία* στον πίνακα Πίνακα 10.6 μπορεί να αλλάξει σε εύρος τιμών 20 έως 30 όπως φαίνεται στον Πίνακα 10.7. Εάν δε η τιμή είναι *κατηγορηματική*, τότε μπορεί να αλλάξει σ' άλλη *κατηγορηματική* τιμή που να δηλώνει όμως μια έννοια ευρύτερη της αρχικής. Παραδείγματος χάρη, η "Πόλη Τρίκαλα" του Πίνακα 10.6 μπορεί να αλλάξει σε "Γεωγραφικό Διαμέρισμα Θεσσαλία" όπως φαίνεται στον Πίνακα 10.7.

Φύλο	Γεωγραφικό Διαμέρισμα	Ηλικία	Ασθένεια
άνδρας	Θεσσαλία	20 έως 30	HIV
άνδρας	Θεσσαλία	20 έως 30	COVID
γυναίκα	Μακεδονία	30 έως 40	HIV
γυναίκα	Μακεδονία	30 έως 40	FLU
άνδρας	Θεσσαλία	30 έως 40	COVID
άνδρας	Θεσσαλία	30 έως 40	HIV

Πίνακας 10.7: Ένας πίνακας με δεδομένα που έχουν *γενικευτεί.*

Ωστόσο, λόγω του ότι η προαναφερθείσα διαδικασία αλλάζει συγκεκριμένες τιμές ενός χαρακτηριστικού σε *γενικευμένες* τιμές, η ανάλυση των δεδομένων έπειτα γίνεται σε υψηλότερο επίπεδο αφαίρεσης, καθώς οι αρχικές *αναλυτικές* τιμές χάνονται. Άρα, η *γενίκευση* των δεδομένων συνήθως οδηγεί στη μείωση της προβλεπτικής αξίας των δεδομένων.

Στην περίπτωσή μας, τώρα, και εξαιτίας του ότι το νοσοκομείο *ανωνυμοποιεί* τον Πίνακα 10.6, είναι αναπόφευκτη η απώλεια κάποιων πληροφοριών στον Πίνακα 10.7. Συνεπώς, είναι επιθυμητό να ποσοτικοποιήσουμε με κάποιον τρόπο αυτήν την απώλεια πληροφοριών κατά την *ανωνυμοποίηση*, για να γνωρίζουμε την *αντιστάθμιση* μεταξύ του βαθμού προβλεπτικής αξίας των δεδομένων και του βαθμού προστασίας της *ιδιωτικότητας* των χρηστών. Στην ακόλουθη υποενότητα θα ορίσουμε και θα περιγράψουμε την *απώλεια πληροφορίας* μ' ένα παράδειγμα.

Απώλεια Πληροφορίας

Η *απώλεια πληροφορίας* ενός *γενικευμένου* πίνακα δεδομένων συνήθως μειώνει τη χρησιμότητά του. Όσο πιο μεγάλη η απώλεια, τόσο πιο μικρή η χρησιμότητα του πίνακα δεδομένων. Εάν η τιμή του χαρακτηριστικού σε μια γραμμή του πίνανα δεδομένων δεν έχει *γενικευτεί*, δεν υπάρχει καμμία παραμόρφωση. Ωστόσο, εάν *γενικευτεί*, υπάρχει παραμόρφωση του χαρακτηριστικού σ' αυτήν τη

γραμμή. Εάν, τώρα, η τιμή γενικευτεί περαιτέρω, τότε και η παραμόρφωση θα αυξηθεί περαιτέρω. Λόγου χάρη, εάν έχουμε την ιεραρχία του χαρακτηριστικού "Χρονική Περίοδος" να αναλύεται σε 4 επίπεδα (ημέρα, μήνας, έτος, δεκαετία), τότε η παραμόρφωση γίνεται ολοένα και μεγαλύτερη όσο η *γενίκευση* αυτού του χαρακτηριστικού μετατρέπεται από το επίπεδο της ημέρας στο επίπεδο της δεκαετίας.

Ορισμός 10.2 Ο βαθμός *παραμόρφωσης (distortion)* d της τιμής ενός χαρακτηριστικού του πίνακα δεδομένων ορίζεται από την απόσταση μεταξύ της αρχικής τιμής του χαρακτηριστικού έως την τρέχουσα διαμορφωμένη τιμή του, διαιρούμενης με το σύνολο των επιπέδων ανάλυσης της ιεραρχίας του χαρακτηριστικού.

Επομένως, εάν μια τιμή *γενικευτεί* κατά ένα επίπεδο ανάλυσης υψηλότερα στην ιεραρχία ανάλυσης ενός χαρακτηριστικού, τότε το κόστος της *γενίκευσης* είναι ίσο με $1/H$ - όπου H (*Height*) το ύψος της ιεραρχίας των επιπέδων στα οποία αναλύεται ένα χαρακτηριστικό. Για παράδειγμα, για την ιεραρχία της διάστασης του χαρακτηριστικού "Χρόνος" μπορούμε να έχουμε 3 επίπεδα ανάλυσης: ημέρα, εβδομάδα, μήνας.

Έστω, λοιπόν, $d_{i,j}$ η *παραμόρφωση (distortion)* της τιμής ενός *χαρακτηριστικού (attribute)* A_j που αφορά στην στήλη j και στη γραμμή i ενός πίνακα δεδομένων. Συνεπώς, η *παραμόρφωση* του πίνακα δεδομένων ισούται με το άθροισμα των *παραμορφώσεων* των επιμέρους τιμών των χαρακτηριστικών στον *γενικευμένο* πίνακα δεδομένων. Η *τρέχουσα (current) παραμόρφωση* που έχουμε επιφέρει στα δεδομένα του αρχικού πίνακα δεδομένων υπολογίζεται ως ακολούθως:

$$D_{current} = \sum_{i,j} d_{i,j} \tag{10.2}$$

Συνάγεται, λοιπόν, ότι το τελικό *ποσοστό παραμόρφωσης* $D\%$ (*Distortion percentage*) του πίνακα δεδομένων ισούται με την *τρέχουσα παραμόρφωση* του *γενικευμένου* πίνακα δεδομένων διαιρεμένη με τη *μέγιστη παραμόρφωση* που μπορεί να υπάρξει στον *πλήρως γενικευμένο* πίνακα δεδομένων ως ακολούθως:

$$\mathrm{D\%} = \frac{D_{current}}{D_{maximum}} * 100 \tag{10.3}$$

όπου ο *πλήρως γενικευμένος* πίνακας δεδομένων $D_{maximum}$ είναι αυτός στον οποίο όλες οι τιμές των χαρακτηριστικών του είναι *γενικευμένες* στο υψηλότερο επίπεδο αφαίρεσης της ιεραρχίας ανάλυσής τους.

Παράδειγμα 10.6 Έστω, λοιπόν, ότι θέλουμε να μετρήσουμε την *παραμόρφωση* για την πρώτη γραμμή του Πίνακα 10.7. Υποθέτουμε στο σημείο αυτό ότι η γεωγραφική διάσταση έχει ύψος μήκους 3 επιπέδων, δηλαδή "Περιοχή", "Πόλη" και "Γεωγραφικό Διαμέρισμα". Δεν υπάρχει *παραμόρφωση* στο πεδίο του "Φύλου" (άντρας ή γυναίκα), αφού δεν υπάρχει *γενίκευση* αυτού του χαρακτηριστικού. Επίσης, δεν υπάρχει *παραμόρφωση* στο πεδίο "Ασθένεια". Ωστόσο,

η *παραμόρφωση* της τιμής για το χαρακτηριστικό "Πόλη" (Λάρισα) είναι 1/3, επειδή αυτό το χαρακτηριστικό *γενικεύεται* από το επίπεδο της "Πόλης" στο επίπεδο του "Γεωγραφικού Διαμερίσματος" (Θεσσαλία). Τονίζουμε ότι υπάρχει μετατόπιση μόνο κατά ένα επίπεδο στην ιεραρχία της γεωγραφικής διάστασης, ενώ το ύψος της είναι μήκους 3. Ομοίως, όπως φαίνεται στον Πίνακα 10.7, σύμφωνα με την ιεραρχία του χαρακτηριστικού της "Ηλικίας", η *παραμόρφωση* της ηλικίας "20 – 30" είναι 1/2. Έτσι, η *παραμόρφωση* για την πρώτη γραμμή του πίνακα του παραδείγματός μας είναι $0 + 1/3 + 1/2 + 0 = 0.83$. Η *συνολική παραμόρφωση* που επιφέραμε στον Πίνακα 10.7 είναι ίση με 4,98, δεδομένου ότι υπάρχουν 6 γραμμές. Είναι εύκολο να υπολογίσουμε ότι η *μέγιστη παραμόρφωση* του *πλήρως γενικευμένου* πίνακα θα ήταν ίση με 6+3=9. Συνεπώς, ο *λόγος παραμόρφωσης* του Πίνακα 10.7 είναι ίσος με $(4.98/9) * 100 = 55,3\%$.

k-Ανωνυμία

Σ' αυτήν την υποενότητα παρουσιάζουμε ένα *ποσοτικοποιημένο μοντέλο προστασίας της ιδιωτικότητας*. Ειδικότερα, για να ποσοτικοποιήσουμε τη δυνατότητα αποτροπής της ταυτοποίησης των ατόμων σ' έναν πίνακα δεδομένων υιοθετούμε την έννοια της *k-ανωνυμίας (k-anonymity)*. Ο ιδιοκτήτης των δεδομένων, λοιπόν, επιδιώκει να τροποποιήσει τα αρχικά δεδομένα ενός πίνακα T και να δημιουργήσει έναν άλλο πίνακα δεδομένων T^* που θα δημοσιευτεί ή θα τον μοιραστεί με τρίτους, έτσι ώστε στον πίνακα T^*, για κάθε άτομο P να υπάρχουν τουλάχιστον $k - 1$ άλλα άτομα που δεν διακρίνονται (δηλαδή, που δεν μπορεί να τα ξεχωρίσει κανείς) από το άτομο P.

Ορισμός 10.3 k-Ανωνυμία (k-anonymity). Ένας πίνακας T ικανοποιεί την *k-ανωνυμία*, εάν κάθε *Quasi-Identifier group (Q-I group)* ικανοποιεί την *k-ανωνυμία*. Ένα *Q-I group*, με τη σειρά του, ικανοποιεί την *k-ανωνυμία* εάν το μέγεθος του *Q-I group (ομάδας)* είναι τουλάχιστον k.

Τονίζεται ότι ως *Q-I group* αναφέρονται εκείνες οι γραμμές του πίνακα δεδομένων που έχουν ακριβώς τις ίδιες τιμές στα πεδία, τα οποία χαρακτηρίζονται ως *ψευδο-προσδιοριστές*, συγκροτώντας μια ομάδα εγγραφών όπου τα άτομα είναι μη διακριτά και συνεπώς μπορούν να παραμείνουν ανώνυμα. Υπενθυμίζουμε ότι στο τρέχον παράδειγμα τα πεδία που είναι *quasi-identifiers* είναι το "Φύλο", η "Πόλη", και η "Ηλικία". Ο στόχος της *k-ανωνυμίας*, επομένως, είναι να διασφαλιστεί ότι κάθε άτομο δεν θα είναι διακριτό από τουλάχιστον $k - 1$ άλλα άτομα του πίνακα δεδομένων. Άρα, ο πίνακας 10.7 ικανοποιεί την *k=2 ανωνυμία*.

Παράδειγμα 10.7 Έστω, λοιπόν, ότι ένας κακόβουλος τρίτος με τη χρήση ενός δημοσίου πίνακα δεδομένων (π.χ. τον εκλογικό κατάλογο εγγεγραμμένων ψηφοφόρων) και τα *Quasi-Identifiers* χαρακτηριστικά ("Φύλο", "Πόλη", και "Ηλικία") του Πίνακα 10.7 αποκαλύπτει την ταυτότητα του ασθενούς -ονομάζεται Πέτρος Πετρίδης- που βρίσκεται στην πρώτη γραμμή

του αρχικού Πίνακα 10.6. Παρόλα αυτά, ο επιτιθέμενος στο σύστημα δεν μπορεί να διακρίνει όμως την πρώτη από τη δεύτερη γραμμή του *γενικευμένου* Πίνακα 10.7. Έτσι, δεν μπορεί να είναι σίγουρος εάν ο Πέτρος πάσχει από HIV επειδή η πρώτη γραμμή του Πίνακα 10.7 συνδέεται με την ασθένεια αυτή ή εάν ο Πέτρος πάσχει από COVID επειδή η δεύτερη γραμμή του Πίνακα 10.7 συνδέεται μ' αυτήν την ασθένεια.

Συνοψίζοντας, θα πρέπει να παρατηρήσουμε ότι σε κάθε διαδικασία *ανωνυμοποίησης* είναι καλό να ορίζεται η έννοια της *ελάχιστης δυνατής παραμόρφωσης των δεδομένων*, έτσι ώστε να διασφαλίζεται ότι τα δεδομένα θα είναι εξίσου χρήσιμα μετά την "αλλοίωσή" τους [LeFevre et al., 2005]. Με άλλα λόγια, η k*-ανωνυμία* δεν θα πρέπει να *γενικεύει* ή να *παραμορφώνει* τα δεδομένα περισσότερο από όσο είναι απαραίτητο. Τέλος, θα πρέπει να αναφέρουμε ότι πολλές φορές, για να προστατεύσουμε τα ευαίσθητα προσωπικά δεδομένα, επιβάλλεται να προχωρήσουμε και σε μετατόπιση των χρονικών στιγμών ή περιόδων συγκεκριμένων γεγονότων. Στο τρέχον παράδειγμα, μπορεί να χρειαστεί να *μετατοπίσουμε χρονικά (time shift)* τις ημερομηνίες διάγνωσης των ασθενειών, διατηρώντας πάντα τη σχετική διαφορά μεταξύ των διαγνώσεων, προκειμένου να μην μπορεί ο τυχών κακόβουλος να τις συνδυάσει με άλλες δημόσια διαθέσιμες πληροφορίες, ώστε να μπορέσει να ταυτοποιήσει απόλυτα έναν ασθενή.

Αλγόριθμοι Διαφοροποιημένης Ιδιωτικότητας

Η *ανωνυμοποίηση* της ταυτότητας των χρηστών μπορεί να είναι μια επιλογή προστασίας των δεδομένων τους, αλλά, όπως ειπώθηκε στην προηγούμενη ενότητα, υπάρχουν τεχνικές *απο-ανωνυμοποίησης* που με τη χρήση δύο ή περισσότερων εξωτερικών, δημόσιων πηγών πληροφοριών μπορούν να αποκαλύψουν τελικά τις πραγματικές ταυτότητες των χρηστών.

Τα τελευταία χρόνια, λοιπόν, έχουν αναπτυχθεί αλγόριθμοι που βασίζονται στη *διαφοροποιημένη προστασία ιδιωτικότητας* **(differential privacy)**, και οι οποίοι τροποποιούν κατάλληλα τα αρχικά δεδομένα, παρέχοντας ταυτόχρονα μετρήσιμες εγγυήσεις προστασίας της *ιδιωτικότητας* των χρηστών. Η *διαφοροποιημένη προστασία ιδιωτικότητας* αποσκοπεί στην παροχή μέσων για τη μεγιστοποίηση της *ακρίβειας* των στατιστικών ερωτημάτων (πχ. μέσος όρος βαθμολογίας χρηστών, μέσος χρόνος παραμονής στην υπηρεσία κ.λπ.) σε μια βάση δεδομένων, ελαχιστοποιώντας παράλληλα τις πιθανότητες ταυτοποίησης των εγγραφών της. Κι αυτό το πετυχαίνει εισάγοντας "θόρυβο" στα πραγματικά δεδομένα έτσι ώστε η προσθήκη ή η αφαίρεση ενός χρήστη στη βάση δεδομένων να μην κάνει αισθητή διαφορά σ' αυτά, αποτρέποντας έτσι την αναγνώριση των προσωπικών του πληροφοριών. Ένας αλγόριθμος εγγυάται τη *διαφοροποιημένη προστασία ιδιωτικότητας*, όταν ένας παρατηρητής που βλέπει την *έξοδό* του δεν μπορεί να καταλάβει εάν οι πληροφορίες ενός συγκεκριμένου ατόμου χρησιμοποιήθηκαν για τον υπολογισμό της.

Ορισμός 10.4 Έστω, λοιπόν, ε ένας θετικός πραγματικός αριθμός και A ένας τυχαίος αλγόριθμος που δέχεται ως *είσοδο* ένα σύνολο δεδομένων D. Ο αλγόριθμος A εγγυάται *ε-διαφοροποιημένη προστασία ιδιωτικότητας (ε-differential privacy)*, εάν για όλα τα δυνατά ζεύγη συνόλων δεδομένων D_1 και D_2, τα οποία διαφέρουν το πολύ σε μία μόνο εγγραφή, και εάν για όλα τα υποσύνολα S του εύρους τιμών *εξόδου* του A ($S \subseteq Range(A)$) -όπου S είναι οι δυνατοί διαφορετικοί συνδυασμοί υποψήφιων στοιχείων που είναι δυνατό να προταθούν στον υπό εξέταση χρήστη- ισχύει:

$$\Pr[A(D_1) \in S] \leq e^{\varepsilon} \times \Pr[A(D_2) \in S]$$

όπου οι πιθανότητες Pr εξαρτώνται από τον βαθμό τυχαιότητας του αλγορίθμου A.

Ο Ορισμός 10.4 εξασφαλίζει ότι, ανεξάρτητα από το εάν ένας χρήστης βρίσκεται στα δεδομένα *εισόδου* ή όχι ενός αλγορίθμου που εγγυάται *ε-διαφοροποιημένη προστασία ιδιωτικότητας*, η κατανομή *εξόδου* του αλγορίθμου θα είναι πάντα η ίδια, έτσι ώστε ο χρήστης να αισθάνεται ασφαλής να συνεισφέρει τα προσωπικά του δεδομένα, δεδομένου ότι αυτά δεν θα μπορούν να παραβιαστούν από κάποιον *επιτιθέμενο (attacker)* τρίτο.

Παράδειγμα 10.8 Λόγου χάρη, ας υποθέσουμε ότι έχουμε έναν πίνακα δεδομένων D_1 όπου κάθε γραμμή του είναι ένα ζευγάρι τιμών ("Όνομα Ασθενούς", "Ασθένεια HIV"), και όπου η στήλη "Ασθένεια HIV" παίρνει δυαδικές τιμές (1 ή 0), οι οποίες δηλώνουν εάν ένας ασθενής πάσχει από αυτήν ή όχι, όπως φαίνεται στον παρακάτω πίνακα:

Όνομα	Ασθένεια HIV
Πέτρος	1
Γιάννης	1
Χριστίνα	0
Βασίλης	1
Γιώργος	0

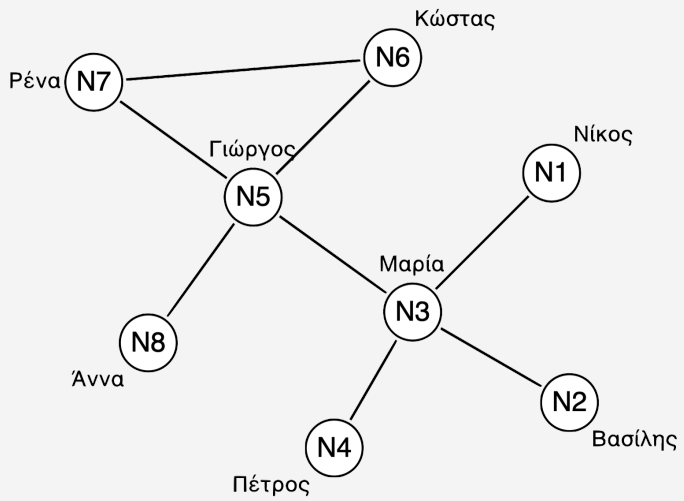
Ας υποθέσουμε, ακόμα, ότι ο *επιτιθέμενος (attacker)* τρίτος θέλει να μάθει εάν ο Βασίλης έχει την ασθένεια HIV ή όχι. Έστω, επίσης, ότι ο *επιτιθέμενος* γνωρίζει σε ποια γραμμή του πίνακα δεδομένων βρίσκεται ο Βασίλης και ότι επιτρέπεται να χρησιμοποιήσει μόνο ένα ερώτημα Q_i, το οποίο μας δίνει το άθροισμα των πρώτων i γραμμών της στήλης "Ασθένεια HIV" από τον πίνακα δεδομένων μας. Για να μάθει την ιατρική κατάσταση του Βασίλη, ο *επιτιθέμενος* εκτελεί τα ερωτήματα $Q_4(D_1)$ και $Q_3(D_1)$, και στη συνέχεια υπολογίζει τη διαφορά τους. Σ' αυτό το παράδειγμα έχουμε $Q_4(D_1) = 3$ και $Q_3(D_1) = 2$, οπότε η διαφορά τους είναι η μονάδα (1). Βάσει των δύο αυτών αποτελεσμάτων μπορούμε να συμπεράνουμε ότι το πεδίο "Ασθένεια HIV" στη γραμμή του Βασίλη είναι η μονάδα (1).

Το παραπάνω παράδειγμα αναδεικνύει τον τρόπο με τον οποίο οι *ευαίσθητες προσωπικές πληροφορίες* μπορούν να *παραβιαστούν (information breach)* ακόμη και χωρίς να αναζητούνται ρητά οι πληροφορίες ενός συγκεκριμένου ατόμου. Παρόλα αυτά, εάν ο επιτιθέμενος λάμβανε τις τιμές Q_i μέσω ενός αλγορίθμου που εγγυάται *ε-διαφοροποιημένη ανωνυμία* για ένα αρκετά μικρό ε, τότε δεν θα έπρεπε να ήταν σε θέση να ξεχωρίσει μεταξύ των αποτελεσμάτων που επιστρέφουν τα δύο προαναφερθέντα ερωτήματα.

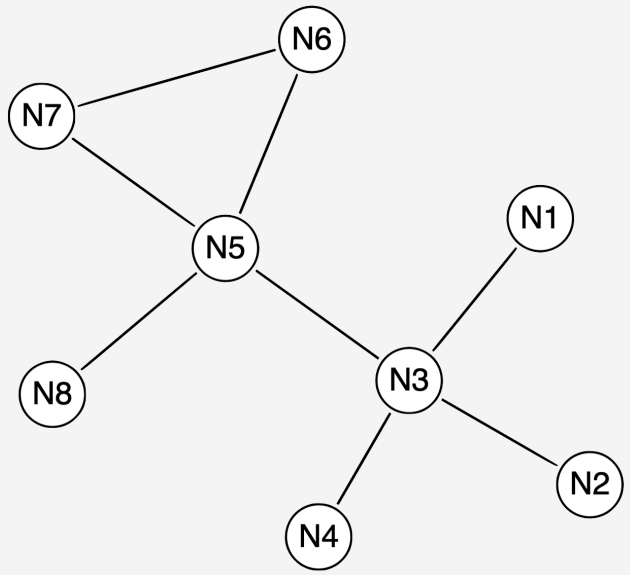
Οι McSherry και Mironov [McSherry and Mironov, 2009] ήταν οι πρώτοι που μελέτησαν την εφαρμογή της *διαφοροποιημένης προστασίας ιδιωτικότητας* στα συστήματα συστάσεων και συγκεκριμένα στο *συνεργατικό φιλτράρισμα*. Χρησιμοποίησαν τον *μηχανισμό Laplace*, ο οποίος μπορεί να προσθέσει "θόρυβο" στον δισδιάστατο πίνακα βαθμολογιών χρήστη-στοιχείου. Στη συνέχεια, πρότειναν την εισαγωγή "θορύβου" στον *πίνακα συνδιακύμανσης στοιχείων (item-item covariance matrix)* -ο οποίος μπορεί να παραχθεί από την πράξη του εσωτερικού γινομένου μεταξύ των διανυσμάτων βαθμολογίας των χρηστών- και ο οποίος εγγυάται την *ε-διαφοροποιημένη προστασία ιδιωτικότητας*. Κατόπιν, ο *πίνακας συνδιακύμανσης*, που έχει πλέον «θόρυβο», μπορεί να χρησιμοποιηθεί για να βρεθεί η *γειτονιά των παρόμοιων στοιχείων (item-based k-nearest neighborhood)* ενός υπό εξέταση στοιχείου και να προταθούν τελικά τα πιο συχνά εμφανιζόμενα στοιχεία μέσα στην προσδιορισμένη γειτονιά του. Επιπροσθέτως, οι McSherry και Mironov [McSherry and Mironov, 2009] μελέτησαν την *αντιστάθμιση (trade off)* μεταξύ της *αποτελεσματικότητας* και της *ιδιωτικότητας*, καταλήγοντας στο ότι με τη *διαφοροποιημένη προστασία ιδιωτικότητας* είναι δυνατόν να εισαχθεί κάποιος "θόρυβος" στα δεδομένα (π.χ. στους μέσους όρους των βαθμολογιών των στοιχείων ή στους μέσους όρους βαθμολογίας των χρηστών) που θα εξασφαλίζει, από τη μία, την ασφάλεια των δεδομένων, χωρίς όμως, από την άλλη, να υπάρχουν μεγάλες απώλειες στην *ακρίβεια* και την *αποτελεσματικότητα* της λίστας προτεινόμενων στοιχείων.

Ο Machanavajjhala και οι λοιποί συν-συγγραφείς [Machanavajjhala et al., 2011] μελέτησαν τις *συστάσεις φιλίας* στα κοινωνικά δίκτυα, τα οποία ταυτόχρονα εγγυώνται την *ιδιωτικότητα* των χρηστών τους. Υποστήριξαν, λοιπόν, ότι οι *τεχνικές ανωνυμοποίησης* που περιγράψαμε στην προηγούμενη ενότητα δεν είναι επαρκείς για να προστατεύσουν τους χρήστες ενός κοινωνικού δικτύου. Παραδείγματος χάρη, έστω ότι δίνεται ένας γράφος που αποτελείται από κόμβους που αντιπροσωπεύουν ανθρώπους, ενώ αυτός αποτυπώνει τις σχέσεις φιλίας τους όπως δείχνει το Σχήμα 10.2.

Παράδειγμα 10.9 Ένας απλός τρόπος για την προστασία των ονομάτων των χρηστών του Σχήματος 10.2(α) είναι η αφαίρεσή τους από τον γράφο, όπως φαίνεται στο Σχήμα 10.2(β). Ωστόσο, ακόμα και ένα τέτοιο κοινωνικό δίκτυο με "ανώνυμους" χρήστες μπορεί να δεχθεί επίθεση και παραβίαση της *ιδιωτικότητας* των χρηστών του.



(α)



(β)

Σχήμα 10.2: Παράδειγμα (α) ενός κοινωνικού δικτύου με τα ονόματα των φίλων και τα ids των κόμβων, (β) ενός κοινωνικού δικτύου φίλων μόνο με τα ids των κόμβων.

Έστω, λόγου χάρη, ότι ο *επιτιθέμενος τρίτος* γνωρίζει ότι η Μαρία έχει τέσσερεις φίλους (δηλαδή γνωρίζει ότι το *node's degree* του "κόμβου Μαρία" ισούται με 4) και ότι κανένας από τους φίλους της δεν έχει σχέση φιλίας με κάποιον άλλον φίλο της. Συνεπώς, ο *επιτιθέμενος τρίτος* μπορεί εύκολα να καταλάβει ότι ο κόμβος N3 του Σχήματος 10.2(β) αντιστοιχεί στη Μαρία. Ομοίως, εάν γνωρίζει ο *επιτιθέμενος τρίτος* ότι ο Γιώργος έχει επίσης τέσσερεις φίλους και ότι κάποιοι από αυτούς γνωρίζονται μεταξύ τους, τότε μπορεί να συμπεράνει ότι ο κόμβος

N5 στο Σχήμα 10.2(β) αντιστοιχεί στον Γιώργο. Έπειτα, παρατηρώντας ο *επιτιθέμενος τρίτος* τον ανώνυμο γράφο μπορεί να συμπεράνει ότι ο Γιώργος και η Μαρία συνδέονται με ακμή φιλίας, παραβιάζοντας έτσι τα προσωπικά τους δεδομένα.

Ο Machanavajjhala και οι λοιποί συν-συγγραφείς [Machanavajjhala et al., 2011], προκειμένου να αντιμετωπίσουν τα προβλήματα που έχουν οι *τεχνικές ανωνυμοποίησης*, όρισαν τη *διαφοροποιημένη προστασία ιδιωτικότητας σε γράφους φιλίας* ως ακολούθως:

Ορισμός 10.5 Ένας *αλγόριθμος συστάσεων* R ικανοποιεί την *ε-διαφοροποιημένη προστασία ιδιωτικότητας*, εάν για κάθε δυνατό ζεύγος γραφημάτων G και G' -τα οποία διαφέρουν ως προς μία ακμή φιλίας (δηλαδή $G = G' + \{e\}$)- και για κάθε πιθανό σύνολο διαφορετικών συστάσεων φιλίας S το οποίο μπορεί να προσφερθεί στους χρήστες ισχύει:

$$\Pr[R(G) \in S] \leq e^{\varepsilon} \times \Pr\left[R(G') \in S\right]$$

όπου οι πιθανότητες Pr εξαρτώνται από τον βαθμό τυχαιότητας του αλγορίθμου συστάσεων R, και όπου S ένα υποσύνολο με τις συστάσεις φιλίας του εύρους τιμών *εξόδου* του R ($S \subseteq Range(R)$).

Ο στόχος, λοιπόν, των συγγραφέων ήταν να σχεδιάσουν έναν αλγόριθμο R για συστάσεις φιλίας σε κοινωνικά δίκτυα με τη μέγιστη δυνατή *ακρίβεια*, υπό τον περιορισμό όμως ότι ο R εγγυάται την *ε-διαφοροποιημένη προστασία ιδιωτικότητας*. Ο Machanavajjhala και οι λοιποί συν-συγγραφείς [Machanavajjhala et al., 2011] απέδειξαν πειραματικά ότι, σε αντίθεση με την εργασία των McSherry και Vironov [McSherry and Mironov, 2009], *αποτελεσματικές συστάσεις φιλίας* επιτυγχάνονται μόνο όταν **δεν υπάρχει ισχυρή διασφάλιση** του *απορρήτου* των δεδομένων των χρηστών. Υπογράμμισαν στην έρευνά τους ότι υπάρχει μια ξεκάθαρη *αντιστάθμιση (trade off)* μεταξύ του *απορρήτου* των δεδομένων και της *αποτελεσματικότητας* των προτεινόμενων στοιχείων. Μ' άλλα λόγια, *ενισχύοντας το απόρρητο των προσωπικών δεδομένων των χρηστών, μειώνεται κατά πολύ η ακρίβεια των παρεχόμενων συστάσεων*. Τέλος, ο Chen και οι λοιποί συν-συγγραφείς [Chen et al., 2018a] πρότειναν τον αλγόριθμο *Joint Differential Privacy Matrix Factorization (Jo-DPMF)*, για να πετύχουν την *παραγοντοποίηση* του πίνακα βαθμολογιών χρηστών-στοιχείων που εγγυάται *ε-διαφοροποιημένη προστασία ιδιωτικότητας*, προσθέτοντας παράλληλα *θόρυβο Laplace* στην κλασική *αντικειμενική συνάρτηση* του *Matrix Factorization* (όπως αυτή αρχικά ορίστηκε στην Εξίσωση 5.27 του 5ου κεφαλαίου ως ακολούθως:

$$\arg\min_{p,q}\left(\sum_{(u,i)}\frac{1}{2}(r_{ui}-\hat{r}_{ui})^2+\frac{\lambda}{2}\left(\sum_{u}\|p_u\|^2+\sum_{i}\|q_i\|^2\right)+\sum_{(u,i)}\eta_{ui}\times p_u^T q_i\right)$$

όπου η_{ui} ο *θόρυβος Laplace* που αντιστοιχεί σε κάθε βαθμολογία r_{ui} ενός χρήστη u για ένα στοιχείο i, και $\eta_{xy} \sim \mathrm{Lap}(0, \Delta/\epsilon)$. Τονίζουμε ότι η παράμετρος Δ εκφράζει την *ευαισθησία* των τιμών βαθ-

μολογίας, δηλαδή τη διαφορά της μέγιστης και της ελάχιστης τιμής βαθμολογίας ($r_{\max} - r_{\min}$). Το πλεονέκτημα του *Jo-DPMF* είναι ότι εγγυάται *ε-διαφοροποιημένη προστασία ιδιωτικότητας* με ταυτόχρονη διατήρηση υψηλής προβλεπτικής αξίας στα δεδομένα, ώστε να παραμένουν χρηστικά και μετά την*παραμόρφωσή* τους.

Συνοψίζοντας, η έρευνα για τα συστήματα συστάσεων που εγγυώνται την *ε-διαφοροποιημένη προστασία ιδιωτικότητας* δείχνει ότι, ενώ σε κάποια περιβάλλοντα (π.χ. στα συστήματα συστάσεων για τα κοινωνικά δίκτυα [Machanavajjhala et al., 2011]) δεν υπάρχουν ταυτόχρονες εγγυήσεις για την *ιδιωτικότητα* και την *ακρίβεια* των συστάσεων, σ' άλλες περιπτώσεις [McSherry and Mironov, 2009] [Chen et al., 2018a] τα συστήματα συστάσεων εγγυώνται ταυτόχρονα και τα δύο. Ωστόσο, στις μελέτες που έχουν διεξαχθεί μέχρι τώρα η εισαγωγή του "θορύβου" στα δεδομένα γίνεται μόνο μία φορά. Δεν πρέπει να αγνοούμε το γεγονός ότι η συνεχής διατήρηση της *ιδιωτικότητας* των δεδομένων απαιτεί την προσθήκη "θορύβου" ανά τακτά χρονικά διαστήματα, πράγμα που είναι πιθανό να οδηγήσει σε περαιτέρω μείωση της *ακρίβειας* και της *αποτελεσματικότητας* της λίστας προτεινόμενων στοιχείων.

Αλγόριθμοι Κρυπτογράφησης

Οι αλγόριθμοι κρυπτογράφησης προσφέρουν εργαλεία κωδικοποίησης των δεδομένων για να τα προστατέψουν. Σε αντίθεση με τις τεχνικές της *ανωνυμοποίησης* και της *διαφοροποιημένης προστασίας ιδιωτικότητας των δεδομένων*, οι αλγόριθμοι κρυπτογράφησης παράγουν ίδιου επιπέδου συστάσεις με τους κοινούς αλγορίθμους που δεν εφαρμόζουν αυτά τα πρωτόκολλα κρυπτογράφησης. Τούτο συνεπάγεται ότι οι αλγόριθμοι κρυπτογράφησης παράγουν *ακριβείς* και *αποτελεσματικές* συστάσεις, διατηρώντας ταυτόχρονα τα στοιχεία του χρήστη απαραβίαστα.

Η πλειονότητα των εργασιών σ' αυτόν τον τομέα βασίζεται στην *πρόσθετη ομο-μορφική κρυπτογράφηση (additive homomorphic encryption)*, όπως είναι το *κρυπτογραφικό σύστημα δημόσιου κλειδιού* [Paillier, 1999]. Το προαναφερθέν σύστημα χρησιμοποιεί ένα *δημόσιο κλειδί* για την κρυπτογράφηση των δεδομένων και επιτρέπει μόνο στο άτομο με το κατάλληλο *ιδιωτικό κλειδί* να τα προσπελάσει. Ο όρος *ομο-μορφικός* περιγράφει τον μετασχηματισμό ενός συνόλου δεδομένων σ' ένα άλλο διατηρώντας παράλληλα τις σχέσεις μεταξύ των στοιχείων και στα δύο σύνολα. Επειδή τα δεδομένα σ' ένα *ομο-μορφικό σύστημα κρυπτογράφησης* διατηρούν την ίδια δομή, πανομοιότυπες μαθηματικές λειτουργίες, είτε αυτές εκτελούνται σε κρυπτογραφημένα δεδομένα είτε σε μη, θα οδηγήσουν σε ισοδύναμα αποτελέσματα. Ένα σύστημα κρυπτογράφησης ονομάζεται *ομο-μορφικό* σε μια πράξη "$\star$", εάν ισχύει η ακόλουθη εξίσωση:

$$E(m_1) \star E(m_2) = E(m_1 \star m_2), \text{για όλα} m_1, m_2 \in M \tag{10.4}$$

όπου E ένας αλγόριθμος κρυπτογράφησης και M το σύνολο των πιθανών μηνυμάτων. Η *πρόσθετη ομο-μορφική κρυπτογράφηση* αποτελείται από τις ακόλουθες βασικές συναρτήσεις:

- KeyGen $\rightarrow (pk, sk)$: διαδικασία δημιουργίας *κλειδιών*, όπου pk το *δημόσιο κλειδί* και sk το *ιδιωτικό (κρυφό-secret).*

- $\mathrm{Enc}(m, pk) \rightarrow c$: διαδικασία κρυπτογράφησης, όπου m το μήνυμα προς κρυπτογράφηση και c το ήδη κρυπτογραφημένο κείμενο.

- $\mathrm{Dec}(c, sk) \rightarrow m$: διαδικασία αποκρυπτογράφησης.

- $\mathrm{Add}\,(c_1, c_2) \rightarrow c_a\,(\mathrm{Enc}\,(c_1 + c_2))$: πράξη πρόσθεσης, όπου c_a το κρυπτογράφημα της πρόσθεσης των δύο απλών κειμένων.

- $\mathrm{DecAdd}\,(c_a, sk) \rightarrow m_a$: διαδικασία αποκρυπτογράφησης c_a.

Το πρόβλημα με τους αλγορίθμους κρυπτογράφησης είναι ότι έχουν πολύ υψηλό κόστος επιπρόσθετων υπολογιστικών επιβαρύνσεων, κάτι που τους καθιστά ακατάλληλους για να εφαρμοστούν σε online συστήματα συστάσεων. Κι αυτό συνεπάγεται ότι ο χρόνος παραγωγής των συστάσεων δεν είναι αρκετά μικρός, ώστε να εκπληρώνονται ικανοποιητικά οι χρόνοι απόκρισης που απαιτούνται για την ομαλή και εύρρυθμη λειτουργία των διαδικτυακών συστημάτων συστάσεων.

10.3.3 Νομικό πλαίσιο για την Προστασία της Ιδιωτικότητας

Για τα συστήματα συστάσεων έγιναν στο παρελθόν δύο βασικές προσπάθειες ρύθμισης του τοπίου, μία από τις Ηνωμένες Πολιτείες και μία από την Ευρωπαϊκή Ένωση. Η πρώτη είναι το *Νομοσχέδιο των Δικαιωμάτων Ιδιωτικότητας του Καταναλωτή* από τις Ηνωμένες Πολιτείες το 2012 (*U.S. Consumer Privacy Bill of Rights*), και η δεύτερη είναι η αναθεωρημένη *Ευρωπαϊκή Οδηγία Απορρήτου* (*European Privacy Directive*), επίσης του 2012. Και οι δύο δίνουν ιδιαίτερη έμφαση στη διαφάνεια και τον έλεγχο των εταιρειών ως προς τη διαχείριση των δεδομένων των χρηστών.

Ειδικότερα, το *Νομοσχέδιο των Δικαιωμάτων Ιδιωτικότητας του Καταναλωτή* των Ηνωμένων Πολιτειών ορίζει ότι «οι εταιρείες πρέπει να περιγράφουν ξεκάθαρα τους λόγους που χρειάζονται τα δεδομένα, όπως και το πώς θα τα χρησιμοποιήσουν, καθώς επίσης και να αιτολογούν επαρκώς τις αυτοματοποιημένες αποφάσεις που λαμβάνουν για λογαριασμό των καταναλωτών». Απαιτεί, επίσης, οι καταναλωτές να έχουν πρόσβαση στα προσωπικά τους δεδομένα (αυτά που συλλέγουν οι εταιρείες) και να τους δίνεται η δυνατότητα να τα διορθώνουν, εφόσον αυτό είναι απαραίτητο. Τέλος, προβλέπει για τη συλλογή δεδομένων αυτή να στοχεύει και να περιορίζεται σ' αυτό που αναμένεται στο πλαίσιο της υπηρεσίας για την οποία παρασχέθηκαν τα δεδομένα από τον καταναλωτή.

Η δε *Ευρωπαϊκή Οδηγία Απορρήτου* απαιτεί από τους ιστοτόπους να ρωτούν σαφώς τους χρήστες εάν αποδέχονται τα cookies. Αυτό είχε ως αποτέλεσμα η online διαφήμιση να γίνει πολύ λιγότερο αποτελεσματική στις χώρες της Ευρωπαϊκής Ένωσης σε σχέση μ' άλλες χώρες. Για να αντιμετωπίσουν το πρόβλημα αυτό, οι περισσότεροι ιστότοποι παρέχουν μόνο δύο επιλογές στους

χρήστες, έτσι ώστε να μη χάνουν χρήματα από την online διαφήμιση: (α) είτε να φύγουν από τον ιστότοπο είτε (β) να αποδεχτούν τα cookies και να συνεχίσουν. Παρόλα αυτά, η συνεχής εμφάνιση ειδοποιήσεων που ζητούν τη συναίνεση στην αποδοχή των cookies έχει δημιουργήσει σύγχυση στους χρήστες, οι οποίοι τα δέχονται πλέον ανεπιφύλακτα, χωρίς δηλαδή να γνωρίζουν επακριβώς τους όρους τους οποίους "συνυπογράφουν". Ακόμη, η ίδια οδηγία ορίζει ότι οι χρήστες πρέπει να μπορούν να έχουν πρόσβαση στα δεδομένα τους, ώστε να έχουν συνακόλουθα και τη δυνατότητα να τα μεταφέρουν σε μια άλλη πλαφόρμα ή υπηρεσία, αλλά και να μπορούν να τα διαγράψουν όποτε αυτοί το θελήσουν.

Κλείνοντας, το 2018 τέθηκε σε ισχύ ο *Γενικός Κανονισμός Προστασίας Προσωπικών Δεδομένων* της Ευρωπαϊκής Ένωσης (*General Data Protection Regulation - GDPR*). Βάσει αυτού ορίστηκε ένα αυστηρότερο πλαίσιο κανόνων εντός των οποίων πρέπει να λειτουργούν οι εταιρείες, περιορίζοντας τη συλλογή, την αποθήκευση και τη χρήση των προσωπικών δεδομένων των χρηστών. Ο *GDPR* απαιτεί τον "περιορισμό της συλλογής προσωπικών δεδομένων" στα απολύτως απαραίτητα και την επεξεργασία μόνο ενός ελάχιστου αριθμού προσωπικών πληροφοριών, και πάντα για έναν σαφώς προσδιορισμένο σκοπό. Επιπλέον, ο χρήστης έχει το δικαίωμα να ζητάει *διευκρινίσεις* (*Right to be explained*) για *αυτοματοποιημένες αλγοριθμικές αποφάσεις* της εκάστοτε εταιρείας. Γι' αυτόν τον λόγο, σε πολλές πλατφόρμες υπάρχει η επιλογή "Γιατί μου προτάθηκε αυτή η διαφήμιση;", με την οποία ο χρήστης ζητάει διευκρινίσεις για τους λόγους που οδήγησαν στη σύσταση μιας διαφήμισης κατά τη δική του *ροή ειδήσεων (news feed)* σε εφαρμογές όπως το *Instagram* και το *Facebook*. Ακόμα, ο *GDPR* απαιτεί από τις εταιρείες να *διαγράφουν τα δεδομένα των χρηστών* αφότου *έχει παρέλθει διάστημα έξι μηνών* από τη συλλογή τους (*Right to be forgotten*). Η νέα αυτή τάση της διαγραφής των προσωπικών δεδομένων επιβεβαιώνεται και από το γεγονός ότι, τα τελευταία χρόνια, οι χρήστες των μέσων κοινωνικής δικτύωσης προτιμούν να δημιουργούν *stories* τα οποία διατηρούνται για 1-2 μέρες και στη συνέχεια διαγράφονται από το να αναρτούν *posts* τα οποία παραμένουν για μεγάλο χρονικό διάστημα.

10.4 Ερωτήσεις Κεφαλαίου

1. Να συγκρίνετε τη στρατηγική της *ελάχιστης δυσαρέσκειας χρήστη* με αυτήν της *μέγιστης ικανοποίησης χρήστη* για τη σύσταση στοιχείων σε ομάδα ατόμων. Ποια από τις δύο στρατηγικές θεωρείτε πιο αποτελεσματική και γιατί;

2. Περιγράψτε τις βασικές αρχές για τον *συγκερασμό διαφορετικών λιστών κατάταξης*. Επιπλέον, τί γνωρίζετε για την *καταμέτρηση κατά Borda* και για τη *σωρευτική άθροιση λιστών κατάταξης κατά Kemeny*; Ποια από τις δύο τεχνικές θεωρείτε πιο αποτελεσματική και γιατί;

3. Πώς μπορούμε να γνωρίζουμε εάν οι αυτοματοποιημένες αλγοριθμικές αποφάσεις που λαμβάνονται από τα συστήματα συστάσεων είναι *δίκαιες* και όχι *μεροληπτικές*;

4. Στα συστήματα συστάσεων υπάρχει *αντιστάθμιση (trade off)* μεταξύ της *αποτελεσματικότητας* της λίστας προτεινόμενων στοιχείων και της *αμεροληψίας* των συστάσεων. Δώστε ένα παράδειγμα που να υποστηρίζει την παραπάνω πρόταση και εξηγήστε τους λόγους που συμβαίνει κάτι τέτοιο.

5. Ποιο είναι το κοινό μειονέκτημα μεταξύ της *γενίκευσης* στην *τεχνική της ανωνυμοποίησης* και της *τεχνικής της διαφοροποιημένης προστασίας ιδιωτικότητας*; Ποια από τις δύο τεχνικές θεωρείτε πιο αποτελεσματική και γιατί;

6. Να ορίσετε την *k-ανωνυμία* στην *τεχνική της ανωνυμοποίησης*. Δώστε ένα παράδειγμα για να δείξετε πώς ο *επιτιθέμενος τρίτος* δυσκολεύεται να ταυτοποιήσει τα στοιχεία ενός χρήστη όταν εφαρμόζεται στα δεδομένα *k-ανωνυμία*.

7. Γιατί σε κάθε διαδικασία *ανωνυμοποίησης* πρέπει να ορίζουμε και την έννοια της *ελάχιστης δυνατής παραμόρφωσης* των δεδομένων;

8. Γιατί η *τεχνική της κρυπτογραφίας* δεν μπορεί να εφαρμοστεί με τρόπο αποδοτικό στα *online συστήματα συστάσεων*;

Βιβλιογραφία

C. Chen, Z. Liu, P. Zhao, J. Zhou, and X. Li. Privacy preserving point-of-interest recommendation using decentralized matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018a.

F. Chen, M. Luo, Z. Dong, Z. Li, and X. He. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2018b.

E. Duriakova, E. Z. Tragos, B. Smyth, N. Hurley, F. J. Peña, P. Symeonidis, J. Geraci, and A. Lawlor. Pdmfrec: a decentralised matrix factorisation with tunable user-centric privacy. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 457--461, 2019.

K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49--60, 2005.

Y. Liao, W. Du, P. Geurts, and G. Leduc. Dmfsgd: A decentralized matrix factorization algorithm for network distance prediction. *IEEE/ACM Transactions on Networking*, 21(5):1511--1524, 2012.

A. Machanavajjhala, A. Korolova, and A. D. Sarma. Personalized social recommendations-accurate or private? *arXiv preprint arXiv:1105.4254*, 2011.

F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627--636, 2009.

P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *International conference on the theory and applications of cryptographic techniques*, pages 223--238. Springer, 1999.

E. Pitoura, K. Stefanidis, and G. Koutrika. Fairness in rankings and recommendations: An overview. *The VLDB Journal*, pages 1--28, 2021.