Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	Διασφάλιση ιδιωτικότητας σε βάσεις δεδομένων σημασιολογικά εμπλουτισμένων τροχιών με έλεγχο ερωτημάτων (Privacy Preservation for semantically enriched trajectory databases using query auditing techniques)
Ονοματεπώνυμο	Γιωτάκης Σπυρίδων
Πατρώνυμο	Βασίλειος
Κατεύθυνση	Δικτυοκεντρικά Πληροφοριακά Συστήματα
Αριθμός Μητρώου	ΜΠΣΠ 12011
Επιβλέπων	Πελέκης Νικόλαος, Επίκουρος Καθηγητής

Πειραιάς, Ιούλιος 2015

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Νικόλαος Πελέκης Επίκουρος Καθηγητής	Ιωάννης Θεοδωρίδης Καθηγητής	Άγγελος Πικράκης Επίκουρος Καθηγητής

Περίληψη

Στις μέρες μας, η χρήση συστημάτων εντοπισμού θέσης έχει εξαπλωθεί τόσο που είναι ενσωματωμένα σε πλήθος συσκευών. Η ευκολία στην συλλογή και αποθήκευση δεδομένων από τέτοιου είδους συσκευές αλλά και οι ενισχυμένες δυνατότητες επεξεργασίας τους, οδήγησαν στην επιθυμία για εξαγωγή χρήσιμων συμπερασμάτων από αυτά. Για παράδειγμα, συγκοινωνιολόγοι, περιβαλλοντολόγοι κ.α., αναλύοντας δεδομένα καταγραφής κινούμενων οντοτήτων, βελτιώνουν τον τρόπο εργασίας τους και τις παρεχόμενες υπηρεσίες τους. Όμως, τα χωροχρονικά «αποτυπώματα» της κάθε κινούμενης και καταγραφόμενης οντότητας αποδεικνύονται ένα επικίνδυνο εργαλείο στα χέρια ενός κακόβουλου χρήστη. Αυτού του είδους τα ζητήματα, σε μεγάλο βαθμό, έχουν αντιμετωπιστεί από την επιστημονική κοινότητα. Τα τελευταία χρόνια, περάσαμε από τις «απλές» τροχιές (*trajectories*) στις σημασιολογικά εμπλουτισμένες τροχιές (*semantic trajectories*) που παρέχουν πλουσιότερη και ουσιαστικότερη καταγραφή/απεικόνιση μιας τροχιάς με αποτέλεσμα την ακόμη μεγαλύτερη επικινδυνότητα των δεδομένων στα χέρια ενός κακόβουλου χρήστη.

Η παρούσα μελέτη αφορά βάσεις δεδομένων που περιέχουν σημασιολογικά εμπλουτισμένες τροχιές και οι οποίες παραμένουν στους κόλπους του οργανισμού που τις δημιούργησε και τις συντηρεί και όχι σε βάσεις που μετά από επεξεργασία των δεδομένων τους, δημοσιοποιήθηκαν στο ευρύ κοινό. Έτσι, βασισμένοι σε αντίστοιχες εργασίες που αφορούσαν «απλές» τροχιές και εμπνεόμενοι αρκετά από παλαιότερες εργασίες πάνω σε στατιστικές βάσεις, αρχικά οριοθετούμε το περιβάλλον εργασίας μας, εντοπίζουμε όλες τις πιθανές παραβιάσεις της ιδιωτικότητας κατά τη χρήση της βάσης και προτείνουμε ένα *συνολικό μηχανισμό ελέγχου ερωτημάτων* ώστε να επιτυγχάνεται η αντιμετώπιση των πιο πάνω κινδύνων. Τέλος, στα πλαίσια αυτού του μηχανισμού, προτείνουμε έναν αυτόνομο αλγόριθμο (*Zoom Out*), ο οποίος βοηθά στην αύξηση της φιλικότητας προς τον χρήστη και της λειτουργικότητας του μηχανισμού. Σκοπός του είναι να τροποποιήσει το αρχικό ερώτημα του χρήστη, αν δεν μπορεί εν πρώτοις να απαντηθεί για λόγους ασφαλείας, στο «πλησιέστερο» δυνατό ερώτημα που μπορεί να απαντηθεί με ασφάλεια.

Abstract

Nowadays, the use of Global Positioning Systems (GPS) has spread so much that they are embedded in many devices. The ease of collecting and storing data from such kind of devices along with the advanced processing capabilities, led to the desire of extracting useful conclusions from them. For example, traffic engineers, environmentalists etc., by analyzing moving entities recorded data, improve the way they work and services provided. Yet, the movement traces in space and time of each recorded entity prove to be a dangerous tool for a malevolent user. Such issues have been largely addressed by the scientific community. In recent years, we have moved from the trajectories on semantically enriched trajectories (semantic trajectories) which provide richer and "deeper" recording / display of a trajectory, leading however to even greater risk of data handed to a malevolent user.

This study deals with databases containing semantically-enriched trajectories which remain in-house to the organization that created and maintains them and not with databases whose an anonymous version of the original dataset has been published. Thus, based on relevant approaches regarding "plain" trajectories and quite inspired by earlier work on statistical databases, we, primarily, define our working framework. Subsequently, we identify all potential privacy breaches and recommend a comprehensive query auditing mechanism in order to address these breaches. Finally, as a distinct component of this mechanism, we propose a standalone algorithm (*Zoom Out*), which helps to improve the user friendliness and functionality of the above mechanism. Its objective is to modify the initial query posed by the user, if it cannot be answered at first for "safety" reasons, to the "nearest" query that can be possibly answered with "safety".

Κατάλογος πινάκων – εικόνων

Κατάλογος σχημάτων

Περιεχόμενα

1 Εισαγωγή

Στις μέρες μας, έχει εξαπλωθεί η χρήση συστημάτων εντοπισμού θέσης (GPS) που μπορεί να είναι ενσωματωμένα σε πλήθος συσκευών (ακόμη και σε ιδιαίτερα απλές όπως ένα κινητό τηλέφωνο) είτε για προσωπικούς, είτε για επαγγελματικούς λόγους. Σε αυτή έρχεται να προστεθεί η γενικότερη πρόοδος στην ταχύτητα επεξεργασίας δεδομένων των υπολογιστικών συστημάτων, η πρόοδος στην ευχρηστία των σημερινών βάσεων δεδομένων, η εκτίναξη της ταχύτητας σύνδεσης στο διαδίκτυο αλλά και η εμφάνιση διαφόρων υπηρεσιών Web που δίνουν τη δυνατότητα για άμεση και πολυποίκιλη χρησιμοποίηση των συσκευών GPS. Η ευκολία στην συλλογή και αποθήκευση δεδομένων από τέτοιου είδους συσκευές αλλά και οι ενισχυμένες δυνατότητες επεξεργασίας τους, οδήγησαν αναπόφευκτα στην επιθυμία για εξαγωγή χρήσιμων συμπερασμάτων από αυτά. Για παράδειγμα, συγκοινωνιολόγοι, περιβαλλοντολόγοι, τουριστικοί οργανισμοί, κοινωφελείς δημόσιες υπηρεσίες, εταιρίες κινητής τηλεφωνίας κ.α., αναλύοντας δεδομένα καταγραφής κινούμενων οντοτήτων, μπορούν να αποκρυσταλλώσουν τις εκτιμήσεις πάνω στους αντίστοιχους κάθε φορά, προβληματισμούς τους και να βελτιώσουν τις παρεχόμενες υπηρεσίες τους. Παράλληλα με αυτή την συντελούμενη πρόοδο όμως, ανέκυψαν θέματα πιθανής παραβίασης της ιδιωτικότητας διότι τα χωροχρονικά «αποτυπώματα» της κάθε κινούμενης και καταγραφόμενης οντότητας (*trajectories*) [12] μπορούν να αποδειχθούν ένα επικίνδυνο εργαλείο στα χέρια ενός κακόβουλου χρήστη. Αυτού του είδους τα ζητήματα, έχουν αντιμετωπιστεί από την επιστημονική κοινότητα σε μεγάλο βαθμό χρησιμοποιώντας ποικίλες προσεγγίσεις [1, 2, 5, 7, 10, 12, 13, 18].

Τα τελευταία χρόνια, οι απαιτήσεις από την επεξεργασία και αναπαράσταση των δεδομένων κίνησης έχουν αυξηθεί. Πιο συγκεκριμένα, καταβλήθηκε προσπάθεια να περάσουμε από τις «απλές» τροχιές (*trajectories*) στις σημασιολογικά εμπλουτισμένες τροχιές (*semantic trajectories*). Πρόκειται για τροχιές που παρέχουν πλουσιότερη και ουσιαστικότερη καταγραφή/ απεικόνιση μιας τροχιάς, αφού έχουν χωριστεί σε τμήματα (*stop & move episodes*) και το κάθε ένα από αυτά έχει εμπλουτιστεί με σχόλια (*annotations*) [14].

Όπως είναι φυσικό, ο σημασιολογικός εμπλουτισμός των τροχιών οδήγησε σε ακόμη μεγαλύτερη επικινδυνότητα των δεδομένων εφόσον αυτά χρησιμοποιηθούν και

αναλυθούν από ένα κακόβουλο χρήστη, γιατί ακόμα και αν δεν περιλαμβάνονται σε αυτά, στοιχεία που προκαλούν την άμεση αναγνώριση/σύνδεση με την καταγραφόμενη οντότητα (*ids*), μέσω των σχολίων (*annotations*) καθίστανται ιδιαίτερα προφανείς οι συνήθειες ενός καταγραφόμενου ατόμου αλλά και άλλες προσωπικές και ενδεχομένως ευαίσθητες πληροφορίες όπως ο τόπος διαμονής, εργασίας κ.α.

Η μελέτη, από την πλευρά μας, επικεντρώθηκε σε βάσεις δεδομένων οι οποίες παραμένουν στους κόλπους του οργανισμού που τις δημιούργησε και τις συντηρεί και όχι σε βάσεις που έχουν τροποποιηθεί και αφού κατέστησαν -θεωρητικά- ασφαλείς, δημοσιοποιήθηκαν τα δεδομένα τους στο ευρύ κοινό. Ουσιαστικά προσπαθήσαμε να βασιστούμε αρχικά στη μελέτη των *Gkoulalas-Divanis & Verykios* [4] και κατά κύριο λόγο στην εργασία των *Pelekis et al* [15] που αφορούσε βάσεις κινούμενων αντικειμένων *όχι σημασιολογικά εμπλουτισμένων* που παρέμεναν στον απόλυτο έλεγχο του εκάστοτε οργανισμού και υπήρχε σε κάθε περίπτωση, ένας μηχανισμός ελέγχου ερωτημάτων προκειμένου να αποφασίζεται δυναμικά κάθε φορά αν το εκάστοτε ερώτημα προς τη βάση θεωρείται ικανοποιητικά ασφαλές να απαντηθεί ή όχι.

Έτσι, με αφορμή αυτές τις εργασίες [4, 15] και εμπνεόμενοι σε διάφορα σημεία από αρκετά παλαιότερες εργασίες πάνω σε στατιστικές βάσεις [3], οριοθετούμε το πεδίο έρευνάς μας σε σημασιολογικά εμπλουτισμένες βάσεις κινούμενων αντικειμένων και προσπαθούμε να προσαρμόσουμε και να προεκτείνουμε τις τεχνικές προστασίας της ιδιωτικότητας που εφαρμόστηκαν στο [15] πάνω στα νέα δεδομένα που ενδεχομένως προκύπτουν. Ουσιαστικά γίνεται προσπάθεια για επέκταση αλγοριθμικά των διαφορετικών κατηγοριών και του τρόπου αντιμετώπισης πιθανών επιθέσεων που παρουσιάστηκαν στο [15], ώστε να καλύπτουν πλέον σημασιολογικά εμπλουτισμένες βάσεις κινούμενων αντικειμένων και για, εν γένει, διατύπωση ορισμένων τεχνικών που μπορεί, πιθανά, να φανούν χρήσιμες ως μέρος ενός διαφορετικού μηχανισμού αντιμετώπισης επιθέσεων σε παρεμφερείς βάσεις.

Ως προς το, υπό μελέτη, περιβάλλον πρέπει να επισημάνουμε τα εξής:

- Διατυπώνοντας όσο το δυνατόν αναλυτικότερα, τί θεωρούμε ως *ερώτημα* που τίθεται στη βάση, μπορούμε να αναφέρουμε τα εξής. Μια σημασιολογικά επαυξημένη βάση κινούμενων αντικειμένων προϋποθέτει ότι περιέχει, εκτός από τις καθαυτό τροχιές με τα δεδομένα που τις περιγράφουν, ένα πλήθος *επεισοδίων* στα οποία κατακερματίστηκαν οι τροχιές αυτές. Κάθε τροχιά

αποτελείται από ένα ή περισσότερα επεισόδια. Αυτά χαρακτηρίζονται το καθένα από μια *χωρική έκταση*, ένα *χρονικό διάστημα*, ένα *είδος επεισοδίου* (Stop|Move) και ένα πλήθος από *tags* τα οποία εμπλουτίζουν σημασιολογικά το επεισόδιο. Αυτές οι ιδιότητες του κάθε επεισοδίου είναι τα δεδομένα επί των οποίων, κατά την αναζήτηση (εκτέλεση ερωτήματος) εγγραφών στη βάση μπορούν να εφαρμοστούν διάφορα κριτήρια. Τα κριτήρια αυτά, εντός του ίδιου ερωτήματος, δεν επικαλύπτονται ούτε χωρικά ούτε χρονικά μεταξύ τους. Τα ερωτήματα αυτά λοιπόν, είναι μια σύνθεση/αλληλουχία από ένα ή περισσότερα ανεξάρτητα υπο-ερωτήματα που το καθένα περιλαμβάνει ως κριτήριο (*where clause*) τουλάχιστον μια από τις 4 κατηγορίες δεδομένων που προαναφέρθηκαν. Τα ερωτήματα αυτά επιστρέφουν ως τελική απάντηση στην διαδικασία που τα εκτελεί, τις κοινές μόνο τροχιές (την τομή τους) μεταξύ των τροχιών που θα επέστρεφε το κάθε υπο-ερώτημα αν εκτελούνταν ξεχωριστά, δηλαδή συνολικά τις διακριτές (*distinct*) σημασιολογικά επαυξημένες τροχιές στις οποίες ανήκουν επεισόδια που βρέθηκαν να πληρούν τα κριτήρια των υπο-ερωτημάτων.

- Ο εξουσιοδοτημένος χρήστης της βάσης *σε κάθε περίπτωση*, οποιουδήποτε είδους ερώτημα και αν θέσει στη βάση, μπορεί να λάβει ως απάντηση το πλήθος μόνο των τροχιών που πληρούν τα κριτήρια του ερωτήματος και τη γραφική απεικόνιση των αντίστοιχων τροχιών στο 2-διάστατο χώρο χωρίς καμία απολύτως επιπλέον πληροφορία για τη κάθε τροχιά.
- Η απάντηση του κάθε ερωτήματος όπως παρουσιάζεται στον τελικό χρήστη παραμένει αποθηκευμένη στη βάση για μελλοντική επαναχρησιμοποίηση. Έτσι αν ζητηθεί στο μέλλον από οποιοδήποτε χρήστη, απάντηση σε ερώτημα με τα ίδια ακριβώς κριτήρια, τότε θα προβληθεί το ίδιο ακριβώς αποτέλεσμα.
- Όλα τα *annotations* (tags) που δύνανται να υπάρξουν στα δεδομένα της βάσης είναι, εν δυνάμει, γνωστά στον οποιονδήποτε χρήστη.
- Ένα επεισόδιο μιας τροχιάς που έχει καθοριστεί αυξημένης «ευαισθησίας» (είτε από τον ίδιο τον καταγραμμένο, είτε από τρίτο), αποκρύπτεται μόνο αυτό και όχι κατ' ανάγκη ολόκληρη η τροχιά.
- Δεν επιστρέφεται ποτέ ως απάντηση ερωτήματος, ένα σετ από τροχιές εφόσον αυτές είναι λιγότερες από *k*, όπου *k* ένας θετικός ακέραιος αριθμός καθορισμένος από τους διαχειριστές της βάσης. Η κάθε τροχιά προσμετρείται

πάντα μία φορά, ασχέτως του πλήθους του επεισοδίων της, που απαντά θετικά στα κριτήρια του εκάστοτε ερωτήματος.

- Σε οποιαδήποτε περίπτωση αναφέρονται στην εργασία έννοιες όπως χώρος, χωρική έκταση, box, περιοχή αναζήτησης, MBB και συναφείς αυτών εκφράσεις, αυτές έχουν το σχήμα ορθογώνιου παραλληλογράμμου και καθορίζονται από 2 ζεύγη συντεταγμένων σύμφωνα με το προεπιλεγμένο κάθε φορά σύστημα αναφοράς.

Ο σκοπός της παρούσας εργασίας είναι

- ο αναλυτικός εντοπισμός όλων των πιθανών παραβιάσεων της ιδιωτικότητας κατά τη χρήση της βάσης και των προβλημάτων που αυτή επιφέρει κάθε φορά, ανάλογα με τα κριτήρια ασφαλείας που έχουν τεθεί.
- η πρόταση για ένα *συνολικό μηχανισμό ελέγχου ερωτημάτων* ώστε να επιτυγχάνεται η αντιμετώπιση των πιο πάνω κινδύνων με αποτέλεσμα την ασφαλέστερη δυνατή απάντηση των ερωτημάτων που τίθενται στη βάση.

Στα πλαίσια, επίσης, του μηχανισμού συνολικής αντιμετώπισης των πιθανών κινδύνων παραβίασης της ιδιωτικότητας που ενσκήπτουν κατά τη χρήση της βάσης, προτείνουμε έναν αυτόνομο αλγόριθμο (*Zoom Out*), ο οποίος βοηθά ουσιαστικά στην περαιτέρω αύξηση της φιλικότητας προς τον χρήστη κυρίως όμως, της λειτουργικότητας του μηχανισμού αυτού. Ο *Zoom Out* καλείται, πρακτικά, να προσπαθήσει να τροποποιήσει το (αρχικό) ερώτημα του χρήστη που δεν μπορεί εν πρώτοις να απαντηθεί για λόγους ασφαλείας από τη βάση, στο «πλησιέστερο» δυνατό ερώτημα που μπορεί με ασφάλεια, αυτή τη φορά, να απαντηθεί.

Το υπόλοιπο της παρούσας εργασίας διαρθρώνεται ως εξής:

Στο 2^{ο} κεφάλαιο καταγράφεται μια συνολική επισκόπηση των σχετικών εργασιών από τη βιβλιογραφία. Το 3^{ο} κεφάλαιο ξεκινά με μια σύντομη ανάλυση σχετικά με το εύρος και είδος της πρότερης γνώσης του κακόβουλου χρήστη που αποτελεί απαραίτητο συστατικό για μια «επιτυχημένη» επίθεση καθώς και μια αναλυτική διατύπωση της σκοπιμότητας που έχει η εκάστοτε επίθεση, ενώ ολοκληρώνεται με την αναλυτική παρουσίαση των τρόπων επίθεσης που διαπιστώσαμε ότι μπορούν να διενεργηθούν. Στο 4^{ο} κεφάλαιο μετά από μια πρότασή μας για το χειρισμό των «ευαίσθητων» επεισοδίων, παρουσιάζεται αναλυτικά ο αλγόριθμος *Zoom Out* ως το 1^{ο} βασικό μέρος του συνολικού μηχανισμού ελέγχου ερωτημάτων και ακολουθεί το 2^{ο} μέρος που

υλοποιεί αλγοριθμικά την διενεργούμενη παρακολούθηση του κάθε ερωτήματος και την εφαρμογή των ανάλογων κριτηρίων προκειμένου να αποφασίζεται δυναμικά κάθε φορά αν πρέπει ή όχι να δίνεται απάντηση στο ερώτημα.

2 Σχετικές εργασίες

2.1 Εισαγωγή

Στο χώρο της Πληροφορικής, οι βάσεις δεδομένων αποτελούσαν εδώ και αρκετές δεκαετίες τη λύση μέσω της οποίας μπορούσε ο οποιοσδήποτε να αποθηκεύει και να επεξεργάζεται από περιορισμένο ως και μεγάλο όγκο πληροφοριών γρήγορα και αποτελεσματικά. Το πρόβλημα της ασφάλειας της πληροφορίας αντιμετωπίστηκε από την επιστημονική κοινότητα από νωρίς και αφορούσε το σύνολο των υποδομών που απαρτίζουν γενικότερα ένα πληροφοριακό σύστημα, μέρος του οποίου ήταν (και είναι) τις περισσότερες φορές ένα σύστημα διαχείρισης βάσεων δεδομένων (σχεσιακού τύπου κατά κανόνα).
Οι βασικές ιδέες που υλοποιήθηκαν ήταν η αρχή της ακεραιότητας, της εμπιστευτικότητας και της διαθεσιμότητας των δεδομένων. Παράλληλα αναπτύχθηκαν τεχνικές που επέτρεπαν την αυθεντικοποίηση του χρήστη, την μη αποφυγή αποποίησης τυχόν ευθυνών από τη χρήση της βάσης κ.α. Όλες αυτές οι μέθοδοι όμως απέτρεπαν ως επί το πλείστον τη χρησιμοποίηση ή και την αλλοίωση στοιχείων από *μη εξουσιοδοτημένους χρήστες*.

2.2 Ιδιωτικότητα και σχεσιακές βάσεις δεδομένων

Ενώ λοιπόν τα προβλήματα της ασφάλειας είχαν αρχίσει να καθίστανται διαχειρίσιμα, εγέρθηκαν ζητήματα που αφορούσαν την προστασία της ιδιωτικότητας των δεδομένων. Πιο συγκεκριμένα, όλη η πληροφορία που συνέλλεγαν τακτικά διάφοροι φορείς και αφορούσε πολλές φορές ευαίσθητα δεδομένα φυσικών προσώπων όπως ιατρικά ή περιουσιακά στοιχεία, κρινόταν σκόπιμο συχνά είτε να αποδοθεί δημοσίως προς χρήση από οποιονδήποτε, είτε τα δεδομένα να παραμείνουν αποκλειστικά στις υποδομές του φορέα και να δοθεί πρόσβαση σε αυτά σε συγκεκριμένες οντότητες. Έτσι, ειδικά όσον αφορά την πρώτη περίπτωση, αφού στα δεδομένα θα είχε πλέον πρόσβαση

οποιοσδήποτε, αυτά έπρεπε να τροποποιηθούν με τέτοιο τρόπο ώστε να καταστεί αδύνατη η σύνδεση στοιχείων που αυτά περιέχουν με συγκεκριμένα φυσικά πρόσωπα και η βασική τεχνική ανωνυμοποίησης ήταν η αφαίρεση δεδομένων από τις εγγραφές του εκάστοτε πίνακα που μαρτυρούσαν ευθέως τη ταυτότητα ενός φυσικού προσώπου, όπως το ονοματεπώνυμό του, τον αριθμό ταυτότητας/διαβατηρίου, το αριθμό κοινωνικής ασφάλισης, τον αριθμό φορολογικού μητρώου και άλλα αντίστοιχα δεδομένα τέτοιου τύπου, όπου αυτά τυχόν υπήρχαν.

Η αφαίρεση, όμως, των δεδομένων αυτών, από μια βάση, που αντιστοιχούν ευθέως και αναμφισβήτητα με κάποιο φυσικό πρόσωπο, δυστυχώς δεν επαρκεί για να θεωρηθεί η βάση αυτή ασφαλής. Με αυτόν τον όρο, εννοείται ότι κανείς και με κανένα τρόπο δεν μπορεί να συνδέσει ούτε μία εγγραφή της βάσης ευθέως με ένα φυσικό πρόσωπο. Η L. Sweeney [17] όμως απέδειξε ότι αυτό δεν είναι επαρκές. Χρησιμοποίησε τον όρο Quasi- Identifiers για να χαρακτηρίσει το συνδυασμό ορισμένων πεδίων μιας βάσης, σε αντιδιαστολή με το ή τα πεδία τα οποία περιέχουν «ευαίσθητη» πληροφορία. Αν κάποιος γνωρίζει τις τιμές των πεδίων που απαρτίζουν τους Quasi- Identifiers για μια εγγραφή, τότε μπορεί να μάθει πολύ εύκολα την «ευαίσθητη» πληροφορία της εγγραφής αυτής. Τις τιμές των Quasi- Identifiers που χρειάζεται ένα κακόβουλος χρήστης της βάσης, είτε τις κατέχει από πρότερη προσωπική γνώση, είτε τις αποκτά αντλώντας τες από άλλη «πηγή» πληροφοριών, όπως για παράδειγμα από στοιχεία δημοσιευμένα στο Διαδίκτυο.

Έτσι, αφού η L. Sweeney απέδειξε την υπάρχουσα αδυναμία στη διατήρηση της ιδιωτικότητας των δεδομένων μιας βάσης, προχώρησε στην πρόταση αντιμετώπισης του προβλήματος με τη χρήση του *k*-anonymity. Πρόκειται για ένα κανόνα που απαγορεύει να υπάρχουν λιγότερες από *k* εγγραφές στη βάση για κάθε πιθανό συνδυασμό των τιμών των quasi-identifiers. Έτσι, όση γνώση και να έχει αποκομίσει ο κακόβουλος χρήστης, γνωρίζοντας ότι μια από τις *k* εγγραφές αφορά το φυσικό πρόσωπο που τον «ενδιαφέρει», δεν μπορεί να συμπεράνει με βεβαιότητα μεγαλύτερη από 1/*k*, ποια είναι αυτή και κατ΄ επέκταση ποια είναι η «ευαίσθητη» πληροφορία που τη συνοδεύει.

Για να είναι, λοιπόν, κοινές οι τιμές των QI, τουλάχιστον ανά *k* εγγραφές , χρησιμοποιούνται διάφορες τεχνικές ανωνυμοποίησης, με βασικότερη τη γενίκευση τιμών [17].

Έτσι στον πίνακα 2.1(α) παρατηρούμε ένα τμήμα από ένα σύνολο αρχικών εγγραφών που αφορούν στατιστικά στοιχεία ασθενών ενός νοσοκομείου για ένα συγκεκριμένο διάστημα. Οι QI είναι η ηλικία, το φύλο και το μορφωτικό επίπεδο, ενώ η ευαίσθητη πληροφορία είναι η διαγνωσθείσα ασθένεια. Στον πίνακα 2.1(β) παρατηρούμε ένα παράδειγμα γενίκευσης των τιμών των QI, έτσι ώστε να επιτευχθεί ένα 3-anonymity.

	Ηλικία	Φύλο	Μόρφωση	Ασθένεια
1	25	Γυναίκα	Πτυχίο	Μηνιγγίτιδα
2	28	Γυναίκα	Msc	Μηνιγγίτιδα
3	31	Γυναίκα	Πτυχίο	Μηνιγγίτιδα
4	35	Άνδρας	Λύκειο	Καρκίνος Πνεύμονα
5	38	Άνδρας	Γυμνάσιο	Αγγειακό Εγκεφαλικό Επ.
6	46	Άνδρας	Λύκειο	Έμφραγμα Μυοκαρδίου
7	53	Γυναίκα	Λύκειο	Καρκίνος Μαστού
8	54	Άνδρας	Δημοτικό	Καρκίνος Προστάτη
9	58	Γυναίκα	Λύκειο	Καρκίνος Μαστού
10	67	Άνδρας	Γυμνάσιο	Καρκίνος Προστάτη

Πίνακας 2.1(α) Πίνακας με τις αρχικές τιμές

	Ηλικία	Φύλο	Μόρφωση	Ασθένεια
1	25-34	Γυναίκα	≥ Πτυχίο	Μηνιγγίτιδα
2	25-34	Γυναίκα	≥ Πτυχίο	Μηνιγγίτιδα
3	25-34	Γυναίκα	≥ Πτυχίο	Μηνιγγίτιδα
4	35-50	Άνδρας	≤ Λύκειο	Καρκίνος Πνεύμονα
5	35-50	Άνδρας	≤ Λύκειο	Αγγειακό Εγκεφαλικό Επ.
6	35-50	Άνδρας	≤ Λύκειο	Έμφραγμα Μυοκαρδίου
7	50-70	Άνθρωπος	≤ Λύκειο	Καρκίνος Μαστού
8	50-70	Άνθρωπος	≤ Λύκειο	Καρκίνος Προστάτη
9	50-70	Άνθρωπος	≤ Λύκειο	Καρκίνος Μαστού
10	50-70	Άνθρωπος	≤ Λύκειο	Καρκίνος Προστάτη

Πίνακας 2.1 (β) Πίνακας με τις γενικευμένες τιμές ώστε να ισχύει 3-anonymity

Η καινοτομία αυτή, όμως, δεν ήταν και πάλι επαρκής. Διαπιστώθηκαν δύο βασικά ευάλωτα σημεία αυτής της μεθόδου. Το πρώτο αφορά την περίπτωση στην οποία και οι *k* εγγραφές όπως διαμορφώθηκαν περιέχουν την ίδια τιμή στο ή στα πεδία της βάσης που χρήζουν «προστασίας» (*homogeneity attack*) [9]. Αυτό παρατηρείται στις 3 πρώτες εγγραφές του πίνακα 2.1(β). Έτσι, ενώ από τη μια δεν μπορούμε να γνωρίζουμε σε ποια εγγραφή ακριβώς αντιστοιχεί η γυναίκα 31 ετών που μας ενδιαφέρει, από την άλλη η «ευαίσθητη» πληροφορία έχει αποκαλυφθεί αφού όλες έχουν την ίδια ασθένεια.

Το δεύτερο αφορά την περίπτωση στην οποία διαφέρουν οι τιμές του «ευαίσθητου» πεδίου, αλλά με τέτοιο τρόπο ώστε αν ο κακόβουλος χρήστης έχει κάποιου είδους επιπλέον γνώση (*background knowledge attack*), μπορεί να εξάγει σίγουρη γνώση παρόλη την αβεβαιότητα που προσφέρει καταρχάς το *k*-anonymity. Η επιπλέον γνώση μπορεί να είναι, ενδεικτικά, εγκυκλοπαιδικού τύπου όπως ότι καμία γυναίκα δεν μπορεί να πάσχει από καρκίνο του προστάτη αν μιλάμε, ενδεικτικά, για ιατρικά δεδομένα. Έτσι, αν ορισμένες εγγραφές με κοινούς QI, όπως οι 4 τελευταίες στο παράδειγμα της Εικ. 1(β), στο «ευαίσθητο» πεδίο (*sensitive attribute*) έχουν τιμή είτε «Καρκίνος Μαστού», είτε «Καρκίνος Προστάτη», η γυναίκα, για την οποία γνωρίζουμε από τους QI ότι συμμετέχει στο σετ αυτών των εγγραφών, πάσχει σίγουρα από την πρώτη ασθένεια.

Μια λύση σε αυτά τα προβλήματα προτάθηκε από τους *A. Machanavajjhala et al* [9] και ονομάστηκε *l*-diversity. Ήταν ένας κανόνας που εισήγαγαν οι συγγραφείς και καθόριζε ότι πρέπει να υπάρχει μια «ποικιλία» διακριτών τιμών ως προς το «ευαίσθητο» πεδίο ή πεδία που επιθυμούμε να προστατευτούν στο πλαίσιο της κάθε ομάδας εγγραφών με κοινή τιμή στους QI. Το πλήθος των απαιτούμενων διακριτών τιμών καθορίζεται από την παράμετρο *l*. Βέβαια, για να επιτευχτεί ένα ανάλογο αποτέλεσμα, ο εκάστοτε πίνακας πρέπει να υποστεί περαιτέρω διαδικασία γενίκευσης των τιμών των QI ώστε *οι νέες ομάδες εγγραφών με κοινές τιμές στους QI (Equivalence Classes)* να ικανοποιούν το *l*-diversity. Είναι αυτονόητο ότι, όσο αυξάνει το *l*, χάνεται πολύτιμη, ίσως, πληροφορία από τη βάση λόγω των περαιτέρω γενικεύσεων των τιμών των QI, αλλά πληθαίνουν οι διακριτές τιμές που περιέχει το κάθε EC με αποτέλεσμα να δυσχεραίνεται η εξαγωγή ασφαλών συμπερασμάτων από τον κακόβουλο χρήστη.

Η προστασία, όμως, των «ευαίσθητων» πεδίων παραμένει σε μη ικανοποιητικά επίπεδα ορισμένες φορές. Σε δυο χαρακτηριστικές περιπτώσεις, μάλιστα, αυτό μπορεί να οδηγήσει σε αποκάλυψη των «ευαίσθητων» δεδομένων.

Η πρώτη προβληματική περίπτωση συμβαίνει όταν υπάρχει μια ασυμμετρία στη παρουσία των *l* διακριτών τιμών εντός ενός ή περισσοτέρων Equivalence Classes σε σχέση είτε με την κατανομή που ακολουθούν οι τιμές αυτού του πεδίου στον αντίστοιχο πίνακα, είτε με πρότερη εγκυκλοπαιδική γνώση του χρήστη (*skewness attack*). Για παράδειγμα, αν το πεδίο περιέχει το «ΝΑΙ» ή το «ΟΧΙ» ψηφοφόρων για ένα δημοψήφισμα με κάθε EC να περιέχει κατά κανόνα 10 με 25 εγγραφές και στην EC που

ενδιαφέρει τον κακόβουλο χρήστη βρεθούν 20 «ΝΑΙ» και 1 «ΟΧΙ», τότε αυτός δικαιούται να πιθανολογήσει έντονα ότι ο άνθρωπος που προσπαθεί να αποκαλύψει κατά πάσα πιθανότητα ψήφισε «ΝΑΙ». Ταυτόχρονα, εδώ συμβαίνει το «τεχνικά» παράδοξο αλλά ανθρωπίνως λογικό ότι αν η συντριπτική πλειοψηφία του συνόλου είχε ψηφίσει «ΝΑΙ», τότε μάλλον δεν παίζει κάποιο ιδιαίτερο ρόλο, ενώ αν είχε ψηφίσει «ΟΧΙ» τότε αποκτά ιδιαίτερη σημασία η προστασία των λίγων και δακτυλοδεικτούμενων που ψήφισαν «ΝΑΙ».

Η δεύτερη περίπτωση, αφορά τον κίνδυνο αποκάλυψης «ευαίσθητης» πληροφορίας όταν, ενώ υπάρχουν *l* διακριτές τιμές σε ένα EC, μπορεί αυτές να απέχουν ελάχιστα ως προς το σημασιολογικό τους περιεχόμενο (*similarity attack*).

	Ηλικία	Φύλο	ΤΚ	Ακίνητη Περιουσία
1	25	Γυναίκα	10431	0
2	28	Γυναίκα	10400	35.000
3	31	Γυναίκα	10535	5.000
4	35	Άνδρας	11145	120.000
5	38	Άνδρας	11380	10.000
6	46	Άνδρας	11154	180.000
7	53	Γυναίκα	15771	1.000.000
8	54	Άνδρας	15772	1.800.000
9	58	Γυναίκα	15771	2.500.000
10	67	Άνδρας	15773	4.500.000

Πίνακας 2.2(α) Πίνακας με τις αρχικές τιμές

	Ηλικία	Φύλο	ΤΚ	Ακίνητη Περιουσία
1	25-34	Γυναίκα	10***	0
2	25-34	Γυναίκα	10***	35.000
3	25-34	Γυναίκα	10***	5.000
4	35-50	Άνδρας	11***	120.000
5	35-50	Άνδρας	11***	10.000
6	35-50	Άνδρας	11***	180.000
7	50-70	Άνθρωπος	1577*	1.000.000
8	50-70	Άνθρωπος	1577*	1.800.000
9	50-70	Άνθρωπος	1577*	2.500.000
10	50-70	Άνθρωπος	1577*	4.500.000

Πίνακας 2.2(β) Πίνακας με τις γενικευμένες τιμές (ισχύει 3-anonymity & 3-divercity)

Ας παρατηρήσουμε τον πίνακα 2.2 ο οποίος περιλαμβάνει στον πίνακα 2.2(α) τις αρχικές εγγραφές και στον πίνακα 2.2(β) τις εγγραφές γενικευμένες ώστε να καλύπτουν

ένα 3-anonymity και ένα 3-divercity. Οι QI είναι η ηλικία, το φύλο και ο ΤΚ, ενώ η ακίνητη περιουσία είναι το «ευαίσθητο» πεδίο.

Στο παράδειγμα αυτό, πόσο προστατεύουμε όταν στην 3^{η}, κατά σειρά, ΕC του πίνακα, η ακίνητη περιουσία όλων φυσικών προσώπων που αυτή περιλαμβάνει, κυμαίνεται μεταξύ 1.000.000 € και 4.500.000 €;

Η λύση που προτάθηκε για τις παραπάνω περιπτώσεις από τους *Li N et al* [8] ήταν το *t-closeness*. Καθόρισαν ένα κανόνα που εφαρμόζοντάς τον, θα πρέπει να υπάρχει παρόμοια κατανομή της συχνότητας εμφάνισης των διακριτών τιμών του πεδίου που επιθυμούμε την προστασία του, μέσα σε κάθε κλάση (ΕC), σε σχέση με την κατανομή που παρουσιάζουν οι ίδιες αυτές τιμές σε όλο τον πίνακα. Η μετρήσιμη «επιτρεπτή» απόκλιση μεταξύ των δύο αυτών μεγεθών πρέπει να φθάνει μέχρι το κατώφλι *t*.

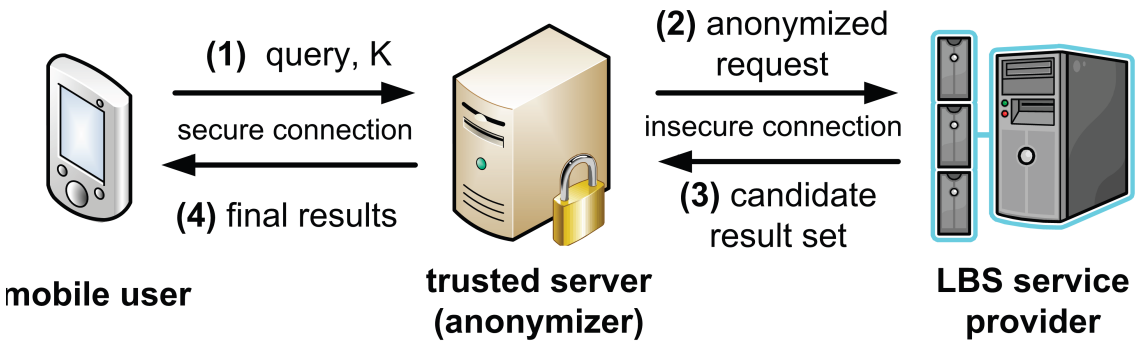
2.3 Η ιδιωτικότητα στις υπηρεσίες γεωγραφικής θέσης (LBS)

Όλες αυτές οι τεχνικές αποτέλεσαν το βασικό ανάχωμα για την αντιμετώπιση εξουσιοδοτημένων αλλά κακόβουλων χρηστών μιας σχεσιακής βάσης δεδομένων. Ανάγκη για προστασία της ιδιωτικότητας, προκύπτει, επίσης, στα πλαίσια της χρήσης *LBS's (Location Based Services)* δηλαδή της χρήσης εφαρμογών υπηρεσιών γεωγραφικής θέσης. Με αυτό τον τρόπο ένα φυσικό πρόσωπο, το οποίο κινείται στο χώρο στέλνει μέσω μιας κινητής συσκευής τα στοιχεία του δηλαδή τη θέση και την ταυτότητα του προκειμένου να του παρασχεθεί μια υπηρεσία θέσης, όπως να του δοθούν πληροφορίες για κοντινότερο φαρμακείο ή πρατήριο υγρών καυσίμων κ.α. Είναι αυτονόητο ότι και σε αυτές τις περιπτώσεις οι φορείς διαχείρισης τέτοιων δεδομένων οφείλουν να σεβαστούν την ιδιωτικότητα των πελατών-χρηστών των εφαρμογών τους και να λειτουργήσουν με τέτοιο τρόπο ώστε να τη διαφυλάξουν. Αυτό σημαίνει ότι θα πρέπει κατά τη χρήση τέτοιων υπηρεσιών από ένα χρήστη, να μην είναι εύκολο να αποκαλυφθεί η ταυτότητα του ακόμα και αν αποστέλλονται στο φορέα παροχής LBS's μόνο οι συντεταγμένες του. Έτσι παράλληλα με την ανάπτυξη και επέκταση τέτοιων υπηρεσιών, εξελίχθηκαν ταυτόχρονα και διάφορες τεχνικές προστασίας της ιδιωτικότητας της θέσης (*location privacy*) με ανώτερο σκοπό να υπάρχει πάντα η «δυνατότητα σε ένα άτομο να καθορίζει μόνος του, πώς και σε ποιο

βαθμό αξιόπιστες πληροφορίες σχετικές με την τρέχουσα ή προηγούμενη τοποθεσία στην οποία βρέθηκε, κοινοποιούνται σε τρίτους».

Η βασική στρατηγική που ακολούθησε η επιστημονική κοινότητα είναι η προστασία της ιδιωτικότητας με τη χρήση και την τροποποιημένη υλοποίηση του *k*-anonymity για δεδομένα χρηστών που ζητούν LBS. Πιο συγκεκριμένα το ζητούμενο είναι, ο χρήστης που ζητά μια υπηρεσία να εστιάζεται σε μια χωρική περιοχή τέτοια που να περιέχει άλλους *k*-1 τουλάχιστον χρήστες της υπηρεσίας ώστε να υπάρχει πιθανότητα να εντοπιστεί κατά το μέγιστο 1/*k*.

Επειδή στην περίπτωση των LBS οι φορείς που παρέχουν αυτές τις υπηρεσίες, χειρίζονται δυναμικές βάσεις δεδομένων, είναι εύλογο να εκτιμώνται πιο συχνές-πιθανές-εύκολες δυνατότητες επιθέσεων στην ιδιωτικότητα των χρηστών-πελατών απ' ότι σε πιο στατικά μοντέλα ΒΔ, με αποτέλεσμα να θεωρούνται, κατά τεκμήριο, μη αξιόπιστοι. Με βάση αυτό, στις πλείστες των περιπτώσεων, παρεμβάλλεται μεταξύ του χρήστη και του παρόχου της υπηρεσίας, ένα ενδιάμεσο επίπεδο επικοινωνίας ως μια Τρίτη ξεχωριστή Οντότητα (*Trusted Third Party*).



Εικ. 2.1 Το συγκεντρωτικό μοντέλο για την ιδιωτικότητα στα LBS – Gkoulalas-Divanis A. et al [6]

Σκοπό έχει να λαμβάνει τα αιτήματα των χρηστών, να αφαιρεί τυχόν πληροφορία συνοδευτική της γεωγραφικής θέσης που μπορεί να προδώσει τον χρήστη, να προκαλέσει μια ανωνυμοποίηση της θέσης ώστε να μην διακρίνεται μεταξύ άλλων *k*-1 τουλάχιστον, θέσεων πραγματικών χρηστών και να στείλει αυτά τα δεδομένα στον πάροχο της υπηρεσίας, ζητώντας έτσι εξυπηρέτηση με βάση μια ολόκληρη περιοχή και όχι για μια και μοναδική θέση. Ο πάροχος απαντά για όλες τις θέσεις που περιλαμβάνει η περιοχή και ο ενδιάμεσος φροντίζει να λάβει ο πραγματικός αιτών της υπηρεσίας τη σωστή απάντηση.

Οι τεχνικές ανωνυμοποίησης που έχουν χρησιμοποιηθεί κατά καιρούς παρουσιάζουν ένα γενικότερο ενδιαφέρον στον τρόπο που αντιμετωπίζουν το πρόβλημα και αποτελούν πάντα μια καλή αφετηρία για αλγοριθμικές λύσεις σε παρεμφερή ζητήματα. Εδώ σημειώνουμε ότι υπάρχουν δύο γενικές κατηγορίες λύσεων. Αυτές που αφορούν τις περιπτώσεις που υπάρχει μια και μοναδική επικοινωνία του χρήστη με τον πάροχο προκειμένου να του παρασχεθεί μια πληροφορία (*Snapshot LBS*) και τις περιπτώσεις που απαιτείται πολλαπλή-συνεχής αποστολή στοιχείων θέσης του χρήστη προκειμένου να λάβει σωστά την παρεχόμενη πληροφορία (*Continuous LBS*).

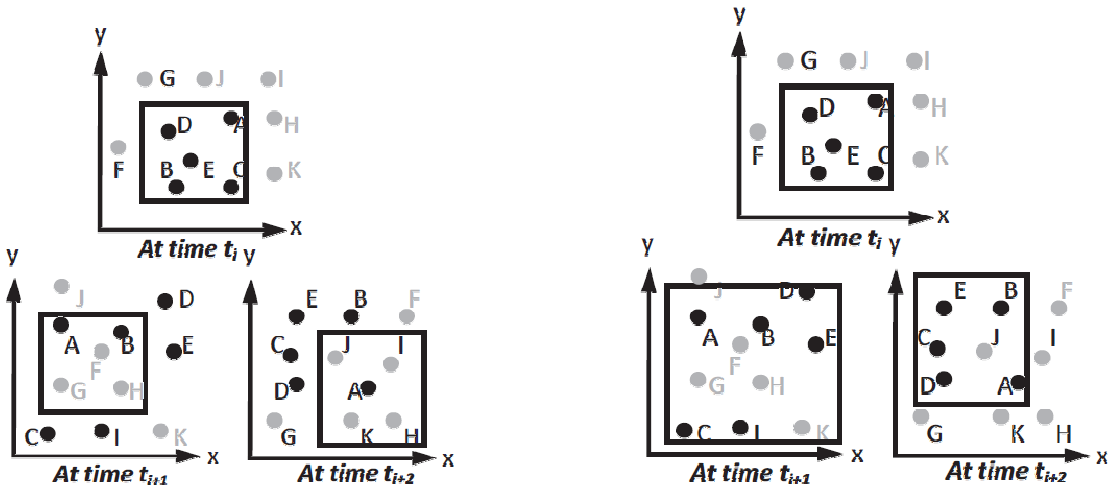
Η βασικότερη μέθοδος ανωνυμοποίησης μιας θέσης, καλείται *απόκρυψη (cloaking).* Γίνεται δηλαδή προσπάθεια να αποκρύπτεται η συγκεκριμένη θέση του ατόμου που ζητά μια υπηρεσία μέσω της αποστολής στην ενδιάμεση υποδομή που όπως προαναφέρθηκε, έχει «στηθεί» ανάμεσα στον χρήστη και τον πάροχο LBS, και είναι υπεύθυνη, εκτός από τη θέση του χρήστη, να συλλέξει και άλλες *k*-1 κοντινότερες θέσεις χρηστών και τότε μόνο να την αναμεταδώσει στον πάροχο LBS. Ουσιαστικά αντί της αποστολής μιας συγκεκριμένης θέσης, αποστέλλεται μια περιοχή που περιέχει *k* θέσεις [6]. Για όλες αυτές τις θέσεις, ο πάροχος LBS's παρέχει τις αντίστοιχες απαντήσεις στην ενδιάμεση υποδομή (*trusted server*) και η οποία με τη σειρά της αναλαμβάνει να εντοπίσει τη σωστή απάντηση στο πραγματικό ερώτημα του χρήστη και να του κοινοποιήσει το περιεχόμενό της.

Για να δημιουργηθεί η περιοχή που θα περιέχει κάθε φορά τουλάχιστον *k* θέσεις, υπάρχουν δύο τρόποι εργασίας. Είτε από τη συγκεκριμένη θέση του αρχικού χρήστη, θα γίνεται μια διεύρυνση της περιοχής γύρω από αυτή τη θέση, δηλαδή να οριοθετείται μια κυκλική περιοχή γύρω από την αρχική θέση, είτε θα έχει κατατμηθεί εξ' αρχής η συνολική περιοχή που παρέχεται η υπηρεσία σε κομμάτια, τα οποία είναι ευκολότερο να τα φανταστούμε ως τμήματα ενός μεγάλου πλέγματος και ξεκινώντας από το κομμάτι που περιλαμβάνει την αρχική θέση του χρήστη θα επιλέγονται ένα-ένα ακέραια γειτονικότερα τμήματα μέχρι να συμπληρωθεί ο αριθμός από τους χρήστες που περιλαμβάνονται σε αυτά.

Πέρα από τη μέθοδο της απόκρυψης, υπάρχουν και άλλες τεχνικές που έχουν εφαρμοστεί και οι οποίες αοριστοποιούν τη θέση του χρήστη, εντάσσοντας την μέσα σε μια ευρύτερη περιοχή, η οποία συνήθως είναι τόσο διευρυμένη όσο μεγαλύτερες απαιτήσεις για ασφάλεια υπάρχουν. Ο πάροχος LBS, πλέον απαντά έχοντας σαν είσοδο

όχι μια (ή έστω *k*) θέση χρήστη, αλλά μια περιοχή που κάπου εντός των ορίων της βρίσκεται ο χρήστης εκείνη τη χρονική στιγμή.

Είναι βέβαια σαφές ότι όσο πιο μεγάλες κατασκευάζονται αυτές οι περιοχές, τόσο περισσότερο ανακριβής μπορεί να καταστεί η παρεχόμενη υπηρεσία. Στην περίπτωση, τώρα, που ο χρήστης αιτείται συνεχώς μια υπηρεσία καθώς κινείται στο χώρο (*continuous LBS*), ως προς την τεχνική απόκρυψης (*cloaking*), η ουσιαστική διαφοροποίηση-διεύρυνση των προηγούμενων τεχνικών είναι ότι πρέπει για να μην υπάρξει αποκάλυψη της ταυτότητας του χρήστη, οι *k*-1 άλλοι χρήστες να λειτουργούν καθ' όλη τη διάρκεια της παροχής αυτής της υπηρεσίας ως μια αδιαίρετη ομάδα. Δηλαδή απαιτείται κάθε φορά που ζητά παροχή υπηρεσίας ο συγκεκριμένος χρήστης, να δημιουργείται ένα MBB *(Minimum Bounding Box)* που θα περιλαμβάνει το σύνολο των *k* θέσεων άσχετα αν πλέον στο χρονικό σημείο αυτό οι *k*-1 άλλες θέσεις πλην της θέσης του χρήστη δεν είναι οι κοντινότερες χωρικά σε αυτήν [6].



(a) The attack scenario. (b) Blocking the attack.

Εικ. 2.2 Η επίθεση μέσω της παρακολούθησης του ερωτήματος και η εξάλειψη της Gkoulalas-Divanis A. et al [6]

Τέλος, διάφορες επιπλέον εναλλακτικές έχουν διατυπωθεί βλέποντας τη γενικότερη έκταση χωρική και χρονική, στην οποία έλαβαν χώρα αιτήματα χρηστών για παροχή συγκεκριμένων, κάθε φορά, LBS. Έτσι έχουν υλοποιηθεί αλγόριθμοι [6] οι οποίοι καταγράφουν την ιστορικότητα των αιτημάτων των χρηστών, έτσι ώστε κάθε φορά που ζητείται μια υπηρεσία από ένα χρήστη να ελέγχονται όλα τα παρελθόντα αιτήματα από εκείνη την περιοχή και όλοι οι διαφορετικοί χρήστες να απαρτίζουν πλέον την

υποτιθέμενη ομάδα χρηστών που αιτούνται μιας υπηρεσίας, ώστε να είναι πλέον ασαφές ποιος το ζητάει εκείνη τη χρονική στιγμή.

Κάνοντας ένα βήμα παραπέρα, οι *Gkoulalas-Divanis & Verykios* [5] πρότειναν την επεξεργασία των αιτημάτων που έχουν διενεργηθεί ιστορικά, προκειμένου να εξαχθούν συγκεκριμένα πρότυπα κίνησης (*mobility patterns*). Έτσι καθορίζεται ως ασφαλής ή μη η τρέχουσα διαδρομή ενός χρήστη, άρα κατά συνέπεια και το κάθε σημείο το οποίο εκπέμπει την θέση του, ανάλογα αν η γενική πορεία του χρήστη είναι μια συχνά ακολουθούμενη πορεία ιστορικά από ικανοποιητικό πλήθος άλλων χρηστών της υπηρεσίας. Αν κριθεί μη ασφαλής τότε και μόνο τότε, ακολουθεί μια τεχνική γενίκευσης που προσπαθεί να εντοπίσει άλλα *k*-1 αιτήματα χρηστών για χρονικό διάστημα από κάποια ώρα πριν το αίτημα μέχρι και τη στιγμή του αιτήματος, τα οποία εκτείνονται σε μια εύλογη απόσταση από τη θέση του αιτούντος, διαφορετικά επέρχεται σημαντική πτώση της ποιότητας της παρεχόμενης υπηρεσίας.

2.4 Βάσεις κινούμενων αντικειμένων

2.4.1 Προστασία της ιδιωτικότητας σε δημοσιοποιημένα δεδομένα κίνησης

Οι τεχνολογικές εξελίξεις όπως η εξάπλωση παντού φορητών GPS συσκευών όπως στο κινητό, το αυτοκίνητο, τα δημόσια μέσα μεταφοράς αλλά και οι επεξεργαστικές και αποθηκευτικές δυνατότητες των κατοπινών υπολογιστικών συστημάτων, οδηγούν σε εύλογη μαζική συγκέντρωση δεδομένων για κινούμενα αντικείμενα. Αυτός ο όγκος αποθηκεύεται σε βάσεις κινούμενων αντικειμένων (*MOD*) με τη μορφή, κατά κανόνα, τροχιών.

Ως τροχιά (*trajectory*) θεωρείται μια αλληλουχία από τριάδες, έστω $T = <(x_1, y_1, t_1), (x_2, y_2, t_2),\ldots, (x_n, y_n, t_n)>$, όπου το t_i $(i = 1\ldots n)$ υποδηλώνει ένα χρονικό σημείο (*timestamp*) τέτοιο ώστε για κάθε $1 \leq i < n$, να ισχύει ότι $t_i < t_{i+1}$ και ότι (x_i, y_i) είναι σημεία στο $\mathbf{R^2}$ [12]. Η αποθήκευση και στατιστική επεξεργασία τέτοιων τροχιών από επιστήμονες διαφόρων ειδικοτήτων μπορεί να αποτελέσει ένα χρήσιμο εργαλείο στα χέρια τους.

Ο συνήθης τρόπος εργασίας είναι ότι ένας οργανισμός αναλαμβάνει τη συλλογή, επεξεργασία, ανωνυμοποίηση και δημοσιοποίηση της διαμορφωθείσας βάσης κινούμενων αντικειμένων (*Moving Object Database*), ώστε είτε περιορισμένος αριθμός φορέων, είτε ακόμα και το ευρύ κοινό να μπορεί να τη χρησιμοποιήσει και να εξάγει διαφόρων ειδών συμπεράσματα. Έτσι όμως, ανοίγεται πάλι ένα παράθυρο παραβίασης της ιδιωτικότητας ενός φυσικού προσώπου ακόμα και αν έχει συμβεί το αυτονόητο, να αφαιρεθούν, δηλαδή, όλα τα στοιχεία από την κάθε τροχιά που υποδηλώνουν ευθέως σε ποιο φυσικό πρόσωπο αντιστοιχεί.

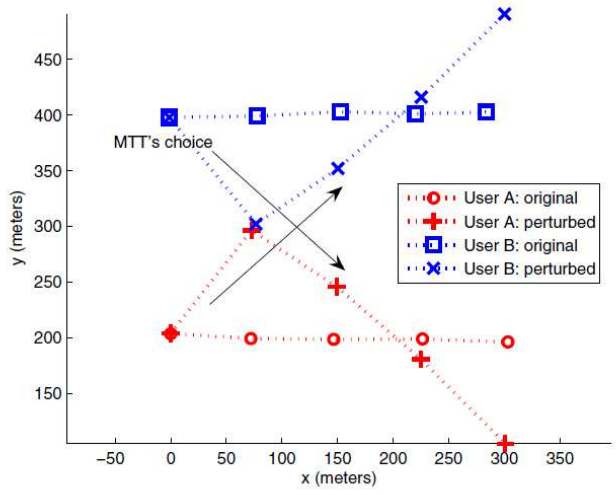
Ας φανταστούμε την περίπτωση που ο κακόβουλος χρήστης γνωρίζει εκ των προτέρων την κατοικία ενός ατόμου που θέλει να παρακολουθήσει ή ακόμη ότι το άτομο βρέθηκε σε ένα συγκεκριμένο μέρος μια συγκεκριμένη χρονική περίοδο, ενώ καταγραφόταν η τροχιά του. Για παράδειγμα, μπορεί ο κακόβουλος να έχει πρόσβαση σε κλήσεις από την Τροχαία οπότε γνωρίζει πού και πότε βρέθηκε ένα άτομο, όταν το κατέγραψε η κάμερα ελέγχου του ορίου ταχύτητας, σε ένα αυτοκινητόδρομο. Με βάση τέτοια στοιχεία μπορεί στη συνέχεια να ανατρέξει στη MOD και να εντοπίσει τη συγκεκριμένη τροχιά. Αν το καταφέρει αυτό, μπορεί να κατοπτεύσει άψογα και την ίδια την τροχιά συνολικά αλλά και να αποκομίσει γνώση και για άλλες τροχιές που ενδεχομένως έχουν καταγραφεί από το ίδιο άτομο, γιατί αν για παράδειγμα στην χρονική αρχή της τροχιάς αναχώρησε το άτομο από το μέρος Α και στο τέλος της ημερήσιας τροχιάς κατέληξε πάλι στο μέρος Α, τότε πολύ πιθανό να είναι η μόνιμη κατοικία του, άρα γνωρίζει μάλλον από πού ξεκινάει το ημερήσιο δρομολόγιό του.

Η θεμελιώδης τεχνική του *k*-anonymity μπορεί να έχει και στο πεδίο των MOD εφαρμογή. Αυτή τη φορά, θεωρούμε ότι εντός της κάθε δημοσιοποιημένης βάσης, κάθε μεμονωμένη τροχιά (που ενδεχομένως «ενδιαφέρει» ένα κακόβουλο χρήστη) έχει καταστεί *δυσδιάκριτη* μεταξύ τουλάχιστον άλλων *k*-1 τροχιών.

Για να επιτευχθεί η ανωνυμοποίηση της βάσης πριν αυτή δοθεί στην δημοσιότητα, έχουν προταθεί πολλές τεχνικές στη βιβλιογραφία. Μία παράμετρος όμως που δεν αποτελούσε αυτονόητα μια σταθερά μεταξύ των διαφορετικών αυτών προσεγγίσεων, ήταν οι διαφορετικές θεωρήσεις για το είδος της πρότερης γνώσης που μπορούσε να έχει ο κακόβουλος χρήστης, αλλά και η διαβάθμιση ή μη της «ευαισθησίας» μεταξύ των σημείων μιας τροχιάς ή ακόμη και η διαφοροποίηση ή μη της «ευαισθησίας» για το

ίδιο σημείο (π.χ. Νοσοκομείο) μεταξύ διαφορετικών καταγεγραμμένων ατόμων (π.χ. εργαζόμενος σε αυτό *vs* νοσηλευόμενος).

Οι *Hoh et al* [7] προτείνουν έναν αλγόριθμο, ο οποίος εξετάζει το σύνολο των τροχιών και σε κάθε περίπτωση όπου 2 τροχιές, των οποίων ορισμένα σημεία τους βρεθούν κοντά μεταξύ τους, τότε παράγονται κατάλληλα ψεύτικα σημεία για την κάθε τροχιά έτσι ώστε να φαίνεται ότι πραγματικά διασταυρώνονται. Η ιδέα είναι απλή. Κάθε φορά που ο κακόβουλος χρήστης ξεκινάει να παρακολουθεί μια τροχιά (γιατί μπορεί να γνωρίζει την ακριβή αφετηρία, για παράδειγμα την κατοικία ενός ατόμου) και συναντά μια διασταύρωση αυτής με μια άλλη, δεν μπορεί να είναι σίγουρος ποια από τις 2 πρέπει να συνεχίζει να παρακολουθεί. Όσες περισσότερες είναι οι διασταυρώσεις (κατασκευασμένες ή μη) τόσο πιο δύσκολο είναι γι' αυτόν να παρακολουθήσει σωστά την αρχική τροχιά. Η εγγύτητα που πρέπει να έχουν 2 τροχιές για να διασταυρωθούν ψεύτικα είναι παραμετροποιήσιμη με γνώμονα ότι όσο μεγαλώνει η ακτίνα τόσο αλλοιώνονται περισσότερο οι αρχικές τροχιές αλλά και τόσο περισσότερες διασταυρώσεις δημιουργούνται και αντίστροφα. Έτσι δημιουργούνται νέες τροχιές που αποκλίνουν σε ορισμένα σημεία από τις αρχικές με τέτοιο τρόπο ώστε να δημιουργείται η μέγιστη δυνατή σύγχυση στον κακόβουλο χρήστη, λόγω των παραγόμενων διασταυρώσεων.



Εικ. 2.3 Δύο χρήστες κινούνται παράλληλα. Ο αλγόριθμος *Path Perturbation* παραλλάσσει το παράλληλο τμήμα τους σε διασταυρούμενο τμήμα - *Hoh et al* [7]

Αυτό βέβαια δεν παύει να σημαίνει ότι έστω και ένα τμήμα της κάθε τροχιάς, εφόσον έχει ταυτιστεί με εξωγενή γνώση του κακόβουλου χρήστη, μπορεί να παραβιάζει «έντονα» την ιδιωτική ζωή ενός ατόμου αν λόγου χάρη η τροχιά αυτή περνά από μια «ευαίσθητη» περιοχή.

Οι *Terrovitis & Mamoulis* [18] εργάστηκαν με ένα διαφορετικό τρόπο πάνω σε μια βάση που ήθελαν να καταστεί ασφαλής σύμφωνα με τη λογική της *k*-anonymity. Θεώρησαν την αρχική βάση ως ένα σύνολο τροχιών, όπου η κάθε τροχιά ήταν ένα σύνολο σημείων στο χάρτη (x_i, y_i) διαδοχικά «τοποθετημένων» μεταξύ τους. Δηλαδή ο χαρτογραφούμενος πέρασε από το σημείο Α, μετά από το σημείο Β και κατέληξε στο Γ. Δηλαδή υπήρξε μια τροχιά του τύπου Α→ Β→ Γ. Από την άλλη υπήρχε μια δεύτερη βάση που ήταν καταχωρημένες όλες οι γνώσεις που είχαν οι κακόβουλοι χρήστες με τη μορφή, επίσης, αλληλουχίας σημείων. Για παράδειγμα ένας κακόβουλος χρήστης γνώριζε ότι κάποιος από τους καταγραφόμενους πήγε πρώτα στο σημείο Α και κατόπιν στο Β. Σκοπός των συγγραφέων ήταν να δημιουργήσουν έναν αλγόριθμο που να κατασκευάζει μια τελική βάση, η οποία να περιέχει με τέτοιο τρόπο τις αρχικές τροχιές ώστε για κάθε γνωστό, στον κακόβουλο χρήστη, τμήμα τροχιάς (π.χ. Α→ Β) να περιλαμβάνονται τουλάχιστον *k* «τελικές» τροχιές που περιέχουν αυτό το τμήμα της τροχιάς.

Η τεχνική που εφαρμόζονταν ήταν η εξάλειψη σημείων από τροχιές της αρχικής βάσης, όταν και για όσο δεν μπορούσε να επιτευχθεί το *k*-anonymity. Αν δηλαδή στο παράδειγμά μας η αρχική ήταν η Α→ Β→ Γ και η γνώση του κακόβουλου Α→ Β, θα έπρεπε τουλάχιστον *k* τροχιές να περιέχουν το τμήμα Α→ Β, αλλιώς θα εξαλείφονταν ένα περισσότερο σημείο της αρχικής τροχιάς.

Η τεχνική τους αφορά κατά βάση χωρικά δεδομένα χωρίς να έχει γίνει σε βάθος μελέτη των επιπτώσεων, βάζοντας στο τραπέζι και τη χρονική παράμετρο. Επίσης δεν είναι ιδιαίτερα επεκτάσιμη, γενικεύσιμη και κατ' επέκταση ρεαλιστική, διότι ο αλγόριθμος τους απαιτεί από τον κάτοχο της βάσης και υπεύθυνο για την δημοσιοποίηση της να γνωρίζει εκ των προτέρων και επ' ακριβώς το σύνολο των γνώσεων του κακόβουλου χρήστη.

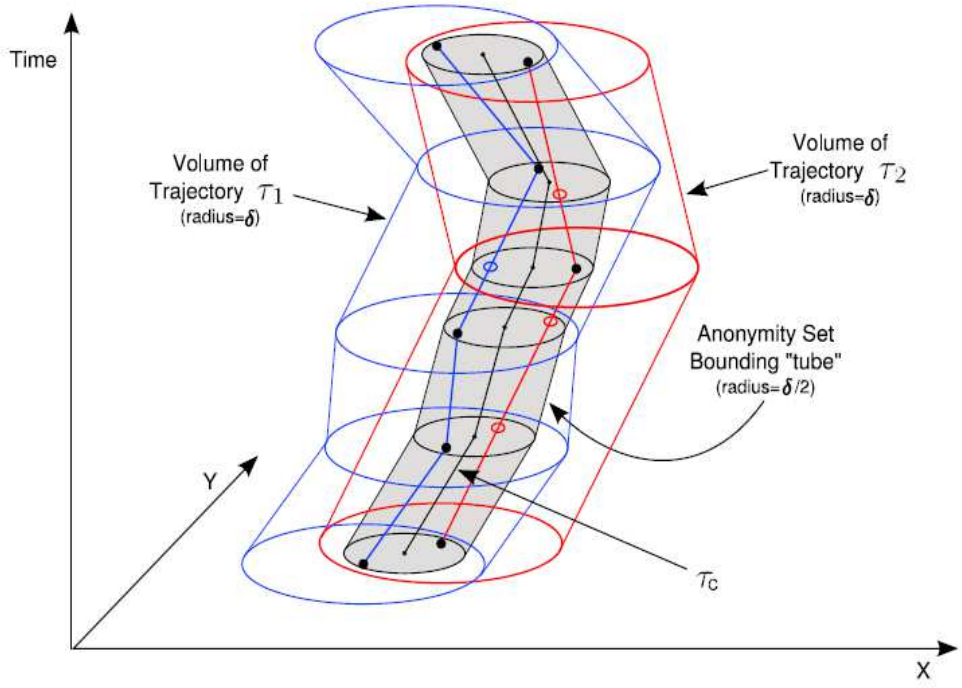
Ένα τελευταίο σημείο συζήτησης είναι τι γίνεται στην περίπτωση που (με βάση το παράδειγμα μας) υπήρχαν *k* τροχιές που οδηγούσαν από το Α στο Β (Α→ Β) αλλά όλες

τους στη συνέχεια οδηγούσαν στο σημείο Γ, το οποίο μπορεί για όλους τους καταγεγραμμένους χρήστες να είναι μια ιδιαίτερα ευαίσθητη τοποθεσία (π.χ. Νοσοκομείο).

Οι *Abul et al* [1] πρότειναν την επέκταση του *k*-anonymity ορμώμενοι από κλασσικούς αλγορίθμους συσταδοποίησης (*clustering*). Με τη χρήση τους, κατασκεύαζαν ομάδες τροχιών ως ξεχωριστές συστάδες τροχιών, εμπνεόμενοι από το γεγονός ότι έτσι και αλλιώς λόγω του περιθωρίου σφάλματος που, εκ των πραγμάτων, έχουν διάφορες συσκευές καταγραφής θέσης (*GPS*) η κάθε τροχιά που εν τέλει καταγράφεται, δεν περνά ακριβώς από τα σημεία που κατέγραψε η συσκευή GPS, αλλά βρίσκεται σίγουρα εντός ενός κυλίνδρου με κεντρικό άξονα την αποτυπωμένη τροχιά και ακτίνα το περιθώριο σφάλματος λόγω της συσκευής ή άλλων εξωγενών παραγόντων. Η διάμετρος αυτού του κυλίνδρου καθορίζει τελικά το μέγεθος της αβεβαιότητας που προκαλείται στον παρατηρητή και την ονόμασαν *δ*. Αν εντός αυτού του κυλίνδρου βρισκόταν και μια δεύτερη, έστω τροχιά, τότε ένας παρατηρητής θα ήταν πραγματικά ανήμπορος να ξεχωρίσει με σιγουριά τις τροχιές μεταξύ τους, συνδέοντάς τες με κάποια πρότερη γνώση (*background knowledge*).
Πιο συγκεκριμένα, εισήγαγαν τον αλγόριθμο *NWA (Never Walk Alone)*. Αυτός, εκτελούμενος, προσπαθεί να επιτύχει αυτό που οι συγγραφείς όρισαν ως *(k-δ)*-anonymity. Ένας συνοπτικός ορισμός που θα μπορούσε να διατυπωθεί είναι ο εξής [1]:
Έστω ένα σύνολο τροχιών (*trajectories*) *D*, ένα κατώφλι αβεβαιότητας *δ* και το (γνωστό) κατώφλι ανωνυμίας *k*. Για να επιτευχθεί το *(k-δ)*-anonymity απαιτείται το *D* να μετασχηματιστεί σε *D΄*, με τέτοιο τρόπο ώστε κάθε τροχιά *τ* του *D΄* να ανήκει πάντα σε ένα σύνολο τροχιών *S* του *D΄*. Τα χαρακτηριστικά του (κάθε) συνόλου *S* είναι: αποτελεί γνήσιο υποσύνολο του *D΄*, το πλήθος των τροχιών που περιέχει είναι τουλάχιστον *k* και όλες μαζί οι τροχιές από τις οποίες απαρτίζεται μπορούν να οριοθετηθούν χωρο-χρονικά εντός ενός κυλινδρικού όγκου (*cylindrical volume*) ακτίνας *δ*/2, με γνώμονα πάντα την ελάχιστη απαιτούμενη διαφοροποίηση του *D΄* σε σχέση με το *D* από το οποίο προήλθε.
Ο αλγόριθμος αποτελείται από 3 βασικά βήματα. Στο πρώτο διενεργείται μια προεργασία πριν την συσταδοποίηση. Στα πλαίσια αυτής, γίνεται μια τυποποίηση των τροχιών ως προς τις χρονικές στιγμές έναρξης και λήξης της καθεμιάς από αυτές,

καθώς και ένας «συγχρονισμός» μεταξύ των τροχιών ως προς το χρονικό τους αποτύπωμα. Δηλαδή γίνεται ένα είδος «δειγματοληψίας» σε συγκεκριμένες χρονικές στιγμές (π.χ. κάθε 10΄) από το σύνολο των διαθέσιμων σημείων των τροχιών, ενώ τα υπόλοιπα σημεία εξαλείφονται.

Στο δεύτερο εκτελείται η συσταδοποίηση (*clustering*) που προαναφέραμε. Οι κεντρικές τροχιές των συστάδων επιλέγονται ως η μακρινότερη τροχιά από την αμέσως προηγούμενη επιλεγμένη κεντρική τροχιά με αφετηρία την μακρινότερη από τον νοητό κεντρικό άξονα που σχηματίζει το σύνολο των τροχιών. Γύρω από κάθε κεντρική τροχιά επιδιώκεται να συγκεντρωθούν k+ άλλες τροχιές αρκεί πάντα να τηρείται ότι η απόσταση της νέας τροχιάς ως προς την κεντρική της είναι μικρότερη από ένα προκαθορισμένο κατώφλι.



Εικ. 2.4 Ένα (2,±)-anonymity σετ που σχηματίζεται από δύο συν-εντοπισμένες τροχιές, ο αντίστοιχος όγκος αβεβαιότητάς τους και ο κεντρικός κυλινδρικός όγκος ακτίνας ±/2, που περιέχει και τις δύο τροχιές Abul et al [1]

Στο τρίτο, αμέσως μετά τη συσταδοποίηση, εφαρμόζουν μια τεχνική μαζικής μετακίνησης σημείων χωρικά (*Space Translation*), ούτως ώστε για την ίδια συστάδα και για συγκεκριμένο διάστημα χρόνου, να τοποθετηθούν όλα τα σημεία με τέτοιο

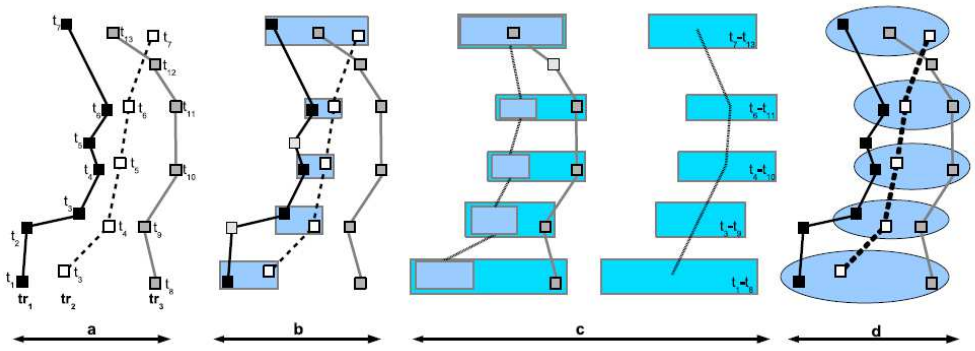
τρόπο, ώστε αν *δ* είναι *η διάμετρος του εικονικού κυλίνδρου*, αυτά να έχουν μέγιστη απόσταση *δ*/2 από τον κεντρικό άξονά του.

Μετά από σύντομο διάστημα, οι συγγραφείς επιχείρησαν μια επέκταση του αλγορίθμου NWA προκειμένου να ξεπεραστούν ορισμένες αδυναμίες του, εκ των οποίων η κυριότερη ήταν ότι βασιζόταν σε μια συνάρτηση που μετρούσε την ευκλείδεια απόσταση μεταξύ τροχιών, ώστε να υπολογιστεί ο βαθμός εγγύτητάς τους και απαιτούσε οι τροχιές να είναι ίδιου μήκους. Αυτός ήταν και ο λόγος ύπαρξης του πρώτου βήματος του NWA. Όμως η ευκλείδεια απόσταση δεν μπορούσε να διαβλέψει σωστά, πολλές φορές το βαθμό εγγύτητας, ειδικά σε ένα πιο αναλυτικό επίπεδο. Η επέκταση ονομάστηκε *W4M (Wait for Me)*, στην οποία ενώ η γενική στρατηγική συσταδοποίησης παραμένει η ίδια, ο τρόπος υπολογισμού και σύγκρισης κατ' επέκταση των αποστάσεων άλλαξε και βασίζεται πλέον στην EDR (*Edit distance function*) απόσταση [2]. Αυτή συνυπολογίζει και τον χρόνο εκτός από την χωρική πληροφορία μεταξύ τροχιών, με αποτέλεσμα να καθίσταται περιττό το πρώτο βήμα του NWA.
Επιπλέον, το σύνολο των τροχιών υπόκειται σε μια μαζική επεξεργασία κατά τη συσταδοποίηση και όχι κομμάτι - κομμάτι όπως στον NWA, και για να ελαφρυνθεί κάπως η υπολογιστική επιβάρυνση που προέρχεται από αυτό, αποφασίστηκε να επιλέγεται τυχαία η νέα κεντρική τροχιά μεταξύ των τροχιών σε κάθε επανάληψη του αλγορίθμου.
Το τρίτο βήμα εκτελείται όπως και στον NWA και οι τροχιές της κάθε συστάδας «πλησιάζουν» την κεντρική, εφόσον είναι απαραίτητο κάτι τέτοιο, με τη διαφορά ότι η κεντρική δεν είναι η μέση τιμή της συστάδας, αλλά η τροχιά που επιλέχθηκε τυχαία προηγουμένως και επειδή οι τροχιές μιας συστάδας δε συμπίπτουν πλέον απόλυτα χρονικά, προσθαφαιρούνται σημεία ώστε στο τέλος όλες οι τροχιές της συστάδας να προσεγγίσουν την τυχαία επιλεγμένη, καθώς και να έχουν όλες το ίδιο πλήθος σημείων (x_i, y_i, t_i) σε κοινές χρονικές στιγμές και πάντα ικανοποιώντας το $(k\text{-}\delta)$-anonymity.

Μια πρόταση αντιμετώπισης των θεμάτων της ανωνυμοποίησης των τροχιών παρατέθηκε από τους *Negriz et al* [13]. Σε αυτή την περίπτωση λαμβάνονταν υπόψη και ο χρόνος, πέραν της χωρικής διάστασης των σημείων που απαρτίζουν την κάθε τροχιά και η προσπάθεια αποσκοπεί στην κατασκευή μιας βάσης από την οποία κάθε

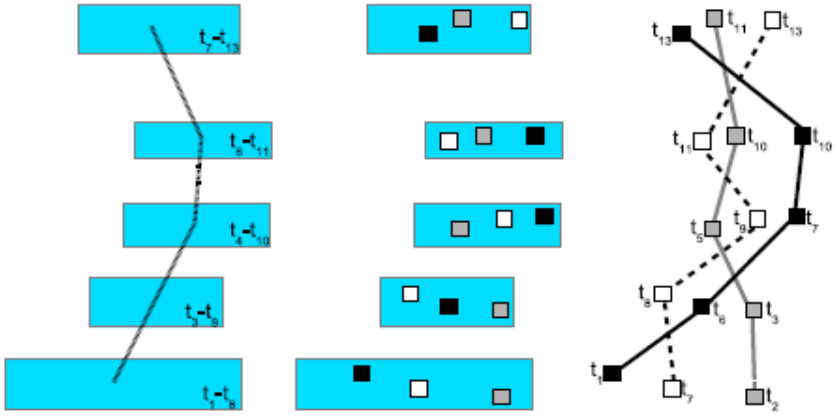
απάντηση σε ερώτημα επιστρέφει τουλάχιστον *k* τροχιές. Θεωρείται ότι ο κακόβουλος χρήστης μπορεί να γνωρίζει τμήμα ή ολόκληρη τροχιά από τις διαθέσιμες στην αρχική βάση.

Η τεχνική αποτελείται από 3 βασικά στάδια. Στο πρώτο γίνεται ένταξη των τροχιών σε ομάδες των *k*, δίνοντας βάρος στον τρόπο που υπολογίζει ο αλγόριθμος την επόμενη πλησιέστερη τροχιά κάθε φορά.



Εικ. 2.5 Διαδικασία ανωνυμοποίησης - Negriz et al [13]
a. οι τροχιές tr1,tr2, and tr3. b. ανωνυμοποίηση σε tr* των tr1 και tr2. c. Ανωνυμοποίηση των tr* και tr3. d. Τα σημεία που χρησιμοποιήθηκαν για την ανωνυμοποίηση των tr1,tr2, and tr3. Η αντιστοίχιση περιέχει συνδέσεις μεταξύ 5 σημείων

Στο δεύτερο στάδιο πραγματοποιείται η καθαυτή ανωνυμοποίηση. Σε κάθε ομάδα από τροχιές και ξεκινώντας με βάση μια από τις *k*, βρίσκεται η «πλησιέστερη» της, ευθυγραμμίζοντας (όσα είναι δυνατόν) τα σημεία μεταξύ των 2 αυτών τροχιών και στη συνέχεια αντικαθίσταται το κάθε ζεύγος σημείων από ένα MBB που τα περικλείει και θεωρώντας το σύνολο των MBB's ως μια νέα συγκεντρωτική τροχιά προχωρά στην αντιστοίχιση και συγχώνευση με την επόμενη μέχρι να εξαντληθούν και οι *k*. Εννοείται ότι όπου δεν επιτυγχάνεται ευθυγράμμιση τα αντίστοιχα σημεία εξαλείφονται. Έτσι καταλήγει να σχηματιστεί μια χρονική αλληλουχία από MBB's που αναπαριστά την αρχική ομάδα από trajectories.



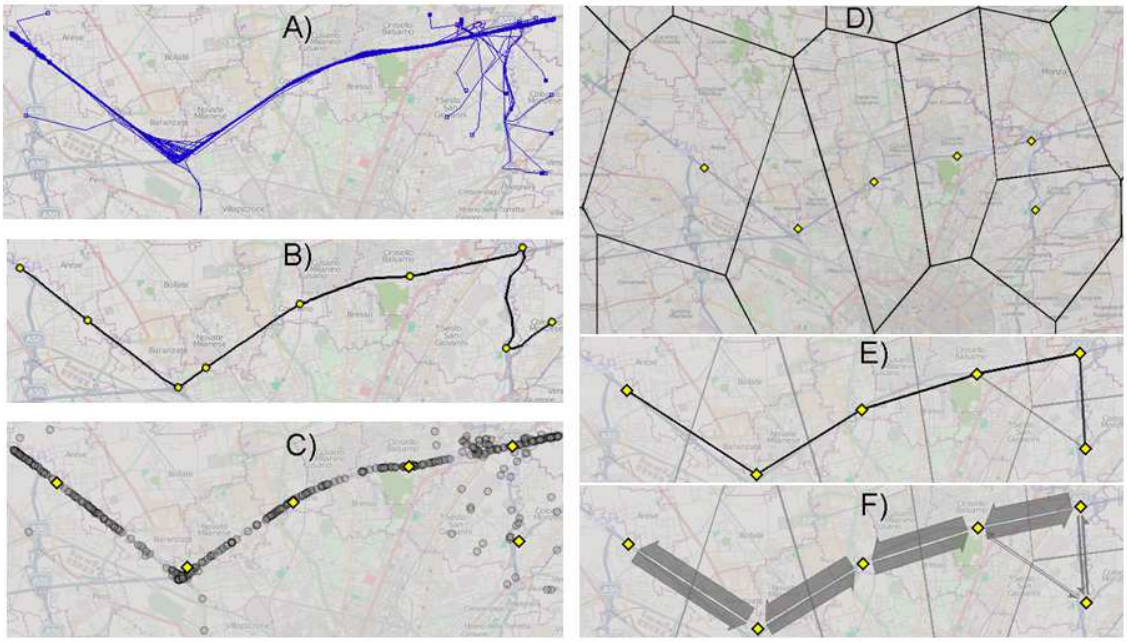
Εικ. 2.6 Διαδικασία ανακατασκευής - Negriz et al [13]

Ως τελευταίο στάδιο, συντελείται μια μαζική ανακατασκευή των επιμέρους τροχιών εκτός της κάθε ομάδας, βασισμένη προφανώς στα MBB's που αντιπροσωπεύουν πλέον την ομάδα.

Αμφίβολο παραμένει βέβαια πόσο αληθοφανές ή μη είναι το τελικό αποτέλεσμα, δηλαδή πόσο ρεαλιστικές είναι οι τροχιές που ανακατασκευάζονται. Αν για παράδειγμα οι τροχιές αφορούν οδηγούς αυτοκινήτων, δεν λαμβάνεται υπόψη το υπάρχον οδικό δίκτυο κατά την ανακατασκευή τους.

Από τους *Monreale et al* [12] προτάθηκε μια διαφορετική αντιμετώπιση, κάνοντας μια μίξη διαδικασιών γενίκευσης δεδομένων και συσταδοποίησης, στη συνέχεια των παραπάνω καθώς και μια τελική αξιολόγηση/αναδιάρθρωση των τροποποιημένων τροχιών βασισμένη στο *k*-anonymity, πριν η βάση δεδομένων να είναι έτοιμη προς δημοσίευση. Θεωρούν ότι ο κακόβουλος χρήστης μπορεί να γνωρίζει επακριβώς όλες τις διαδικασίες που εφαρμόζονται για να ανωνυμοποιηθούν οι τροχιές όπως και συγκεκριμένα τμήματα τροχιών που καταγράφηκαν για «αντικείμενα» (ανθρώπινη κίνηση) που τους ενδιαφέρουν.

Οι βασικοί άξονες των ενεργειών που εκτελούνται είναι οι εξής: Αρχικά για κάθε μια τροχιά, επιλέγονται τα αντιπροσωπευτικότερα σημεία αυτής, όπως το πρώτο και το τελευταίο χρονικό σημείο, σημεία στα οποία αλλάζει πάνω από «ένα βαθμό» η κατεύθυνση της τροχιάς ή απλά σημεία ανά κάποια απόσταση που διαγράφουν μεγάλες ευθείες και αυτό είναι ένα πρώτο μέτρο γενίκευσης των δεδομένων.



Εικ. 2.7 : Απεικόνιση της διαδικασίας γενίκευσης των τροχιών. - Monreale et al [12]
A) Υποσύνολο των τροχιών.
B) Μια από τις τροχιές και τα χαρακτηριστικά σημεία που εξήχθησαν από αυτή.
C) Χαρακτηριστικά σημεία από όλες τις τροχιές (αναπαριστώνται με κύκλους) και τα κεντροειδή των χωρικών συστάδων των σημείων (αναπαριστώνται με ρόμβους).
D) Κατάτμηση της περιοχής σε κελιά Voronoi .
E) Γενικευμένες τροχιές.
F) Μια συνοπτική αναπαράσταση των τροχιών. Το πάχος του κάθε συμβόλου ροής είναι ανάλογο με τον αριθμό των τροχιών που πηγαίνουν εντός του αντίστοιχου ζεύγους περιοχών στη συγκεκριμένη κατεύθυνση.

Στη συνέχεια εκτελείται μια συσταδοποίηση στα δεδομένα αυτά που λαμβάνει υπόψη μόνο τη χωρική διάσταση των σημείων και δεν την ενδιαφέρει σε ποια τροχιά ανήκει το κάθε σημείο. Έτσι ακολούθως είναι εύκολο να οριοθετηθεί-κατασκευαστεί η κάθε περιοχή που περικλείει τα σημεία της κάθε συστάδας και στη συνέχεια να βρεθεί το κεντροειδές αυτής. Ουσιαστικά χρησιμοποιείται η λογική των κελιών Voronoi με τη χαρακτηριστική τους ασυμμετρία που καλύπτουν πλέον όλη τη χωρική περιοχή και εφάπτονται μεταξύ τους.

Κατόπιν, «εξαναγκάζονται» όλα τα σημεία της συστάδας να ταυτιστούν με το κεντροειδές της. Μετά περνάμε σε ένα δεύτερο επίπεδο γενίκευσης που στοχεύει στην συνένωση γειτονικών περιοχών, ώστε να μεγαλώνει το πλήθος των σημείων που περιέχει όταν αυτό κρίνεται απαραίτητο.

Έχοντας δηλαδή πλέον διαμορφώσει τις νέες τροχιές που αποτελούν όλες μια αλληλουχία «επισκέψεων» στα κεντροειδή διαφόρων περιοχών, προκειμένου να έχει εφαρμογή το *k*-anonymity, ελέγχονται όλες οι γειτονικές μεταξύ τους περιοχές, έτσι

ώστε οι μεταβάσεις τροχιών από τη μια συγκεκριμένη περιοχή σε μια άλλη, να ξεπερνούν τις *k*. Διαπιστώθηκε ότι καλύτερο είναι να μην εκτελείται μέχρι τέλους το συγκεκριμένο τμήμα του αλγορίθμου, διότι οδηγεί σε μεγάλες γενικεύσεις και τεράστια απώλεια χρήσιμης πληροφορίας. Αντιθέτως μάλιστα, υπάρχει πρόβλεψη, ώστε οι τροχιές που έχουν απορριφθεί προσωρινά, να επανεξετάζονται συνολικά και εφόσον υπάρχουν κοινά τμήματα μεταξύ τους (*subtrajectories*) να επανεισάγονται (μόνο αυτά) πλέον στα δεδομένα της βάσης. Για να υλοποιηθεί ο γενικός έλεγχος του *k*-anonymity που περιγράφηκε, χρησιμοποιήθηκε μια δομή δεδομένων καλούμενη *prefix tree* που είναι ένα σύνολο ιεραρχικών δομημένων κόμβων που ο καθένας περιγράφει μια περιοχή και πόσες κοινές τροχιές «έρχονται» από κάθε γειτονική περιοχή (αν υπάρχουν). Με βάση αυτό αποφασίζεται εύκολα ποια τμήματα της δενδροειδούς δομής δεν πρέπει να εμφανίζονται στην τελική δημοσιεύσιμη έκδοση της βάσης, προκειμένου να ισχύει το *k*-anonymity.

Μια πρόταση που διαφοροποιεί την ομοιογένεια με την οποία αντιμετωπίζονται στις προηγούμενες προσεγγίσεις, οι τροχιές στη φάση της συσταδοποίησης διατυπώθηκε από τους *Mahdavifar et al* [10]. Το βασικό σκεπτικό ήταν ότι δεν είναι λογικό ούτε πρέπει να είναι υποχρεωτικό, οι απαιτήσεις ιδιωτικότητας μεταξύ των χρηστών/αντικειμένων να είναι ίδιες. Με βάση αυτό, η κάθε τροχιά έχει διαφορετικό επίπεδο ιδιωτικότητας και όσο πιο υψηλό το επίπεδο τόσο μεγαλύτερο το πλήθος των τροχιών εντός της συστάδας που ανήκει η συγκεκριμένη τροχιά. Εξετάζονται πρώτα οι τροχιές «υψηλής ιδιωτικότητας» και ελέγχεται μονίμως να μην ξεπερνιέται μια μέγιστη επιτρεπτή απόσταση του κεντροειδούς της συστάδας από την υποψήφια τροχιά. Όλες οι συγκρίσεις αυτές μεταξύ αποστάσεων, γίνονται με τη χρήση της συνάρτησης *EDR*.

Μετά από αυτή την διαδικασία, ακολουθεί η φάση της ανωνυμοποίησης των τροχιών εντός της κάθε συστάδας με τη βοήθεια του αλγορίθμου FLTP. Είναι βέβαιο πάντως ότι η προσπάθεια για διαφοροποιημένη αντιμετώπιση μεταξύ τροχιών ή αλλιώς το να προσθέτει κανείς ειδικό βάρος σε ορισμένες από αυτές, είναι ένα σκεπτικό που εύλογα συνάδει με τη φύση των βάσεων δεδομένων από κινούμενα αντικείμενα και ειδικά όταν πρόκειται για τροχιές ανθρώπων.

2.4.2 Στατιστικές βάσεις

Όπως ήδη διατυπώθηκε στην εισαγωγή, η παρούσα εργασία, σκοπό έχει να παρουσιάσει κινδύνους και τρόπους αντιμετώπισης αυτών που προκύπτουν από ερωτήματα που τίθενται προς μια βάση δεδομένων που «φιλοξενεί» σημασιολογικά επαυξημένες τροχιές, δηλαδή στη πράξη μιλάμε για αλληλουχίες επεισοδίων, τα οποία απαρτίζουν την κάθε τροχιά και επί των οποίων μπορούν να τεθούν διάφορα ερωτήματα. Όπως λοιπόν προαναφέραμε, η βάση επιστρέφει δύο είδη πληροφορίας που είναι μια αναπαράσταση τροχιών στο χώρο (παραλλαγμένες / εμπλουτισμένες σε σχέση με τις πραγματικές) *και* ένα στατιστικό μέγεθος, εν προκειμένω, το *πλήθος* των τροχιών που καλύπτουν τα κριτήρια ενός ερωτήματος.

Η δυνατότητα / πιθανότητα / τεχνική παραβίασης της ιδιωτικότητας οντοτήτων που έχουν καταγραφεί σε μια βάση διενεργούμενων διαδοχικών στατιστικών ερωτημάτων όπως το *πλήθος* είναι ένα επιστημονικό πεδίο το οποίο ως προς τις παραδοσιακές μορφές βάσεων δεδομένων (σχεσιακές και άλλες) έχει μελετηθεί επαρκώς εδώ και αρκετά χρόνια.

Έτσι, μελέτες όπως των *Adam & Wortmann* [3], η οποία αναλύει όλα τα καταγεγραμμένα προβλήματα και τις πιθανές αντιμετωπίσεις στις στατιστικές βάσεις δεδομένων της εποχής, μπορούν να αποτελέσουν μια επιπλέον πλατφόρμα συζήτησης και, γιατί όχι, έμπνευσης πάνω στα τρέχοντα ζητήματα παραβίασης της ιδιωτικότητας στο σκέλος τουλάχιστον που οι μοντέρνες βάσεις κινούμενων αντικειμένων επιστρέφουν στους χρήστες τους, στατιστικά μεγέθη ως απάντηση στα ερωτήματα τους. Αυτό ουσιαστικά επιτρέπουμε να συμβεί αν παραλληλίσουμε, θεωρητικά το κάθε επεισόδιο μιας τροχιάς αποθηκευμένης σε μια MOD με ένα απλό μη αριθμητικό δεδομένο σε μια σχεσιακή βάση δεδομένων. Μια γενική παρατήρηση, κατ' αρχάς, αφορά το γεγονός ότι εξ' αιτίας των περιορισμένων υπολογιστικών δυνατοτήτων της υποδομής εκείνης της εποχής, διαφαίνεται ότι ενώ σε αλγοριθμικό επίπεδο υπήρχε πληθώρα σκέψεων και απόψεων, λόγω της αυξημένης υπολογιστικής πολυπλοκότητας που απαιτούσαν, είτε δεν προχωρούσε η υλοποίηση της, είτε πολλές φορές εφαρμόζονται εξ' αρχής ορισμένου βαθμού εκπτώσεις – συμβιβασμοί προκειμένου να διευκολυνθεί η υλοποίηση των αλγόριθμων αυτών. Έτσι, κατέγραψαν 4 μεγάλες κατηγορίες αντιμετώπισης των προβλημάτων παραβίασης της ιδιωτικότητας.

Η πρώτη κατηγορία ονομάστηκε εννοιολογική προσέγγιση (*conceptual approach*) και αποτελείται από δύο κατηγορίες με τη σειρά της. Η πρώτη εισήγαγε την έννοια του A-population. Το A-population ήταν η εξ' ορισμού μικρότερη αδιαίρετη πληροφορία που μπορούσε να απαντηθεί στον χρήστη της βάσης, ανεξαρτήτως του ερωτήματος. Για κάθε «ευαίσθητο» συνδυασμό πεδίων της βάσης που μπορούσε να αποκαλύψει την ύπαρξη ή όχι μιας εγγραφής με χαρακτηριστικά που ενδιαφέρουν τον κακόβουλο χρήστη, λαμβάνονταν μέριμνα ώστε να μην είναι ποτέ μοναδικός, αλλά σε ομάδα των δύο. Υπήρχε πρόβλεψη για τις περιπτώσεις των delete και update στη βάση, μέσω κατά κανόνα είτε γενικεύσεων, είτε απαλοιφής των δεδομένων που θα δημιουργούσαν πρόβλημα στην ασφάλεια, είτε ακόμα και εισαγωγές ψεύτικης πληροφορίας για τον αποπροσανατολισμό του κακόβουλου χρήστη. Αυτή η τεχνική θυμίζει έντονα το *k*-anonymity, όπως το εισήγαγε η L. Sweeney μια δεκαετία αργότερα και για την ακρίβεια το 2-anonymity.

Η δεύτερη υποκατηγορία της συγκεκριμένης προσέγγισης κάνει μια διαφορετική θεώρηση του τρόπου που μπορεί κανείς να εποπτεύσει μια στατιστική βάση δεδομένων. Θεωρεί ότι την πληροφορία μπορεί να την προσαρμόζει σε διαφορετικά επίπεδα συνάθροισης κάθε φορά, έτσι ώστε να ελέγχεται από το μηχανισμό αν σε ένα πιο αναλυτικό επίπεδο παρουσίασης της πληροφορίας υπάρχει περίπτωση μια μόνο εγγραφή να έχει ένα συγκεκριμένο συνδυασμό χαρακτηριστικών. Σε αυτή την περίπτωση δίνεται η δυνατότητα γενίκευσης της παρεχόμενης πληροφορίας μέσω της αφαίρεσης κατά την παρουσία ενός, κάθε φορά, χαρακτηριστικού μέχρι να επιτευχθεί το απαιτούμενο επίπεδο ασφάλειας. Όλη αυτή η λογική δείχνει να είναι ένας προάγγελος του ευρέως πλέον χρησιμοποιούμενου *κύβου* δεδομένων. (Κάθε χαρακτηριστικό είναι ουσιαστικά μια διάσταση του κύβου. Μπορεί να παρουσιαστεί ως προς οποιοδήποτε συνδυασμό διαστάσεων και με εφαρμογή διαφόρων κριτηρίων επί των χαρακτηριστικών, αρκεί να μην υπάρχει περίπτωση προβολής μιας και μόνης εγγραφής που πληροί τα κριτήρια που θέτει ο χρήστης.

Η δεύτερη μεγάλη κατηγορία αφορά την περίπτωση κατά την οποία αντιμετωπίζουμε τις διάφορες επιθέσεις στη βάση με διατάραξη των δεδομένων της (*data perturbation*). Ουσιαστικά πρόκειται για σύνολο από μεθοδολογίες αλλοίωσης / τροποποίησης των δεδομένων με σκοπό το καλύτερο δυνατό αποτέλεσμα κάθε φορά τόσο από τη σκοπιά της διατήρησης της ιδιωτικότητας όσο και από τη σκοπιά της διατήρησης της

στατιστικής χρησιμότητας της βάσης. Αυτή, κατά κανόνα, αφορά τις περιπτώσεις που δίνεται στη δημοσιότητα η βάση και δεν εμπλουτίζεται με νέα δεδομένα συνεχώς. Η μια κατηγορία τεχνικών υλοποίησης βασίζεται στην αναδημιουργία των δεδομένων της βάσης από την αρχή, βασισμένα στα στατιστικά μεγέθη που «απεικονίζουν» την πραγματική βάση, δηλαδή με τέτοιο τρόπο ώστε να ακολουθούνται οι ίδιες πιθανότητες κατανομής των τιμών του κάθε πεδίου στην αρχική αλλά και στην κατασκευασμένη βάση. Η πλήρης ανακατασκευή των δεδομένων με τρόπο ώστε να τηρείται στη νέα βάση η «τάση» της εν γένει πληροφορίας, αποτελεί ένα πρωτογενές υπόβαθρο αλγοριθμικού προβληματισμού πάνω στα σύγχρονα ανοιχτά προβλήματα που αφορούν πολύ πιο σύνθετες δομές δεδομένων, όπως η γραφική απεικόνιση τροχιών κινούμενων αντικειμένων εξ' ολοκλήρου ανακατασκευασμένων από τις πραγματικές για να επέρχεται η επιθυμητή ανωνυμοποίηση τους. Μάλιστα αντίστοιχα περίπου προβλήματα με το συστηματικό σφάλμα που παρουσίαζαν πολλές φορές αυτές οι μέθοδοι στις παραδοσιακές ΒΔ, έχουν εντοπιστεί και αντιμετωπίζονται. Η δεύτερη κατηγορία εφάρμοζε μια σταθερή αλλοίωση των τιμών του «ευαίσθητου» πεδίου. Όταν αφορούσε αριθμητικά πεδία, είναι εύκολο να το φανταστεί κανείς, ενώ όταν αφορούσε αριθμητικά πεδία με δύο μόνο είδη αποδεκτών τιμών (έστω x και y) η τροποποίηση των δεδομένων εφαρμοζόταν ως εξής. Ο διαχειριστής της βάσης (*DBA*) όριζε την τιμή μιας σταθερής παραμέτρου *p*, με τυπικές τιμές μεταξύ 0,6 και 0,8. Έτσι, με πιθανότητα p η αρχική τιμή (x) παρέμενε αμετάβλητη (x), ενώ με πιθανότητα 1-p η αρχική τιμή (x) άλλαζε σε (y) και αντίστροφα. Η προσπάθεια να διερευνηθεί αυτή η μεθοδολογία για περισσότερα από 2 είδη τιμών του κάθε πεδίου οδήγησε σε τεράστια απώλεια χρήσιμης πληροφορίας.

Η τρίτη μεγάλη κατηγορία αφορά την διατάραξη της απάντησης που δίνεται κάθε φορά στο χρήστη και όχι όλης της βάσης εκ προοιμίου όπως προηγουμένως. Μια προσέγγιση υλοποιείται με την τυχαία επιλογή από το answer set του ερωτήματος ενός δείγματος εγγραφών. Εντοπίστηκαν ποικίλα προβλήματα όπως το γεγονός ότι σε μικρού μεγέθους αρχικό answer set, υπάρχει πρόβλημα στην «εξαγωγή» ασφαλούς δείγματος καθώς και το φαινόμενο ότι το ίδιο ερώτημα αν επαναληφθεί δεν δίνει ακριβώς την ίδια απάντηση. Μια άλλη προσέγγιση στρογγυλοποιεί τα εκάστοτε αποτελέσματα με σκοπό να γενικεύσει με αυτόν τον τρόπο την παραγόμενη πληροφορία. Γενικά δεν αποδείχθηκε

ιδιαίτερα δόκιμη μέθοδος, γιατί πολλές φορές με κατάλληλη χρησιμοποίηση απλών μαθηματικών κατέληγε κανείς να γνωρίζει το αρχικό answer set.

Η τέταρτη και τελευταία μεγάλη κατηγορία αφορά την προστασία της βάσης, στηριζόμενη σε τεχνικές ελέγχου των ερωτημάτων κατά τη διενέργεια τους. Οι κυριότερες τεχνικές από αυτές ήταν:

α) Η τεχνική ελέγχου του μεγέθους της απάντησης στο ερώτημα. Αυτή θεωρούσε ότι αν το πλήθος των οντοτήτων που απαρτίζουν την απάντηση είναι C τότε πρέπει να ισχύει $K \leq C \leq L\text{-}K$, όπου L είναι το συνολικό μέγεθος της βάσης και να ισχύει επιπλέον ότι $0 \leq K \leq 1/2$. Αυτή είναι ξεκάθαρα ένας προπομπός του *k*-anonymity. Ως εκ τούτου διαπιστώνεται ότι ως γενικότερη λογική, η συγκεκριμένη τεχνική, *έχει και θα συνεχίσει να έχει εφαρμογή και στις μοντέρνες βάσεις* δεδομένων με το σκεπτικό ότι μπορεί διαρκώς να αποτελεί ένα πρώτο στοιχειώδες και ταυτόχρονα *ουσιαστικό επίπεδο* ελέγχου του, προς απάντηση, ερωτήματος.

β) Η τεχνική ελέγχου επικάλυψης των ερωτημάτων. Αυτή ελέγχει αν και σε ποιο βαθμό μια απάντηση που ετοιμάζεται να δοθεί στον χρήστη περιλαμβάνει, μεταξύ άλλων, εξ' ολοκλήρου ένα σύνολο οντοτήτων της βάσης που δόθηκε σε προγενέστερο χρόνο ως ενιαία απάντηση σε αυτόν. Τέτοιου είδους επιθέσεις μπορούν να εφαρμοστούν πολύ εύκολα ειδικά χρησιμοποιώντας στη σύνταξη των ερωτημάτων το λογικό τελεστή OR, χωρίς να αποκλείονται «επιτυχείς» επιθέσεις βασισμένες μόνο σε AND τελεστές.

Είναι αυτονόητο ότι με τη ματιά στραμμένη σε σύνθετα ερωτήματα που μπορούν να τεθούν σε μια βάση κινούμενων αντικειμένων, είναι λογικό ότι, *κατά κανόνα,* δεν χρειάζεται ο τελεστής OR (δηλαδή μας ενδιαφέρει π.χ. πόσοι βρέθηκαν το πρωί στο Α μέρος ΚΑΙ το μεσημέρι στο Β μέρος ΚΑΙ το βράδυ ξανά στο Α κ.ο.κ.). Είναι αυτονόητο, βέβαια, ότι η ήδη αποδεδειγμένη, στις απλές σχεσιακές βάσεις, *αυξημένη επικινδυνότητα του OR τελεστή* σε σχέση με τον AND, παραμένει ως μια stand by διαπίστωση στις περιπτώσεις MOD, όσο διευρύνονται οι δυνατότητες των ερωτημάτων που αυτές εκτελούν.

Μια προσέγγιση στην περίπτωση που διαπιστώνεται ότι ένα ερώτημα επικαλύπτει ένα προηγούμενο είναι να ελεγχθεί ο *βαθμός επικάλυψης*. Πιο συγκεκριμένα, ελέγχεται το μέγεθος του λόγου k/r όπου k είναι το πλήθος των στοιχείων που συμβάλλουν στην εκάστοτε σύνθεση της απάντησης του στατιστικού ερωτήματος (όπως ο μέσος όρος ή το άθροισμα) και r το μέγιστο προκαθορισμένο επιτρεπτό πλήθος κοινών στοιχείων από

τα δεδομένα της βάσης που συνθέτουν τις απαντήσεις μεταξύ δύο διαδοχικών ερωτημάτων του ίδιου χρήστη. Για παράδειγμα, αν *k*=4 και *r*=2, αυτό σημαίνει ότι η απάντηση στο κάθε στατιστικό ερώτημα πρέπει να υπολογίζεται από τουλάχιστον 4 αριθμητικά δεδομένα. Έστω, επίσης ότι 2 διαδοχικά ερωτήματα του χρήστη συντίθενται από 4 αριθμητικά δεδομένα το καθένα. Σ' αυτή την περίπτωση, αφού *r*=2, τα κοινά αριθμητικά δεδομένα από τη βάση επί των οποίων υπολογίστηκαν οι 2 απαντήσεις, πρέπει να είναι το πολύ τα 2 από τα 4.

Τέλος οι *Adam & Wortmann* επισημαίνουν ένα σημείο που δείχνει να είναι διαχρονικό, ότι, δηλαδή, αν συνεργαστούν πολλοί κακόβουλοι χρήστες έχοντας ένα κοινό στόχο, ακόμη και αν παρακολουθούνται από το σύστημα ξεχωριστά ο καθένας, είναι ιδιαίτερα δύσκολο να διαπιστωθεί τεκμηριωμένα η επίθεσή τους και να αποτραπεί.

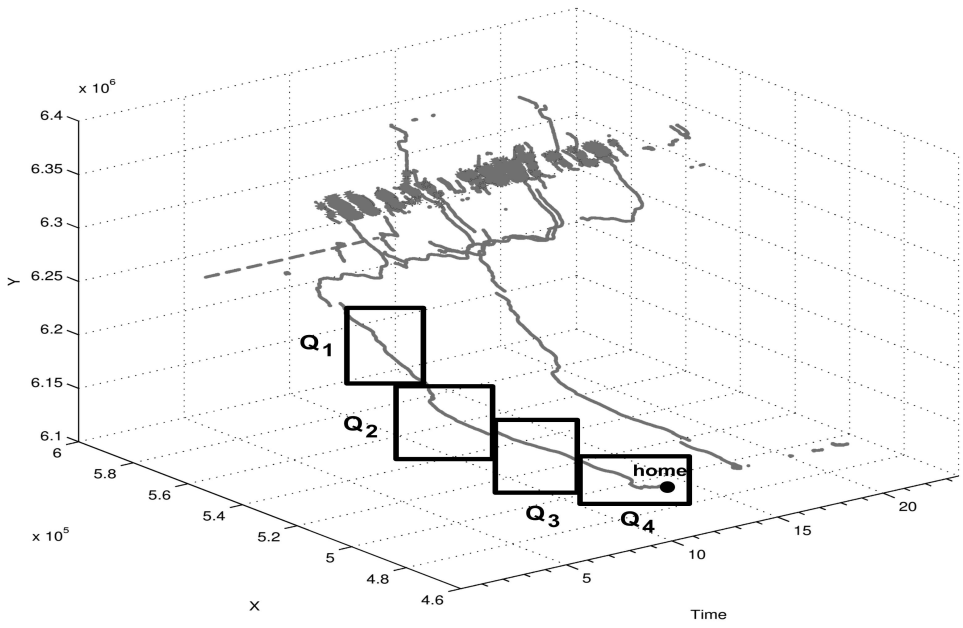
2.4.3 Προστασία της ιδιωτικότητας από ερωτήματα σε βάσεις κινούμενων αντικειμένων

Πέρα από τα LBS's τα οποία αντιμετώπιζαν με ένα ιδιαίτερο τρόπο τη μεταφορά και αποθήκευση των δεδομένων της θέσης των χρηστών και με μόνη ίσως εξαίρεση χρονολογικά την τεχνική της «εσωτερικής» παρακολούθησης ερωτημάτων (μεταξύ πολλών άλλων) σε στατιστικές βάσεις, αυτό που είχε καθιερωθεί ήταν η προσπάθεια δημοσίευσης σε συγκεκριμένο (ή και στο ευρύ) κοινό βάσεων δεδομένων σχεσιακού τύπου (*RDBMS*) και στη συνέχεια βάσεων κινουμένων αντικειμένων (*MOD*), με γνώμονα πάντα το δικαίωμα στην προστασία της ιδιωτικότητας και χρησιμοποιώντας διάφορες μεθοδολογίες.

Παρόλα αυτά, μια διαφορετική οπτική, ώθησε την επιστημονική κοινότητα να εξετάσει μια άλλη λογική πρόσβασης του κοινού σε μια βάση δεδομένων. Αυτή προέβλεπε ότι η βάση θα παρέμενε «φιλοξενούμενη» στην υποδομή του εκάστοτε οργανισμού-φορέα και θα είχαν πρόσβαση σε αυτή ένα πλήθος εξουσιοδοτημένων χρηστών. Με τον τρόπο αυτό υπήρχαν διάφορα πιθανά οφέλη όπως:

α) ο οργανισμός αντιμετώπιζε επιτυχώς διάφορους πιθανούς νομικούς περιορισμούς
β) διατηρούσε μια συνεχή εποπτεία – καταγραφή των οντοτήτων που χρησιμοποιούσαν τη βάση

γ) ανανέωνε τη βάση με νέα δεδομένα τα οποία ήταν άμεσα διαθέσιμα στο κοινό
δ) σε κάθε περίπτωση που διαπιστώνονταν οποιαδήποτε πρόβλημα στην προστασία της ιδιωτικότητας, υπήρχε πάντα η ευχέρεια διορθωτικών παρεμβάσεων.



Εικ. 2.8 Διαδοχική παρακολούθηση - Gkoulalas-Divanis & Verykios [4]

Βάσει αυτού του μοντέλου εργασίας, πρώτοι οι *Gkoulalas-Divanis & Verykios* [4], πρότειναν ένα τρόπο υλοποίησης ενός μηχανισμού προστασίας της ιδιωτικότητας. Έτσι, επιτρέπουν την πρόσβαση σε μια βάση κινούμενων αντικειμένων και τη διενέργεια ερωτημάτων από τους χρήστες με γνώμονα πάντα ότι κάθε απάντηση που δίνεται πρέπει να περιλαμβάνει τουλάχιστον *k* τροχιές βασισμένη στη συλλογιστική του *k*-anonymity. Στην περίπτωση που δεν καλύπτεται ο αριθμός-όριο, ένας μηχανισμός παράγει τις υπολειπόμενες τροχιές μέχρι να γίνουν *k* βασιζόμενος στα δεδομένα των πραγματικών τροχιών που περιλαμβάνονται στην απάντηση. Ο ίδιος μηχανισμός παρακολουθεί τα ερωτήματα που τίθενται ανά χρήστη και ελέγχει τους πιθανούς κινδύνους που δημιουργούνται. Αυτοί αφορούσαν δύο βασικές περιπτώσεις επίθεσης από κάποιο κακόβουλο χρήστη. Η πρώτη ονομάστηκε User identification attack και συντελούνταν όταν τα ερωτήματα επικαλύπτονται χωρικά με αποτέλεσμα οι ψεύτικες τροχιές που παράγονταν κατά περίπτωση να είναι εύκολο να αναγνωριστούν. Η δεύτερη ονομάστηκε Sequential tracking attack και αφορούσε την περίπτωση στην

οποία με συνεχή ερωτήματα που εφάπτονταν χωρο-χρονικά το ένα στο άλλο μπορεί κάποιος να παρακολουθήσει την πραγματική πορεία ενός αντικειμένου, αφού για τις ψεύτικες τροχιές που δημιουργούνταν κατά περίπτωση δεν λαμβάνονταν αρχικά υπ' όψιν οι ήδη δημιουργηθείσες σε γειτονικές χωρο-χρονικά περιοχές.
Η λύση που προτάθηκε για την πρώτη περίπτωση ήταν η άρνηση απάντησης αντίστοιχων ερωτημάτων, ενώ για τη δεύτερη η δημιουργία (και αποθήκευση στη βάση) των ψεύτικων τροχιών, έτσι ώστε να εκτείνονται και πέρα από τα χωρο-χρονικά όρια του εκάστοτε ερωτήματος και έτσι να δίνουν ένα πιο αληθοφανές αποτέλεσμα.
Σε κάθε περίπτωση αυτή ήταν μια πρώτη ιδιαίτερα αξιόλογη, προσπάθεια προς αυτή τη νέα κατεύθυνση «χειρισμού» αντίστοιχων βάσεων δεδομένων και άνοιγε ένα νέο πεδίο προβληματισμού και βελτιωμένων υλοποιήσεων. Χαρακτηριστικό μειονέκτημα ήταν ότι ο αλγόριθμος κατασκευής των ψεύτικων τροχιών δεν λάμβανε υπ' όψη του τη χρονική παράμετρο της κάθε τροχιάς που ενέπλεκε προκειμένου να βασιστεί για την κατασκευή της κάθε καινούριας ψεύτικης τροχιάς. Εξάλλου, η βασική πρόκληση ήταν ότι εισήλθαν σε μια λογική αντιμετώπισης των προβλημάτων ανωνυμοποίησης των δεδομένων – on the fly- γεγονός που υποχρέωνε σε ανάλυση των πιθανών τρόπων επιθέσεων από μια διαφορετική οπτική γωνία.

Οι *Pelekis et al* [15] επέκτειναν την προηγούμενη εργασία. Υλοποίησαν τον Hermes++, ένα μηχανισμό που δημιουργούσε ένα layer πάνω από την αρχική βάση, που επέτρεπε καταρχήν τη διαχείριση τροχιών κινούμενων αντικειμένων, τη διενέργεια διαφόρων τύπων ερωτημάτων στη βάση και τον έλεγχο των απαντήσεων αυτών πριν προωθηθούν οριστικά στον τελικό χρήστη. Σε σχέση με την προηγούμενη εργασία,
α) βελτιώθηκε η ποικιλία στις δυνατότητες ερωτημάτων,
β) ελέγχθηκε ένας επιπλέον κίνδυνος, αυτός της αποκάλυψης μιας «ευαίσθητης» θέσης που επισκέφθηκε ένας χρήστης. Τέτοιες θέσεις είναι η αρχή και το τέλος της κάθε τροχιάς, καθώς και οποιαδήποτε από τις θέσεις καθορίζει ο χρήστης ως ευαίσθητη και
γ) βελτιώθηκε ο τρόπος κατασκευής των ψεύτικων τροχιών.
Η βασικότερη βελτίωση του μηχανισμού κατασκευής ψεύτικων τροχιών (*FTG*) είναι η ενσωμάτωση και του χρόνου, πλέον ως παράμετρο υπολογισμού για την κατασκευή πιο ρεαλιστικών ψεύτικων τροχιών. Έτσι μετά από μια τμηματοποίηση των τροχιών, δημιουργούνται συστάδες των διάφορων υποτροχιών και καθορίζεται για κάθε μια από

αυτές μια υποτιθέμενη αντιπροσωπευτική, της γενικής κίνησης, τροχιά. Με βάση αυτές, κατασκευάζονται οι τυχόν ψεύτικες.

Έτσι ο μηχανισμός, από την αρχή που θα τεθεί ένα ερώτημα,

α) ελέγχει αν επικαλύπτει το παρόν ερώτημα, απαντήσεις προηγούμενων ερωτημάτων. (Σε αυτή την περίπτωση αρνείται να συνεχίσει)

β) Παραλλάσσονται οι τροχιές σε τυχόν τμήματά τους, που περιέχουν «ευαίσθητες» γεωγραφικά θέσεις

γ) υπολογίζονται οι τυχόν επιπλέον τροχιές που απαιτούνται για να καλύπτεται ως προς το *k*-anonymity threshold η απάντηση.

2.4.4 Σημασιολογικά εμπλουτισμένες τροχιές κινούμενων αντικειμένων

Όπως ήδη αναφέρθηκε στην εισαγωγή, το επόμενο βήμα μετά την συλλογή, επεξεργασία και στατιστική (ή άλλου είδους) μελέτη που αφορούσε τροχιές κινούμενων αντικειμένων, ήταν ο σημασιολογικός εμπλουτισμός τους και η (κατά κανόνα) τμηματοποίησή τους σε επεισόδια (*episodes*). Σκοπός είναι να δίνονται πιο άμεσα στον τελικό χρήστη χρήσιμες πληροφορίες σχετικά με τις τροχιές αυτές, ώστε να φύγουμε πλέον από το *τί* διαδρομή ακολουθεί ημερησίως ο παρακολουθούμενος, αλλά με *τί* ασχολείται ημερησίως αυτός. (Δηλαδή ότι εργάζεται για x ώρες, μετακινείται κατά μέσο όρο με το αυτοκίνητο του y χλμ κάθε μέρα κ.λ.π.) Ουσιαστικά μια σημασιολογικά εμπλουτισμένη τροχιά (*semantic trajectory*) είναι μια τροχιά που

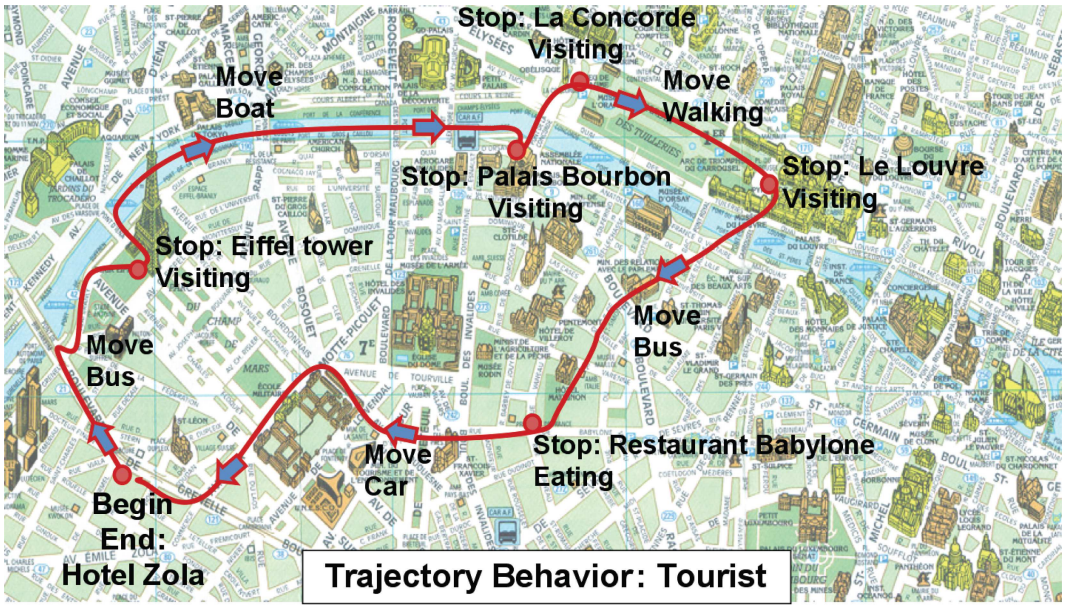
α) έχει ενισχυθεί από ένα πλήθος σχολίων (*annotations*) που την αφορούν είτε συνολικά, είτε μέρος(η) της και/ή

β) έχει υποστεί διάφορες τμηματοποιήσεις (*segmentations*) [14]

Με αυτόν τον τρόπο όμως, οι πιθανοί κίνδυνοι παραβίασης της ιδιωτικότητας αυξάνονται γιατί κάθε καταγεγραμμένη στάση του χρήστη, πλέον μπορεί να περιέχει τυχόν χαρακτηρισμό όπως σπίτι, εργασία, διασκέδαση κ.α. ο οποίος αυτόματα προκαλεί μια κατακόρυφη αύξηση στο επίπεδο «ευαισθησίας» της συγκεκριμένης πληροφορίας.

Τα νέα αυτά χαρακτηριστικά των δεδομένων απαιτούν εύλογα ένα αρκετά διαφορετικό τρόπο αντιμετώπισης σε σχέση με τους προαναφερθέντες. Οι *Monreale et al* [11] πρότειναν μια μεθοδολογία που αντιμετωπίζει σημασιολογικά πλέον το ζήτημα της

ιδιωτικότητας. Ουσιαστικά είναι μια τεχνική που προκαλεί γενίκευση των δεδομένων για να επιτευχθεί το απαιτούμενο επίπεδο ανωνυμοποίησης.



Εικ. 2.9 Μια σημασιολογικά εμπλουτισμένη τροχιά με σχόλια σε κάθε επεισόδιο – C. Parent et al. [14]

Πιο συγκεκριμένα, θεωρεί ότι ο κακόβουλος χρήστης γνωρίζει τη διαδικασία της ανωνυμοποίησης που ακολουθείται, την παρουσία τροχιάς ατόμου που τον ενδιαφέρει στη βάση, την ταξινόμηση της ονοματολογίας που χρησιμοποιείται στη βάση για να χαρακτηρίσει το κάθε μέρος (δηλαδή την κάθε στάση του παρακολουθούμενου) καθώς και ένα συνδυασμό ορισμένου αριθμού στάσεων του παρακολουθούμενου που θα χρησίμευε ενδεχομένως να τον ξεχώριζε κανείς ανάμεσα στις υπόλοιπες καταγεγραμμένες οντότητες. Με βάση αυτό, σκοπός είναι να τροποποιηθεί η βάση έτσι ώστε ο χρήστης ακόμη και αν γνωρίζει μια σειρά τοποθεσιών που επισκέφθηκε ο παρακολουθούμενος να μην είναι σίγουρος με πιθανότητα μεγαλύτερη από *C* (*C-safety*) ποια «ευαίσθητη» τοποθεσία επισκέφθηκε στη συνέχεια. Έτσι γενικεύεται και γίνεται πιο αφαιρετική-γενική η διατύπωση της ευαίσθητης τοποθεσίας του παρακολουθούμενου. Για παράδειγμα αντί για «Νοσοκομείο» ή «Κλινική» ή «Φαρμακείο» εισάγεται ενδεχομένως ο όρος «Υγεία» και έτσι ο κακόβουλος χρήστης δεν μπορεί να είναι σίγουρος αν ο παρακολουθούμενος επισκέφτηκε το νοσοκομείο ή απλά ένα φαρμακείο της γειτονιάς. Δύο είναι τα βασικά μειονεκτήματα της μεθοδολογίας αυτής. Το πρώτο είναι ότι δεν λάμβανε υπ' όψη τη χρονική πληροφορία των τροχιών και το δεύτερο είναι ότι δύο τοποθεσίες, για παράδειγμα, που απέχουν

πολύ χωρικά μεταξύ τους εντάσσονται κάτω από το ίδιο όνομα, επειδή έχουν ίδια ή παρεμφερή χαρακτηριστικά. Ορισμένες επίσης βελτιώσεις μπορούσαν να διευκολύνουν την εν γένει λειτουργικότητα-αποτελεσματικότητα του αλγορίθμου. Για παράδειγμα η αρχική ομαδοποίηση των τροχιών γίνεται μεταξύ τροχιών ίδιου μήκους, γεγονός που ενδεχομένως προκαλεί μεγαλύτερα επίπεδα γενίκευσης της πληροφορίας από τα απολύτως απαραίτητα ώστε να καταστεί, εν τέλει, ασφαλής η βάση.

3 Τύποι επιθέσεων

Αυτό που επιδιώκεται σε αυτή την ενότητα είναι η καταγραφή όλων των διαφορετικών τύπων επιθέσεων που μπορούν να στοιχειοθετηθούν με βάση την πρότερη γνώση του επιτιθέμενου, τον σκοπό που μπορούν να έχουν αυτές αλλά και τους τρόπους με τους οποίους μπορούν να είναι υλοποιήσιμες. Αναλύοντας έτσι καθένα από αυτά τα σημεία θα έχουμε μια συνολική εικόνα των δυνατοτήτων των επιτιθέμενων ώστε να ληφθούν όλα τα απαραίτητα μέτρα για την αντιμετώπισή τους.

3.1 Πρότερη γνώση του κακόβουλου χρήστη

Σκοπός της παρούσας ενότητας είναι η καταγραφή, ανάλυση και ομαδοποίηση όλων των πιθανών συνδυασμών πρότερης γνώσης που μπορεί να έχει ένας κακόβουλος χρήστης και με την οποία θα επιδιώξει τυχόν επίθεση στα περιεχόμενα μιας βάσης. Αυτό που καταγράφουμε είναι οτιδήποτε αποτελεί ένα συνδυασμό γνώσης *που* και γνώσης *πότε* χωρίς να αποκλείεται μια προσπάθεια επίθεσης χρησιμοποιώντας μόνο το ένα από τα δύο είδη γνώσης. Έτσι, ενώ γνωρίζουμε ότι στα raw data, το *πού* σημαίνει μια γεωγραφική περιοχή, στα semantic trajectories προκύπτουν 3 διαφορετικές υποκατηγορίες γνώσης.
Σε ποιο box; ➔ [(Xmin, Ymin), (Xmax, Ymax)]
Τι είδος επεισοδίου; ➔ [STOP | MOVE]
Ποιο είδος tag(s); ➔ [Home| Work| Fun| Visit| Car| (user defined….)| POI's..]
Έτσι στην ερώτηση : Ξέρεις πού ήταν; Όλες οι απαντήσεις του τύπου
-Ναι, στη περιοχή του Συντάγματος
-Ναι, είχε meeting σε ένα νέο πελάτη του
-Ναι, στο Athens Mall
συνιστούν ένα είδος χωρικής πληροφορίας χωρίς να είναι απαραίτητο να γνωρίζουμε πλέον πάντα και τις αντίστοιχες συντεταγμένες.
Κάποιος θα μπορούσε να θεωρήσει ότι σε τεχνικό επίπεδο, δηλαδή ως προς τον τρόπο αποθήκευσης στη ανάλογη βάση, ο χαρακτηρισμός STOP | MOVE είναι ανάλογος των

άλλων tags που εμπλουτίζουν την συνολική εικόνα του κάθε επεισοδίου και για αυτό τον λόγο δεν προκύπτει η ανάγκη για διαφορετική μεταχείριση μεταξύ τους. Παρόλα αυτά, υπάρχουν ποιοτικές και ποσοτικές διαφορές που μας προτρέπουν να τα διαχειριστούμε ξεχωριστά.

Οι λόγοι είναι ότι η πρωταρχική ουσία και η αιτία, κατ' επέκταση, που ένα semantic trajectory διαχωρίστηκε σε ένα συγκεκριμένο αριθμό επεισοδίων είναι η αναγνώριση, μέσα στα raw data, τμημάτων χωροχρονικών συντεταγμένων που βάσει των χαρακτηριστικών που απαιτούνται από τους αντίστοιχους αλγορίθμους κάθε φορά, πληρούν τις προϋποθέσεις έτσι ώστε να χαρακτηριστούν επεισόδια στάσης (STOP) ή κίνησης (MOVE). Η ύπαρξη λοιπόν του annotation STOP | MOVE, είναι δεδομένη σε κάθε περίπτωση όταν κάποιος κατασκευάζει semantic trajectories σε αντίθεση με την ύπαρξη διαφόρων άλλων tags που, ανά επεισόδιο, μπορούμε να την θεωρήσουμε προαιρετική. Εξ' άλλου, σε κάθε περίπτωση υπάρχει η δυνατότητα με τη χρησιμοποίηση κατάλληλων εργαλείων/αλγορίθμων να μπορεί να εξαχθεί η πληροφορία τί είδους επεισόδιο είναι κάθε φορά, ενώ δεν μπορούμε να έχουμε καμία ιδέα ποια θα μπορούσαν να είναι τα υπόλοιπα tags.

Από ποσοτικής άποψης, έχουμε να παρατηρήσουμε ότι όσο μικρή ή μεγάλη βάση έχουμε στα χέρια μας κάθε φορά, ο αριθμός του είδους των επεισοδίων θα παραμένει 2, ενώ ο αριθμός του πλήθους των διακριτών τιμών των υπόλοιπων tags που εμπεριέχονται κάθε φορά, θα αυξάνει -πιθανά- συνεχώς.

Αυτό που γνωρίζει λοιπόν, ο κακόβουλος χρήστης είναι έναν ή περισσότερους από τους πιθανούς συνδυασμούς που προκύπτουν από τη γνώση των 4 αυτών ειδών πληροφορίας ήτοι το χρόνο, το χώρο, το είδος επεισοδίου και/ή το tag.

Μεταξύ των 4 αυτών ειδών πληροφορίας, τα tags είναι αυτά που αποτελούν την ουσιαστική καινοτομία στις σημασιολογικά εμπλουτισμένες τροχιές. Κατ' επέκταση δεν είναι βέβαιο ότι έχει παγιωθεί ακόμη ένας καθιερωμένος τρόπος χρησιμοποίησής τους γενικότερα, με αποτέλεσμα να δημιουργείται πολλές φορές μια σύγχυση σχετικά με το τί σημαίνουν αυτά (ως πληροφορία για τον χρήστη) στην εκάστοτε βάση. Έτσι, κρίνουμε σκόπιμο να οριοθετήσουμε καλύτερα τον τρόπο λειτουργίας της βάσης ως προς τα tags, ώστε τα συμπεράσματα που τυχόν θα εξάγουμε να είναι όσο γίνεται πιο συγκεκριμένα και ασφαλή.

Θέτουμε, λοιπόν, μια σύγκριση του πλήθους των tags στο σύνολο μιας βάσης, με το πλήθος των δημιουργημένων επεισοδίων της βάσης. «Πολλά» tags στη βάση σημαίνει ότι αυξάνει η πιθανότητα για κάθε ένα ξεχωριστά, να αντιστοιχεί *προφανέστατα* σε ένα χώρο *αρκετά ως πάρα πολύ* συγκεκριμένο χωρικά. Επειδή είναι αυτονόητο ότι θεμελιώδης στόχος των tags είναι να προσδώσουν μια πιθανή σημασία, ένα πιθανό νόημα σε κάθε ένα επεισόδιο και όχι να υποκαθιστούν την αμιγώς χωρική πληροφορία, προχωρούμε στις εξής 2 παραδοχές.

α) Όταν το κάθε χρησιμοποιούμενο tag σε μια βάση δεδομένων σχετίζεται/αφορά μια γενική κατά κοινή ομολογία έννοια όπως Home, Car, Work, Fun, Rest, Visitor, Patient κ.α., τότε θα *πρέπει* να έχει μια *επαρκή αντιπροσώπευση* επί του συνολικού πλήθους των δεδομένων της βάσης αυτής που αφορούν tags. Με άλλα λόγια, κάθε τέτοιου είδους tag καλό είναι να χρησιμοποιείται από τουλάχιστον ένα συγκεκριμένο αριθμό φορών στο σύνολο των επεισοδίων.

Για παράδειγμα, ένας τρόπος που θα μπορούσε, ενδεικτικά, να καθορίσει το κατώφλι μιας επαρκούς αντιπροσώπευσης, θα ήταν ο εξής. Έστω:

k: το anonymity threshold που ισχύει τη δεδομένη χρονική στιγμή στο μηχανισμό της βάσης

m: ο μέσος όρος του πλήθους των tags ανά επεισόδιο

Te: ο αριθμός των συνολικών επεισοδίων στη βάση

Ne: ο ελάχιστος αριθμός των επεισοδίων στα οποία συμμετέχει το κάθε προαναφερθέν tag.

Έτσι η εξίσωση διαμορφώνεται ως εξής:

$$\boldsymbol{Ne = k * m * Te * 10^{-3}}$$

Για παράδειγμα, σε βάση που έχουν καταγραφεί 10.000 άνθρωποι για ένα χρονικό διάστημα και που έχουν προκύψει 9 επεισόδια ανά τροχιά κατά Μ.Ο., έστω ότι *k*=5 και *m*=2. Κάθε tag λοιπόν, θα πρέπει να έχει χρησιμοποιηθεί τουλάχιστον *Ne* φορές όπου *Ne* = 5 * 2 * (9 * 10.000) * 0,001 = 900. Παρατηρούμε λοιπόν ότι επιθυμούμε από μια βάση με 90.000 επεισόδια και 2 tags ανά επεισόδιο, το κάθε tag να έχει παρουσία τουλάχιστον 900 φορές.

β) Όταν το κάθε χρησιμοποιούμενο tag σε μια βάση δεδομένων σχετίζεται/αφορά ένα POI (point of interest) ή μία ROI (region of interest), τότε θα *πρέπει* να μπορεί, εν δυνάμει, *να αντιστοιχεί σε τουλάχιστον άλλα Np POI's ή ROI's* μέσα από τη συνολική γεωγραφική περιοχή στην οποία εκτείνεται η καταγραφή των τροχιών, ώστε να είναι αποδεκτή η χρησιμοποίηση του, εν λόγω, tag.

Ένας απλός τρόπος ώστε να καθοριστεί η τιμή του *Np*, είναι να *ισούται* με το *k-anonymity*.

Είναι λοιπόν αντιληπτό ότι με αυτό τον τρόπο, πετυχαίνουμε να «επιβάλλουμε» μια πιο αφαιρετική χρήση των tags και να δημιουργήσουμε μία *αρχική προληπτική προστασία* απέναντι σε κακόβουλες επιθέσεις. Πέραν αυτού όμως, η αφαιρετική αυτή λογική που επιδιώκουμε να πετύχουμε έχει επίσης ως σκοπό *την αύξηση της ποιότητας της παρεχόμενης υπηρεσίας*. Αυτό συμβαίνει γιατί στην αντίθετη περίπτωση, αν υπάρχει πολυκερματισμός της ίδιας έννοιας, τότε δεν αντανακλάται σωστά στα στατιστικά ερωτήματα, το πραγματικό πλήθος των επεισοδίων (tags home, myhome, residence κ.ο.κ). Επιπλέον αποφεύγοντας τον πολυκερματισμό αυτό, θα είναι λιγότερες, κατ' αναλογία, οι φορές που δεν θα μπορεί να απαντηθεί ένα ερώτημα επειδή το πλήθος των τροχιών που επιστρέφει είναι μικρότερο του *k*-anonymity.

Θεωρώντας έτσι ως δεδομένο ότι η εκάστοτε βάση στο περιβάλλον μελέτης της παρούσας εργασίας λειτουργεί με αυτές τις προϋποθέσεις, μπορούμε να διενεργήσουμε μια επισκόπηση πάνω στους πιθανούς συνδυασμούς των 4 ειδών γνώσης.

Οι πιθανοί συνδυασμοί μεταξύ των ειδών γνώσης, που προκύπτουν, είναι 15. Ουσιαστικά διαπιστώνουμε ότι σχηματίζονται 3 κατηγορίες. Η $1^{η}$ κατηγορία δεν περιλαμβάνει, σε κάθε περίπτωση, αμιγώς χωρική πληροφορία, η $2^{η}$ κατηγορία δεν περιλαμβάνει γνώση χρονικού διαστήματος και η $3^{η}$ κατηγορία είναι αυτή που περιλαμβάνει και τα δύο προηγούμενα σε όλους τους πιθανούς συνδυασμούς τους.

Όταν ένας (κακόβουλος ή μη) χρήστης της βάσης γνωρίζει στοιχεία για περισσότερα του ενός επεισόδια, τότε υπάρχουν 2 περιπτώσεις. Είτε ο χρήστης γνωρίζει ταυτόχρονα πληροφορίες για, περισσότερα του ενός, επεισόδια στα πλαίσια μιας τροχιάς και γνωρίζει τη χρονική αλληλουχία μεταξύ τους, είτε όχι.

Έτσι λοιπόν, καθώς οριοθετείται το είδος της πρότερης γνώσης που μπορεί να κατέχει ένας κακόβουλος χρήστης της βάσης, διαπιστώνουμε εύλογα ότι οι πιο «επικίνδυνες» επιθέσεις που μπορούν να διενεργηθούν, είναι αυτές που γίνονται βασισμένες στην

κατοχή της πλέον εξειδικευμένης/συγκεκριμένης πληροφορίας/γνώσης. Η «επικινδυνότητα» αυξάνεται όσο μικραίνει το ποσοστό, επί του συνόλου, των επεισοδίων της βάσης που περιέχουν τη συγκεκριμένη πληροφορία. Με βάση αυτό, η κατοχή πληροφοριών που περιέχουν χωροχρονική γνώση (με ή χωρίς tags) γύρω από ένα ή περισσότερα επεισόδια μιας τροχιάς, καθίσταται το πλέον επικίνδυνο εργαλείο στα χέρια ενός κακόβουλου χρήστη και εκεί είναι που θα εστιάσουμε την προσοχή μας, ώστε να αποκαλυφθούν όλες οι δυνατές επιθέσεις που μπορεί αυτός να διενεργήσει στη βάση.

3.2 Σκοπός της επίθεσης

Ο βασικός σκοπός κάθε διενεργούμενης επίθεσης είναι *η διεύρυνση της γνώσης του επιτιθέμενου σχετικά με ένα άτομο ή την κατάσταση που τον ενδιαφέρει, για ιδιοτελείς σκοπούς. Η διεύρυνση της γνώσης συντελείται όταν ο επιτιθέμενος αυξάνει την πεποίθηση του σχετικά με ένα γεγονός*.

Το γεγονός αυτό μπορεί να αφορά
α) ένα άτομο που αποτελεί τον «στόχο» για τον επιτιθέμενο ή
β) μια κατάσταση για την οποία επιθυμεί ο επιτιθέμενος να αποκτήσει πιο συγκεκριμένη γνώση γι' αυτή.

Η αύξηση της πεποίθησης συντελείται όταν ο επιτιθέμενος γνωρίζει *μετά* την επίθεση
α) πιο αναλυτικά στοιχεία ενός επεισοδίου το οποίο, γνώριζε εκ των προτέρων, ότι αφορούσε το άτομο – «στόχο» ή
β) στοιχεία άλλων επεισοδίων πέραν του ή των επεισοδίων τα οποία γνώριζε εκ των προτέρων ο επιτιθέμενος για το άτομο – «στόχο», στα πλαίσια της ίδιας τροχιάς ή
γ) περισσότερα στοιχεία για μια κατάσταση από αυτά που γνώριζε εκ των προτέρων.

Στοιχεία για το άτομο – «στόχο», όπως διατυπώσαμε στην ενότητα 3.1, μπορούν να είναι ο *χώρος*, ο *χρόνος*, το *είδος του επεισοδίου* ή/και η ύπαρξη συγκεκριμένων *tags*. Ενδεικτικά για την περίπτωση γ, *στοιχείο* μπορεί να είναι ότι σε ένα συγκεκριμένο χωροχρονικό πλαίσιο, ο κακόβουλος χρήστης εξήγαγε το συμπέρασμα ότι κανένα

επεισόδιο δεν έλαβε χώρα ή έλαβαν χώρα συγκεκριμένα επεισόδια/τροχιές, λιγότερα από το όριο του *k*-anonymity.

Το τελικό όριο της *διεύρυνσης της γνώσης* είναι η *βεβαιότητα* που αποτελεί μια ουσιαστικά εκφυλισμένη μορφή της *αύξησης της πεποίθησης*.

3.3 Τρόποι επιθέσεων

3.3.1 Ολικώς επικαλυπτόμενα ερωτήματα

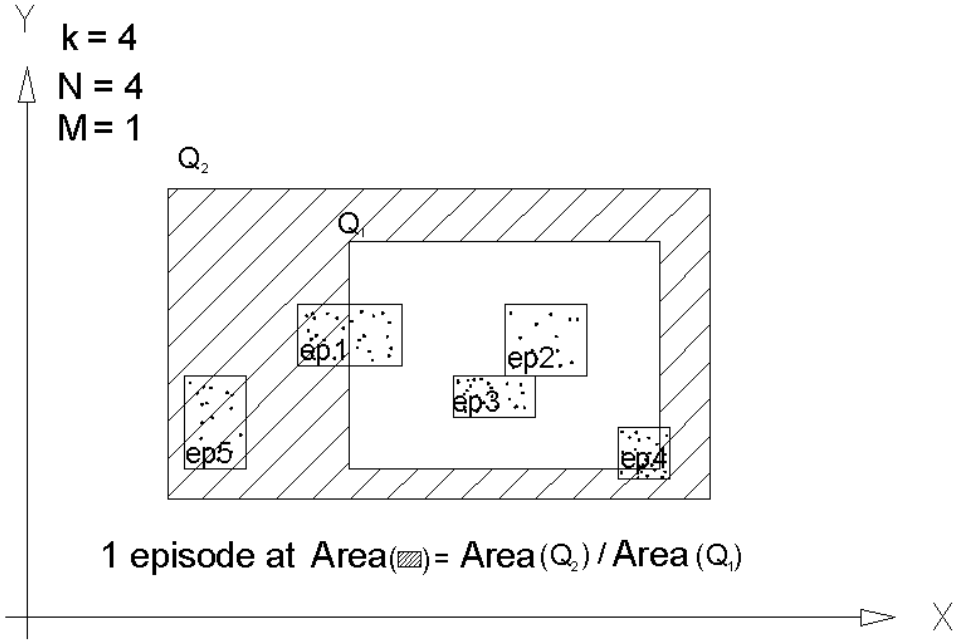
Αρχικά θα αναλύσουμε τις τεχνικές που μπορούν να ακολουθηθούν από την πλευρά του κακόβουλου χρήστη όσον αφορά μια μεγάλη, γενική κατηγορία που ονομάζεται Overlapping Queries. Αυτή η περίπτωση έχει απασχολήσει στο παρελθόν την επιστημονική κοινότητα [4],[15] όμως αφορούσε μη σημασιολογικά επαυξημένες τροχιές. Τα Overlapping Queries είναι μια ακολουθία τουλάχιστον 2 ερωτημάτων που θέτει ο χρήστης με χαρακτηριστικό γνώρισμα ότι υπάρχει επικάλυψη ανάμεσα στα κριτήρια των διαδοχικών αυτών ερωτημάτων. Για να εξάγονται, όμως πιο βέβαια συμπεράσματα κάθε φορά, η λογική υπαγορεύει να ισχύει κάθε φορά *είτε* ότι τα κριτήρια των 2 ερωτημάτων θα διαφέρουν μόνο ως προς ένα είδος (χώρος/χρόνος/tags), *είτε* ότι θα διαφέρουν αριθμητικά κατά ένα ή περισσότερα ολόκληρα υπο-ερωτήματα (διατηρώντας εντελώς κοινά τα, μεταξύ τους, υπόλοιπα υπο-ερωτήματα).

Πρέπει να διευκρινίσουμε ότι όταν μιλούμε για *ερώτημα* εννοούμε αυτό που με λεπτομέρεια ορίζουμε στο κεφάλαιο 1 (*Εισαγωγή*). Ουσιαστικά δηλαδή για ερωτήματα που έχουν το καθένα τους, ένα ή περισσότερα υπο-ερωτήματα. Για λόγους απλότητας όμως, επεξηγούμε πολλές φορές το σκεπτικό μας χρησιμοποιώντας για παράδειγμα, ερωτήματα που έχουν ένα μόνο υπο-ερώτημα (εκφυλισμένη μορφή ερωτήματος).

Έτσι στη συνέχεια αναλύεται κάθε μια διακριτή περίπτωση της γενικής αυτής κατηγορίας.

Χωρικά ολικώς επικαλυπτόμενα ερωτήματα Έστω ότι σε μια βάση υπάρχει ένα προκαθορισμένο όριο ασφαλείας βάσει της αρχής του *k*-anonymity, ώστε να μην απαντώνται ερωτήματα όταν το πλήθος των τροχιών είναι μικρότερο του *k*. Σκοπός του

επιτιθέμενου αυτή τη φορά είναι να αποκτήσει περισσότερη γνώση πάνω σε μια *κατάσταση*, όπως αυτή έχει διατυπωθεί στην ενότητα 3.2 (περίπτωση γ). Στη περίπτωση αυτή, για να επιτευχθεί αυτό, ο επιτιθέμενος διενεργεί μια σειρά (τουλάχιστον 2) ερωτημάτων τα οποία μεταξύ τους είναι απολύτως όμοια εκτός από ένα *είδος γνώσης* κάθε φορά (βλ. 3.1). Αν μεταβάλλεται η χωρική διάσταση του ερωτήματος, μιλάμε για *spatial overlapping query*. Σε αυτή τη περίπτωση, ο επιτιθέμενος κάνει το πρώτο ερώτημα και εφόσον το πλήθος των τροχιών είναι τουλάχιστον *k*, συνεχίζει με ένα ή περισσότερα ερωτήματα αλλάζοντας κάθε φορά τη χωρική διάσταση με τέτοιο τρόπο ώστε κάθε φορά το επόμενο box να *περιέχει πλήρως* το αμέσως προηγούμενο.

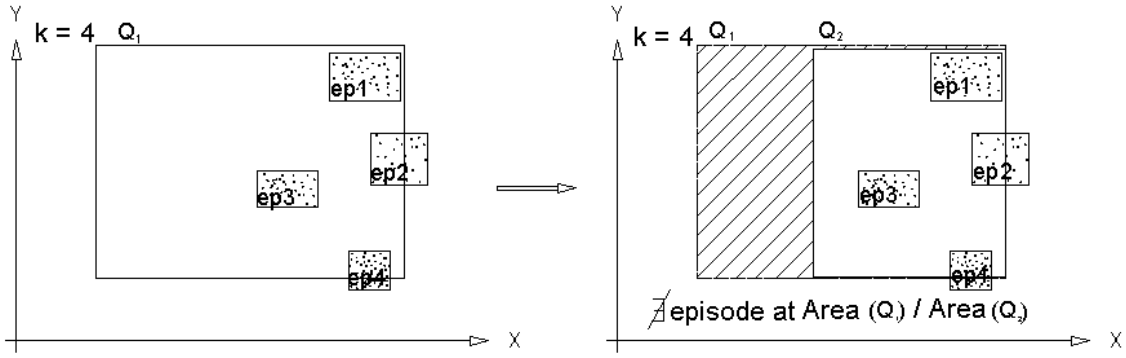


Σχήμα 3.1 – Χωρικώς ολικά επικαλυπτόμενα ερωτήματα (1)

Έτσι, αν υποθέσουμε ότι το πρώτο ερώτημα περιέχει N τροχιές, με N>=k, και το επόμενο ερώτημα επιστρέψει N+M, όπου M<k, τότε «δικαιούται» να εξάγει το συμπέρασμα ότι η χωρική περιοχή που αφορά το $2^{ο}$ ερώτημα και δεν αποτελεί μέρος της χωρικής περιοχής του $1^{ου}$ ερωτήματος, περιέχει M τροχιές οι οποίες όμως είναι λιγότερες από k και έχουμε παραβίαση της ιδιωτικότητας σύμφωνα με ότι έχουμε ορίσει στην ενότητα 3.2 (περίπτωση γ). Ο επιτιθέμενος μπορεί να παραμείνει σε αυτή τη διαδικασία όσο συνεχίζει να εξάγει χρήσιμα συμπεράσματα.

Αυτό που πρακτικά συμβαίνει είναι ότι με αυτή τη διαδικασία ένας κακόβουλος χρήστης μπορεί να οριοθετεί σε πολύ πιο συγκεκριμένες χωρικές περιοχές από το επιθυμητό, την ύπαρξη επεισοδίων. Όταν δε, από το προηγούμενο ερώτημα στο επόμενο, δεν μεταβάλλεται το πλήθος των τροχιών, ακόμη και αυτό (δηλ. η ανυπαρξία επεισοδίων στην «επιπλέον» περιοχή) αποτελεί προσθήκη στις γνώσεις του επιτιθέμενου για τη συγκεκριμένη περιοχή.

Στα πλαίσια όμως, αυτού του τύπου των επιθέσεων, υπάρχει η δυνατότητα για τον επιτιθέμενο να μεταβαίνει από το «γενικότερο» στο «ειδικότερο» ερώτημα, δηλαδή να εκτελεί την κάθε επόμενη φορά το ίδιο ερώτημα με τη χωρική διάστασή του, να *περιέχεται*, αυτή τη φορά, στο box του αμέσως προηγούμενου ερωτήματος.



Σχήμα 3.2 - Χωρικώς ολικά επικαλυπτόμενα ερωτήματα (2)

Έτσι στο σχήμα 3.2, παρατηρούμε ότι μετά από την εκτέλεση και του 2^{ου} ερωτήματος, συμπεραίνεται ότι η γραμμοσκιασμένη περιοχή δεν περιλαμβάνει *κανένα* επεισόδιο, που αποτελεί γνώση που δεν θα έπρεπε να λάβει ο κακόβουλος χρήστης.

Τα παραπάνω παραδείγματα που αναλύσαμε, αφορούν ερωτήματα που αποτελούνται από ένα μόνο υπο-ερώτημα, δηλαδή δεν είναι η κλασσική περίπτωση ερωτημάτων. Στην περίπτωση που εκτελείται ένα ερώτημα με τουλάχιστον 2 υπο-ερωτήματα, για να προκληθεί «επιτυχής» επίθεση με επίκεντρο το χώρο από τη πλευρά του κακόβουλου χρήστη, πρέπει να υφίστανται μια από τις 2 περιπτώσεις.

α) Κάθε επόμενο ερώτημα του χρήστη να έχει ακριβώς τον ίδιο αριθμό υπο-ερωτημάτων και κάθε MBB του αντίστοιχου χωρικού κριτηρίου του κάθε υπο-

ερωτήματός του, να περιέχει εντός των ορίων του, το αντίστοιχο MBB του 1ου ερωτήματος για το ανάλογο υπο-ερώτημα.

β) Κάθε επόμενο ερώτημα του χρήστη να έχει ακριβώς τον ίδιο αριθμό υπο-ερωτημάτων και κάθε MBB του αντίστοιχου χωρικού κριτηρίου του κάθε υπο-ερωτήματός του, να περιέχεται εντός των ορίων του αντίστοιχου MBB του 1ου ερωτήματος για το ανάλογο υπο-ερώτημα.

Είναι εύλογο ότι με την έννοια «περιέχει» ή «περιέχεται», εννοούμε ότι μπορεί ορισμένα MBB του 2ου ερωτήματος να είναι απολύτως ίσα με τα αντίστοιχα του 1ου, αλλά πρέπει τουλάχιστον ένα «ζεύγος» από MBB's να διαφέρει.

Χρονικά ολικώς επικαλυπτόμενα ερωτήματα Μία παραλλαγή στον προηγούμενο τρόπο επίθεσης, είναι η αντικατάσταση, συνολικά στο σκεπτικό, του χώρου με τον χρόνο. Έτσι μεταξύ των διαδοχικών ερωτημάτων θα διατηρούνται σταθερές όλες οι παράμετροι του ερωτήματος εκτός από το χρονικό διάστημα αυτή τη φορά. Έτσι κάθε «επιπλέον» τροχιά που κάνει πρώτη φορά την εμφάνισή της, ο επιτιθέμενος την εντάσσει χρονικά σε πολύ πιο συγκεκριμένα πλαίσια από το επιθυμητό όριο ασφαλείας που ορίζεται από το *k*-anonymity. Επίσης, αν δεν διαφέρει το πλήθος των τροχιών είτε μεγαλώνει το χρονικό διάστημα στο 2ο ερώτημα, είτε μικραίνει, ο κακόβουλος χρήστης αποκτά γνώση πέρα από την επιτρεπτή.

Ο τρόπος που ενεργεί ένας κακόβουλος χρήστης χρησιμοποιώντας ερωτήματα με πολλά υπο-ερωτήματα αυτή τη φορά, είναι ίδιος όπως τον αναλύσαμε για το χωρικό Overlapping Query.

Ολικώς επικαλυπτόμενα ερωτήματα ως προς τα tags Η έννοια των αλληλεπικαλυπτόμενων ερωτημάτων μπορεί να επεκταθεί και στο επίπεδο των tags. Έτσι ενδέχεται ένας κακόβουλος χρήστης να επιχειρήσει μια σειρά ερωτημάτων που διαφέρουν μεταξύ τους μόνο ως προς τα tags.

Πρέπει να διευκρινιστεί ότι παρόλο που τα tags, από τη μια, και το είδος επεισοδίου (Stop|Move), από την άλλη, αποτελούν στην δική μας γενική προσέγγιση, δύο διαφορετικά είδη πληροφορίας, στη συγκεκριμένη ενότητα θα τα αντιμετωπίσουμε ενιαία, επειδή δημιουργούν το ίδιο, ποιοτικά, πρόβλημα στην ασφάλεια της βάσης.

Πιο συγκεκριμένα, ας φανταστούμε την περίπτωση στην οποία σε βάση όπου ισχύει ένα *k*-anonymity = 5, τεθεί ένα ερώτημα για συγκεκριμένο χώρο και χρόνο και επιστραφεί -έστω- πλήθος 6 τροχιών. Αν στη συνέχεια τεθεί ένα δεύτερο ερώτημα που περιλαμβάνει επακριβώς τα κριτήρια του πρώτου ερωτήματος και επιπλέον περιλαμβάνει το κριτήριο να πρόκειται για Move επεισόδιο και έστω ότι έτσι επιστραφούν 5 τροχιές, τότε αυτόματα εξάγεται το συμπέρασμα ότι ένας -και μόνο ένας- καταγραφόταν ως σταματημένος στη συγκεκριμένη περιοχή.

Ένα αντίστοιχο παράδειγμα είναι το εξής. Σε μια συγκεκριμένη χωροχρονική περιοχή καταγράφονται 7 Stop επεισόδια. Αν στη συνέχεια προσθέσει ο χρήστης στο ερώτημά του το κριτήριο να έχουν οι στάσεις αυτές χαρακτηριστεί με tag = 'work' και επιστρέψει το ερώτημα 5 τροχιές, τότε είναι αντιληπτό ότι 2 παρακολουθούμενοι, στην «ευαίσθητη» -έστω- περιοχή αυτή δεν ήταν εργαζόμενοι αλλά επισκέπτες.

Μια ελαφρά παραλλαγή του προηγούμενου παραδείγματος, μπορεί να υλοποιηθεί ως εξής. Έστω ότι γνωρίζει κάποιος ότι ένας παρακολουθούμενος βρέθηκε σε συγκεκριμένη περιοχή και είχε tag= 'fun', ο οποίος ισχυρίστηκε ότι συνοδευόταν από άλλο ένα παρακολουθούμενο. Αν το πρώτο ερώτημα (χωροχρόνος + Stop episode) φέρει 10 τροχιές και το δεύτερο (χωροχρόνος + Stop episode + tag = 'work') φέρει 9, τότε ο παρακολουθούμενος ψεύδεται.

Ολικώς επικαλυπτόμενα ερωτήματα ως προς το πλήθος των υπο-ερωτημάτων

Η τελευταία κατηγορία των overlapping queries αφορά την περίπτωση στην οποία ένας κακόβουλος χρήστης γνωρίζει στοιχεία ενός ή περισσοτέρων επεισοδίων μιας τροχιάς και επιδιώκει –για άλλη μια φορά- να αυξήσει τη γνώση του πάνω σε μια κατάσταση. Στις 3 προηγούμενες περιπτώσεις ο χρήστης εκτελούσε μια σειρά ερωτημάτων τα οποία μεταξύ τους διέφεραν κάθε φορά αποκλειστικά είτε ως προς τη χωρική ή τη χρονική διάσταση, είτε το σημασιολογικό περιεχόμενό τους. Τα ερωτήματα αυτά μπορούν να περιέχουν ένα ή περισσότερα υπο-ερωτήματα.

Στη προκειμένη περίπτωση ο χρήστης εκτελεί σειρά από ερωτήματα τα οποία διαφέρουν μεταξύ τους ως προς το πλήθος των υπο-ερωτημάτων που το καθένα περιέχει. Είναι αυτονόητο ότι μεταξύ των 2 διαδοχικών ερωτημάτων, αν το πρώτο περιέχει λ υπο-ερωτήματα, το δεύτερο περιέχει $\lambda+\mu$, όπου $\mu>=1$. Όσο, βέβαια, μεγαλώνει το μ τόσο μειώνονται οι πιθανότητες να λάβει τέτοιες απαντήσεις από τα 2

αλλεπάλληλα ερωτήματα που να τον «βολεύουν» ώστε να εξάγει «χρήσιμα» συμπεράσματα. Βασικό προαπαιτούμενο για να επιτευχθεί επίθεση από το χρήστη, είναι να γνωρίζει ότι ο παρακολουθούμενος για τον οποίο θέλει να διευρύνει τις γνώσεις του, συμμετέχει σίγουρα στα λ υπο-ερωτήματα με αντίστοιχο πλήθος επεισοδίων και εξετάζεται η συμμετοχή του στα πλαίσια του επιπλέον υπο-ερωτήματος. Ας καταλάβουμε καλύτερα το σκεπτικό με το εξής απλό παράδειγμα: Έστω ότι καθημερινά καταγράφονται τροχιές για να αποτυπώσουν διάφορες συμπεριφορές και συνήθειες στη συγκεκριμένη περιοχή. Έστω ότι ένας χρήστης γνωρίζει πού διαμένει μόνιμα ο κινούμενος «στόχος» και πού εργάζεται. Έστω ότι επιθυμεί να μάθει αν ένα συγκεκριμένο βράδυ κοιμήθηκε σπίτι του ή όχι. Θεωρούμε ότι έχει προφανώς πρόσβαση στη βάση η οποία «προστατεύεται» από ένα *k*-anonymity=4. Θέτει το πρώτο ερώτημα με υπο-ερωτήματα που σίγουρα «συμμετείχε» ο καταγραφόμενος (δηλ. ποιοι διέμειναν το βράδυ στην χ περιοχή και το μεσημέρι βρέθηκαν να εργάζονται στην ψ) και το πλήθος των απαντήσεων ήταν 16. Θέτει το δεύτερο ερώτημα που περιλαμβάνει επακριβώς τα δύο πρώτα υπο-ερωτήματα ενώ προσθέτει και ένα τρίτο με το οποίο αναζητά ποιοι βρέθηκαν ξανά τις απογευματινές/βραδινές ώρες στην χ περιοχή. Αν η απάντηση αυτή τη φορά είναι –έστω- 15, τότε ο χρήστης μπορεί να εξάγει το συμπέρασμα ότι ο παρακολουθούμενος κατά 94% κοιμήθηκε στην περιοχή που βρίσκεται το σπίτι του (χωρίς να μπορεί να αποκλειστεί το γεγονός να επέστρεψε στην περιοχή που διαμένει αλλά να μην κοιμήθηκε σπίτι του ακριβώς). Εννοείται ότι αν το αποτέλεσμα ήταν 16 ή <*k*, τότε θα είχε τη *βεβαιότητα ότι κοιμήθηκε σίγουρα* ή *ότι δεν κοιμήθηκε με πιθανότητα 75%* αντίστοιχα. Έτσι, αντιλαμβανόμαστε εύλογα ότι αν χρήστης εκτελούσε από την αρχή ευθέως το δεύτερο ερώτημα, το αποτέλεσμα θα ήταν το ίδιο, όμως δεν θα αύξανε τη βεβαιότητά του ως προς τη συμπεριφορά του παρακολουθούμενου.

Ένα δεύτερο παράδειγμα που προσεγγίζει κάπως διαφορετικά τον τρόπο επίθεσης είναι το εξής (*k*-anonymity=5). Ένας χρήστης γνωρίζει ότι ο παρακολουθούμενος βρισκόταν ένα συγκεκριμένο πρωινό στο σπίτι του (γνωστή τοποθεσία), επέστρεψε σε αυτό το βράδυ και εξετάζει αν πήγε ή όχι στην εργασία του την ημέρα αυτή, θεωρώντας ότι είναι γνωστός ο τόπος εργασίας του. Σε αυτή τη περίπτωση ο χρήστης διενεργεί ένα πρώτο ερώτημα με δύο υπο-ερωτήματα. Ποιοι ήταν στη περιοχή χ κατά τις πρώτες πρωινές ώρες και στην ίδια περιοχή τις βραδινές ώρες. Έστω ότι η απάντηση είναι 10.

Στη συνέχεια προσπαθεί να δημιουργήσει ερωτήματα με ένα επιπλέον υπο-ερώτημα κάθε φορά και κοινά τα 2 του πρώτου ερωτήματος.

Δηλαδή, έστω ότι ρωτάει:

Ποιοι ήταν στη περιοχή χ κατά τις πρώτες πρωινές ώρες, στην ίδια περιοχή τις βραδινές ώρες και τις μεσημβρινές ώρες *οπουδήποτε με είδος επεισοδίου = Move*;

Έστω ότι η απάντηση είναι 5. Προφανώς σε αυτούς δεν περιλαμβάνεται ο παρακολουθούμενος γιατί θα ήταν σε Stop επεισόδιο εκείνο το χρονικό διάστημα.

Στη συνέχεια έστω ότι ρωτάει:

Ποιοι ήταν στη περιοχή χ κατά τις πρώτες πρωινές ώρες, στην ίδια περιοχή τις βραδινές ώρες και τις μεσημβρινές ώρες *οπουδήποτε με είδος επεισοδίου = Stop και tag= 'Fun'*;

Έστω ότι η απάντηση είναι 5. Προφανώς από τη μια δεν είναι οι ίδιοι με τους 5 του προηγούμενου ερωτήματος και από την άλλη σε αυτούς δεν περιλαμβάνεται πάλι ο παρακολουθούμενος γιατί αν ήταν σε Stop επεισόδιο εκείνο το χρονικό διάστημα θα είχε (μάλλον) tag= 'Work'.

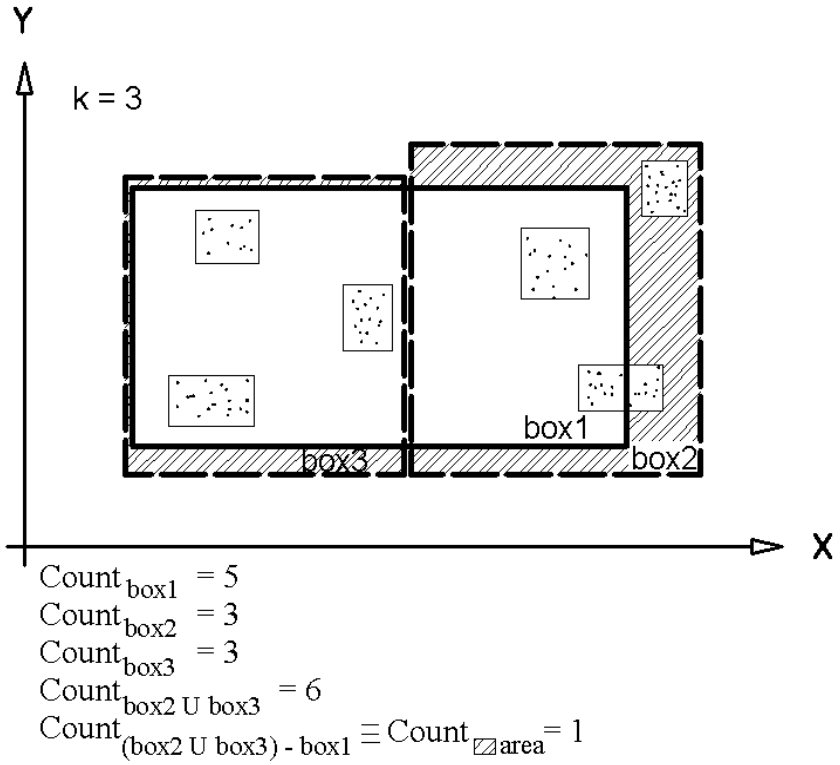
Όμως το άθροισμα των απαντήσεων είναι 10, γεγονός που μας οδηγεί αυτόματα στο συμπέρασμα ότι ο παρακολουθούμενος δεν πήγε στην εργασία του εκείνη την ημέρα.

Παρατηρούμε ότι με αυτή τη τεχνική ο χρήστης προσπαθεί να ελέγξει το NOT του υπο-ερωτήματος που τον ενδιαφέρει σε συνδυασμό πάντα με τα στοιχεία που γνωρίζει γιατί με αυτό τον τρόπο μπορεί να ξεπεράσει το «σκόπελο» του k-anonymity. Δηλαδή ερευνά πόσες (τουλάχιστον) 5αδες θα μπορούσαν να βρεθούν να κάνουν *οτιδήποτε άλλο πέρα* από το να βρίσκονται στον χώρο εργασίας του παρακολουθούμενου το συγκεκριμένο χρονικό διάστημα με τα ανάλογα tags.

3.3.2 Πολλαπλώς τεμνόμενα ερωτήματα

Πέρα από την μεγάλη κατηγορία των Overlapping Queries που αναλύθηκε προηγουμένως, μια επόμενη κατάσταση που δημιουργεί απειλές και «ρήγματα» στην ασφάλεια, είναι ένα είδος ερωτημάτων τα οποία θα ονομάσουμε *πολλαπλώς τεμνόμενα ερωτήματα*. Αυτά συνιστούν μια ακολουθία τουλάχιστον 3 ερωτημάτων που θέτει ο χρήστης με την εξής λογική. Αρχικά, ας υποθέσουμε για λόγους απλότητας της επεξήγησης, ότι μιλάμε για ερωτήματα που απαρτίζονται μόνο από ένα υπο-ερώτημα. Τα κριτήρια των υπο-ερωτημάτων με τα οποία μπορεί να επιτευχθεί επίθεση με αυτή τη

«τεχνική» είναι μόνο ο χώρος και ο χρόνος. Θεωρούμε ότι η τριάδα (τουλάχιστον) των ερωτημάτων που τίθεται, διαφοροποιείται είτε ως προς τον χώρο, είτε ως προς τον χρόνο κάθε φορά, ώστε να εξάγονται πιο συγκεκριμένα (άρα και επικίνδυνα) συμπεράσματα κάθε φορά από τον κακόβουλο χρήστη, όπως και στη περίπτωση των Overlapping Queries.



Σχήμα 3.3 - Πολλαπλώς τεμνόμενα ερωτήματα

Ας αναφερθούμε πιο συγκεκριμένα, μελετώντας τη περίπτωση του χώρου αυτή τη φορά. Έστω ότι ο χρήστης θέτει ένα ερώτημα που το χωρικό κριτήριο αντιστοιχεί σε μια περιοχή με όνομα box1. Στη συνέχεια, θέτει το 2^{ο} ερώτημα που καλύπτει μέρος της περιοχής box1 *μαζί* με κάποια επιπλέον περιοχή. Έστω ότι η συνολική περιοχή που καλύπτει αυτό το ερώτημα καλείται box2. Το 3^{ο} ερώτημα που τίθεται, περιλαμβάνει *το εναπομείναν* μέρος της περιοχής του box1 μαζί με κάποια άλλη επιπλέον περιοχή.

Παρατηρούμε ότι καμία χωρική έκταση από τις τρεις, δεν επικαλύπτει (ή δεν επικαλύπτεται) πλήρως (από) κάποια από τις άλλες δύο, δηλαδή δεν είναι μεταξύ τους *overlapping areas*. Όμως στην περίπτωση που η ένωση των περιοχών box2 και box3 περιλαμβάνει πλέον *εξ' ολοκλήρου* το box1, τότε *η περιοχή* που προκύπτει από την

ένωση των 2 αυτών περιοχών και δεν ταυτίζεται με την περιοχή του box1, είναι μια περιοχή για την οποία ο κακόβουλος χρήστης μπορεί να λάβει επιπλέον γνώση που ίσως να μην έπρεπε να είναι επιτρεπτή, ως εξής.

Έστω ότι το πλήθος των τροχιών που περιλαμβάνονται στα box2 & box3 μαζί, είναι μ και το πλήθος των τροχιών εντός του box1 είναι λ. Ο αριθμός μ-λ εκφράζει το πλήθος των τροχιών που εντοπίζονται στην περιοχή που μόλις περιγράψαμε και φαίνεται γραμμοσκιασμένη στο σχήμα 3.3.

Αν αυτός είναι μικρότερος του ορίου που θέτει το *k*-anonymity, τότε υπάρχει παραβίαση του προβλεπόμενου επιπέδου ασφαλείας.

Με ακριβώς αντίστοιχο τρόπο, όταν μεταβάλλεται ο χρόνος μεταξύ των 3 ερωτημάτων και παραμένει ο χώρος σταθερός, προκύπτουν ανάλογες καταστάσεις. Αντίθετα, επειδή πρόκειται για διακριτές τιμές, δεν μπορεί να υπάρξει «εν μέρει» επικάλυψη πληροφορίας όσον αφορά τα tags, με αποτέλεσμα η λογική των Multiple Intersection Queries να μην έχει εφαρμογή.

Ουσιαστικά πρόκειται για ένα έμμεσο τρόπο «κατασκευής» Overlapping Queries, από την πλευρά του κακόβουλου χρήστη που φέρει 2 επιπλέον χαρακτηριστικά:

1) Δίνεται η δυνατότητα για οριοθέτηση πιο ακανόνιστων, γεωμετρικά, χωρικών περιοχών σε σχέση με τα Overlapping Queries.

2) Γίνεται άμεσα αντιληπτό ως αρχική διαπίστωση ότι αυτού του είδους τα ερωτήματα είναι πιο «δυσδιάκριτα» να εντοπιστούν, γεγονός που υπαγορεύει στον όποιο μηχανισμό αντιμετώπισης παρόμοιων επιθέσεων, μια αρκετά εξειδικευμένη και «έξυπνη» αντιμετώπιση.

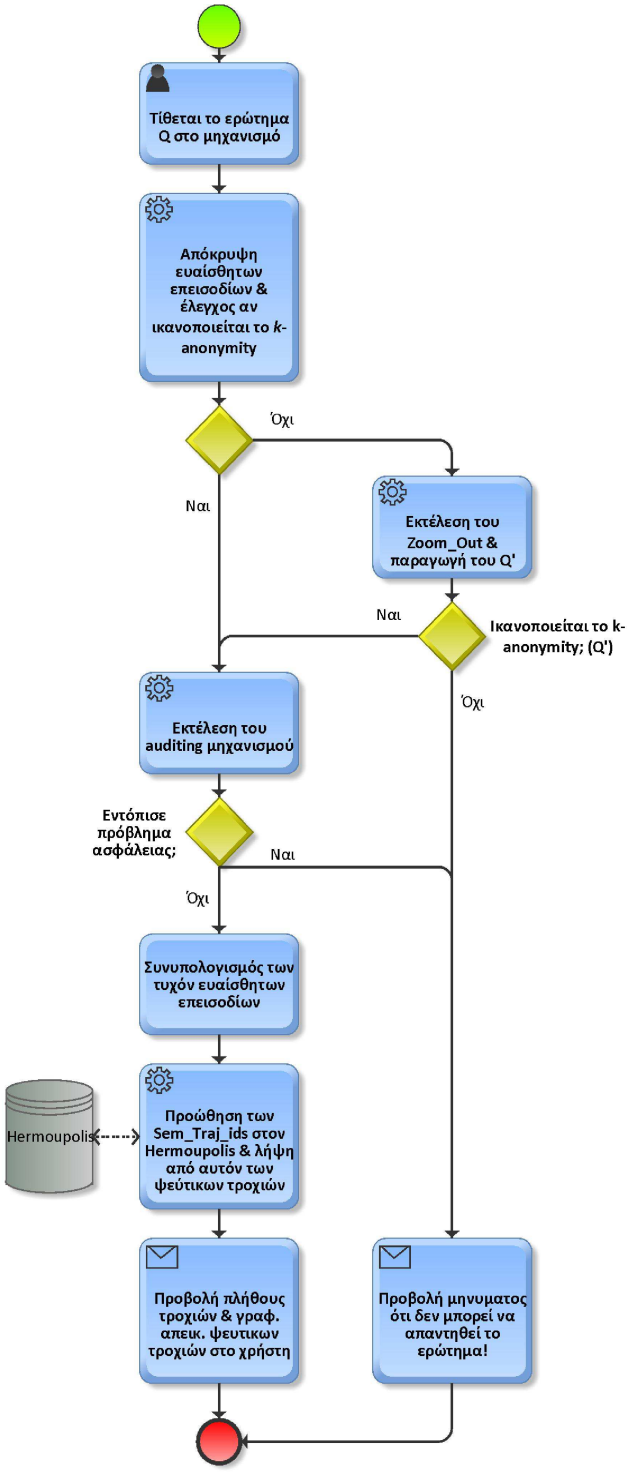
4 Αντιμετώπιση των επιθέσεων

4.1 Εισαγωγικές διατυπώσεις

Το επόμενο βήμα μετά την ανάλυση και καταγραφή της *γνώσης* που μπορεί να έχει ένας κακόβουλος χρήστης, του *σκοπού* για τον οποίο διενεργεί την επίθεση και κυρίως του *τρόπου* με τον οποίο μπορεί να το επιτύχει αυτό, περνούμε στο στάδιο της αντιμετώπισης αυτών των επιθέσεων.
Η αντιμετώπιση είναι πολυδιάστατη με 4 βασικούς άξονες δράσης.
1) Αυστηρός έλεγχος και τήρηση της αρχής του *k*-anonymity πριν από κάθε απάντηση ερωτήματος από την βάση.
2) Καθορισμός και ιδιαίτερη διαχείριση επεισοδίων τα οποία έχουν χαρακτηριστεί με οποιονδήποτε τρόπο ως επεισόδια που ενδέχεται να αποκαλύπτουν ιδιαίτερα ευαίσθητη πληροφορία για τις συνήθειες μιας οντότητας καταγραμμένης στη βάση.
3) Ενεργοποίηση ενός μηχανισμού (Zoom_out) που επιτρέπει να γίνεται προσπάθεια απάντησης σε κάθε ερώτημα ακόμη και αν δεν τηρείται το *k*-anonymity. Αυτό επιτυγχάνεται με την τροποποίηση/μετάλλαξη του ερωτήματος ώστε το πλήθος των απαντήσεων να φθάνει, πλέον, τουλάχιστον τις *k* με γνώμονα την ελάχιστη δυνατή αλλοίωση του αρχικού ερωτήματος.
4) Λειτουργία ενός συνεχούς μηχανισμού παρακολούθησης/καταγραφής όλου του ιστορικού των ερωτημάτων ανά χρήστη και ανάλυσή τους αντιπαραβάλλοντάς τα με κάθε νέο ερώτημα που τίθεται, για τον εντοπισμό πιθανών παραβιάσεων της αρχής του *k*-anonymity.

Hermoupolis Εφόσον δεν έχει διακοπεί η διαδικασία προώθησης της απάντησης στον τελικό χρήστη για λόγους διαφύλαξης της ασφάλειας των δεδομένων, οι τροχιές που συνιστούν την απάντηση, προωθούνται με τη σειρά τους στον μηχανισμό *Hermoupolis* [19]. Αυτός είναι ένας μηχανισμός ο οποίος λαμβάνει ως είσοδο ορισμένες τροχιές και παράγει ως έξοδο την 2-διάστατη γραφική απεικόνιση αντίστοιχου πλήθους προσεκτικά κατασκευασμένων ψεύτικων τροχιών. Έτσι, ο τελικός χρήστης ενημερώνεται για τον

αριθμό των τροχιών που απαντούν στο ερώτημά του, καθώς και για τις τροχιές που κατασκεύασε ο *Hermoupolis*.



Σχήμα 4.1 - Η γενική εικόνα

4.2 Καθορισμός «ευαίσθητων» επεισοδίων από το χρήστη

Ένα μέτρο βελτίωσης της θωράκισης μιας βάσης κινούμενων αντικειμένων απέναντι σε κακόβουλες επιθέσεις είναι η θεώρηση ότι ορισμένα επεισόδια από το σύνολο, περιέχουν ιδιαίτερα «ευαίσθητη» πληροφορία. Αυτό σημαίνει ότι αυτά τα επεισόδια θεωρείται ότι, είτε η τοποθεσία στην οποία αναφέρονται, είτε το χρονικό διάστημα, είτε οποιοδήποτε από το ή τα tags που συνοδεύουν το κάθε επεισόδιο, είτε οποιοσδήποτε συνδυασμός των παραπάνω δεδομένων, έχουν τη δυνατότητα να βλάψουν περισσότερο από ότι ένα συνηθισμένο επεισόδιο την ιδιωτικότητα μιας οντότητας. Αυτή η θεώρηση και ο καθορισμός κατ' επέκταση αυτών των επεισοδίων δεχόμαστε ότι επιτυγχάνεται μέσω ενός μέτρου ταξινόμησης/αξιολόγησης του κινδύνου είτε από τους ίδιους τους καταγραμμένους στη βάση ή από τους διαχειριστές της.

Στη βιβλιογραφία, όταν πρόκειται για διαχείριση απλών τροχιών (*trajectories*) μη σημασιολογικά επαυξημένων, υπάρχουν προτάσεις που έχουν διατυπωθεί για τη προστασία «ευαίσθητων» περιοχών [15], αλλά διερευνώντας για καθορισμό και ειδική μεταχείριση «ευαίσθητων» επεισοδίων δημιουργημένα από σημασιολογικά επαυξημένες τροχιές, δεν καταφέραμε να εντοπίσουμε παρόμοιες εργασίες.

Το αυτονόητο είναι, τα επεισόδια αυτού του είδους, να μην λαμβάνονται καθόλου υπόψη κατά το σχηματισμό της απάντησης στον τελικό χρήστη.

Η πρότασή μας συνίσταται στην *κατ' αρχήν απόκρυψη* των εν λόγω επεισοδίων από το εν δυνάμει πλήθος επεισοδίων που συγκροτεί μια απάντηση σε ένα ερώτημα. Αν όμως τίθενται στη βάση ερωτήματα, που επιστρέφουν στον τελικό χρήστη απαντήσεις που φθάνουν ή ξεπερνούν σε πλήθος, το όριο του *k*-anonymity, προκύπτει το εξής ζήτημα. Τί πραγματικά εξυπηρετεί η απόκρυψη ενός αριθμού «ευαίσθητων» επεισοδίων από την τελική απάντηση όταν οι *μη* «ευαίσθητες» τροχιές ξεπερνούν από μόνες τους, το προκαθορισμένο όριο ασφαλείας; Από τη μια, τα επεισόδια αυτά μπορούν να θεωρηθούν καλά «κρυμμένα» στην απάντηση, λόγω της ύπαρξης «επαρκούς» πλήθους μη «ευαίσθητων» επεισοδίων και από την άλλη το μόνο που «προσφέρει» η απόκρυψή τους είναι μια αλλοίωση της πραγματικής απάντησης στο ερώτημα χωρίς ιδιαίτερα συγκεκριμένο όφελος.

Η λύση που προτείνουμε είναι ο συνυπολογισμός των επεισοδίων αυτών, *μόνο αν* ισχύει ότι τα μη «ευαίσθητα» επεισόδια που συμμετέχουν στην απάντηση, ξεπερνούν

από μόνα τους, το όριο του *k*-anonymity. Η διαδικασία αυτή επιτελείται κατά το *τελευταίο στάδιο* ελέγχου και κατασκευής της απάντησης.

4.3 Η τεχνική Zoom Out

4.3.1 Η ιδέα & ο σκοπός του μηχανισμού

Ένας εξουσιοδοτημένος χρήστης της βάσης επιθυμεί να θέσει ένα ερώτημα αν υπάρχουν ή όχι τροχιές βάσει ενός συνόλου κριτηρίων. Αν το πλήθος των τροχιών που επιστρέφει αυτό, είναι μικρότερο από το *k*-anonymity τότε το ερώτημα δεν θα πρέπει να απαντηθεί, διασφαλίζοντας έτσι την πρώτου επιπέδου, προστασία που έχουμε καθορίσει.

Η ιδέα είναι η εξής: ο μηχανισμός απάντησης των ερωτημάτων, αντί να μην απαντήσει καθόλου το ερώτημα, επιχειρεί να δώσει απάντηση *σε κάθε περίπτωση* σεβόμενος πάντα την αρχή του *k*-anonymity. Με άλλα λόγια, ο μηχανισμός προσπαθεί να απαντήσει τηρώντας το ίδιο επίπεδο ασφάλειας ένα, όσο γίνεται, πιο *παρεμφερές ερώτημα* με αυτό που έθεσε ο χρήστης σε πρώτη φάση.

Το αποτέλεσμα αυτής της διαδικασίας φαντάζει ως ένα αναγκαστικό *zoom out* σε ένα γεωγραφικό χάρτη που ενώ ο χρήστης ζητά να «δει» μια περιοχή, το πρόγραμμα του επιστρέφει αυτό το οποίο είναι δυνατόν να προβάλει «σωστά» μέχρι δηλαδή να φθάνει μόνο μέχρι μια, «επιτρεπτού» βαθμού, ανάλυση.

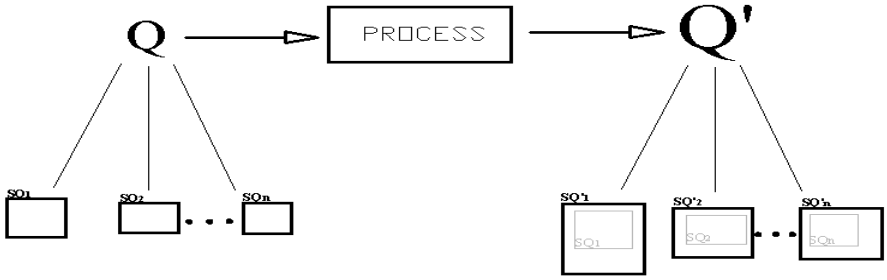
Σκοπός της υλοποίησης μιας τέτοιας ιδέας είναι η αύξηση της φιλικότητας προς τον χρήστη και η βελτίωση της, εν γένει, λειτουργικότητας της βάσης. Ουσιαστικά βοηθά τον χρήστη να λάβει μια πληροφορία με τα στοιχεία που τον ενδιαφέρουν και όχι να διενεργεί συνεχώς ερωτήματα διευρύνοντας διαρκώς το εύρος των κριτηρίων που θέτει, μέχρι να του δοθεί μια απάντηση. Αυτό δημιουργεί και άσκοπη επιβάρυνση στη χρήση της βάσης και την πιθανότητα να μην λάβει ο χρήστης την καλύτερη *δυνατή* απάντηση.

Για να μπορέσει να συμβεί αυτό, ο μηχανισμός πρέπει να επιτρέπει την διεύρυνση του ενός ή των περισσότερων από όσα κριτήρια είναι απαραίτητο, του κάθε υπο-ερωτήματος στα πλαίσια πάντα του ενιαίου και συνολικού ερωτήματος που έθεσε ο

χρήστης. Η διεύρυνση μπορεί να είναι χωρική, χρονική ή ακόμα και σε επίπεδο σημασιολογικών αναφορών (*tags*).

4.3.2 Καθορισμός εισόδου & εξόδου της διαδικασίας

Προκειμένου να είναι απόλυτα σαφής ο τρόπος λειτουργίας του αλγορίθμου που θα προτείνουμε, ας διευκρινίσουμε τί ακριβώς δέχεται ως είσοδο και τί παράγει ως έξοδο προς το επόμενο στάδιο του συνολικού μηχανισμού.
Η είσοδος είναι ένα ερώτημα, έστω Q, όπως ακριβώς ορίστηκε στο κεφάλαιο 1. Αυτό το ερώτημα, αν εκτελεστεί, περιέχει λιγότερες από *k* τροχιές.



Σχήμα 4.2 – Είσοδος / Έξοδος της διαδικασίας

Μετά την επεξεργασία στην οποία υπόκειται, παράγεται ως έξοδος της διαδικασίας *ένα νέο ερώτημα*, έστω Q'. Αν η διαδικασία μετασχηματισμού του Q σε Q' είναι επιτυχής, τότε το Q', εκτελούμενο, περιέχει τουλάχιστον *k* τροχιές. Αν όχι, τότε το Q' σημαίνει ότι δεν περιέχει τουλάχιστον *k* τροχιές. Το Q' αποτελεί ένα σύνολο από υπο-ερωτήματα ίσα στο πλήθος με όσα περιείχε το Q. Κάθε υπο-ερώτημα του Q' προκύπτει να είναι είτε το ίδιο είτε διευρυμένο σε σχέση με το αντίστοιχο υπο-ερώτημα που περιλαμβάνεται στο Q. Όταν το υπο-ερώτημα του Q' έχει διευρυνθεί, εννοούμε ότι ένα ή περισσότερα από τα κριτήρια που αυτό περιλαμβάνει έχουν τροποποιηθεί ώστε να αποτελούν υπερσύνολο των αρχικών κριτηρίων.
Κριτήριο είναι μια χωρική έκταση (δηλ. ένα MBB) ή/και ένα χρονικό διάστημα ή/και ένα ή περισσότερα tags που χαρακτηρίζουν επεισόδια που αναζητούμε μέσω του κάθε ερωτήματος.

Έτσι, το αποτέλεσμα της διαδικασίας δεν είναι τροχιές αλλά ένα νέο ερώτημα που *αν απαντηθεί τελικά* επιστρέφει τουλάχιστον *k* τροχιές.

4.3.3 Πού εντάσσεται σε σχέση με συνολικό μηχανισμό ελέγχου

Ως συνολικό μηχανισμό ελέγχου, εννοούμε το σύνολο των διαδικασιών αυτών, οι οποίες υλοποιούν την καταγραφή, τον ιστορικό έλεγχο, τη βελτίωση/τροποποίηση των ερωτημάτων και άλλες λειτουργίες έτσι ώστε να ισχύουν οι 4 βασικοί άξονες δράσης που καταγράψαμε στην εισαγωγή αυτής της ενότητας.
Ο αλγόριθμος που θα περιγράψουμε εκτελείται μετά την απόκρυψη τυχόν «ευαίσθητων» επεισοδίων από τα διαθέσιμα επεισόδια μιας βάσης και εφόσον το ερώτημα που τέθηκε από τον χρήστη, δεν καλύπτει σε πλήθος τροχιών τις *k*.
Μετά την εκτέλεση του αλγορίθμου, επανεξετάζεται αν το τροποποιημένο ερώτημα που προέκυψε ως αποτέλεσμα, εκτελούμενο καλύπτει σε πλήθος τουλάχιστον *k* τροχιές. Αν η απάντηση είναι αρνητική, τότε ο μηχανισμός δεν επιτρέπει την εκτέλεση του ερωτήματος και ενημερώνεται κατάλληλα ο χρήστης. Αν η απάντηση είναι θετική, τότε το τροποποιημένο ερώτημα οδηγείται προς την διεκπεραίωση της διαδικασίας ελέγχου ερωτημάτων και άλλων συναφών διαδικασιών (auditing μηχανισμός). Να σημειώσουμε ότι εφόσον εκτελείται ο αλγόριθμος zoom out για ένα ερώτημα, ο auditing μηχανισμός παραλαμβάνει, ελέγχει, αντιπαραβάλει και αποθηκεύει αν χρειαστεί το *τροποποιημένο* ερώτημα που ο zoom out παρήγαγε.

4.3.4 Πότε πρέπει να εκτελείται

Αυτό που θα φάνταζε αρχικά λογικό είναι η μέθοδος αυτή, να έχει εφαρμογή μόνο όταν *οριακά* δεν μπορεί να απαντηθεί ένα ερώτημα σε σχέση με το προκαθορισμένο *k*-anonymity της βάσης. Άλλωστε αυτός είναι και ο βασικός λόγος που ένας χρήστης μπορεί να ζητούσε μεγαλύτερη «ευελιξία» από τον μηχανισμό απάντησης των ερωτημάτων. Ας φανταστούμε ερώτημα που τίθεται σε περιοχή 20.000 m^2 με k=10 και περιέχει μόλις 9 επεισόδια. Αν επεκτεινόμενο κατά π.χ. μόλις 500 m^2 μπορεί να περιλάβει 1 επιπλέον επεισόδιο, τότε δίνει κατά ένα μεγάλο ποσοστό στο

(καλοπροαίρετο) χρήστη της βάσης την πληροφορία που ουσιαστικά ζητάει για την περιοχή. Υπάρχει βέβαια και η πιθανότητα, σε μια άλλη ακραία περίπτωση, να μην περιλαμβάνεται κανένα επεισόδιο στο χώρο που αρχικά ζητήθηκε και για να κατασκευαστεί το τροποποιημένο ερώτημα που θα συμπεριλαμβάνει τον απαραίτητο αριθμό επεισοδίων σύμφωνα με το *k*-anonymity, να χρειάζεται ο αλγόριθμος να το διευρύνει χωρικά τόσο πολύ που να ξεφεύγουμε από το νόημα του αρχικού ερωτήματος.

Η υλοποίηση μιας τέτοιας συλλογιστικής θα απαιτούσε από την πλευρά μας, να καθορίσουμε ένα κατώφλι με βάση το οποίο θα ενεργοποιούταν ή όχι αυτός μηχανισμός κάθε φορά που δεν θα μπορούσε κατευθείαν να απαντηθεί το αρχικό ερώτημα με το σκεπτικό ότι καθίσταται λιγότερο χρηστική η τυχόν ενεργοποίηση του μηχανισμού κάθε φορά που ο αριθμός των επεισοδίων που υπολείπεται, μέχρι την ικανοποίηση του *k*-anonymity, είναι μεγάλος.

Αν η πρότασή μας όμως, κινηθεί σε αυτή τη κατεύθυνση, τότε προκύπτει σοβαρό πρόβλημα ασφάλειας. Ας υποθέσουμε ότι θα υπήρχε ένα κατώφλι που θα εξέφραζε τον μέγιστο αριθμό των επεισοδίων που αν υπολείπονταν σε σχέση με το *k*, θα ενεργοποιούταν ο μηχανισμός. Θεωρούμε επίσης τον αριθμό αυτό γνωστό στον κακόβουλο χρήστη. Έστω λοιπόν ότι το κατώφλι είναι ο αριθμός 1 και μετά από ένα αρχικό ερώτημα του χρήστη για την περιοχή Α, του επιστράφηκαν ως απάντηση *k* τροχιές και μια περιοχή Β + Α. Άρα για την περιοχή Β, δηλαδή για τον χώρο που προστέθηκε για να περιλάβει τις νέες τροχιές, ο χρήστης γνωρίζει ότι περιέχει μία τροχιά ακριβώς, δηλαδή ότι *αυτή και μόνο αυτή* βρίσκεται στην περιοχή Β, ενώ αν είχε ρωτήσει τη βάση απ' ευθείας για την περιοχή Β, δεν θα είχε λάβει απάντηση λόγω του *k*-anonymity. Υπάρχει λοιπόν περίπτωση, όταν υπάρχει *συγκεκριμένο όριο* στην ενεργοποίηση του μηχανισμού, να οδηγούμαστε σε παραβίαση της ιδιωτικότητας.

Κατά συνέπεια, η λύση είναι να εκτελείται *πάντα* ο συγκεκριμένος αλγόριθμος ανεξάρτητα από πόσες τροχιές υπολείπονται κάθε φορά, ώστε ο χρήστης να μην γνωρίζει πόσες περισσότερες περιλαμβάνονται στην επιπλέον περιοχή που τυχόν επιστρέφει το ερώτημα κατά την εκτέλεσή του.

4.3.5 Ορισμός του distortion Unit, ο συντελεστής βαρύτητας λ & η έννοια του «κοντινότερου» γείτονα

Όπως ήδη διατυπώθηκε, ο αλγόριθμος προσπαθεί να δημιουργήσει ένα, όσο γίνεται, πιο παρεμφερές ερώτημα με το αρχικό. Αυτό σημαίνει ότι αναγκάζεται να τροποποιήσει ένα ή περισσότερα από τα υπο-ερωτήματα που αυτό περιλαμβάνει. Η τροποποίηση γίνεται προκειμένου να αντιστοιχούν/περιλαμβάνονται περισσότερα επεισόδια ανά υπο-ερώτημα ώστε να μην υπάρχει πρόβλημα με το *k*-anonymity. Για να μπορέσει όμως ο αλγόριθμος να αποφασίζει ποιο επεισόδιο είναι προτιμητέο να περιλαμβάνεται στην απάντηση του τροποποιημένου πλέον ερωτήματος (άρα ουσιαστικά προς ποια «κατεύθυνση» θα τροποποιηθεί το ερώτημα) πρέπει να του δοθεί η δυνατότητα να συγκρίνει την αλλοίωση που δύναται να προκύψει μεταξύ 2 ή περισσότερων υποψηφίων προς ένταξη επεισοδίων. Υπενθυμίζουμε ότι το κάθε υπο-ερώτημα έχει τη δυνατότητα να διευρυνθεί είτε χωρικά, είτε χρονικά αλλά είτε και ως προς το πλήθος των tags που πλέον θα περιλαμβάνει ως κριτήρια.

Έτσι, για να λειτουργήσει ο αλγόριθμός είναι απαραίτητη μια νέα μονάδα μέτρησης που θα καλούμε από εδώ και στο εξής *distortion unit*. Είναι ένας καθαρός αριθμός και προκύπτει ως αποτέλεσμα διαίρεσης μεταξύ ομοειδών δεδομένων.

Έτσι έστω ότι
Ep1 : το υπό εξέταση επεισόδιο
SQ : το υπο-ερώτημα προς τροποποίηση
SQ' : το υπο-ερώτημα όπως θα διαμορφωθεί για να περιλάβει το Ep1.
Area(): Μια συνάρτηση που δέχεται ως είσοδο το MBB του υπο-ερωτήματος και υπολογίζει το εμβαδό σε τ.μ.
Duration(): Μια συνάρτηση που δέχεται ως είσοδο το χρονικό διάστημα που αναφέρεται το υπο-ερώτημα και το υπολογίζει σε λεπτά.
Common_tags_between(): Μια συνάρτηση που δέχεται ως είσοδο δύο υπο-ερωτήματα και υπολογίζει το πλήθος των κοινών tags μεταξύ τους.

Distinct_tags_in(): Μια συνάρτηση που δέχεται ως είσοδο δύο υπο-ερωτήματα και υπολογίζει το σύνολο των διακριτών τιμών των tags που περιλαμβάνουν αθροιστικά τα 2 αυτά υπο-ερωτήματα.

Για παράδειγμα, αν *SQ = {A, B, C}* & *SQ' = {B, C, D}*, τότε ισχύει ότι

Common_tags_between (SQ, SQ') = {B, C} & *Distinct_tags_in(SQ, SQ') = {A, B, C, D}*

Η χωροχρονική αλλοίωση υπολογίζεται ως εξής:

$$D\_Unit\_st=(((Area(SQ')-Area(SQ))/Area(SQ))+((Duration(SQ')-Duration(SQ))/Duration(SQ)))/2$$

Η αλλοίωση ως προς τα tags υπολογίζεται ως εξής:

$$D\_Unit\_tags=1-(Common\_tags\_between(SQ, SQ')/Distinct\_tags\_in(SQ, SQ'))$$

Προσπαθώντας να συνδυάσουμε τα 2 προηγούμενα μεγέθη, εισάγουμε μια συνολική αλλοίωση που μπορεί να προκύψει με κατανομή του βάρους μεταξύ της χωροχρονικής αλλοίωσης και της αλλοίωσης λόγω των tags, βάσει ενός συντελεστή λ. Έτσι ισχύει ότι

$$D\_Unit=\lambda * D\_Unit\_st + (1-\lambda)*D\_Unit\_tags \quad \text{με } 0\leq\lambda\leq1$$

Αν δεν είναι επιθυμητή η λειτουργία του αλγορίθμου με συνυπολογισμό/ευελιξία των tags, τίθεται $\lambda=1$.

Άλλωστε οφείλουμε να τονίσουμε ότι αυτή η μεθοδολογία ως προς τα tags είναι αρκετά απλοποιημένη με σκοπό να είναι άμεσα και παντού εφαρμόσιμη. Είναι προφανές ότι ένα ερώτημα με tag= 'bike' είναι πολύ κοντινότερο σημασιολογικά σε ένα επεισόδιο με tag= 'bicycle' από ότι σε ένα επεισόδιο με tag= 'fun'. Παρόλα αυτά, η *D_Unit_tags* θα ισούται και για τις 2 περιπτώσεις με 1.

Αναφορικά με τις συγκρίσεις ως προς τα tags, υπάρχει η περίπτωση ένα, υπό εξέταση, επεισόδιο να μην έχει καθόλου τιμή στα tags (null value). Σε αυτή τη περίπτωση είναι, πλέον, ζήτημα προτίμησης αν το null value θεωρείται «τιμή» διαφορετική από οποιοδήποτε άλλο tag με το οποίο μπορεί να συγκριθεί ή αν το null value θεωρείται «τιμή» ανεκτή κατά τη σύγκρισή της με όλες τις άλλες τιμές. Με άλλα λόγια, καλό είναι να δίνεται η δυνατότητα εξατομίκευσης (σε επίπεδο βάσης), αν η τομή μεταξύ του «τίποτα» με ένα tag value, είναι μηδέν (*D_Unit_tags=1*) ή είναι το ίδιο το tag value *(D_Unit_tags=0)*.

Σε κάθε περίπτωση, ζήτημα προκύπτει ως προς την τελική διαμόρφωση του κάθε υπο-ερωτήματος στις περιπτώσεις που διευρύνθηκε ως προς τα tags, προκειμένου να περιλάβει επαρκή αριθμό επεισοδίων. Δεν είναι δυνατόν να τροποποιηθεί το υπο-ερώτημα με τρόπο τέτοιο που να παρέχεται συγκεκριμένη *επιπλέον* πληροφορία. Δεν μπορεί, για παράδειγμα, να ζητά κανείς πληροφορία για Space1, Time1 & Tag = 'bike' και να λαμβάνει πληροφορία για Space1, Time1 & Tag = 'bike', 'bicycle' ή πιο σωστά Space1, Time1 & (Tag = 'bike' or Tag = 'bicycle'). Αυτή η διατύπωση αφενός, οδηγεί σε αποκάλυψη συγκεκριμένης πληροφορίας προς το χρήστη και αφ' ετέρου απαιτεί μια εντελώς διαφορετική συλλογιστική χειρισμού των tags από το πλαίσιο εργασίας μας. Η μόνη ασφαλής λύση σε αυτή τη περίπτωση, είναι το υπο-ερώτημα που διευρύνεται να διαμορφωθεί πλέον χωρίς συγκεκριμένα tags.

Ο αλγόριθμος Zoom Out, προκειμένου να είναι εφαρμόσιμος απαιτεί μια εντελώς διευρυμένη λογική πάνω στα ερωτήματα *κοντινότερου γείτονα.* Αυτό που παραδοσιακά ισχύει είναι ότι τέτοιου τύπου ερωτήματα εντοπίζουν τη χωρική απόσταση μεταξύ δύο γεωμετρικών σχημάτων στο επίπεδο ή και στο χώρο με τη χρήση ευκλείδειας γεωμετρίας. Στη δική μας περίπτωση, ο κοντινότερος γείτονας στα κριτήρια ενός υπο-ερωτήματος μπορεί να είναι κάποιο επεισόδιο που ενδέχεται να απέχει και χωρικά, και χρονικά αλλά και ως προς το σημασιολογικό του περιεχόμενο (tags) από τα κριτήρια αυτά. Έτσι πρέπει αναγκαστικά να συντελείται μια «ταυτόχρονη» διεύρυνση προς όλες τις παραμέτρους ενός υπο-ερωτήματος, ώστε να βρεθούν οι κοντινότεροι γείτονές του.
Από δω και στο εξής, κοντινότερος γείτονας προς τα κριτήρια ενός υπο-ερωτήματος, είναι, εκείνο το επεισόδιο που έχει την μικρότερη *D_Unit* βαθμολογία μεταξύ των υπολοίπων.

4.3.6 Περιγραφή του αλγορίθμου

Ο αλγόριθμος 1 είναι αυτός που υλοποιεί την τεχνική Zoom Out (σχήμα 4.4). Αποτελείται από 3 γενικά στάδια. Το πρώτο είναι όλες οι διαδικασίες αρχικοποίησης, προετοιμασίας μεταβλητών και λοιπών δομών δεδομένων (γραμμές 1-3) και μια αρχική διαδικασία γρήγορης εκκίνησης (γραμμές 4-6). Το δεύτερο είναι μια επαναληπτική διαδικασία, χωρισμένη σε 3 μέρη, που αποτελεί και την καρδιά του αλγορίθμου (γραμμές 7-25). Το τρίτο αποτελείται από διαδικασίες που προκαλούν ένα «φινίρισμα» πριν την τελική αποδέσμευση του ερωτήματος από τον αλγόριθμο (γραμμές 26-27).

Ένα βασικό στοιχείο που χρειάζεται ο αλγόριθμος για να λειτουργήσει, είναι μια δομή δεδομένων (για παράδειγμα ένας ή περισσότεροι επιπλέον βοηθητικοί πίνακες στη βάση) που θα λειτουργεί σαν μια λίστα με όλα τα υποψήφια για ένταξη, κάθε φορά, επεισόδια.

Κατά το 1^{ο} στάδιο, ο αλγόριθμος, μετά τις διαδικασίες αρχικοποίησης, κάνει αυτό που αναφέραμε ως γρήγορη εκκίνηση. Εκτελείται το κάθε υπο-ερώτημα υποθετικά σαν να ήταν ένα αυτόνομο ερώτημα και επιλέγεται το υπο-ερώτημα αυτό το οποίο επέστρεψε το μεγαλύτερο πλήθος επεισοδίων βάσει των κριτηρίων του (γραμμή 4). Αν υπάρχουν περισσότερα από ένα υπο-ερωτήματα με το μεγαλύτερο (ίδιο) αριθμό επεισοδίων, τότε επιλέγεται τυχαία ένα από αυτά. Αν δεν υπάρχει κανένα υπο-ερώτημα που να περιλαμβάνει έστω και ένα επεισόδιο, τότε, επίσης, επιλέγεται τυχαία, ένα απ' όλα. Στο υπο-ερώτημα που επιλέχθηκε, αν το πλήθος του answer set είναι μικρότερο του *k* (γραμμή 5), τότε εκτελείται ένα ερώτημα κοντινότερου γείτονα όπως καθορίστηκε στην ενότητα 4.3.5 και εντοπίζονται οι *k* - *n* κοντινότεροι γείτονες (όπου *n* το πλήθος του answer set του επιλεγμένου υπο-ερωτήματος). Με άλλα λόγια, εντοπίζονται τόσα επεισόδια όσα απαιτούνται ώστε προστιθέμενα στο αρχικό answer set του υπο-ερωτήματος, να πληρείται το *k*-anonymity για το ένα μόνο αυτό υπο-ερώτημα. Ακολούθως, τροποποιείται/μεταβάλλεται ανάλογα το επιλεγμένο υπο-ερώτημα ώστε να εμπεριέχει πλέον στην, εν δυνάμει, απάντησή του και τα *k* - *n* κοντινότερα επεισόδια που εντοπίστηκαν (γραμμή 6). Αν το πλήθος του answer set είναι μεγαλύτερο του *k*, τότε η εκτέλεση του αλγορίθμου προχωρά κατευθείαν στο 2^{ο} τμήμα.

Στη συνέχεια, το *δεύτερο στάδιο* του αλγορίθμου είναι μια επανάληψη χωρισμένη σε 3 διακριτά μέρη.

1ο μέρος Στην αρχή κάθε επανάληψης, λαμβάνει χώρα μια μίνι επανάληψη (γραμμές 9-13) όπου εκτελείται κάθε υπο-ερώτημα του ερωτήματος μεμονωμένα (γραμμή 11) και εντοπίζονται οι τροχιές που αυτό περιλαμβάνει. Οι τροχιές αυτές συμπληρώνουν έναν πίνακα (*H*) ο οποίος περιέχει 2 πεδία καταρχάς,

α) την κάθε *τροχιά (tr_id)* που βρέθηκε και

β) τη *συχνότητα εμφάνισής* της (*freq*), στο σύνολο των υπο-ερωτημάτων (γραμμή 12). Αυτό πρακτικά σημαίνει ότι αυτό το πεδίο είναι ένας μετρητής. Ξεκινά με αρχική τιμή 1, την πρώτη φορά που εισέρχεται μια νέα τροχιά στον πίνακα (*H*) και κάθε φορά που ένα επόμενο υπο-ερώτημα εκτελείται και επιστρέφει επεισόδιο που ανήκει στην *ίδια τροχιά*, αυξάνεται κατά 1 κ.ο.κ.. Με άλλα λόγια, ο αλγόριθμος, είτε εντάσσει στο πίνακα για πρώτη φορά τη συγκεκριμένη τροχιά και ο μετρητής γίνεται 1 (*insert*), είτε εντοπίζει εντός του πίνακα την τροχιά να υπάρχει ήδη σε μία από τις εγγραφές, οπότε αυξάνει κατά ένα τον μετρητή (*update*).

Να υπενθυμίσουμε ότι αφού μόλις προηγουμένως έχει εκτελεστεί το 1ο τμήμα του κώδικα (*γρήγορη εκκίνηση*), ένα τουλάχιστον από τα υπο-ερωτήματα θα περιέχει *k* τουλάχιστον τροχιές. Έτσι, η εικόνα του πίνακα σε αυτό το σημείο θα είναι η εξής. Θα περιέχει τουλάχιστον *k* εγγραφές δηλαδή *k* διαφορετικές τροχιές, με τιμή στον μετρητή (*συχνότητα εμφάνισης*) ίση ή μεγαλύτερη από 1 στην κάθε μια εγγραφή. Η μέγιστη τιμή που μπορεί εκ των πραγμάτων να λάβει ο μετρητής σε κάθε εγγραφή είναι ένας αριθμός που *πάντα θα ισούται* με το πλήθος των υπο-ερωτημάτων που απαρτίζουν το ερώτημα που τέθηκε ως είσοδος στον *Zoom_Out*. Έτσι, αν για παράδειγμα, ένα ερώτημα αποτελείται από 3 υπο-ερωτήματα οι μετρητές της κάθε τροχιάς θα έχουν τιμή από 1 μέχρι 3. Αν ο μετρητής μιας τροχιάς έχει το μέγιστο αριθμό, σημαίνει ότι εντοπίστηκαν επεισόδιά της σε ΟΛΑ τα υπο-ερωτήματα του ερωτήματος, γεγονός που σημαίνει ότι είναι μια τροχιά που επιστρέφεται ως απάντηση όταν τίθεται το (συνολικό) ερώτημα.

Αφού συμπληρωθεί με αυτόν τον τρόπο ο πίνακας, η διαδικασία ολοκληρώνεται με μια *ταξινόμησή* του ως προς τη συχνότητα εμφάνισης (γραμμή 14).

$2^{ο}$ μέρος Μέσα από τον πίνακα (*H*) που μόλις συμπληρώθηκε και ταξινομήθηκε, ο αλγόριθμος εντοπίζει ποια ή ποιες είναι οι τροχιές αυτές του πίνακα που από τη μια δεν έχουν στη συχνότητά τους τη μέγιστη δυνατή τιμή (δηλ. συχνότητα=πλήθος υπο-ερωτημάτων), αλλά έχουν τη μεγαλύτερη τιμή μεταξύ όλων των υπόλοιπων τροχιών του πίνακα (γραμμή 15). Στη συνέχεια, ξεκινά μια, μικρού εύρους, επανάληψη (γραμμή 17-21) εντός της οποίας βρίσκεται η βασική διαδικασία της $2^{ης}$ ενότητας (γραμμή 18). Η επανάληψη αυτή διασφαλίζει ότι αν, για διάφορους λόγους που θα περιγράψουμε στη συνέχεια, δεν μπορεί να θεωρηθεί κατάλληλο κανένα επεισόδιο από την τροχιά που εντόπισε μόλις προηγουμένως η διαδικασία *Find_Freq_Position*, να μπορεί να συνεχιστεί η διερεύνηση στην αμέσως επόμενη (ή επόμενες αν ισοβαθμούν) τροχιά, δηλαδή την τροχιά που έχει την αμέσως μικρότερη συχνότητα στον πίνακα *H*. Η επανάληψη αυτή ολοκληρώνεται είτε όταν βρεθεί επιτρεπτό επεισόδιο να ενταχθεί, είτε αν διερευνηθούν όλες οι εναπομείνασες τροχιές του πίνακα και δεν έχει βρεθεί επεισόδιο (γραμμή 21).

Αρχικά, ας δούμε τι συμβαίνει εντός της διαδικασίας *Compute_Distortion_Units* για τις πιο απλές περιπτώσεις. Πιο συγκεκριμένα, αν υπάρχει μια τροχιά που έχει συχνότητα κατά 1 μικρότερη από τη μέγιστη δυνατή (άρα έχει και τη μεγαλύτερη τιμή από τις υπόλοιπες), τότε είναι αυτονόητο ότι όλα τα υπο-ερωτήματα του ερωτήματος που εξετάζουμε, περιέχουν στις απαντήσεις τους, επεισόδια αυτής της τροχιάς, πλην ενός υπο-ερωτήματος. Για αυτό το υπο-ερώτημα το οποίο δεν «περιέχει» επεισόδιο αυτής της τροχιάς, υπολογίζεται η αλλοίωση που θα προκαλείτο στο ίδιο, ώστε (τροποποιημένο πλέον) να συμπεριλαμβάνει στην απάντησή του, επεισόδιο και της εν λόγω τροχιάς. Αν, από την άλλη, η συχνότητα αυτής της τροχιάς, υπολείπεται κατά 2 ή περισσότερο από τη μέγιστη δυνατή τιμή, τότε σημαίνει ότι υπάρχουν 2 ή περισσότερα υπο-ερωτήματα που δεν περιλαμβάνουν επεισόδια αυτής της τροχιάς. Έτσι, για κάθε ένα από αυτά τα υπο-ερωτήματα υπολογίζεται (πάντα με όρους *distortion units*) ξεχωριστά η εν δυνάμει αλλοίωση σε αυτά ώστε να περιλάβουν επεισόδια και αυτής της τροχιάς. Αν υπάρχουν 2 ή περισσότερες τροχιές που έχουν την ίδια συχνότητα εμφάνισης, τότε υπολογίζονται για όλες αυτές τις τροχιές και για όλα τα υπο-ερωτήματα στα οποία η καθεμιά δεν «συμμετέχει» με αντίστοιχο επεισόδιο, οι αντίστοιχες αλλοιώσεις επί των υπο-ερωτημάτων. Είναι αυτονόητο, τέλος, ότι αν, για παράδειγμα, 2 τροχιές ισοβαθμούν στη συχνότητα και υπολείπονται από τη κάθε μία, 2

υπο-ερωτήματα που δεν «συμμετέχουν», αυτό δεν σημαίνει σε καμία περίπτωση ότι από το σύνολο των υπο-ερωτημάτων, τα 2 υπο-ερωτήματα που αφορούν την $1^{η}$ τροχιά, είναι τα ίδια 2 που αφορούν την $2^{η}$.

Έτσι, η διαδικασία *Compute_Distortion_Units* χρειάζεται να εκτελέσει ερωτήματα στη βάση τα οποία να αναζητούν πού βρίσκεται το κοντινότερο επεισόδιο για την κάθε τροχιά προς το αντίστοιχο υπο-ερώτημα που μας απασχολεί εκείνη τη στιγμή. Κατόπιν, υπολογίζεται η αλλοίωση που θα προέκυπτε στο κάθε υπο-ερώτημα αν ενσωμάτωνε το αντίστοιχο επεισόδιο που μόλις βρέθηκε, μέσω του υπολογισμού των distortion units και ενημερώνεται ο πίνακας (γραμμή 18).

Ο πίνακας λοιπόν, εκτός από τα 2 πρώτα πεδία που αναφέραμε ότι περιέχει, πρέπει να μπορεί να καταχωρεί τα στοιχεία των *distortion units* ανά υπο-ερώτημα και ανά τροχιά. Με αυτό τον τρόπο, όπως φαίνεται και στο σχήμα 4.3, καταχωρούνται επιπλέον

α) τα στοιχεία του εκάστοτε επεισοδίου που βρέθηκε να είναι το κοντινότερο (*ep_id*) και

β) η αλλοίωση που θα προκληθεί στο αντίστοιχο υπο-ερώτημα αν ο αλγόριθμος αποφασίσει να επιλέξει το εν λόγω, επεισόδιο.

Tr_id	**Freq**	**SubQuery$_1$**		**SubQuery$_2$**		**SubQuery$_3$**		**…**	**SubQuery$_n$**	
		Ep_id	**Distortion Units**	**Ep_id**	**Distortion Units**	**Ep_id**	**Distortion Units**		**Ep_id**	**Distortion Units**

Σχήμα 4.3 – Γενικό υπόδειγμα γραμμογράφησης του βοηθητικού πίνακα

Στο σημείο αυτό και πριν προχωρήσουμε στο πώς επ’ ακριβώς επιλέγει ο αλγόριθμος ποιο επεισόδιο από τα υποψήφια θα εντάξει στο αντίστοιχο υπο-ερώτημα *εντός της κάθε επανάληψης*, πρέπει να αναφερθούμε στις εξής καταστάσεις.

Το *distortion unit*, εξ’ ορισμού δεν έχει κάποιο περιορισμό στις τιμές που μπορεί να λάβει. Πρέπει όμως να υπάρχει ένα όριο στην αλλοίωση που μπορεί να προκληθεί στο οποιοδήποτε υπο-ερώτημα (*distortion limit*) και αυτό διότι σε κάθε άλλη περίπτωση υπάρχει ο κίνδυνος να έχει τροποποιηθεί στο πέρας της εκτέλεσης του αλγορίθμου, τόσο πολύ το τελικό ερώτημα που να μην είναι πλέον *παρεμφερές* του ερωτήματος που

έθεσε αρχικά ο χρήστης και για το οποίο επιθυμούσε μια απάντηση. Το όριο αυτό το θέτουμε να είναι ίσο με 0,1 και αφορά το όριο στην αλλοίωση που μπορεί να προκαλέσει ένα επεισόδιο που εντάσσεται στα πλαίσια της *κάθε* επανάληψης και όχι τη συνολική αλλοίωση που ενδεχομένως θα έχει ένα τελικό (τροποποιημένο) υπο-ερώτημα σε σχέση με το αρχικό που δόθηκε σαν είσοδος στο αλγόριθμο. Για να το αντιληφθούμε πρακτικά, *distortion limit*=0,1 σημαίνει ότι (χωρίς διεύρυνση σε επίπεδο *tags*) επιτρέπουμε να αυξηθεί κατά 10% και η χωρική έκταση και η χρονική διάρκεια των κριτηρίων του *κάθε* υπο-ερωτήματος σε *κάθε* επανάληψη.

Στην περίπτωση που ο αλγόριθμος υπολογίζει ένα *distortion unit* μεγαλύτερο από το *distortion limit*, τότε στο πίνακα καταχωρείται ένας ξεχωριστός συμβολισμός ως «άπειρο» (*INF*), ώστε να μη ληφθεί υπ' όψη το συγκεκριμένο επεισόδιο στη συγκεκριμένη επανάληψη.

Με βάση λοιπόν τα μόλις προαναφερθέντα, ας εξετάσουμε ορισμένες πιο σύνθετες καταστάσεις, αυτή τη φορά, που μπορούν να λάβουν χώρα εντός της διαδικασίας *Compute_Distortion_Units*. Αν η πρώτη τροχιά χρειάζεται να διευρυνθεί σε 1 υπο-ερώτημα και με την εκτέλεση των κατάλληλων ερωτημάτων, αυτό βρεθεί με βαθμολογία *INF*, καταχωρείται στο πίνακα αυτός ο χαρακτηρισμός και συνεχίζει ο αλγόριθμος στη αμέσως επόμενη τροχιά ιεραρχικά και αυτό συνεχίζεται είτε μέχρι να βαθμολογηθεί υπο-ερώτημα χωρίς *INF*, είτε να τελειώσουν οι τροχιές του πίνακα όπως προαναφέραμε. Αν μια τροχιά που εξετάζεται εκείνη τη δεδομένη στιγμή (είτε είναι η πρώτη, είτε κάποια επόμενη κ.ο.κ) έχει πάνω από ένα υπο-ερώτημα που πρέπει να τροποποιηθεί και η βαθμολογία έστω και στο ένα από τα δύο, είναι *INF*, τότε ο αλγόριθμος συνεχίζει στη επόμενη τροχιά και δεν επιλέγει να σταματήσει εκεί και να διευρύνει αυτό που δεν είχε βαθμολογία *INF*. Αυτό συμβαίνει διότι αφού εν τέλει σκοπός είναι να φθάσει η εκάστοτε τροχιά να έχει «συμμετοχή» σε όλα τα υπο-ερωτήματα, ο αλγόριθμος επιλέγει να μην ασχοληθεί με τροχιές που εκείνη τη δεδομένη στιγμή δεν έχουν «καλή» εικόνα ως προς όλα τα υπο-ερωτήματα στα οποία δεν «συμμετέχουν». Με άλλα λόγια, αν επιλέγαμε να διευρύνουμε το υπο-ερώτημα (εκ των δυο) εκείνο που δεν είχε βαθμολογία *INF* και ολοκληρωνόταν με αυτό τον τρόπο η συγκεκριμένη επανάληψη (γραμμές 7-25), είναι σχεδόν βέβαιο ότι στην επόμενη επανάληψη θα εξεταζόταν ξανά πρώτη η συγκεκριμένη τροχιά και το αποτέλεσμα για το εναπομείναν υπο-ερώτημα θα ήταν το ίδιο, δηλαδή *INF*. Άρα θα έπρεπε να

παρακαμφθεί η τροχιά αυτή και να αναζητηθεί η επόμενη σύμφωνα με τα προαναφερθέντα. Έτσι, αποδεικνύεται προφανέστατα *άσκοπη* η διεύρυνση του πρώτου (εκ των δυο) υπο-ερωτήματος και γι' αυτό το λόγο ο αλγόριθμος λειτουργεί κατ' αυτόν τον τρόπο. Με ακριβώς ανάλογο σκεπτικό αντιμετωπίζεται και η περίπτωση που εξετάζονται 2 τροχιές που ισοβαθμούν στη συχνότητα και δεν «συμμετέχουν» σε τουλάχιστον 2 υπο-ερωτήματα η καθεμιά. Αν έστω και ένα υπο-ερώτημα κάποιας, βαθμολογηθεί με *INF*, τότε σε αυτή την επανάληψη αποκλείεται να ενταχθεί άλλο επεισόδιο αυτής της τροχιάς σε οποιοδήποτε υπο-ερώτημα, όσο καλό *distortion unit* και να έχει λάβει σε σχέση με την άλλη τροχιά.

Με βάση λοιπόν αυτές τις συνθήκες που πρέπει να ισχύουν, ο αλγόριθμος επιλέγει ως προτιμότερο επεισόδιο προς ένταξη, κάθε φορά, αυτό που έχει τη μικρότερη *distortion unit* βαθμολογία (αν υπάρχουν περισσότερα από ένα), αρκεί, την ίδια στιγμή, η ίδια τροχιά να μη έχει σε άλλο υπο-ερώτημα το χαρακτηρισμό *INF* (γραμμή 19).

3^{ο} μέρος Καθώς ολοκληρώνεται η κάθε επανάληψη και εφόσον έχει υπάρξει από την προηγούμενη διαδικασία, επιλογή κατάλληλου επεισοδίου (γραμμή 22), γίνεται διεύρυνση-τροποποίηση του αντίστοιχου υπο-ερωτήματος ώστε να το ενσωματώνει και με αυτό τον τρόπο ολοκληρώνεται η τρέχουσα επανάληψη (γραμμή 24).

Η επανάληψη σταματά (γραμμή 25)
α) όταν βρεθούν *k* τροχιές με συχνότητα ίση με το πλήθος των υπο-ερωτημάτων του αρχικού ερωτήματος ή

β) αν στο τέλος οποιασδήποτε επανάληψης του αλγορίθμου, δεν εντάχθηκε νέο επεισόδιο.

Αφού στο τέλος της κάθε επανάληψης, αλλάζει ένα υπο-ερώτημα επειδή εντάσσεται ένα επεισόδιο σε αυτό, συνεπάγεται ότι στην αρχή της κάθε επόμενης επανάληψης, κατά την εκ νέου συμπλήρωση του πίνακα *H*, υπάρχει η πιθανότητα να συμβούν επιπλέον 2 τινά σε σχέση με την αμέσως προηγούμενη εικόνα του πίνακα:

α) αύξηση της συχνότητας παρουσίας των άλλων τροχιών του πίνακα, επειδή κατά τη διεύρυνση των κριτηρίων του υπο-ερωτήματος ενδέχεται αυτά να «περιλαμβάνουν» διάφορα επιπλέον επεισόδια

β) εμφάνιση νέων τροχιών στον πίνακα κατά την επόμενη επανάληψη.

Έτσι, σε κάθε νέα επανάληψη ως ερώτημα του χρήστη, θεωρούμε πλέον το τρέχον (δηλ. όπως έχει τροποποιηθεί από την προηγούμενη επανάληψη) και όσο δεν απαντά αυτό με *k* τουλάχιστον τροχιές συνεχίζουμε να αναζητούμε το, από κει και πέρα, προσφορότερο επεισόδιο προς ένταξη.

Το *τρίτο* και τελευταίο *στάδιο* του αλγορίθμου αφορά την εκτέλεση ενός πλήθους διαδικασιών που αποσκοπούν στη δημιουργία μιας ασφαλούς απάντησης που επιλύει κατά το δυνατόν, όλα τα τυχόν επιπλέον προβλήματα ασφάλειας που δημιουργήθηκαν από τη λειτουργία του Zoom Out (γραμμές 26-27). Τέτοιου είδους προβλήματα με τις αντίστοιχες προτεινόμενες λύσεις περιγράφονται αναλυτικά στην ενότητα 4.3.8.
Παρατηρούμε, ότι πρόκειται για υλοποίηση ενός *ευρετικού αλγορίθμου* με την έννοια ότι ο αριθμός των περιπτώσεων προς εξέταση είναι απαγορευτικά μεγάλος (όλες οι τροχιές της βάσης) και έτσι δεν είναι δυνατόν να εξετασθούν εξ' αρχής εξαντλητικά όλες οι υποψήφιες λύσεις προκειμένου να επιλεγεί κάποια, που να ικανοποιεί τις προϋποθέσεις που θέτουμε. Δηλαδή, δεν υπάρχει εξ' αρχής εγγύηση ότι βρίσκει γενικά τη βέλτιστη λύση, θεωρώντας ως βέλτιστη λύση να βρεθούν οι τροχιές αυτές που προκαλούν *αποδεδειγμένα* την *συνολικά* μικρότερη αλλοίωση στο αρχικό ερώτημα (με όρους distortion units) προκειμένου αυτό με τη σειρά του να «περιλαμβάνει» στο answer set του τουλάχιστον *k* τροχιές.

Algorithm 1 – Zoom Out

Input: k // *k*-anonymity threshold
Q = <SQ_1,SQ_2, SQ_3,…, SQ_n> // αρχικό ερώτημα με τα υπο-ερωτήματά του (n: αρ. υπο-ερωτ)
D = < > // DataBase
H[tr_id, freq, SubQuery$_1$, SubQuery$_2$, …, SubQuery$_n$] // ο βοηθητικός πίνακας

Output: F_Q = < F_SQ_1, F_SQ_2, F_SQ_3,…, F_SQ_n > // τελικό ερώτημα με τα υπο-ερωτήματά του

```
1:  F_Q ← Q                 // αρχικοποίηση
2:  H ← 0                   // αρχικοποίηση
3:  Ntr ← Count(F_Q)        // συνολικός αριθμός αρχικών τροχιών
4:  Find_most_numerous_SQ(in F_Q, out F_SQi)
    // βρίσκει ποιο υπο-ερώτημα έχει τις περισσότερες τροχιές
5:  if Count(F_SQi) < k then
6:      SQ_Expansion(in k, in out F_SQi)
        // ενσωματώνει τυχόν επεισόδια στο πολυπληθέστερο υπο-ερώτημα
        // ώστε να περιλαμβάνει τουλάχιστον k τροχιές
7:  repeat
8:    Something_Changed ← False
9:    for i=1 to n do
10:   begin
11:     Execute_Query(in F_SQi out tr_ids)
        // επιστρέφει όλα τα tr_ids που αντιστοιχούν στην εκτέλεση του κάθε υπο-ερωτήματος
12:     Fill_Help_Table(in H, tr_ids out H)
        // προσθέτει στον πίνακα τα tr_ids & ανανεώνει το πλήθος των συχνοτήτων
13:   end
14:   Sort_Help_Table(in out H)
15:   Find_Freq_Position(in H, n out i)
      // βρίσκει τη θέση του πίνακα που να ισχύει ότι freq = max(frequencies) και freq < n
16:   episode_found ← False
17:   repeat
18:     Compute_Distortion_Units(in out episode_found, H, i)
        // συμπληρώνεται ο πίνακας με τους υπολογισμούς των distortion units
        // και επιλέγονται το ή τα κατάλληλα υποψήφια επεισόδια
19:     Select_best_candidate_episode(in H, i out tr_id, ep_id)
        // γίνεται η επιλογή του πλέον συμφέροντος επεισοδίου (αν είναι πάνω από ένα)
20:     i ← i + 1
21:   until episode_found or EOF  // End Of File
      // η επανάληψη τελειώνει αν βρεθεί κατάλληλο επεισόδιο από την προηγ. διαδικασία ή
      // αν τελειώσουν τα δεδομένα (τροχιές) του πίνακα
22:   if episode_found then
23:     Embed_New_Episode(in H, tr_id, ep_id, in out F_SQi, Something_Changed)
        // τροποποίηση των κριτηρίων του υπο-ερωτήματος ως προς χώρο, χρόνο και/ή tags,
        // βασιζόμενη στο επεισόδιο που επιλέχθηκε
24:   Ntr ← Count(F_Q)
25: until (not Something_Changed) or (Ntr=k)
26: Compute_Random_Number(in Rmin, Rmax out R)
    //
27: Compute_New_Episodes(in F_Q, R out F_Q)
    //
28: return F_Q
```

Σχήμα 4.4 – Ο αλγόριθμος Zoom Out

4.3.7 Παράδειγμα λειτουργίας

Στη συνέχεια, παρατίθεται ένα παράδειγμα στο οποίο επεξηγείται πιο συγκεκριμένα η εξωτερική επανάληψη του αλγορίθμου (γραμμές 7-25). Για λόγους απλότητας και σαφήνειας του παραδείγματος, προτιμήσαμε να μην συμπεριλάβουμε στην συγκεκριμένη αναζήτηση, επεισόδια με διαφορετικά tags από τα κριτήρια του κάθε υπο-ερωτήματος, δηλαδή ότι λ=1. Επίσης για λόγους απλότητας και επειδή είναι αδιάφορο στην επεξήγηση που θα ακολουθήσει, δεν αναφέρονται στους πίνακες που αποτυπώνουν το παράδειγμα (πίνακες 4.1 ως 4.5) οι κωδικοί (ids) των εκάστοτε επεισοδίων.

Έστω ότι ισχύει ένα *k*-anonymity=3 και ένας χρήστης θέτει ένα ερώτημα Q_1 στη βάση και ζητάει τροχιές που ικανοποιούν τα χωροχρονικά κριτήρια 4 υπο-ερωτημάτων, ότι δηλαδή αναζητούνται τροχιές των οποίων αντίστοιχα επεισόδια, βρίσκονταν κατά το Δt_1 στο $Region_1$, κατά το Δt_2 στο $Region_2$, κατά το Δt_3 στο $Region_3$ και κατά το Δt_4 στο $Region_4$. Το πλήθος των τροχιών που ικανοποιούν τα πιο πάνω κριτήρια είναι μικρότερο του *k*-anonymity και έτσι ο μηχανισμός δεν απαντά το ερώτημα αλλά το δίνει ως είσοδο στον zoom_out. Έστω ότι το καθένα από τα 4 υπο-ερωτήματα εκτελούμενα μεμονωμένα «περιέχουν» ένα αριθμό επεισοδίων το καθένα που αντιστοιχούν σε ορισμένες τροχιές, δηλαδή σε ορισμένα tr_ids.

Σε κάθε επανάληψη, ενσωματώνεται τουλάχιστον ένα νέο επεισόδιο διευρύνοντας κάποιο από τα υπο-ερωτήματα είτε χωρικά είτε χρονικά είτε ταυτόχρονα χωρικά και χρονικά.

Πιο συγκεκριμένα ένα υποθετικό αρχικό στιγμιότυπο του πίνακα θα μπορούσε να είναι το εξής: (k=3, Number_of_subqueries=4)

Tr_id	**Freq**	**Distortion Units**								
		Sp	**Temp**	**Total**	**Sp**	**Temp**	**Total**	**Sp**	**Temp**	**Total**
A	**4**	Ok								
B	2	INF	INF	INF	-	-	-	0.2	0	0.2
C	2	0.25	0	**0.25**	0.3	0.7	1.0	-	-	-
D	1	-	-	-	-	-	-	-	-	-
E	1	-	-	-	-	-	-	-	-	-
F	1	-	-	-	-	-	-	-	-	-

Πίνακας 4.1 – $1^{ο}$ στιγμιότυπο του παραδείγματος

Βλέπουμε ότι η τροχιά Α ήδη εμφανίζεται 4 φορές, όσο και το πλήθος των υπο-ερωτημάτων. Άρα αναζητούνται άλλες 2, αφού θέλουμε να φθάσουμε τις 3 τροχιές (*k*=3). Παρατηρούμε ότι οι τροχιές Β, C έχουν την αμέσως μικρότερη συχνότητα, δηλαδή 2. Αυτό σημαίνει ότι υπάρχουν άλλα 2 υπο-ερωτήματα των οποίων τα κριτήρια δεν ικανοποιούνται, δηλαδή οι τροχιές Β και C δεν συμμετέχουν με αντίστοιχο επεισόδιό τους. Η τρίτη στήλη τηρεί τα *Distortion units* που υπολογίστηκαν και ουσιαστικά είναι για κάθε τροχιά ένας ξεχωριστός υπο-πίνακας τόσων θέσεων όσο είναι η μέγιστη δυνατή συχνότητα εμφάνισης – 1. Έτσι, επειδή *k*-1=3, στους πίνακες του παραδείγματος παρατηρούμε 3 υπο-πίνακες εντός του πεδίου *Distortion Units.*

Εσωτερικά ο κάθε υποπίνακας χωρίζεται σε 3 πεδία. Το πρώτο αφορά τη μέτρηση της χωρικής αλλοίωσης βάσει του κοντινότερου επεισοδίου (αν υπάρχει), το δεύτερο τη μέτρηση της χρονικής αλλοίωσης που προκαλείται από το ίδιο και το τρίτο, το άθροισμα αυτών. Έτσι, για την τροχιά Β (πίνακας 4.1), ως προς το πρώτο υπο-ερώτημα δεν βρέθηκε να έχει κοντινό επεισόδιο (INF), ενώ για το επόμενο βρήκε ένα με βαθμολογία 0,2. Αντίστοιχα, το επεισόδιο της τροχιάς C, για το πρώτο από τα 2 υπολειπόμενα υπο-ερωτήματα, βρέθηκε με βαθμολογία 0,25. Ενώ το επεισόδιο της τροχιάς Β έχει μικρότερη βαθμολογία, *δεν επιλέγεται*, διότι εκείνη τη στιγμή έχει βρεθεί άπειρη η απόστασή του από ένα από τα 2 υπο-ερωτήματα και επιλέγεται να ενταχθεί το επεισόδιο της τροχιάς C (βαθμολογία 0,25), χωρίς να είναι απαραίτητο ότι στις επόμενες επαναλήψεις θα παραμείνει άπειρη η, εν λόγω, απόσταση.

Ας παρακολουθήσουμε πώς θα μπορούσε υποθετικά να συνεχίσει ο αλγόριθμος στον πίνακα 4.2.

Tr_id	**Freq**	**Distortion Units**								
		Sp	**Temp**	**Total**	**Sp**	**Temp**	**Total**	**Sp**	**Temp**	**Total**
A	**4**	Ok								
C	3	-	-	-	0.3	0.7	**1.0**	-	-	-
B	2	-	-	-	-	-	-	-	-	-
E	2	-	-	-	-	-	-	-	-	-
D	1	-	-	-	-	-	-	-	-	-
F	1	-	-	-	-	-	-	-	-	-
G	1	-	-	-	-	-	-	-	-	-

Πίνακας 4.2 – 2^{ο} στιγμιότυπο του παραδείγματος

Εδώ, η τροχιά C καταλαμβάνει πλέον τη 2η θέση και έχει πλέον μόνο ένα υπο-ερώτημα στο οποίο δεν έχει «παρουσία», αλλά υπάρχει ένα επεισόδιό της σε μια «αποδεκτή» απόσταση και χωρίς να πρέπει/μπορεί να συγκριθεί με επεισόδιο άλλης τροχιάς.
Επίσης, χαρακτηριστικό είναι ότι, με την προηγούμενη διεύρυνση, εμφανίστηκε μια νέα τροχιά (G) καθώς και ότι αυξήθηκε η συχνότητα της τροχιάς E.
Η συνέχεια θα μπορούσε να διαμορφωθεί ως εξής (πίνακας 4.3):

Tr_id	**Freq**	**Distortion Units**								
		Sp	**Temp**	**Total**	**Sp**	**Temp**	**Total**	**Sp**	**Temp**	**Total**
A	**4**	Ok								
C	**4**	Ok								
B	2	0.8	0	0.8	-	-	-	0.2	0	**0.2**
E	2	0.2	0.3	0.5	INF	INF	INF	-	-	-
D	1	-	-	-	-	-	-	-	-	-
F	1	-	-	-	-	-	-	-	-	-
G	1	-	-	-	-	-	-	-	-	-

Πίνακας 4.3 – 3ο στιγμιότυπο του παραδείγματος

Παρατηρούμε ότι συγκρίνονται πιθανά επεισόδια των τροχιών B και E ως προς τα αντίστοιχα υπο-ερωτήματα και ότι, ενώ αρχικά το ένα εκ των δύο επεισοδίων της B, βρέθηκε άπειρο, τώρα είναι πλέον εντός των αποδεκτών ορίων και συγκρίνεται ισότιμα με τα υπόλοιπα επεισόδια.
Διαπιστώνουμε και πρακτικά ότι όταν αλλάζει ένα υπο-ερώτημα επειδή εντάσσεται σε αυτό ένα επεισόδιο, υπάρχει η πιθανότητα να αυξηθεί η συχνότητα εμφάνισης και άλλων τροχιών πέραν από αυτή στην οποία ανήκει το επεισόδιο ή ακόμα και να εμφανιστούν νέες υποψήφιες τροχιές στη λίστα.

Έστω ότι το επόμενο στιγμιότυπο είναι το εξής (πίνακας 4.4):

Tr_id	Freq	Distortion Units								
		Sp	Temp	Total	Sp	Temp	Total	Sp	Temp	Total
A	**4**	Ok								
C	**4**	Ok								
B	3	0.8	0	0.8	-	-	-	-	-	-
E	2	-	-	-	-	-	-	-	-	-
D	1	-	-	-	-	-	-	-	-	-
F	1	-	-	-	-	-	-	-	-	-
G	1	-	-	-	-	-	-	-	-	-

Πίνακας 4.4 – 4ο στιγμιότυπο του παραδείγματος

Εντάσσεται ένα επεισόδιο της τροχιάς B, με τη συχνότητά της να ανεβαίνει στο 3 και να είναι η μοναδική που εξετάζεται αρχικά.

Αφού υπάρχει επεισόδιο στο υπολειπόμενο υπο-ερώτημα που δεν έχει τιμή «άπειρο», τότε εντάσσεται και αυτό με τη σειρά του στο υπο-ερώτημα που αντιστοιχεί και το διευρύνει ανάλογα.

Η τελική εικόνα είναι αυτή (πίνακας 4.5):

Tr_id	Freq	Distortion Units								
		Sp	Temp	Total	Sp	Temp	Total	Sp	Temp	Total
A	**4**	Ok								
C	**4**	Ok								
B	**4**	Ok								
E	2	-	-	-	-	-	-	-	-	-
D	1	-	-	-	-	-	-	-	-	-
F	1	-	-	-	-	-	-	-	-	-
G	1	-	-	-	-	-	-	-	-	-

Πίνακας 4.5 – 5ο στιγμιότυπο του παραδείγματος

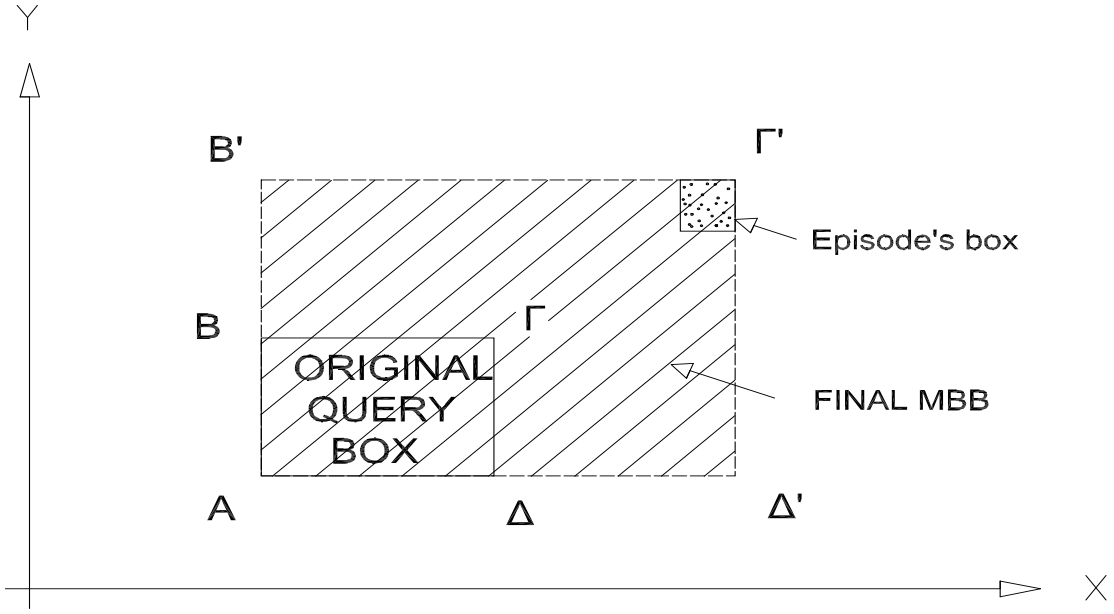
Αξίζει να αναφέρουμε ότι, αν στο προτελευταίο στιγμιότυπο (πίνακας 4.4), η τροχιά Β είχε πάρει τιμή «άπειρο» για το τελευταίο της υπο-ερώτημα, τότε ο αλγόριθμος θα αναζητούσε υποψήφια επεισόδια για τη τροχιά Ε και αν δεν βρίσκονταν ούτε γι' αυτή, τότε θα αναζητούσε μαζικά για τις τροχιές D, F και G μέσω της μίνι επαναληπτικής διαδικασίας που προβλέπεται στον αλγόριθμο (γραμμές 17-21).

Ουσιαστικά, η συνολική «παραγωγή» του *κάθε* στιγμιότυπου αντιστοιχεί σε *μία* επανάληψη (γραμμές 7-25) του αλγορίθμου όπως αυτή έχει διατυπωθεί, όμως εντός της κάθε επανάληψης ο αλγόριθμος εξετάζει την/τις τροχιές με τη μεγαλύτερη εκείνη τη στιγμή συχνότητα και όσο δεν βρίσκει μη «άπειρες» τιμές, συνεχίζει στις τροχιές με την αμέσως μικρότερη συχνότητα κ.ο.κ μέχρι να εξαντληθούν οι τροχιές του πίνακα, μέσω της εσωτερικής επανάληψης (γραμμές 17-21).

4.3.8 Προβλήματα ασφαλείας

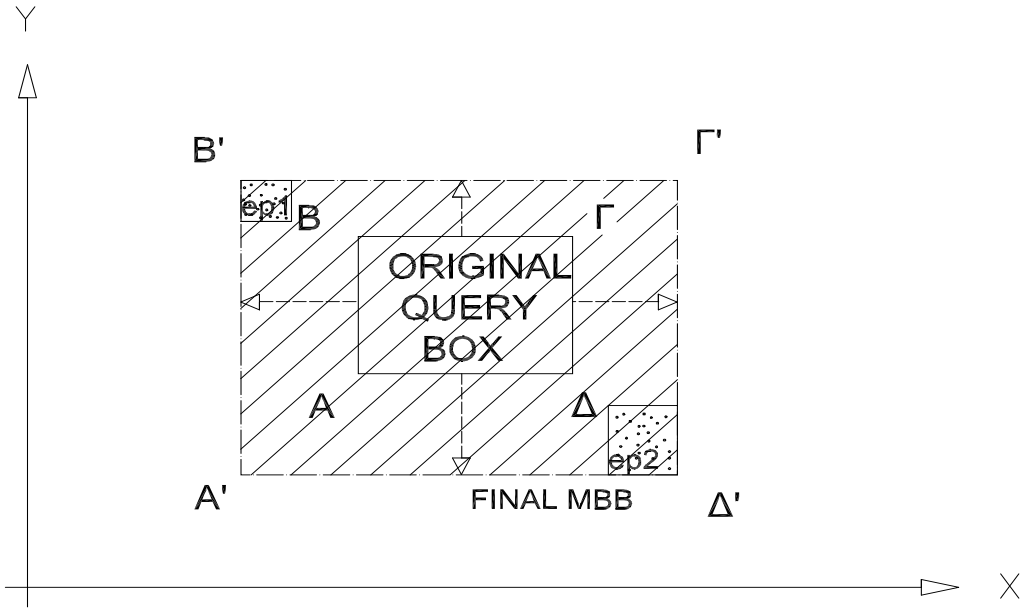
Χωρική επέκταση Ένα πρόβλημα που ενδέχεται να προκύψει αφορά την περίπτωση που ισχύει ένα 2-anonymity. Στην περίπτωση αυτή, ο μηχανισμός (Zoom Out) ενεργοποιείται όταν, το αρχικό ερώτημα που τίθεται, δύναται να επιστρέψει ως απάντηση βάσει των κριτηρίων του, καμία ή μόλις μία τροχιά. Για λόγους απλότητας και κατανόησης του προβλήματος, θεωρούμε ότι το ερώτημα αποτελείται από ένα μόνο υπο-ερώτημα. Ας δούμε πιο προσεκτικά τι συμβαίνει στην περίπτωση που υπάρχει καταγεγραμμένο *ένα* επεισόδιο στον αρχικό χώρο που ορίζεται από το αντίστοιχο κριτήριο του ερωτήματος και εκτελείται ο αλγόριθμος. Έστω ότι αυτός διενεργεί μια χωρική διεύρυνση και εντοπίζει ένα επιπλέον επεισόδιο που είναι αναγκαίο προκειμένου να καλύπτεται η απαίτηση του 2-anonymity. Ανακατασκευάζει το χωρικό κριτήριο του ερωτήματος ώστε να περιλαμβάνει εκτός από το αρχικό κριτήριο και το νέο επεισόδιο που βρέθηκε. Στην περίπτωση όμως αυτή, το νέο ορθογώνιο παραλληλόγραμμο που σχηματίζεται είναι μια επέκταση του αρχικού προς μια συγκεκριμένη κατεύθυνση, έτσι ώστε να συμπεριλαμβάνει το νέο επεισόδιο. Παρατηρώντας το σχήμα 4.5, έστω ότι ΑΒΓΔ το χωρικό κριτήριο του αρχικού ερωτήματος και ΑΒ'Γ'Δ' το τελικό χωρικό κριτήριο μετά την εκτέλεση του *Zoom Out*. Από τα 4 σημεία που προσδιορίζουν το αρχικό κριτήριο (δηλ. τα Α, Β, Γ και Δ), στο

σημείο Α δεν μεταβάλλονται οι συντεταγμένες, στα Β και Δ μεταβάλλεται μόνο η μία εκ των δύο συντεταγμένων κάθε φορά, ενώ στο σημείο Γ μεταβάλλονται και οι δύο συντεταγμένες. Είναι έτσι προφανές για τον χρήστη ότι σίγουρα το σημείο του οποίου μεταβλήθηκαν και οι 2 συντεταγμένες (στο παράδειγμα, το σημείο Γ), συμπεριλαμβάνεται σίγουρα στο νέο επεισόδιο που προστεθήκε ως απάντηση στο τροποποιημένο, από τον *Zoom Out*, ερώτημα του χρήστη. Αυτό προκαλεί παραβίαση της ιδιωτικότητας γιατί ο χρήστης γνωρίζει πλέον με ακρίβεια την περιοχή ενός εκ των δύο επεισοδίων.



Σχήμα 4.5 – Αποκάλυψη θέσης του επεισοδίου στη περίπτωση του *2*- anonymity

Ένα ακόμη πιο δυσμενές σενάριο προκύπτει όταν κανένα επεισόδιο δεν περιλαμβάνεται στο αρχικό ερώτημα και ο αλγόριθμος αναζητά 2 επεισόδια να προσθέσει ως απάντηση. Στη περίπτωση αυτή κατασκευάζεται από τον αλγόριθμο ένα μεγαλύτερο ορθογώνιο. Αν αυτό *έχει προεκταθεί προς κάθε μια από τις 4 κατευθύνσεις (πλευρές)* και αφού ο χρήστης γνωρίζει ότι αυτό πλέον περιλαμβάνει 2 επεισόδια, είναι προφανές ότι αυτά είναι τοποθετημένα σε απέναντι γωνίες. Χαρακτηριστικά στο σχήμα 4.6 παρατηρούμε ότι πρόκειται είτε για τις γωνίες Α' – Γ', είτε για τις γωνίες Β' – Δ' γεγονός που αποτελεί παραβίαση της ιδιωτικότητας.



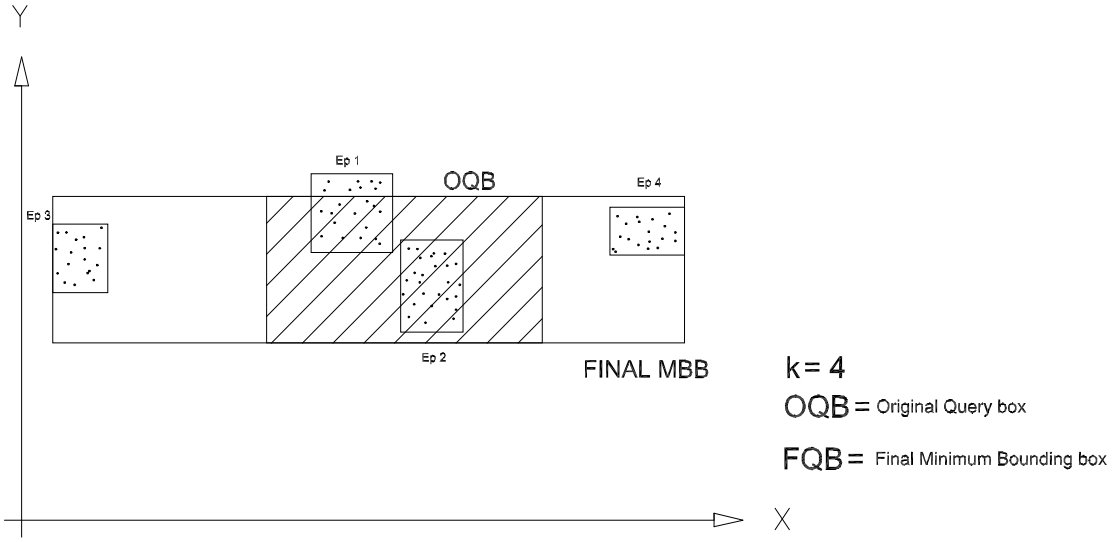
Σχήμα 4.6 – Αποκάλυψη θέσεων επεισοδίων στη περίπτωση του *2*- anonymity

Μια τρίτη περίπτωση που χρήζει προσοχής αφορά κατά κανόνα Stop episodes, των οποίων τα MBB's που τα οριοθετούν χωρικά έχουν δύο διαφορές σε σχέση με αυτά των Move episodes, γι' αυτό και μας απασχολούν στη συγκεκριμένη περίπτωση. Αυτές είναι ότι είναι πολύ μικρότερα σε έκταση σε σχέση με των Move episodes αλλά και ότι οι πλευρές του MBB τους, μεταξύ τους, δεν απέχουν κατ' αναλογία σε μήκος ιδιαίτερα. Δηλαδή αν η μια πλευρά έχει μήκος a, τότε η άλλη δεν ξεπερνά τις περισσότερες φορές το 2*a με 3*a.

Αυτό που μελετούμε είναι να συμβεί το ενδεχόμενο τα επεισόδια που θα πρέπει προστεθούν από τον αλγόριθμο για να φθάσουν όλα μαζί *k* τον αριθμό, να σχηματίζουν μια παράλληλη, σε γενικές γραμμές, ευθεία με οποιονδήποτε από τους δύο άξονες συντεταγμένων.

Αυτό θα είχε σαν αποτέλεσμα, το νέο MBB που σχηματίζεται για να τα περιλάβει, όπως βλέπουμε και στο σχήμα 4.7, να έχει πολύ μεγαλύτερη σε μήκος την μια του πλευρά σε σχέση με την άλλη. Αυτό από μόνο δεν αποτελεί απαραίτητα πρόβλημα. Ο ιδιαίτερος προβληματισμός δημιουργείται αν η μικρή σε μήκος πλευρά βρίσκεται πολύ κοντά στο μέσο μήκος των πλευρών αντίστοιχων επεισοδίων ως τάξη μεγέθους, τότε κάποιος θα δικαιούται να υποθέσει ότι δίπλα ακριβώς και στις δύο μικρές πλευρές του παραλληλογράμμου, υπάρχουν επεισόδια που προστέθηκαν για να επιτευχθεί το k-anonymity. Ως εκ τούτου, αποκαλύπτεται η σχεδόν ακριβής θέση ενός ή δύο

επεισοδίων από τα *k* εντός του νέου MBB που κατασκεύασε ο αλγόριθμος και έχουμε πάλι παραβίαση της ιδιωτικότητας.



Σχήμα 4.7 - Αποκάλυψη θέσεων επεισοδίων στη περίπτωση υπερβολικής ανομοιογένειας μεταξύ του μήκους των 2 πλευρών του σχηματιζόμενου MBB

Η πρότασή μας για αντιμετώπιση των απειλών που μόλις περιγράψαμε είναι εμπνευσμένη από διάφορες τεχνικές που έχουν χρησιμοποιηθεί για την προστασία της ιδιωτικότητας στα LBS [6]. Ενδιαφερόμαστε ειδικότερα για τη περίπτωση που στα LBS γίνεται προσπάθεια απόκρυψης (cloaking) της θέσης του αιτούντος μιας υπηρεσίας, μέσω της μετατροπής της σε μια γενικευμένη χωρική περιοχή.

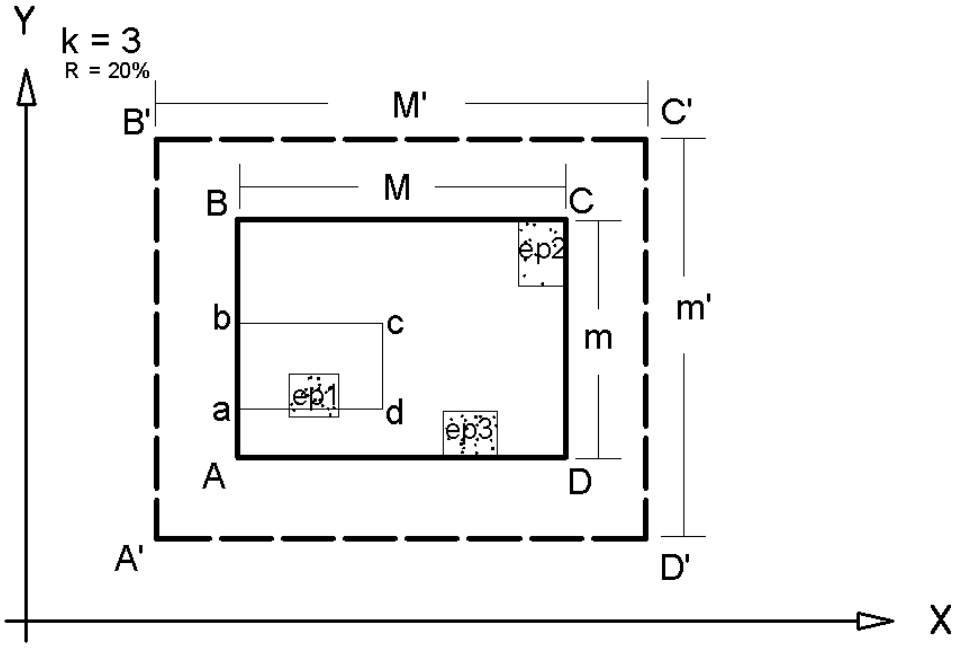
Κατ' αναλογία, προσπαθούμε να γενικεύσουμε το MBB που έχει δημιουργήσει ο αλγόριθμος και που, πλέον, περιλαμβάνει τουλάχιστον *k* επεισόδια, προεκτείνοντάς το προς *κάθε* κατεύθυνση, με σκοπό καθαρά τη δημιουργία μιας *«ζώνης αβεβαιότητας»* περιμετρικά του MBB. Επιπλέον, για να περιοριστεί περισσότερο το πρόβλημα που αφορούσε τα MBB με τις ιδιαίτερα άνισες πλευρές (Σχήμα 4.7), επιλέγουμε η διεύρυνση των πλευρών του MBB να μην γίνεται κατά το ίδιο ποσοστό αλλά να μεγαλώνει πάντα κατ' αναλογία περισσότερο η, μικρότερη σε μήκος, πλευρά. Το ποσοστό αύξησης του μήκους των πλευρών είναι ένας αριθμός που επιλέγεται τυχαία κάθε φορά μεταξύ 2 αριθμών που δίνονται ως παράμετροι από τους διαχειριστές του συστήματος. Ο λόγος είναι απλά ότι αφού θεωρούμε τον αλγόριθμο γνωστό στο καθένα, δεν είναι δυνατό για κάθε ερώτημα να τροποποιούνται οι πλευρές του MBB

πάντα κατά το ίδιο ποσοστό. Η κάθε πλευρά προεκτείνεται και προς τα δύο άκρα της και μάλιστα ισόποσα.

Η διαδικασία αυτή λαμβάνει χώρα εντός της διαδικασίας *Compute_New_Episodes* (γραμμή 27) και όταν αυτή εκτελείται για να επεξεργαστεί ένα επεισόδιο το οποίο διευρύνθηκε (τουλάχιστον) χωρικά από τις προηγούμενες διεργασίες, εκτελεί μια σειρά βημάτων ώστε να δημιουργηθεί αυτό που ορίσαμε ως *«ζώνη αβεβαιότητας»*.

Πιο συγκεκριμένα (σχήμα 4.8), έστω ένα MBB διαστάσεων M x m, όπου M > m. Ο αλγόριθμος εκτελεί τη διαδικασία *Compute_Random_Number* (γραμμή 26) και υπολογίζει ένα τυχαίο αριθμό **R** ως το ποσοστό διεύρυνσης, μεταξύ των 2 αριθμών $\mathbf{R}_{min}$ και $\mathbf{R}_{max}$ οι οποίοι εισάγονται ως παράμετροι στη, εν λόγω, ,διαδικασία.

Στη συνέχεια εκτελεί τη διαδικασία *Compute_New_Episodes* η οποία διευρύνει το ζεύγος των μεγαλύτερων πλευρών κατά το ποσοστό αυτό και ώστε να ισχύει ότι M’ = M + M***R**, κατόπιν διευρύνει το ζεύγος των μικρότερων πλευρών τόσο όσο διευρύνθηκαν οι μεγαλύτερες πλευρές σε απόλυτες τιμές ώστε να ισχύει ότι m’ = m + M***R** και τέλος κατασκευάζει το τελικό MBB διαστάσεων M’ x m’.



Σχήμα 4.8 – Γενίκευση του MBB που κατασκευάστηκε ώστε να δημιουργηθεί η «ζώνη αβεβαιότητας»

Το αποτέλεσμα είναι ότι αν μια μόνο πλευρά του αρχικού box είχε μετατοπιστεί χωρικά για να κατασκευαστεί το MBB, δεν σημαίνει *πλέον* ότι αυτή *αναγκαστικά* εφάπτεται ή τέμνεται οπουδήποτε κατά μήκος της με την περίμετρο ενός επεισοδίου από αυτό ή αυτά που *περιλάβαμε* για τη δημιουργία του τελικού MBB.

Χρονική επέκταση Στην περίπτωση της χρονικής επέκτασης μπορεί επίσης να υπάρξει παραβίαση της ιδιωτικότητας. Αυτό συμβαίνει, διότι ο χρήστης δύναται να αποκτήσει περισσότερη γνώση για μια κατάσταση. Πιο συγκεκριμένα, όταν ενσωματώνεται ένα επεισόδιο, δηλαδή αλλάζει η δομή του ανάλογου υπο-ερωτήματος, τότε ο χρήστης αυτόματα γνωρίζει ότι το τροποποιημένο(α) σκέλος(η) του χρονικού διαστήματος είναι η χρονική αφετηρία ή η χρονική κατάληξη σίγουρα ενός συγκεκριμένου επεισοδίου.
Για παράδειγμα, αν ο χρήστης ζητήσει τροχιές για μια περιοχή, με συγκεκριμένο(α) tag(s) και για το χρονικό διάστημα 10:00 – 11:00 και έλθει απάντηση που όμως πλέον αφορά το διάστημα 10:00 – 11:40, τότε ο χρήστης γνωρίζει *σίγουρα ότι τουλάχιστον ένας και κατά πάσα πιθανότητα μόνο ένας* από τους *k* παρακολουθούμενους τελείωνε η καταγραφή του δικού του επεισοδίου στις 11:40! Αυτό δημιουργεί μια *αύξηση της γνώσης* που αποκτά ένας, ενδεχομένως, κακόβουλος χρήστης της βάσης, για *μια συγκεκριμένη τροχιά* γεγονός που αντιβαίνει στην αρχή του *k*-anonymity.
Μια λύση που μπορεί να προταθεί και να διεκπεραιωθεί εντός της διαδικασίας *Compute_New_Episodes* (γραμμή 27) είναι να μην συμπεριλαμβάνεται ολόκληρο το χρονικό διάστημα του, υπό ένταξη επεισοδίου, στο υπο-ερώτημα, αλλά μόνο το ήμισυ του χρονικού διαστήματος αυτού. Έτσι, επιστρέφοντας στο προηγούμενο παράδειγμα και εφόσον 10:00 – 11:40 είναι το διευρυμένο χρονικό διάστημα, δεν μπορεί κανείς να πει με σιγουριά τίποτα για τη μια από τις n, ενδεχομένως, τροχιές που προστέθηκαν και αυτό γιατί η στιγμή 11:40 θα αποτελεί το χρονικό *επίκεντρο* του επεισοδίου δηλαδή μπορεί να εκφράζει τα πιθανά επεισόδια 11:10 – 12:10 ή 11:20 – 12:00 ή 11:30 – 11:50 κ.ο.κ Όμως το να γνωρίζει κανείς το επίκεντρο, είναι σαφώς πιο ασαφής πληροφορία για τον κακόβουλο χρήστη από ότι να γνωρίζει την έναρξη ή τη λήξη συγκεκριμένου επεισοδίου.

4.3.9 Μελλοντικές βελτιώσεις

Ο αλγόριθμος αυτός μπορεί να παραμετροποιηθεί στο μέλλον, εφόσον κριθεί απαραίτητο, ως προς τα εξής σημεία:

α) Η μέγιστη επιτρεπτή διεύρυνση κάθε φορά ενός υπο-ερωτήματος να μην είναι ένας προκαθορισμένος και συγκεκριμένος αριθμός αλλά να υπολογίζεται δυναμικά προσαρμοζόμενη στα τρέχοντα δεδομένα της εκάστοτε βάσης
β) να υπάρχει μια διαφοροποίηση υπέρ της χρονικής ή της χωρικής αλλοίωσης που υπολογίζεται για ένα, υπό ένταξη, επεισόδιο μέσω ενός επιπλέον συντελεστή βαρύτητας.

4.4 Μηχανισμός ελέγχου ερωτημάτων

4.4.1 Εισαγωγή

Στην παρούσα ενότητα παρουσιάζονται το πλαίσιο και οι τεχνικές αντιμετώπισης επιθέσεων σε μια βάση δεδομένων που φιλοξενεί σημασιολογικά εμπλουτισμένες τροχιές κινούμενων αντικειμένων. Το είδος των ερωτημάτων που μπορούν να τεθούν σε αυτή τη βάση περιγράφονται επακριβώς στο κεφ. 1. Οι επιθέσεις που καλούμαστε να αντιμετωπίσουμε είναι ακριβώς αυτές που περιγράφονται αναλυτικά στην ενότητα 3.3.
Ο συγκεκριμένος μηχανισμός που παρουσιάζουμε αποτελεί το τελευταίο στάδιο του ευρύτερου μηχανισμού χειρισμού / έλεγχου των ερωτημάτων των χρηστών σε βάσεις δεδομένων αυτού του είδους (σχήμα 4.1) και εκτελείται πάντα πριν απαντηθεί τελικά ένα ερώτημα και όχι όπως στην περίπτωση του *Zoom Out* που εκτελείται κατά περίπτωση.
Σκοπός του είναι να μην επιτρέπει την προώθηση της απάντησης ενός ερωτήματος στον χρήστη που έθεσε το ερώτημα όταν διαπιστώνεται ότι, με αυτό τον τρόπο, αυτός θα αποκτήσει (περαιτέρω) γνώση για μια κατάσταση ή για μια καταγεγραμμένη οντότητα στη βάση όπως περιγράφεται στη ενότητα 3.2. Ο τρόπος για να συμβεί αυτό είναι ο

χρήστης να υποστεί *άρνηση της παρεχόμενης υπηρεσίας*, δηλαδή να του μεταδίδεται ένα μήνυμα ότι δεν μπορεί να δοθεί απάντηση στο ερώτημα του.

Όπως περιγράφεται στην ενότητα 3.3, ένας χρήστης προκειμένου να «υλοποιήσει» μια, εν δυνάμει, επίθεση, διενεργεί μια σειρά 2 ή περισσοτέρων ερωτημάτων με κάποια κοινά χαρακτηριστικά μεταξύ τους. Από αυτό συμπεραίνουμε ότι για τη λειτουργία του μηχανισμού που θα παρουσιάσουμε, μας είναι απαραίτητες *όλες* οι εγγραφές της βάσης που αφορούν ερωτήματα που έχει θέσει ο *ίδιος* χρήστης, θεωρώντας βέβαια διαφορετικούς χρήστες το κάθε διαφορετικό username/user id που χρησιμοποιείται για την πρόσβαση στη βάση.

Έτσι, ως είσοδος του παρόντος μηχανισμού είναι η ίδια η βάση, το όριο *k*-anonymity που έχει επιλεγεί, το user id του χρήστη που διενεργεί το ερώτημα και βέβαια το ίδιο το ερώτημα που αυτός έθεσε. Η έξοδος είναι η βάση όπως αυτή μπορεί να έχει εμπλουτιστεί (και θα εξηγήσουμε στην ενότητα 4.4.3 γιατί μπορεί να συμβεί αυτό) και το βασικότερο, μια Boolean μεταβλητή που εκφράζει αν τελικά κρίνεται ότι μπορεί ή όχι να απαντηθεί με ασφάλεια το ερώτημα που τέθηκε ως είσοδος.

Η στόχευσή μας στη παρούσα προσέγγιση, αλγοριθμικά, όπως θα φανεί και στη συνέχεια, ήταν να ελαχιστοποιηθούν οι «τελεσίδικες» αρνήσεις στην απάντηση ερωτημάτων, η μεγιστοποίηση της φιλικότητας προς τον χρήστη καθώς και η προσπάθεια για γενικότερη οικονομία στους πόρους του υπολογιστικού συστήματος που θα κληθεί να υποστηρίξει την αλγοριθμική αυτή προσέγγιση.

Με βάση την ανάλυση επί των επιθέσεων, αποφασίζεται ο αλγόριθμος να προσπαθεί να εντοπίσει

1) Χωρικά ολικώς επικαλυπτόμενα ερωτήματα
2) Χρονικά ολικώς επικαλυπτόμενα ερωτήματα
3) Ολικώς επικαλυπτόμενα ερωτήματα ως προς τα tags
4) Μερικώς επικαλυπτόμενα ερωτήματα ως προς το χώρο ή το χρόνο
5) Ολικώς επικαλυπτόμενα ερωτήματα ως προς το πλήθος των υπο-ερωτημάτων

Συνυπολογίζουμε πάντα ότι στις περιπτώσεις (1) ως (4) ο αριθμός των υπο-ερωτημάτων μεταξύ των, υπό σύγκριση, ερωτημάτων, πρέπει να είναι ίδιος, ενώ για την περίπτωση (5) το αντίθετο. Επίσης, όταν αναφερόμαστε σε επικάλυψη (μερική ή ολική) ως προς το

ένα κριτήριο, θεωρούμε ότι εκείνη τη στιγμή, όλα τα άλλα πρέπει να ταυτίζονται. Όλες αυτές οι περιπτώσεις συνοψίζονται στον πίνακα 4.6.

Περίπτωση	Κριτήριο	Επικάλυψη		Πλήθος υπο-ερωτημάτων	
		Ολική	Μερική	Ίδιο	Διαφορετικό
a	Χώρος	✓	✓	✓	-
b	Χρόνος	✓	✓	✓	-
c	Tags	✓	-	✓	-
d	Υπο-ερωτήματα	✓	-	-	✓

Πίνακας 4.6 – Συγκεντρωτική απεικόνιση των περιπτώσεων προς διερεύνηση

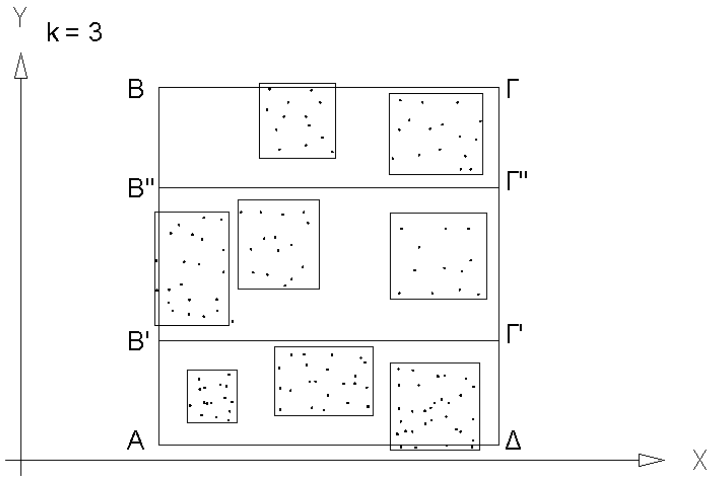
Θυμίζουμε, τέλος, ότι όταν μιλούμε για ολική επικάλυψη μεταξύ ίδιου αριθμού υπο-ερωτημάτων (περιπτώσεις a, b, c του πίνακα 4.6), εννοούμε ότι δεν μπορεί (δηλ. δεν το θεωρούμε δυνατότητα επίθεσης), για παράδειγμα, το ένα υπο-ερώτημα του $1^{ου}$ ερωτήματος να επικαλύπτει ένα υπο-ερώτημα του $2^{ου}$ ερωτήματος και ταυτόχρονα ένα άλλο υπο-ερώτημα του $1^{ου}$ ερωτήματος να επικαλύπτεται από ένα υπο-ερώτημα του $2^{ου}$ ερωτήματος.

4.4.2 Ελαστικοποίηση του μηχανισμού

Η ιδέα Εμπνεόμενοι από διάφορα σημεία της εργασίας των *Adam & Wortmann* [3], συνδυαστικά, (ιδιαίτερα από τις ενότητες §3.2 και §3.3) και ενώ είναι δεδομένο ότι είναι επικίνδυνο να επιτρέπονται ολικώς επικαλυπτόμενα, μεταξύ τους, ερωτήματα, διαπιστώνουμε ότι ίσως δεν είναι παράλογο να συναρτάται η «τελεσίδικη» απαγόρευση εκτέλεσης ενός ερωτήματος από το *βαθμό που η απάντησή του επικαλύπτει την απάντηση ενός από τα προηγούμενα ερωτήματα του ίδιου χρήστη*. Με την κατάλληλη διαχείριση αυτού του ζητήματος που φέρνουμε στην επιφάνεια, θεωρούμε ότι μπορούν να προκύψουν ιδιαίτερα οφέλη στην φιλικότητα και τη λειτουργικότητα του συγκεκριμένου μηχανισμού, διότι είναι θεμιτό να μπορεί ο χρήστης να διατυπώνει ενίοτε ερωτήματα που επικαλύπτονται με «ασφαλή» τρόπο και που εντέλει είναι καλοπροαίρετα και ιδιαίτερα χρήσιμες οι απαντήσεις τους για τον ερωτώντα.

Πιο συγκεκριμένα, αφού ο κάθε χρήστης έχει το δικαίωμα να λάβει απάντηση σε ένα ερώτημα αν περιέχει τουλάχιστον *k* επεισόδια, γιατί να μην μπορεί να λάβει απάντηση όταν το ερώτημα που θέτει, ναι μεν από την μια επικαλύπτει ολικώς ένα προηγούμενο, αλλά από την άλλη η εκμαιευόμενη αριθμητική διαφορά που συμπεραίνεται, αυτόματα ξεπερνά και αυτή το *k*.

Στο σχήμα 4.9 παρατηρούμε το εξής: Έστω ότι *k*=3 και έστω ότι τίθενται ερωτήματα αποτελούμενα (για λόγους απλότητας) από 1 υπο-ερώτημα. Ο χρήστης, αρχικά, ρωτά ένα ερώτημα με χωρικό κριτήριο που αντιστοιχεί στο ΑΒΓΔ και λαμβάνει πλήθος τροχιών = 8. Στη συνέχεια θέτει $2^{ο}$ ερώτημα με χωρικό κριτήριο που αντιπροσωπεύεται από το ΑΒ΄Γ΄Δ. Προφανώς πρόκειται για *ολικώς επικαλυπτόμενο* χωρικό ερώτημα και θα έπρεπε να απαγορεύεται η εκτέλεση του με το σκεπτικό ότι αποκαλύπτεται έμμεσα το πλήθος των τροχιών που υπάρχουν στο Β΄ΒΓΓ΄, δηλ. οι 5 τροχιές. Όμως γιατί να είναι αυτό επικίνδυνο αφού αν κάποιος ρωτήσει εξ' αρχής Β΄ΒΓΓ΄ θα λάβει εκ των πραγμάτων απάντηση; Έτσι, για τους λόγους που προαναφέραμε, προτείνουμε τέτοια ερωτήματα να απαντώνται.



Σχήμα 4.9 – Διαδοχικά ολικώς επικαλυπτόμενα ερωτήματα

Προβλήματα που δημιουργούνται στο χώρο/χρόνο Παρόλα αυτά ας εξετάσουμε τί μπορεί να συμβεί στη συνέχεια. Αν ο χρήστης μετά τα 2 πρώτα ερωτήματα (ΑΒΓΔ & ΑΒ΄Γ΄Δ) ρωτήσει Β΄Β΄΄Γ΄΄Γ΄, τότε το ερώτημα δεν επικαλύπτει ολικά κανένα άλλο πέραν του ΑΒΓΔ. Αν συγκριθεί το πλήθος των 2 αυτών απαντήσεων (ΑΒΓΔ vs Β΄Β΄΄Γ΄΄Γ΄), προκύπτει μια διαφορά 5 τροχιών που ξεπερνά και πάλι το *k*-anonymity.

Όμως με το ερώτημα ΑΒ΄Γ΄Δ, ήδη απαντημένο, εύκολα κανείς συμπεράνει ότι στην εναπομείνασα χωρική Β΄΄ΒΓΓ΄΄, περιέχονται 2 μόλις τροχιές που αυτό *αποτελεί παραβίαση* του συγκεκριμένου *k*-anonymity. Ακριβώς παρόμοια είναι η αντιμετώπιση των περιπτώσεων που προκύπτουν με διαφορετικό κριτήριο τον χρόνο και κοινά τα υπόλοιπα.

Η προτεινόμενη λύση με εφαρμογή για χώρο/χρόνο Η λύση που προτείνουμε είναι η δημιουργία *πλασματικών ερωτημάτων*, δηλ. ερωτημάτων που θα είναι καταχωρημένα στη βάση μαζί με τα πραγματικά ερωτήματα που έχει θέσει στο παρελθόν ο κάθε χρήστης. Με αυτό τον τρόπο, όταν η βάση πρέπει να ελέγχεται για όλο το ιστορικό του κάθε χρήστη, τα ερωτήματα αυτά θα συνυπολογίζονται ως υπαρκτά.

Πιο συγκεκριμένα, έστω Q_1 ένα ερώτημα που θέτει, εκείνη τη στιγμή, ο χρήστης και που αποτελείται από 3 υπο-ερωτήματα. Ας τα αναπαραστήσουμε σχηματικά ως εξής: Α → Β → Γ. Έστω ότι κατά τον πρώτο έλεγχο στη βάση διαπιστωθεί μια χωρική ή χρονική ολική επικάλυψη του Γ με ένα Γ΄ που ανήκει, έστω, σε ένα ερώτημα Q_2 του ίδιου χρήστη και αναπαρίσταται σχηματικά ως εξής: Α → Β → Γ΄. Πρόκειται προφανώς για επικίνδυνη κατάσταση γιατί ταυτόχρονα τα υπόλοιπα 2 υπο-ερωτήματά τους, ταυτίζονται σε όλα μεταξύ τους. Αντί να αποτρέψουμε κατευθείαν την εκτέλεση του Q_1, συγκρίνουμε το πλήθος των τροχιών μεταξύ τους και, έστω, ότι προκύπτει μεγαλύτερο του ορίου *k*-anonymity, άρα «ασφαλές». Θεωρούμε λοιπόν, ότι καλό είναι να απαντηθεί, δημιουργώντας παράλληλα μια *πλασματική καταχώρηση* ενός υποθετικού ερωτήματος, έστω ονόματι Q_3. Αυτό *πρέπει* να είναι το Α → Β → [Γ-Γ΄] (θεωρώντας ότι Γ>Γ΄, αλλιώς θα ήταν το Α → Β → [Γ΄-Γ]).

Έτσι, αν ο ίδιος χρήστης επιχειρήσει ένα $3^{ο}$ ερώτημα ($4^{ο}$, μετρώντας και το πλασματικό πλέον), ονόματι Q_4, το Α → Β → Γ΄΄, με το Γ΄΄ να επικαλύπτεται ολικώς (δηλ. περιέχεται) στο [Γ-Γ΄], θα γίνει η σύγκριση του πλήθους των τροχιών μεταξύ Q_3 και Q_4 και αν είναι μικρότερη, αυτή τη φορά, από το *k*-anonymity δεν θα απαντηθεί. Αν είναι μεγαλύτερη από το *k*-anonymity, τότε θα απαντηθεί αλλά θα δημιουργηθεί ταυτόχρονα για χρήση στο μέλλον μια νέα πλασματική που θα αφορά το [(Γ-Γ΄) – Γ΄΄] ως προς το επίμαχο υπο-ερώτημα κ.ο.κ. Αυτή η τεχνική, όπως προαναφέραμε, έχει εφαρμογή μόνο όταν το κριτήριο που μεταβάλλεται μεταξύ των ερωτημάτων είναι ο χώρος ή ο χρόνος.

Η πρόταση ειδικά ως προς τα tags Η προσέγγιση για μια φιλικότερη αντιμετώπιση των ερωτημάτων στην περίπτωση των tags είναι η εξής: Ολική επικάλυψη έχουμε αποφασίσει, εκ των πραγμάτων, ότι έχουμε όταν ένα (υπό)ερώτημα έχει ίδια χωροχρονικά κριτήρια με ένα άλλο και το ένα από τα δύο, δεν έχει τιμή στα tags (*is null*), ενώ το άλλο έχει μια συγκεκριμένη (*is not null*). Όταν δεν έχει τιμή το tag ενός οποιουδήποτε υπο-ερωτήματος, θεωρούμε *πάντα* ότι το, εν λόγω, κριτήριο ικανοποιείται με οποιαδήποτε τιμή, τυχόν, περιέχει το κάθε επεισόδιο στη βάση.

Ο κανόνας που πρέπει να ισχύει *σε κάθε περίπτωση*, για να απαντηθεί οποιοδήποτε από τα ερωτήματα αυτού του είδους, ακόμη κι αν υφίσταται ολική επικάλυψη μεταξύ τους, είναι ότι πρέπει *πάντα να ισχύει* ότι:

$$Count(Q[tag\ is\ null]) - \Sigma\, Count(Qs[tag\ is\ not\ null]) \geq k$$

Αυτή η ανισότητα σημαίνει ότι, αρχικά συγκεντρώνουμε το σύνολο των ερωτημάτων που έχουν ήδη απαντηθεί στο παρελθόν μαζί με το τρέχον που εξετάζουμε αν πρέπει να απαντηθεί. Όλα, μεταξύ τους, έχουν μια απόλυτη ταύτιση σε όλα τα κριτήρια πλην των tags. Υπάρχει η περίπτωση, σε αυτό το σημείο, είτε να μην έχει απαντηθεί στο παρελθόν, είτε να μην είναι το τρέχον ερώτημα αυτό που δεν έχει καθόλου tags (*Q[tag is null]),* δηλ. κατά κάποιο τρόπο το ερώτημα – υπερσύνολο. Σε αυτή την περίπτωση το ερώτημα απαντάται χωρίς περαιτέρω διερεύνηση. Αν όμως υφίσταται αυτού του είδους το ερώτημα, τότε θα συγκριθεί το πλήθος των απαντήσεων που περιέχει αυτό, με το άθροισμα του πλήθους των απαντήσεων που αντιστοιχούν σε όλα τα άλλα ερωτήματα που έχουν ήδη απαντηθεί και έχουν συγκεκριμένο *tag*. Σε αυτά συμπεριλαμβάνουμε όπως προείπαμε *και* το εκάστοτε ερώτημα που εξετάζουμε αν πρέπει να απαντηθεί.

Για παράδειγμα, έστω ότι τίθενται ερωτήματα του ίδιου χρήστη που αποτελούνται από ένα μόνο υπο-ερώτημα πάντα. Όλα έχουν τα ίδια κριτήρια ως προς χώρο και χρόνο. Έστω ότι το Α εκφράζει το συγκεκριμένο χωροχρονικό πλαίσιο και ότι ισχύει k=3.

$1^{ο}$ ερώτημα Q_1: A + tag= ‘home’ → Count = 3

Καμία σύγκριση, ερώτημα απαντάται.

$2^{ο}$ ερώτημα Q_2: A + tag is null → Count = 10

Πρόκειται για το ερώτημα-υπερσύνολο. Σύγκριση: Count(Q_2)- Count(Q_1)=10-3=7> k. Το ερώτημα απαντάται.

$3^{ο}$ ερώτημα Q_3: A + tag= ‘work’ → Count = 3

Σύγκριση: Count(Q_2) - (Count(Q_1) + Count(Q_3))=10-3-3=4> *k*. Το ερώτημα απαντάται.

4^o ερώτημα Q_4: A + tag= 'fun' → Count = 3

Σύγκριση: Count(Q_2)- (Count(Q_1) + Count(Q_3) + Count(Q_4))=10-3-3-3=1< *k*. Το ερώτημα *δεν* απαντάται.

Με αυτό τον τρόπο δεν απορρίπτεται κατευθείαν ένα, τέτοιας φύσης, ερώτημα, όμως από την άλλη δεν μπορεί να κατασκευαστεί ένα «συμπληρωματικό» ερώτημα όπως στις περιπτώσεις χωρικής ή χρονικής ολικής επικάλυψης.

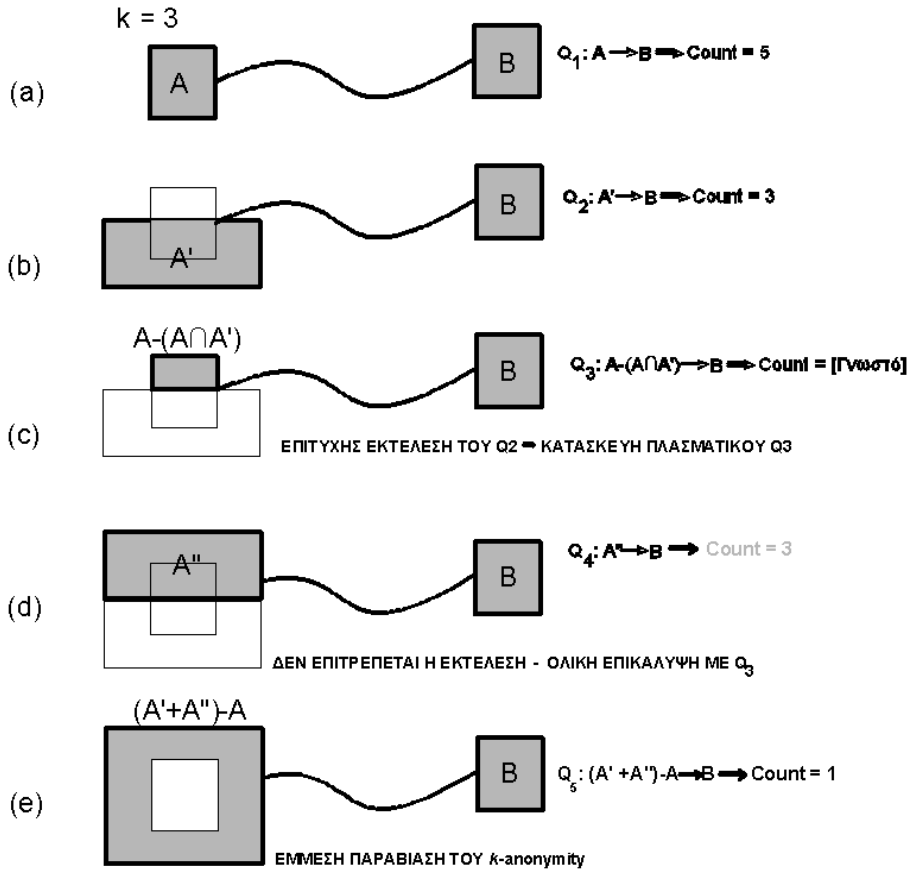
Η πρόταση για τα ολικώς επικαλυπτόμενα ερωτήματα ως προς το πλήθος των υπο-ερωτημάτων Η προσέγγιση για την περίπτωση που έχουμε ολικώς επικαλυπτόμενα ερωτήματα του ίδιου χρήστη κατά την εξέταση του ιστορικού των ερωτημάτων είναι να μην απορρίπτεται η απάντηση στο τρέχον ερώτημα αρκεί η απόλυτη τιμή της διαφοράς του πλήθους των τροχιών που πρόκειται να επιστρέψει το τρέχον ερώτημα μείον το πλήθος των τροχιών του προγενέστερου ερωτήματος να είναι μεγαλύτερη ή ίση του *k*-anonymity. Με αυτό τον τρόπο θεωρούμε ότι δεν εκμαιεύεται πληροφορία πιο συγκεκριμένη/επικίνδυνη από ότι επιτρέπουν οι προδιαγραφές λειτουργίας της βάσης. Για παράδειγμα, έστω Q_1 ένα ερώτημα που θέτει, εκείνη τη στιγμή, ο χρήστης και που αποτελείται από 3 υπο-ερωτήματα και αναπαρίστανται σχηματικά ως εξής: A → B → Γ και Q_2 ένα ερώτημα που έχει τεθεί στο παρελθόν από τον χρήστη και που αποτελείται από 2 υπο-ερωτήματα και αναπαρίστανται σχηματικά ως εξής: A → B.

Αν |Count(Q_2) - Count(Q_1)| >= *k*, τότε το ερώτημα Q_1 μπορεί με ασφάλεια να απαντηθεί. Στη συγκεκριμένη περίπτωση, αυτό που εκμαιεύεται εμμέσως, πέρα από την απάντηση του ερωτήματος Q_1 είναι ότι |Count(Q_2) - Count(Q_1)| είναι το πλήθος των παρακολουθούμενων οντοτήτων που ενώ αρχικά ακολούθησαν την πορεία A → B, στη συνέχεια είτε η πορεία/καταγραφή τους διακόπηκε, είτε βρέθηκαν οπουδήποτε αλλού πέρα από την χωροχρονική περιοχή (με ή χωρίς tags) που εντοπίζεται από το υπο-ερώτημα Γ.

Η μεταφορά της ιδέας των πλασματικών εγγραφών από τα ολικώς επικαλυπτόμενα στα πολλαπλώς τεμνόμενα ερωτήματα Ανάλογη αντιμετώπιση με την πρόταση για τον χώρο/χρόνο υπάρχει και στην περίπτωση των *πολλαπλώς τεμνόμενων ερωτημάτων*, όπως αυτά αναλύονται στην ενότητα 3.3.2. Θυμίζουμε ότι όταν ένα ερώτημα τέμνεται

απλώς με ένα άλλο ως προς το χρονικό ή χωρικό κριτήριο με τα υπόλοιπα κριτήρια να ταυτίζονται, αυτό από μόνο του δεν αποτελεί επίθεση. Πρέπει να λάβουμε όμως ορισμένα μέτρα διότι ένα $3^{ο}$ ερώτημα μπορεί, κατά περίπτωση, να προκαλέσει τη «έμμεση» δημιουργία ενός ολικώς επικαλυπτόμενου ερωτήματος.

Πιο συγκεκριμένα στο σχήμα 4.10 αναπαριστούμε το εξής παράδειγμα. Έστω ότι εφαρμόζεται ένα *k*-anonymity = 3 και εκτελείται ένα ερώτημα αποτελούμενο από 2 υπο-ερωτήματα (A → B). Η απάντηση είναι 5 και το ερώτημα απαντάται (σχήμα 4.10(a)). Στη συνέχεια, διενεργείται ένα ερώτημα A΄ → B, του οποίου η απάντηση είναι 3 και το οποίο έχει απλώς μια τομή στο ένα από τα δυο υπο-ερωτήματα με το προηγούμενο (υπο)ερώτημα του ίδιου χρήστη, οπότε με τη σειρά του απαντάται και αυτό (σχήμα 4.10(b)). Αν κατόπιν επιχειρηθεί ένα ερώτημα A΄΄ → B όπως αυτό απεικονίζεται στο σχήμα 4.10(d), έχει ήδη αναλυθεί στην ενότητα 3.3.2 ότι αυτό δεν πρέπει να απαντηθεί. Αυτό ισχύει διότι αν υποθέσουμε ότι η απάντηση στο A΄΄ → B ήταν 3, τότε η συνολική πληροφορία από τα 3 ερωτήματα, εκτός από το επιθυμητό, παρέχει την επιπλέον γνώση ότι η απάντηση στο υποθετικό ερώτημα ((A΄ + A΄΄) – A) → B είναι 1, κάτι που δεν το επιτρέπουμε λόγω του *k*-anonymity (σχήμα 4.10(e)). Η λύση είναι να εφαρμόσουμε ξανά την ίδια λογική περί πλασματικών ερωτημάτων στη βάση. Κάθε φορά που τίθεται ένα ερώτημα που τέμνει ένα άλλο (με τις ανάλογες υπόλοιπες προϋποθέσεις) ενόσω αυτό απαντάται, την ίδια στιγμή δημιουργείται και καταχωρείται στη βάση ένα επιπλέον πλασματικό ερώτημα που δημιουργείται από όλα τα κοινά (που ταυτίζονται) υπο-ερωτήματα των 2 αυτών ερωτημάτων και από ένα νέο υπο-ερώτημα με τρόπο τέτοιο όπως απεικονίζεται στο σχήμα 4.10(c). Ουσιαστικά, το κριτήριο στο νέο υπο-ερώτημα κατασκευάζεται ως εξής. Αν μιλούμε για χωρικό κριτήριο, τότε επιλέγεται αυτό το υπο-ερώτημα από τα δυο, το οποίο έχει τη μικρότερη σε μήκος πλευρά στην τομή τους και αφαιρούμε από αυτό, τον κοινό τόπο των 2 αυτών υπο-ερωτημάτων. Αυτό που απομένει (στο παράδειγμά μας, το (A-(A∩A΄))) συνθέτει το νέο κριτήριο του συγκεκριμένου υπο-ερωτήματος του πλασματικού ερωτήματος (σχήμα 4.10(c)). Το τελικό αποτέλεσμα είναι ότι αν πλέον επιχειρηθεί ένα ερώτημα όπως Q_4 : A΄΄ → B, δεν θα επιτραπεί η απάντηση διότι θα εντοπιστεί ως ολικώς επικαλυπτόμενο ερώτημα επί του πλασματικού Q_3 : A-(A∩A΄) → B (σχήμα 4.10(d)).



Σχήμα 4.10 – Η αντιμετώπιση για τα πολλαπλώς τεμνόμενα ερωτήματα

Προφανώς στη γενική περίπτωση που το 2^{ο} ερώτημα δεν τέμνει εντελώς (από τη μια άκρη μέχρι την άλλη) τη μία από τις 2 πλευρές που οριοθετούν το χωρικό κριτήριο του 1^{ου} ερωτήματος, δεν μπορεί να δημιουργηθεί πλασματική εγγραφή στη βάση, γιατί θα προϋπόθετε τη δυνατότητα της βάσης να κατασκευάζει και πολύγωνα ως κριτήρια των εκάστοτε ερωτημάτων και ταυτόχρονα βέβαια να αποθηκεύει τις χωρικές εκτάσεις των εκάστοτε επεισοδίων των αποθηκευμένων τροχιών, επίσης και ως πολύγωνα αν ήταν απαραίτητο. Παρόλα αυτά, αν στο μέλλον μελετηθεί αυτή η δυνατότητα, είναι σίγουρα σκόπιμο και απόλυτα εφικτό να ελέγχονται και αυτά τα ερωτήματα που, *συνδυαστικά*, δημιουργούν μια τομή σε ένα προγενέστερο ερώτημα από τη μια άκρη μέχρι την άλλη και τότε θα απαιτούνταν επίσης η δημιουργία κάποιου τύπου πλασματικής εγγραφής.

Σχετικά με το χρονικό κριτήριο, ως μονοδιάστατο μέγεθος, η υλοποίηση τυχόν πλασματικών τροχιών ακολουθεί ακριβώς την ίδια λογική/τεχνική χωρίς να υπάρχουν μάλιστα οι ειδικές περιπτώσεις που αναφέραμε στη αμέσως προηγούμενη παράγραφο για το χωρικό κριτήριο.

Τονίζουμε ότι αυτή η τεχνική ως προς τα *πολλαπλώς τεμνόμενα ερωτήματα*, εφαρμόζεται επιτυχώς όταν *ένα* μόνο υπο-ερώτημα τέμνεται σε ένα μόνο κριτήριό του με ένα άλλο (υπο)ερώτημα στη βάση και όλα τα υπόλοιπα στοιχεία τους ταυτίζονται.

4.4.3 Περιγραφή αλγορίθμου

Η είσοδος του αλγορίθμου είναι η βάση δεδομένων που είναι καταχωρημένο το ιστορικό των ερωτημάτων όλων των χρηστών, το προκαθορισμένο από τους διαχειριστές της βάσης, *k*-anonymity, το αρχικό ερώτημα με τα υπο-ερωτήματά του (με n συμβολίζεται ο αριθμός των υπο-ερωτημάτων του) και το *uid* που είναι το κωδικοποιημένο όνομα του χρήστη στη βάση. Η έξοδος είναι πάλι η βάση με ενημερωμένο (αν κρίθηκε απαραίτητο) το ιστορικό των ερωτημάτων όλων των χρηστών και μια Boolean μεταβλητή (*is_ok*) που απαντά αν κρίνεται «αθώο» το ερώτημα ή όχι.

$1^{ο}$ μέρος – $1^{ο}$ τμήμα Ο αλγόριθμος χωρίζεται σε 2 κύρια μέρη, το πρώτο μέρος χωρίζεται με τη σειρά του σε 2 διακριτά τμήματα. Στην αρχή του αλγορίθμου (δηλ. το $1^{ο}$ τμήμα του $1^{ου}$ μέρους) (γραμμές 2-12) γίνεται μια σύγκριση του $1^{ου}$ υπο-ερωτήματος του ερωτήματος που θέτει, εκείνη τη στιγμή, ο χρήστης με το σύνολο των (υπο)ερωτημάτων που έχει θέσει ο ίδιος στο παρελθόν μέσω μιας επαναληπτικής διαδικασίας επί όλων αυτών των ερωτημάτων. Για κάθε ερώτημα (Q_i) (γραμμή 2) που υπάρχει στη βάση και για κάθε υπο-ερώτημα αυτού (Q_{iSQj}) (γραμμή 3), αντιπαραβάλλεται το χωρικό κριτήριό του με το χωρικό κριτήριο του $1^{ου}$ υπο-ερωτήματος (Q_{SQ1}**)** του ερωτήματος (Q) που μόλις έχει θέσει ο χρήστης και εξετάζουμε αν υπάρχει, οποιουδήποτε είδους, τομή μεταξύ τους (γραμμή 4). Αν βρεθεί οποιουδήποτε είδους χωρική επικάλυψη (είτε μερική, είτε ολική) προχωρά στη σύγκριση μεταξύ των αντίστοιχων χρονικών κριτηρίων (γραμμή 5). Αν υπάρχει ταυτόχρονα και χρονική (είτε μερική, είτε ολική) επικάλυψη, τότε το ερώτημα θεωρείται «ύποπτο» και αποφασίζεται ότι χρήζει περαιτέρω διερεύνησης. Γι' αυτό το λόγο αποθηκεύεται το ερώτημα σε ένα προσωρινό βοηθητικό πίνακα (*Hst*) καθώς και ο αριθμός του υπο-ερωτήματός του, που τεμνόταν με το υπο εξέταση υπο-ερώτημα (γραμμές 6-8). Έτσι, ο αλγόριθμος ολοκληρώνει μια πλήρη σάρωση στο σύνολο των παρελθόντων (υπο)ερωτημάτων του χρήστη και μπορεί να εστιάσει σε ένα (κατά πάσα

πιθανότητα) πολύ μικρότερο αριθμό από αυτά αποφεύγοντας να αντιπαραβάλει εξ' αρχής εξαντλητικά όλο το ερώτημα που θέτει ο χρήστης με τα αποθηκευμένα στη βάση. Με το τέλος λοιπόν της επανάληψης αυτής (γραμμή 12), όλες οι πιθανές περιπτώσεις τύπου a, b ή c του πίνακα 4.6 και μέρος της περίπτωσης d έχουν εντοπιστεί και καταχωρηθεί στον πίνακα *Hst* (Το μέρος της περίπτωσης d που έχει εντοπιστεί αφορά περιπτώσεις όπου το ερώτημα που εξετάζεται είναι της μορφής Α → Β και στη βάση υπάρχει αποθηκευμένο παλαιότερο ερώτημα της μορφής Α → Β → Γ, όπου Α, Β, Γ τα υπο-ερωτήματα που συνθέτουν τα αντίστοιχα ερωτήματα). Εννοείται ότι ανάμεσά σε αυτά που έχουν εντοπιστεί είναι και ερωτήματα που, εν συνεχεία, θα βρεθούν «αθώα». Προτιμούμε όμως ένα ταχύ τρόπο ελέγχου όλων των υπο-ερωτημάτων στη βάση με αποτέλεσμα την εισαγωγή στον πίνακα *Hst* και μερικών ερωτημάτων που θα «κριθούν» αθώα στη συνέχεια της διαδικασίας, παρά το αντίθετο.

1^{ο} μέρος – 2^{ο} τμήμα Στη συνέχεια, ακολουθεί το 2^{ο} τμήμα του 1^{ου} μέρους (γραμμές 13-45). Σ' αυτό γίνεται μια επανάληψη που σαρώνει τον πίνακα *Hst*, πλέον, και για κάθε ερώτημα που αυτός περιέχει γίνονται στοχευμένες συγκρίσεις μεταξύ των κριτηρίων αυτού του υπο-ερωτήματος που βρέθηκε να τέμνεται με οποιοδήποτε τρόπο με το 1^{ο} υπο-ερώτημα του ερωτήματος που μόλις έχει θέσει ο χρήστης. Με αυτές τις συγκρίσεις επιδιώκουμε να εξάγουμε τα ανάλογα συμπεράσματα, έτσι ώστε να ληφθούν 2 αποφάσεις.

α) Συνεχίζει να θεωρείται ύποπτο το ερώτημα (Q_i) μετά την αναλυτική σύγκριση του 1^{ου} υπο-ερωτήματος (Q_{SQ1}**)** του ερωτήματος (Q) με το υπο-ερώτημά του (Q_{iSQm})**;**

β) Αν ναι, με ποιες προϋποθέσεις θα συνεχιστεί η περαιτέρω διερεύνηση των 2 αυτών ερωτημάτων; Πιο συγκεκριμένα, ποιο είδος επίθεσης από τα 4 είδη θεωρούμε ότι μπορεί να λαμβάνει χώρα εκείνη τη στιγμή (πίνακας 4.6);

Για να απαντήσουμε στις 2 αυτές ερωτήσεις, διενεργούμε αρχικά συγκρίσεις ώστε αν τεκμηριωθεί *πιθανότητα ολικώς επικαλυπτόμενου* ερωτήματος ως προς το χώρο, το χρόνο ή τα tags (γραμμές 16-26), τότε καλείται η διαδικασία *Check_suspicious_query* για περαιτέρω διερεύνηση ολόκληρων πλέον των ερωτημάτων (γραμμή 27).

Κατόπιν, γίνονται συγκρίσεις ώστε να τεκμηριωθεί *πιθανότητα μερικώς επικαλυπτόμενου* ερωτήματος ως προς το χώρο, το χρόνο (γραμμές 28-33) και αν υφίσταται, τότε καλείται η διαδικασία *Check_suspicious_query*, ώστε να ληφθούν από

τώρα μέτρα, για να αποφευχθεί ένα μελλοντικό *πολλαπλώς τεμνόμενο ερώτημα* από τον ίδιο χρήστη (γραμμή 34).

Τέλος, όταν το πλήθος των υπο-ερωτημάτων, μεταξύ των, υπό σύγκριση, ερωτημάτων, είναι διαφορετικό, γίνονται συγκρίσεις ώστε να τεκμηριωθεί *πιθανότητα ολικώς επικαλυπτόμενου* ερωτήματος ως προς το πλήθος των υπο-ερωτημάτων (γραμμές 36-40). Αν υπάρχει τέτοια πιθανότητα, τότε γίνεται μια σύγκριση μεταξύ του πλήθους των απαντήσεων των 2 ερωτημάτων (γραμμή 37) και εφόσον η διαφορά τους βρεθεί μικρότερη από *k* (μόνο) τότε καλείται η διαδικασία *Check_suspicious_query* ξανά για περαιτέρω διερεύνηση αυτών των ερωτημάτων (γραμμή 42). Αν η διαφορά είναι μεγαλύτερη από *k*, όπως έχουμε ήδη αναλύσει στην ενότητα 4.4.2, ακόμα και να υπήρχε ταύτιση των υπόλοιπων υπο-ερωτημάτων, το ερώτημα του χρήστη κρίνεται «ασφαλές».

Ενδεικτικές περιπτώσεις οι οποίες «αποχαρακτηρίζουν» το ερώτημα από ύποπτο και δεν συνεχίζεται η διερεύνηση, υπάρχουν όταν τα 2 συγκρινόμενα υπο-ερωτήματα:

α) έχουν ίδιο χωρικό και χρονικό κριτήριο αλλά διαφορετικά tags.

β) έχουν ίδιο χωρικό κριτήριο και το χρονικό κριτήριο του 1^{ου} περιέχει το χρονικό κριτήριο του 2^{ου} αλλά τα tags είναι διαφορετικά.

γ) ενώ τα 2 ερωτήματα περιέχουν διαφορετικό αριθμό από υπο-ερωτήματα, δεν υπάρχει πλήρης ταύτιση των 2 υπό σύγκριση υπο-ερωτημάτων, αλλά τέμνονται απλώς χωρικά (και ας ταυτίζονται τα άλλα 2 κριτήρια).

Έτσι, ολοκληρώνεται το 1^{ο} μέρος του αλγορίθμου (γραμμές 2-45) που βασίστηκε στην καταρχήν εξέταση μόνο του 1^{ου} SQ με όλα τα (υπο)ερωτήματα στη βάση και περαιτέρω διερεύνηση μόνο μικρού μέρους εξ' αυτών. Τον τελικό χαρακτηρισμό αν πρέπει ή όχι το ερώτημα να απαντηθεί τον προσδίδει η διαδικασία *Check_suspicious_query* η οποία περιγράφεται αναλυτικά στη συνέχεια αυτής της ενότητας.

2^{ο} μέρος Το 2^{ο} μέρος (γραμμές 46-60) εκτελείται, καταρχάς, στην περίπτωση που το ερώτημα του χρήστη (Q) περιέχει πάνω από 1 υπο-ερώτημα και ταυτόχρονα δεν έχει ληφθεί από τα προηγούμενα βήματα του αλγορίθμου, απόφαση για *άρνηση εκτέλεσης του ερωτήματος* (γραμμή 46). Αυτό το τμήμα αλγορίθμου ελέγχει τις υπόλοιπες υποκατηγορίες της περίπτωσης d του πίνακα 4.6 που δεν ήταν εφικτό να εντοπιστούν

από τη μέχρι τώρα διαδικασία. Πρόκειται για την περίπτωση όπου υπάρχουν *Ολικώς επικαλυπτόμενα ερωτήματα ως προς το πλήθος των υπο-ερωτημάτων* και το $1^{ο}$ υπο-ερώτημα του χρήστη δεν ταίριαζε με κανένα από τα υπο-ερωτήματα στη βάση. Αυτό που φοβόμαστε όμως, είναι, μήπως τα υπόλοιπα υπο-ερωτήματα του τρέχοντος ερωτήματος αποτελούν όλα μαζί ένα παλιότερο αυτόνομο ερώτημα στη βάση. Με άλλα λόγια, όταν ο χρήστης θέτει ερώτημα με x υπο-ερωτήματα και στη βάση υπάρχει ήδη ένα «υπερσύνολο» των x αυτών υπο-ερωτημάτων, η επίθεση θα είχε ήδη εντοπιστεί (γραμμή 36).

Για παράδειγμα αν υπάρχει στη βάση το ερώτημα Α → Β → Γ και προσπαθεί να εκτελεστεί τώρα το Α → Β, θα είχε ήδη εντοπιστεί γιατί όποιο και να είναι το «λεγόμενο» $1^{ο}$ υπο-ερώτημα (Q_{SQ1}) σίγουρα θα «ταίριαζε» απόλυτα με κάποιο από τα αποθηκευμένα υπο-ερωτήματα (είτε το Α, είτε το Β). Αν όμως ο χρήστης θέσει ένα ερώτημα – υπερσύνολο προηγούμενου αποθηκευμένου ερωτήματος, τότε το $1^{ο}$ μέρος του αλγορίθμου (γραμμές 2-45) δεν μπορεί να αντιληφθεί τον κίνδυνο και γι' αυτό χρειάζεται το $2^{ο}$ μέρος. Για παράδειγμα, έστω ότι τίθεται το Γ → [Α → Β] με το Q_{SQ1} να είναι το Γ και υπάρχει στη βάση το [Α → Β]. Τέτοια περίπτωση δεν εντοπίζεται από το $1^{ο}$ μέρος του αλγορίθμου.

Έτσι, ξεκινά μια επανάληψη που διατρέχει όλη τη βάση ξανά και που εκτελείται για κάθε υπο-ερώτημα από το $2^{ο}$, πλέον, μέχρι το n-οστό. Κάθε υπο-ερώτημα συγκρίνεται με όλα τα αποθηκευμένα υπο-ερωτήματα *μόνο* όταν αυτά ανήκουν σε ερωτήματα που *έχουν πλήθος υπο-ερωτημάτων n-j+1* (γραμμή 47) όπου

n: πλήθος υπο-ερωτημάτων του υπό εκτέλεση ερωτήματος

j: τρέχον αύξων αριθμός υπο-ερωτήματος του υπό εκτέλεση ερωτήματος

Έτσι, περιορίζουμε σημαντικά τις άσκοπες συγκρίσεις, γιατί ενώ διατρέχουμε όλη τη βάση, στεκόμαστε μόο σε ερωτήματα που έχουν συγκεκριμένο, κάθε φορά, αριθμό υπο-ερωτημάτων. Για παράδειγμα, αν το αρχικό μας ερώτημα έχει 3 υπο-ερωτήματα και εξετάζουμε/συγκρίνουμε το $2^{ο}$, σε εκείνη την επανάληψη, τότε μας ενδιαφέρουν *μόνο* όσα ερωτήματα έχουν 3-2+1=2 υπο-ερωτήματα συνολικά. Αν βρεθεί να ταυτίζεται (γραμμή 51) το τρέχον υπο-ερώτημα με κάποιο στη βάση, τότε γίνεται μια σύγκριση μεταξύ του πλήθους των απαντήσεων των 2 ερωτημάτων (γραμμή 52) και εφόσον η διαφορά τους βρεθεί μικρότερη από *k* (μόνο) τότε καλείται η διαδικασία *Check_suspicious_query* ξανά για περαιτέρω διερεύνηση αυτών των ερωτημάτων

(γραμμή 53). Αν η διαφορά είναι μεγαλύτερη από *k*, όπως ξαναείπαμε, ακόμα και να υπήρχε ταύτιση των υπόλοιπων υπο-ερωτημάτων, το ερώτημα του χρήστη κρίνεται «ασφαλές». Εν τέλει, αυτή η επανάληψη εξαντλεί όλα τα υπο-ερωτήματα του Q διότι θυμίζουμε ότι η επίθεση με χρήση υπο/υπερσυνόλου προηγούμενων ερωτημάτων δεν επιτυγχάνεται μόνο όταν μεταξύ τους διαφέρουν κατά ένα.

Για παράδειγμα, αν ο χρήστης ρωτά Α → Β → Γ και στη βάση υπάρχει το Γ αυτόνομο, τότε αφού το 1^{ο} μέρος του αλγορίθμου έχει ελέγξει το Α με τη βάση και δεν έχει βρεθεί, έστω, καμία τομή, έρχεται το 2^{ο} μέρος του αλγορίθμου και εξετάζει με τη σειρά όλα τα υπόλοιπα υπο-ερωτήματα μόνο για πλήρη ταύτιση όλων των κριτηρίων και όλων των υπο-ερωτημάτων μεταξύ τους, αντίστοιχα. Δηλαδή το Β ταυτίζεται με κάποιο; Αν ναι, ουσιαστικά αναζητούμε ένα ενδεχόμενο συνδυασμό Β → Γ μήπως υπάρχει στη βάση. Αν όχι, εξετάζεται το Γ (μόνο του ως τελευταίο) μήπως αυτό έχει τεθεί στο παρελθόν.

Θυμίζουμε ότι περιπτώσεις που ο χρήστης ρωτάει Α → Β → Γ και στη βάση υπάρχει Α → Β → Δ ή Β → Γ → Δ *δεν αποτελούν κίνδυνο* σύμφωνα με την ενότητα 3.3.1. Με αυτό τον τρόπο, καθώς ολοκληρώνεται ο αλγόριθμος (γραμμές 61-64), επιστρέφει μια Boolean μεταβλητή αν θα επιτραπεί ή όχι να εκτελεστεί το ερώτημα και αν ναι τότε ενσωματώνει στη βάση ό,τι πλασματικές εγγραφές έχουν δημιουργηθεί.

Η διαδικασία Check_suspicious_query Η διαδικασία που διερευνά περαιτέρω 2 ερωτήματα και καλείται από τον αλγόριθμο Auditor, λέγεται *Check_suspicious_query* και παρουσιάζεται στο σχήμα 4.12. Αυτή δέχεται ως είσοδο 2 ερωτήματα, το *k*-anonymity και ένα flag που προσδιορίζει τις προϋποθέσεις κάτω από τις οποίες κλήθηκε η διαδικασία ώστε αυτή να προβεί στις ανάλογες ενέργειες κάθε φορά. Η έξοδος της είναι μια Boolean μεταβλητή που προσδιορίζει την ύπαρξη πιθανής επίθεσης ή όχι και ένας βοηθητικός πίνακας που συμπληρώνεται με τα πλασματικά ερωτήματα που τυχόν χρειαστεί να δημιουργηθούν. Επί της ουσίας, αποτελείται από 3 διακριτά τμήματα και μόνο ένα τμήμα εκτελείται σε κάθε κλήση της, ανάλογα με την τιμή του flag.

Flag=0 Όταν το flag είναι 0, τότε έχει κληθεί όταν έχουν βρεθεί χωρικά, χρονικά ή ως προς τα tags ολικώς επικαλυπτόμενα ερωτήματα μεταξύ ερωτημάτων με ίσο πλήθος υπο-ερωτημάτων. Το 1^{ο} βήμα είναι ένας πλήρης έλεγχος των 2 ερωτημάτων (procedure *Full_Comparison*) σε σχέση (και) με τα υπόλοιπα υπο-ερωτήματά τους (γραμμή 5). Αν

τα υπόλοιπα υπο-ερωτήματα ταυτίζονται ή επικαλύπτονται στο ίδιο κριτήριο μόνο, κατά τον ίδιο πάντα τρόπο, τότε η μεταβλητή *there_is_danger* παίρνει τιμή TRUE. Σημειώνουμε εδώ, ότι αν πρόκειται για επικάλυψη ως προς τα tags, τότε η διαδικασία *Full_Comparison* που έχει κληθεί, έχει ήδη εξετάσει την ανισότητα που περιγράφηκε στην *πρότασή μας ειδικά ως προς τα tags* (ενοτ. 4.4.2, σελ. 90) και που αν δεν είναι αληθής τότε τελεσίδικα υπάρχει κίνδυνος από το τρέχον ερώτημα. Στη συνέχεια, αν όντως διαπιστωθεί κίνδυνος και οφείλεται σε tags, τότε διαπιστώνεται τελεσίδικος κίνδυνος (γραμμή 15), ενώ αν πρόκειται για χωρική ή χρονική επικάλυψη (γραμμή 7) δίνεται μια ευκαιρία να επιτραπεί η απάντηση, μόνο αν η διαφορά του πλήθους των απαντήσεων είναι τουλάχιστον *k* (γραμμή 8) δημιουργώντας ταυτόχρονα μια πλασματική εγγραφή, ενώ στην αντίθετη περίπτωση δεν πρέπει να απαντηθεί το αρχικό ερώτημα (γραμμή 12). Η διαδικασία που δημιουργεί τις πλασματικές εγγραφές λέγεται *Create_dummy_values* και παίρνει ως είσοδο το μέτρο της διαφοράς πάνω στο κριτήριο που υφίσταται η επικάλυψη μαζί με όλα τα υπόλοιπα κοινά χαρακτηριστικά των 2 ερωτημάτων και ενημερώνει ένα βοηθητικό πίνακα (*Help_table*) με την παραγόμενη εγγραφή (γραμμή 9-10).

Flag=1 Όταν το flag είναι 1, πρόκειται για ερωτήματα που απλώς τέμνονται μεταξύ τους σε ένα κριτήριο (χωρικό ή χρονικό). Δεν αποκλείουμε την απάντηση *ποτέ* σε αυτή την περίπτωση, όμως φροντίζουμε αν διαπιστωθεί περίπτωση *πολλαπλώς τεμνόμενων ερωτημάτων* (βλ. ενότητα 3.3.2) να κατασκευαστεί μια πλασματική εγγραφή ώστε να αποτραπεί τυχόν μελλοντικό ολικώς επικαλυπτόμενο ερώτημα. Για να τεκμηριωθεί περίπτωση πιθανού μελλοντικού *πολλαπλώς τεμνόμενου ερωτήματος*, πρέπει *όλα* τα υπόλοιπα υπο-ερωτήματα (πλην των αρχικώς συγκρινόμενων) να ταυτίζονται (γραμμές 22-29) και μόνο τότε δημιουργείται η πλασματική εγγραφή (γραμμή 30).

Flag=2 Όταν το flag είναι 2, πρόκειται για την περίπτωση όπου το ερώτημα που εξετάζεται, έχει περισσότερα επεισόδια από το άλλο με το οποίο συγκρίνεται και ενδεχομένως υπάρχει περίπτωση *ολικώς επικαλυπτόμενων ερωτημάτων ως προς το πλήθος των υπο-ερωτημάτων*, αφού βρέθηκε ήδη από την κύρια διαδικασία να έχουν από ένα υπο-ερώτημα που ταυτίζεται μεταξύ τους. Αυτό που εξετάζει η διαδικασία, εν προκειμένω, είναι αν όλα τα υπόλοιπα επεισόδια ταυτίζονται επίσης μεταξύ τους (γραμμές 37-44) και αν αυτό ισχύει δεν πρέπει να απαντηθεί το αρχικό ερώτημα.

Να επισημάνουμε ότι ως είσοδος στη διαδικασία *Check_suspicious_query*, πλην της περίπτωσης όπου flag=0, μπαίνουν το κατάλληλο υποσύνολο (σε επίπεδο υπο-ερωτημάτων) των αρχικών ερωτημάτων κάθε φορά που χρειάζεται να συγκριθούν (γραμμές 34,41,50 στο σχήμα 4.11) και όχι αυτούσια τα ερωτήματα όπως τα επεξεργάζεται η κύρια διαδικασία.

Algorithm 2 – Auditor

Input: Db = < > // Η βάση δεδομένων που είναι καταχωρημένο το ιστορικό των ερωτημάτων όλων των χρηστών
*k*_anon // *k*-anonymity threshold
Q = <SQ_1,SQ_2, SQ_3,…, SQ_n> // αρχικό ερώτημα με τα υπο-ερωτήματά του (**n**: αρ. υπο-ερωτημάτων)
uid // User identification number στη βάση

Output: Db = < > // Η βάση με ενημερωμένο (αν χρειαστεί) το ιστορικό των ερωτημάτων όλων των χρηστών
is_ok // Boolean μεταβλητή που απαντά αν κρίνεται «αθώο» το ερώτημα ή όχι

```
1:  p ← 1; m ← 0 ; is_ok ← true ; Hst[] ← 0 ; Hstsq[] ← 0 ; Help_table[] ← 0; k ← 1
2:  for each Qi ∈ Db(where user_id=uid) do   //εξέταση όλης της βάσης με τα ερωτήματα
3:      for each SQj ∈ Qi do                 // εξέταση κάθε υπο-ερωτήματος
4:          if Spatial_overlaps(QSQ1, QiSQj) then     // επιστρέφει true όταν βρεθεί χωρική τομή
5:             if Time_overlaps(QSQ1, QiSQj) then     // επιστρέφει true όταν βρεθεί χρονική τομή
6:                Hst[k]  ← Qi                        // γέμισμα βοηθητικού πίνακα με το ύποπτο ερώτημα
7:                Hstsq[k] ← j                        // γέμισμα 2ου βοηθ. πίνακα με το index του συγκ/νου SQ
8:                k ← k + 1
9:             end if
10:         end if
11:     end for
12: end for
13: for each Qi ∈ Hst do                    // σάρωση του βοηθητικού πίνακα
14:     m ← Hstsq[i]
15:     if Num_of_SQs(Qi) = n then          // τι κάνουμε όταν έχουμε ίσο πλήθος SQ μεταξύ Q και Qi
16:        if  ( ( ( Area(QSQ1) = Area(QiSQm))                                  and
17:              ( Time(QSQ1) = Time(QiSQm))                                    and
18:              (  (Tags(QSQ1) = Tags(QiSQm))                                  or
19:                 (Tags(QSQ1) is null and Tags(QiSQm) is not null) or
20:                 (Tags(QSQ1) is not null  and Tags(QiSQm) is null)  )   )   OR  // tags Overlap
21:           ( ( Area(QSQ1) = Area(QiSQm))                                    and
22:             ( Tags(QSQ1) = Tags(QiSQm))                                    and
23:             ( Is_contained(Time(QSQ1), Time(QiSQm)))                       )   OR // time Overlap
24:           ( ( Time(QSQ1) = Time(QiSQm))                                    and
25:             ( Tags(QSQ1) = Tags(QiSQm))                                    and
26:             ( Is_contained(Area(QSQ1), Area(QiSQm)))  )                    ) then // Space Overlap
27:           Check_suspicious_query(in Q, Qi, 0 , k_anon, in out is_ok, Help_table, p)
28:        else if (  ( Area(QSQ1) = Area(QiSQm))                              and
29:                   ( Tags(QSQ1) = Tags(QiSQm))                              and
30:                   ( Is_intersected(Time(QSQ1), Time(QiSQm)))               )   OR // time intersection
31:                ( ( Time(QSQ1) = Time(QiSQm))                               and
32:                  ( Tags(QSQ1) = Tags(QiSQm))                               and
33:                  ( Is_intersected(Area(QSQ1), Area(QiSQm)))                )   then // space intersection
34:                Check_suspicious_query(in Q(SQ2...SQn), Qi(All SQs - SQm), 1 , k_anon, in out is_ok, Help_table, p)
35:        end if
36:     else if Is_identical(QSQ1, QiSQm) then // με άνισο πλήθος SQ μεταξύ τους, αναζητούμε ταύτιση υπολοίπων
37:           if |Count(Q) - Count(Qi)| < k then      // προχωρούμε όταν η διαφορά < k
38:              Qtemp1 ← Q(SQ2...SQn) ; Qtemp2 ← Qi(All SQs - SQm)
39:              if  Num_of_SQs(Q temp1) > Num_of_SQs(Q temp2) then  // απαιτείται συγκεκριμένη σειρά..
40:                 Qtemp ← Q temp1 ; Qtemp1 ← Q temp2 ; Qtemp2 ← Q temp // των Qs, πριν τη κλήση της…
41:              end if                                              // Check_Suspicious_query
42:              Check_suspicious_query(in Q temp1, Q temp2, 2, k_anon, in out is_ok, Help_table, p)
43:           end if
44:     end if
45: end for
```

```
46: if n > 1 and is_ok then  // αρχή 2ου μέρους. Αφορά Q με SQs περισσότερα από 1
47:    for j =2 to n do        // τόσες επαναλήψεις όσα τα αρχικά SQs - 1
48:       for each Qi ∈ Db(where user_id=uid) do // σάρωση της βάσης…
49:          if Num_of_SQs(Qi) = (n – j + 1) then // …για ερωτήματα του χρήστη με συγκ/νο πλήθος από SQs
50:             for k=1 to Num_of_SQs(Qi) do    // αντιπαραβολή του j-οστού SQ του Q με κάθε ένα από τα SQs…
51:                if Is_identical(QSQj, QiSQk) then // …του Qi
52:                   if |Count(Q) - Count(Qi)| < k then  // προχωρούμε όταν η διαφορά < k
53:                      Check_suspicious_query(in Q(SQj...SQn), Qi, 2, k_anon, in out is_ok, Help_table, p)
54:                   end if
55:                end if
56:             end for
57:          end if
58:       end for
59:    end for
60: end if
61: if (is_ok) and (Help_table  is not null ) then
62:    Db ← Db + Help_table  // αν το ερώτημα είναι «αθώο», εντάσσονται στο ιστορικό της βάσης ...
63: end if                          // …οι τυχόν πλασματικές εγγραφές
64: return Db, is_ok
```

Σχήμα 4.11 – Αλγοριθμική προσέγγιση της γενικής τεχνικής αντιμετώπισης επιθέσεων

Procedure Check_suspicious_query(in Q_a , Q_b , flag, k **in out** is_ok, Help_table, p **)**

```
 1: begin
 2:    case
 3:       flag     : 0   // 1η περίπτωση
 4:       begin
 5:          Full_Comparison(in Qa , Qb out there_is_danger, kind_of_overlap) // πλήρης αντιπαραβολή 2 ερ/των
 6:          if there_is_danger then
 7:             if (kind_of_overlap = 'AREA') or (kind_of_overlap = 'TIME') then
 8:                if |Count(Qa) - Count(Qb)| >= k then  // όταν η διαφορά μεταξύ του πλήθους απαντήσεων >=k…
 9:                   Create_dummy_values(in ||Qa - Qb|| out Help_table[p])  //…δημιουργούμε πλασμ. εγγραφή…
10:                   p ← p + 1
11:                else                                                    //…αλλιώς δεν απαντάται
12:                   is_ok ← false
13:                end if
14:             else                  //  Αν kind_of_overlap = 'TAGS' και διαπιστώθηκε κίνδυνος, το ερώτημα δεν…
15:                is_ok ← false   // …απαντάται
16:             end if
17:          end if
18:       end
19:       flag     : 1   // 2η περίπτωση
20:       begin
21:          x ← 0
22:          for i=1 to Num_of_SQs(Qa) do
23:             for j=1 to Num_of_SQs(Qb) do
24:                if Is_identical(QaSQi, QbSQi) then
25:                   x ← x + 1   // για κάθε SQ του ενός που ταυτίζεται με ένα SQ του άλλου…
26:                end if           //…αυξάνεται κατά 1 ο αθροιστής x
27:             end for
28:          end for
29:          if x = Num_of_SQs(Qa) then // αν όλα ταυτίζονται, τότε απαντάται το ερώτημα αφού πρώτα…
30:            Create_dummy_values(in ||Qa - Qb|| out Help_table[p]) // …κατασκευαστεί πλασματική εγγραφή
31:            p ← p + 1
32:          end if
33:       end
34:       flag     : 2   // 3η περίπτωση
35:       begin
36:          x ← 0
37:          for i=1 to Num_of_SQs(Qa) do
38:             for j=1 to Num_of_SQs(Qb) do
39:                if Is_identical(QaSQi, QbSQi) then
40:                   x ← x + 1  // για κάθε SQ του ενός που ταυτίζεται με ένα SQ του άλλου…
41:                end if          //…αυξάνεται κατά 1 ο αθροιστής x
42:             end for
43:          end for
44:          if x = Num_of_SQs(Qa) then  // αν όλα ταυτίζονται, τότε δεν απαντάται το ερώτημα
46:            is_ok ← false
49:          end if
47:       end
48: end
```

Σχήμα 4.12 – Αλγοριθμική προσέγγιση της διαδικασίας *Check_suspicious_query*

5 Συμπεράσματα

Στις μέρες μας, λόγω της επέκτασης της χρήσης συστημάτων εντοπισμού θέσης (GPS) που μπορεί να είναι ενσωματωμένα σε πλήθος συσκευών και σε συνδυασμό με την γενικότερη πρόοδο στην ταχύτητα επεξεργασίας δεδομένων των υπολογιστικών συστημάτων, στην ευχρηστία των σημερινών βάσεων δεδομένων, στη ταχύτητα σύνδεσης στο διαδίκτυο αλλά και στην εμφάνιση διαφόρων, σχετικών με GPS, υπηρεσιών Web, υπήρξε αναπόφευκτα η επιθυμία για εξαγωγή χρήσιμων συμπερασμάτων από όλα αυτά τα δεδομένα.

Η παρούσα εργασία ασχολήθηκε με ενδεχόμενα ζητήματα πιθανής παραβίασης της ιδιωτικότητας σε βάσεις κινούμενων αντικειμένων σημασιολογικά εμπλουτισμένων τροχιών (*semantic trajectories*) οι οποίες κατασκευάστηκαν και (συν)τηρούνται στην υποδομή ενός οργανισμού, ο οποίος έχει και το αποκλειστικό δικαίωμα του ελέγχου της πρόσβασης στα δεδομένα της βάσης μέσω ενός κατάλληλου μηχανισμού.

Αφού οριοθετήθηκε το πλαίσιο λειτουργίας/εργασίας (ενότητες 3.1, 3.2), προχωρήσαμε στη μελέτη αυτού του περιβάλλοντος και διαπιστώθηκε ότι *υπάρχει* κίνδυνος παραβίασης της ιδιωτικότητας, καθώς ο χρήστης διενεργεί ορισμένα διαδοχικά ερωτήματα και ορισμένες φορές μάλιστα συνδυάζοντάς τα με πρότερη γνώση που τυχόν αυτός μπορεί να έχει, *παρόλο* που οι εγγραφές στη βάση δεν περιέχουν κανένα στοιχείο που να προκαλεί την άμεση αναγνώριση/σύνδεσή τους με την εκάστοτε καταγεγραμμένη οντότητα (*ids*) και ταυτόχρονα δεν απαντώνται ποτέ ερωτήματα όταν το πλήθος των απαντήσεων είναι μικρότερο του *k* (αριθμός καθορισμένος από τους διαχειριστές του συστήματος). Διαπιστώθηκε επίσης ότι, ενώ ο βασικός άξονας αντιμετώπισης πιθανής επίθεσης είναι, διαχρονικά, η άρνηση απάντησης στο ερώτημα, είναι ανοιχτές οι επιλογές σχετικά με το επίπεδο «αυστηρότητας» του μηχανισμού ελέγχου που περιέχει η βάση, γιατί όσο αυτό ανεβαίνει, τόσο μειώνεται η φιλικότητα και η λειτουργικότητα της βάσης με αποτέλεσμα να δυσχεραίνεται η εργασία του κάθε καλόπιστου χρήστη της βάσης.

Η συνεισφορά της παρούσας εργασίας είναι *η διακρίβωση* όλων των πιθανών τεχνικών επίθεσης στην ιδιωτικότητα καταγεγραμμένων οντοτήτων στη βάση που μπορεί να χρησιμοποιήσει ένας κακόβουλος χρήστης, *η πρόταση* για ένα συνολικό μηχανισμό παρακολούθησης, σύγκρισης και απόρριψης, αν κριθεί απολύτως απαραίτητο, των

ερωτημάτων, ιδωμένο από την οπτική γωνία της προστασίας της ιδιωτικότητας και τέλος, *η πρόταση* για ένα αυτόνομο αλγόριθμο (*Zoom Out*) που δύναται να «συμμετέχει» σε ένα από τα στάδια του προαναφερθέντος συνολικού μηχανισμού και ο οποίος βοηθά από τη μια, στην βελτίωση της φιλικότητας του περιβάλλοντος εργασίας προς τον χρήστη κυρίως όμως, της λειτουργικότητας του συνολικού αυτού μηχανισμού. Ο *Zoom Out* καλείται, πρακτικά, να προσπαθήσει να τροποποιήσει το (αρχικό) ερώτημα του χρήστη που δεν μπορεί εν πρώτοις να απαντηθεί για λόγους ασφαλείας από τη βάση, στο «πλησιέστερο» δυνατό ερώτημα που μπορεί με ασφάλεια, αυτή τη φορά, να απαντηθεί.

Μια μελλοντική εργασία θα μπορούσε να επικεντρωθεί στην προγραμματιστική ανάπτυξη των αλγορίθμων που περιέχονται στη παρούσα, στα πλαίσια της κατασκευής ενός συνολικού μηχανισμού παρακολούθησης ερωτημάτων για σημασιολογικά εμπλουτισμένες βάσεις κινούμενων αντικειμένων καθώς και, ενδεχομένως, της προγραμματιστικής υλοποίησης ορισμένων τύπων ερωτημάτων που είναι απαραίτητοι για τη λειτουργία των πιο πάνω αλγορίθμων.

Βιβλιογραφία

[1] Abul O, Bonchi F, and Nanni M. Never walk alone: Uncertainty for anonymity in moving objects databases. In Proceedings of ICDE, pages 376-385, 2008.

[2] Abul O, Bonchi F, and Nanni M. Anonymization of moving objects databases by clustering and perturbation. Information Systems, 35(8):884-910, 2010.

[3] N. R. Adam and J. C. Wortmann. Security-control methods for statistical databases: A comparative study. ACM Computing Surveys, 21(4):515–556, 1989.

[4] Gkoulalas-Divanis A, Verykios VS. A privacy–aware trajectory tracking query engine. SIGKDD Explorations, 10(1), pp.40-49, 2008

[5] Gkoulalas-Divanis A, Verykios VS. A Free Terrain Model for Trajectory K-Anonymity, Proceedings of the 19th international conference on Database and Expert Systems Applications, Turin, Italy, 2008

[6] Gkoulalas-Divanis A, Kalnis P, Verykios VS. Providing k-anonymity in location based services. ACM SIGKDD Explorations Newsletter, 12(1), pp.3-10, 2010.

[7] Hoh B. and Gruteser M.. Protecting location privacy through path confusion. In SECURECOMM, pages 194-205, 2005.

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

[9] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. *l*-diversity: Privacy beyond *k*-anonymity. In Proceedings of ICDE, pp.24, 2006.

[10] Mahdavifar S, Abadi M, Kahami M, Mahdikhani H. A clustering-based approach for personalized privacy preserving publication of moving object trajectory data. Network and System Security. Springer Berlin Heidelberg, 149-165, 2012.

[11] Monreale A, Trasarti R, Pedreschi D, Renso C, Bogorny V. C-safety: a framework for the anonymization of semantic trajectories. Transactions on Data Privacy, vol. 4, no. 2, pp. 73–101, 2011.

[12] Monreale A, Andrienko G, Andrienko N, Giannotti F, Pedreschi D, Rinzivillo S, Wrobel S. Movement Data Anonymity through Generalization. Transactions on Data Privacy 3(2), 91-121, 2010.

[13] Nergiz M. E, Atzori M. and Saygin Y. Towards trajectory anonymization: A generalization-based approach. Transactions on Data Privacy 2(1): 47-75 (2009)

[14] Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., Theodoridis, Y., and Yan, Z. 2013. Semantic trajectories modeling and analysis. ACM Comput. Surv. 45, 4, Article 42 (August 2013).

[15] Pelekis N, Gkoulalas-Divanis A, Vodas M, Kopanaki D, Theodoridis Y. Privacy Aware Querying over Sensitive Trajectory Data. In Proceedings of CIKM, pp.895-904, 2011.

[16] Pelekis N, Gkoulalas-Divanis A, Vodas M, Plemenos A, Kopanaki D, Theodoridis Y. Private-HERMES: A Benchmark Framework for Privacy-Preserving Mobility Data Querying and Mining Methods. In Proceedings of EDBT, 2012.

[17] Sweeney L. *k*-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 : 557-570, 2002.

[18] Terrovitis M, Mamoulis N. Privacy preservation in the publication of trajectories. In MDM, pp.65-72, 2008.

[19] N. Pelekis, C. Ntrigkogias, P. Tampakis, S. Sideridis, Y. Theodoridis: “Hermoupolis: A Trajectory Generator for Simulating Generalized Mobility Patterns”, demo paper, In Proceedings of ECML/PKDD'13, Prague, Czech Republic, September 2013.