Εισαγωγή

Ο οδηγός εύρεσης ενός κατάλληλου μοντέλου ανωνυμοποίησης παραθέτει ερωτήσεις με κριτήριο την διαφορετική προσέγγιση του κάθε μοντέλου και επιστρέφει πληροφορίες για αυτόν καθώς και την βιβλιογραφία που τον συνοδεύει για λόγους πληρότητας. Το παρακάτω κείμενο απευθύνεται στους χρήστες εκείνους που δεν έχουν χρησιμοποιήσει ποτέ το εργαλείο ή δεν έχουν γενικά ασχοληθεί ξανά με το αντικείμενο της προστασίας των προσωπικών δεδομένων.

Προσέγγιση στην ανωνυμοποίηση

Η ανωνυμοποίηση δεδομένων στοχεύει στην προστασία της ταυτότητας και των ευαίσθητων πληροφοριών των ιδιοκτητών τους. Ακόμα και με την διαγραφή αναγνωριστικών ταυτότητας μπορεί κανείς να προσδιορίσει μονοσήμαντα ένα άτομο. Με τον συνδυασμό γνωρισμάτων της εγγραφής (που αποκαλούνται ψευδό-αναγνωριστικά) μπορεί μοναδικά να προσδιοριστεί η ταυτότητα ενός ανθρώπου. Η διαδικασία εύρεσης ευαίσθητων πληροφοριών ονομάζεται **“αποκάλυψη”.** Για παράδειγμα αποδείχτηκε πως το 87.1% δηλαδή τα 216 από τα 248 εκατομμύρια του πληθυσμού των Ηνωμένων Πολιτειών της Αμερικής, μπορούν να προσδιοριστούν μονοσήμαντα με τη χρήση μόνο τριών χαρακτηριστικών γνωρισμάτων: του 5-ψήφιου ταχυδρομικού κώδικα, του φύλου και της ημερομηνίας γεννήσεως.

Για να αποτραπούν αυτές οι επιθέσεις ο εκδότης δεδομένων παρέχει έναν ανωνυμοποιημένο πίνακα:

T(QID', ευαίσθητα γνωρίσματα, μη-ευαίσθητα γνωρίσματα)

όπου QID' είναι η ανωνυμοποιημένη έκδοση της QID (ψευδό-αναγνωριστικά) και προέκυψε με επεξεργασία των χαρακτηριστικών γνωρισμάτων της QID του αρχικού πίνακα.

Επιλογή ψευδό-αναγνωριστικών

Μια μεγάλη πρόκληση που καλείται να αντιμετωπίσει ο ιδιοκτήτης δεδομένων είναι η ταξινόμηση των χαρακτηριστικών γνωρισμάτων ενός πίνακα σε τρεις κατηγορίες : στα ψευδο-αναγνωριστικά, στα ευαίσθητα γνωρίσματα και στα μη ευαίσθητα γνωρίσματα. Στα ψευδο-αναγνωριστικά πρέπει να περιλαμβάνεται ένα γνώρισμα Α εάν υπάρχει πιθανότητα ένας αντίπαλος να το έχει αποκτήσει από εξωτερικές πηγές. Έπειτα αφού έχουν καθοριστεί τα ψευδο-αναγνωριστικά, τα υπόλοιπα γνωρίσματα ομαδοποιούνται

στα ευαίσθητα και μη ευαίσθητα με βάση το επίπεδο της ευαισθησία τους. Δεν υπάρχει μια ξεκάθαρη απάντηση στον τρόπο με τον οποίο ένας κάτοχος δεδομένων μπορεί να προσδιορίσει ακριβώς ποια γνωρίσματα προέρχονται από εξωτερικές πηγές, αλλά θα πρέπει να γνωρίζει ποιες είναι οι συνέπειες της εσφαλμένης ταξινόμησης. Στην περίπτωση που γίνει λάθος ταξινόμηση, ένα ευαίσθητο γνώρισμα S μπορεί να τεθεί σε κίνδυνο γιατί μπορεί ένας αντίπαλος με το γνώρισμα A που δεν εντάχθηκε όπως θα έπρεπε στα ψευδο-αναγνωριστικά να κάνει σύνδεση πινάκων. Από την άλλη πλευρά, αν ένα ευαίσθητο χαρακτηριστικό γνώρισμα S κατηγοριοποιηθεί εσφαλμένα στα ψευδο-αναγνωριστικά, θα υπήρχε άσκοπη απώλεια πληροφοριών. Η σωστή επιλογή των ψευδό-αναγνωριστικών πάντως παραμένει ακόμα ανοιχτό ζήτημα.

Κατηγορίες Επιθέσεων

Με βάση λοιπόν τις επιθέσεις που μπορούν να δεχτούν οι βάσεις δεδομένων τα μοντέλα ιδιωτικότητας χωρίζονται σε τρεις κατηγορίες: *αποκάλυψη εγγραφής, αποκάλυψη χαρακτηριστικού γνωρίσματος και αποκάλυψη παρουσίας στον πίνακα.* Υποθέτουμε πως και στις τρεις κατηγορίες ο επιτιθέμενος γνωρίζει τα ψευδο-αναγνωριστικά του *θύματος* (με τον όρο θύμα εννοούμε τον άνθρωπο του οποίου οι ευαίσθητες πληροφορίες δέχονται την επίθεση). Υποθέτουμε επιπλέον πως στις αποκαλύψεις εγγραφής και χαρακτηριστικού γνωρίσματος γνωρίζει πως η εγγραφή και τα ευαίσθητα χαρακτηριστικά του θύματος βρίσκονται στον δημοσιευμένο πίνακα. Στην αποκάλυψη παρουσίας στον πίνακα, ο επιτιθέμενος ουσιαστικά ψάχνει να βρει αν η εγγραφή του ανθρώπου που τον ενδιαφέρει υπάρχει ή όχι ανάμεσα στις εγγραφές. Στην **αποκάλυψη εγγραφής** ο επιτιθέμενος προσπαθεί να απομονώσει και να ταυτοποιήσει μονοσήμαντα ευαίσθητες πληροφορίες ανθρώπων. Στην **αποκάλυψη χαρακτηριστικού γνωρίσματος** αν οι περισσότερες εγγραφές σε μια κλάση ισοδυναμίας έχουν ίδιες ή και παρόμοιες τιμές στα ευαίσθητα χαρακτηριστικά τότε ο επιτιθέμενος μπορεί να συνδέσει έναν άνθρωπο με τις ευαίσθητες τιμές του χωρίς να απαιτείται να προσδιορίσει την ακριβή εγγραφή του. Σε ορισμένες περιπτώσεις όπως στην **αποκάλυψη παρουσίας στον πίνακα** η διαπίστωση της παρουσίας (ή της απουσίας) μιας εγγραφής στον πίνακα αποκαλύπτει ευαίσθητες πληροφορίες του θύματος.

Παρακάτω παρατίθεται ένας πίνακας που απεικονίζει περιληπτικά τα μοντέλα ιδιωτικότητας που μπορεί να δώσει ως έξοδο ο οδηγός και που ταιριάζουν στο κάθε είδος επίθεσης:

Μοντέλο ιδιωτικότητας	Μοντέλο επίθεσης		
	Αποκάλυψη Εγγραφής	Αποκάλυψη Χαρακτηριστικού Γνωρίσματος	Αποκάλυψη παρουσίας Πίνακα
k-ανωνυμία	✓		
Πολυσχεσιακή k-ανωνυμία	✓		
(c,t)-απομόνωση	✓		
k^m-ανωνυμία	✓		
l-ποικιλομορφία	✓	✓	
Οριοθέτηση δύναμης		✓	
(X,Y)-ιδιωτικότητα	✓	✓	
(a, k)-ανωνυμία	✓	✓	
LKC-ιδιωτικότητα	✓	✓	
(k, e)-ανωνυμία		✓	
(ε,m)-ανωνυμία		✓	
t-εγγύτητα		✓	
Εξατομικευμένη ιδιωτικότητα		✓	
FF-ανωνυμία		✓	
m-αμεταβλητότητα	✓	✓	
δ-παρουσία			✓
ε-διαφορική ιδιωτικότητα			✓
(d-g)-ιδιωτικότητα			✓

Πίνακας: Μοντέλα ιδιωτικότητας

Μοντελοποίηση πληροφοριών αντιπάλου

Παρ' όλες τις διάφορες τεχνικές που έχουν αναπτυχθεί, και τις διάφορες εγγυήσεις που προτείνονται, παραμένουν πολλοί κίνδυνοι για την ιδιωτικότητα χωρίς αποτελεσματική αντιμετώπιση. Η διαθέσιμη πληροφορία που μπορεί να κατέχει ο επιτιθέμενος μπορεί να έχει πολλές μορφές. Παράλληλα, τα μοντέλα των δημοσιευμένων δεδομένων μπορεί να διαφέρουν κάθε φορά, με αποτέλεσμα η κάθε περίπτωση να απαιτεί διαφορετική επεξεργασία προκειμένου να εξασφαλίζεται η ιδιωτικότητα των βάσεων δεδομένων. Συγκεκριμένα ένας μεγάλος αριθμός μοντέλων υποθέτει πως οι γνώσεις του αντιπάλου περιορίζονται στα ψευδο-αναγνωριστικά. Όλες οι μελέτες έχουν δείξει πως είναι ιδιαίτερα σημαντικό για τον κάτοχο των δεδομένων να γνωρίζει τι γνώσεις έχει ο αντίπαλος ώστε να επεξεργαστεί κατάλληλα την βάση δεδομένων. Οι γνώσεις των αντιπάλων προέρχονται από την κοινή λογική, εκλογικούς καταλόγους, δημογραφικές βάσεις, κοινωνικά δίκτυα και άλλες προσωπικές πληροφορίες. Αναπτύχθηκε το μοντέλο skyline ιδιωτικότητα [38] το οποίο υποστηρίζει ότι, αφού είναι ανέφικτο για έναν εκδότη δεδομένων να προβλέψει το ακριβές γνωστικό υπόβαθρο που κατέχει ο αντίπαλος, θα πρέπει να μελετήσει τις πληροφορίες που είναι φυσικό να έχει και που μπορεί να διαχειριστεί. Ειδικότερα ορίζει τρεις τύπους γνωστικού υπόβαθρου, γνωστές και ως τριών-διαστάσεων γνώσεις τις:

- γνώσεις για το θύμα
- γνώσεις για άλλους κατόχους εγγραφών του πίνακα
- γνώσεις για άλλους κατόχους εγγραφών στην κλάση με το ίδιο ευαίσθητο στοιχείο του θύματος

Το γνωστικό υπόβαθρο εκφράζεται ποσοτικά από τα γράμματα l, k, m που υποδεικνύουν πως ο αντίπαλος ξέρει:

1. l ευαίσθητες τιμές που το θύμα t δεν έχει,
2. τις ευαίσθητες τιμές από k ανθρώπους και
3. m ανθρώπους που έχουν την ίδια ευαίσθητη τιμή όπως ο t

Έπειτα το μοντέλο αξιοποιεί τις παραπάνω τιμές και δέχεται κάποιες επιπλέον παραμέτρους που εισάγει ο κάτοχος των δεδομένων για μεγαλύτερη ακρίβεια και ευελιξία στην ανωνυμοποίηση. Η επιτυχής ανωνυμοποίηση των δεδομένων είναι εφικτή εάν οριστούν οι τιμές l, k, m σωστά