Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακό Πρόγραμμα Σπουδών *Πληροφοριακά και Επικοινωνιακά Συστήματα*

Μεταπτυχιακή Διατριβή

Ανωνυμοποίηση γράφων κοινωνικής δικτύωσης

Μαργαρίτα Καραχρήστου

Επιβλέπων Καθηγητής
Κωνσταντίνος Λιμνιώτης

Δεκέμβριος 2016

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακό Πρόγραμμα Σπουδών *Πληροφοριακά και Επικοινωνιακά Συστήματα*

Μεταπτυχιακή Διατριβή

Ανωνυμοποίηση γράφων κοινωνικής δικτύωσης

Μαργαρίτα Καραχρήστου

Επιβλέπων Καθηγητής
Κωνσταντίνος Λιμνιώτης

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων για απόκτηση μεταπτυχιακού τίτλου σπουδών Στα Πληροφοριακά και Επικοινωνιακά Συστήματα από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών του Ανοικτού Πανεπιστημίου Κύπρου.

Δεκέμβριος 2016

ΛΕΥΚΗ ΣΕΛΙΔΑ

Περίληψη

Με την ανάπτυξη της τεχνολογίας τα τελευταία χρόνια οι ανάγκες για τη χρήση και τη χρησιμοποίηση δεδομένων μέσω των κοινωνικών δικτύων αυξάνεται μέρα με τη μέρα. Υπάρχει τεράστιος όγκος πληροφοριών που οι χρήστες παρέχουν, ηθελημένα ή μη, με τη χρήση των κοινωνικών δικτύων. Η πληροφορία αυτή μπορεί να αποτελέσει χρήσιμο εργαλείο για επιστημονική ανάλυση. Κρίσιμο ζήτημα ωστόσο είναι το γεγονός ότι μέσω των κοινωνικών δικτύων υπάρχει μεγάλη επεξεργασία προσωπικών δεδομένων που πολλές φορές και οι ίδιοι οι χρήστες δεν επιθυμούν.

Ειδικότερα, τα τελευταία χρόνια πλήθος δεδομένων προερχόμενα από κοινωνικά δίκτυα δημοσιεύονται με διάφορους τρόπους για περαιτέρω αξιοποίησή τους για λοιπούς επιστημονικούς σκοπούς. Ωστόσο, η δημοσίευση των δεδομένων αυτών, ακόμα και αν έχουν απαλειφτεί τα στοιχεία ταυτοποίησης των χρηστών για σκοπούς διατήρησης της ανωνυμίας τους, δε διασφαλίζει τελικά πλήρως την προστασία των προσωπικών τους δεδομένων. Ένας επιτιθέμενος, γνωρίζοντας κάποια πρόσθετη πληροφορία (π.χ. πληροφορία για «φιλίες» μεταξύ των «φίλων» ενός χρήστη του κοινωνικού δικτύου), μπορεί να είναι σε θέση να αναγνωρίσει κάποιον χρήστη από τα, φαινομενικά, ανωνυμοποιημένα δημοσιοποιημένα δεδομένα του κοινωνικού δικτύου.

Αντικείμενο της παρούσας διατριβής είναι η μελέτη τεχνικών ανωνυμοποίησης γράφων κοινωνικών δικτύωσης (social networks graphs). Συγκεκριμένα, γίνεται μελέτη των πλεονεκτημάτων που παρέχουν οι γράφοι κοινωνικής δικτύωσης για την ανάλυση δεδομένων, σε αντιπαραβολή όμως με τους κινδύνους παραβίασης της ιδιωτικότητας των χρηστών των κοινωνικών δικτύων. Μελετώνται και καταγράφονται οι γνωστές μέχρι σήμερα τεχνικές ανωνυμοποίησης των γράφων αυτών, με συγκριτική αποτίμηση των πλεονεκτημάτων και μειονεκτημάτων τους. Στο πλαίσιο αυτό, αναπτύχθηκε μία νέα τεχνική ανωνυμοποίησης με σκοπό την κατά το δυνατόν ελάχιστη τροποποίηση ενός γράφου κοινωνικού δικτύου έτσι ώστε να επιτυγχάνεται ανωνυμία με – κατά το δυνατόν – μικρή απώλεια πληροφορίας.

Λέξεις Κλειδιά:

Ιδιωτικότητα, ανωνυμοποίηση δεδομένων, Ασφάλεια, γράφοι, επιθέσεις ιδιωτικότητας, κοινωνικά δίκτυα, k-ανωνυμία, l-ποικιλομορφία

Summary

In recent years, with the development of technology, the need for the use and utilization of the data via social networks is increasing day by day. There is a vast amount of information that users provide, intentionally or not, through the use of social networks. This information can be a useful tool for scientific analysis. The critical issue however is the fact that through social networks there is great processing of personal data, which most of the time is performed without the user's consent. .

In particular, in recent years numerous data coming from social networks are published in various ways for their further utilization for several scientific purposes. However, the publication of such data, even if the users' identification information are hidden to maintain anonymity, does not succeed to protect their personal data. An attacker having knowledge of some additional background information (eg. information about "friendships" among the "friends" of a user's social network) may be able to recognize a user by the seemingly anonymized data, publicly available through social networks.

The subject of this thesis is the technical study of social networking graphs anonymization (social networks graphs). More presicely, this thesis focuses on the advantages offered by social networking graphs used for data analysis, in conjunction with the privacy risks that arise. All known anonymization techniques, until today, are being discussed via a comparative studyof their advantages and disadvantages. In this context, a new anonymization technique is proposed, that aims to achieve minimal modification of a social networking graph in order to maintain anonymity of the users with the smallest possible information loss.

Keywords:
Privacy, anonymizing data, Security, graphs, privacy attacks, social networks, k-anonymity, l-diversity

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή μου και επιβλέπων κ. Κωνσταντίνο Λιμνιώτη για την εξαιρετική συνεργασία, την συνεχή υποστήριξη και καθοδήγηση του καθ' όλη την διάρκεια της εκπόνησης της εργασίας αυτής.

Τέλος θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους φίλους μου για την ανεκτίμητη στήριξή τους.

Περιεχόμενα

Κεφάλαιο 1

Τα προσωπικά δεδομένα και η προστασία τους

Στις μέρες μας, με την εξέλιξη των τεχνολογιών αλλά και των τεχνικών εξόρυξης δεδομένων, υπάρχει υπερπληθώρα διαθέσιμων πληροφοριών (συχνά αναφερόμαστε σε αυτά με τον όρο «μεγάλου όγκου δεδομένα» - Big Data) οι οποίες, αν αξιοποιηθούν κατάλληλα, μπορούν να οδηγήσουν σε επιστημονικές και στατιστικές αναλύσεις για την εξαγωγή συμπερασμάτων που παλαιότερα δεν ήταν εφικτό να πραγματοποιηθούν. Ειδικότερα οι πληροφορίες που λαμβάνουμε από τα διάφορα μέσα κοινωνικής δικτύωσης αποτελούν έναν μεγάλο όγκο δεδομένων - λόγω ακριβώς της ευρείας χρήσης των κοινωνικών δικτύων - που η καθεμία από αυτές τις πληροφορίες μπορεί να είναι διαθέσιμη σε καλοπροαίρετους αλλά, δυστυχώς, και κακοπροαίρετους χρήστες.

Λόγω των ανωτέρω, τίθεται συνεχώς το ζήτημα της προστασίας της ιδιωτικότητας των χρηστών. Για παράδειγμα, πληροφορίες που μπορούν να αντληθούν από μεγάλες βάσεις δεδομένων ή από χρήστες κοινωνικών δικτύων ενδεχομένως να είναι αφενός εξαιρετικά χρήσιμες για επιστημονικούς σκοπούς (π.χ. οι βάσεις δεδομένων ενός νοσοκομείου, με στοιχεία υγείας, διαδικασίας θεραπείας κτλ. έχουν έντονο ιατρικό ενδιαφέρον και μπορούν να συμβάλλουν και στη διαμόρφωση πολιτικών για τη δημόσια υγεία), αλλά αφετέρου ενδέχεται να αποκαλύψουν κρίσιμα και ευαίσθητα προσωπικά δεδομένα (π.χ. την πάθηση κάποιου προσώπου). Αυτό το φαινόμενο έχει επομένως οδηγήσει σε αναζήτηση από διάφορους ειδικούς στον χώρο της πληροφορικής να προβούν σε τεχνικές ανωνυμοποίησης, δηλαδή σε τεχνικές που προσπαθούν, κατά το δυνατόν, να ανωνυμοποιήσουν τα δεδομένα χωρίς όμως να χάνεται η χρήσιμη, προς περαιτέρω επιστημονική ανάλυση, πληροφορία. Ταυτόχρονα, υπάρχει και κατάλληλο νομικό πλαίσιο για την προστασία των προσωπικών δεδομένων, με σκοπό τη συμμόρφωση όλων σε συγκεκριμένες αρχές ώστε να προστατεύονται τα προσωπικά δεδομένα με τον καλύτερο δυνατό τρόπο.

1.1 Η έννοια των «Μεγάλων Δεδομένων» (Big Data)

Ως όρος τα Μεγάλα Δεδομένα ή Big Data χρησιμοποιείται για να περιγράψει τα σύνολα δεδομένων που είναι σύνθετα και ραγδαία μεταβαλλόμενα. Τα δεδομένα αυτά κάθε φορά αλλάζουν και η ανάλυσή τους εξαρτάται άμεσα από τη βελτίωση των υπολογιστών με την πάροδο των χρόνων. Η πληροφορία που υπάρχει στα μέσα κοινωνικής δικτύωσης, ως προς τις δραστηριότητες και συνήθειες των χρηστών, ανήκει σαφέστατα στην κατηγορία των Big Data.

Ταυτόχρονα, πέρα από τα Big Data, ζούμε στην εποχή και των «Ανοιχτών Δεδομένων» (Open Data). Τα ανοιχτά δεδομένα έχουν αυτήν την ονομασία γιατί είναι ευκόλως προσβάσιμα οπότε τόσο οι άνθρωποι όσο και οι επιχειρήσεις και οι οργανισμοί μπορούν να τα χρησιμοποιήσουν για διάφορους σκοπούς – π.χ. για να ξεκινήσουν νέες επιχειρήσεις, να αναλύσουν πρότυπα και τάσεις με τη λήψη αποφάσεων που θα βασίζονται σε αυτά τα δεδομένα, καθώς επίσης και να επιλύουν διαφόρων τύπων σύνθετα προβλήματα.

Οι ορισμοί που έχουν διατυπωθεί για τα ανοιχτά δεδομένα περιλαμβάνουν δύο βασικά χαρακτηριστικά. Πρώτον ότι τα δεδομένα πρέπει να είναι διαθέσιμα στο κοινό για οποιονδήποτε θελήσει να τα χρησιμοποιήσει και πρέπει να διαθέτουν άδεια κατά τέτοιο τρόπο που να επιτρέπει την εκ νέου χρησιμοποίησή τους. Δεύτερον τα Ανοιχτά Δεδομένα θα πρέπει να είναι σχετικά εύκολα στη χρήση, αν και υπάρχουν διαβαθμίσεις στην ελεύθερη πρόσβαση σε αυτά. Τα Ανοιχτά Δεδομένα θα πρέπει να διατίθενται δωρεάν ή με ελάχιστο κόστος.

Αν και οι δύο όροι πολλές φορές χρησιμοποιούνται, κακώς, ως συνώνυμες, υπάρχουν ουσιώδεις διαφορές μεταξύ των Big Data και των Ανοιχτών Δεδομένων: καταρχάς, τα Big data που δεν είναι ανοιχτά, δηλαδή δεν είναι ελεύθερα προσβάσιμα. Για παράδειγμα, δεδομένα μεγάλου όγκου από ιατρικά δεδομένα μπορούν να είναι διαθέσιμα μόνο σε αρμόδια ιατρικά κέντρα και όχι στο ευρύ κοινό. Στον αντίποδα, τα Ανοιχτά Δεδομένα δεν είναι απαραίτητα «μεγάλα». Τόσο βέβαια τα Big data, όσο και τα Ανοιχτά Δεδομένα αποτελούν πολύτιμο εργαλείο για πολλούς φορείς. Τα Big Data μας βοηθούν να κατανοήσουμε, να αναλύσουμε και να διαφυλάξουμε κάθε πληροφορία στον κόσμο τον

οποίο ζούμε. Τα Ανοιχτά Δεδομένα, στενά συνυφασμένα με τη διαφάνεια, διασφαλίζουν ότι η εξουσία θα πρέπει να μοιράζεται και υποστηρίζει κάθε άποψη με δημοκρατικό περιεχόμενο. (http://www.theguardian.com)

1.2 Η σπουδαιότητα και το νομικό πλαίσιο της προστασίας προσωπικών δεδομένων

Τα προσωπικά δεδομένα ή αλλιώς τα δεδομένα προσωπικού χαρακτήρα είναι, σύμφωνα με Οδηγία 95/46/ΕΚ, κάθε πληροφορία, είτε είναι άμεση είτε έμμεση, που αναφέρεται σε φυσικό πρόσωπο και χαρακτηρίζει το υποκείμενο από φυσική, βιολογική, ψυχολογική, οικονομική, πολιτιστική ή κοινωνική άποψη. Σε αυτά δεν υπολογίζονται τα προσωπικά δεδομένα που είναι στατιστικής φύσεως ή ακόμη και συγκεντρωτικά στοιχεία.

Για παράδειγμα, προσωπικά δεδομένα μπορεί να είναι η κάθε πληροφορία που μας χαρακτηρίζει ως άτομα όπως το όνομά μας η διεύθυνσή μας είτε ταχυδρομική είτε ηλεκτρονική (e-mail), το τηλέφωνό μας, τα ενδιαφέροντά μας, οι απόψεις μας, η εικόνα μας σε μορφή εικόνας ή βίντεο κ.α. Όμως ως προσωπικά δεδομένα νοούνται και άλλες πληροφορίες ακόμα και αν δεν είναι προφανές σε ποιο άτομο αναφέρονται. Για παράδειγμα, το ψευδώνυμό μας (nickname) σε μία διαδικτυακή υπηρεσία, ακόμα και αν δεν παραπέμπει στο πραγματικό μας ονοματεπώνυμο, όπως επίσης και η IP διεύθυνση του υπολογιστή μας από τον οποίο εισερχόμαστε στο διαδίκτυο.

Με το θεσμικό πλαίσιο της προστασίας προσωπικών δεδομένων, τίθενται προϋποθέσεις νομιμότητας της επεξεργασίας προσωπικών δεδομένων, καθώς επίσης αναγνωρίζονται συναφή δικαιώματα και υποχρεώσεις, στο πλαίσιο προστασίας του θεμελιώδους αγαθού της ιδιωτικότητας. Η σχετική Οδηγία 95/46/ΕΚ έχει κατάλληλα ενσωματωθεί στα Κράτη-Μέλη (π.χ. στην Ελλάδα είναι σε ισχύ ο σχετικός νόμος 2472/1997). (Cormode G, Srivastava D, Yu T, Zhang Q., 2008: 833-844)

1.3 Η σημασία της ανωνυμοποίησης

Δεδομένα τα οποία δεν μπορούν να οδηγήσουν σε αναγνώριση (ταυτοποίηση) κάποιου προσώπου, αποκαλούνται ανώνυμα δεδομένα. Τα ανώνυμα δεδομένα δεν θεωρούνται προσωπικά δεδομένα. Στην Οδηγία 95/46/ΕΚ αναφέρεται ρητώς ότι οι αρχές της

προστασίας δεν εφαρμόζονται σε δεδομένα που έχουν καταστεί ανώνυμα κατά τρόπο ώστε να μην μπορεί να εξακριβωθεί πλέον η ταυτότητα του προσώπου στο οποίο αναφέρονται. Ωστόσο, είναι σημαντικό να σημειωθεί ότι για να εκτιμηθεί αν τα δεδομένα μπορούν να οδηγήσουν σε αναγνώριση, θα πρέπει να ληφθεί υπόψη κάθε δυνατό μέσο που μπορεί να διαθέτει ένας ο οποίος θέλει να άρει την ανωνυμία: αυτό σημαίνει ότι είναι, δυστυχώς, εύκολο να θεωρούμε κάποια δεδομένα ανώνυμα χωρίς πραγματικά να είναι, γιατί δεν έχουμε σκεφτεί όλους τους πιθανούς τρόπους και μέσα που μπορεί να εφαρμόσει κάποιος για να αναγνωρίσει κάποιο πρόσωπο από τα «ανωνυμοποιημένα» δεδομένα. Η παρούσα διατριβή πραγματεύεται ζητήματα ανωνυμοποίησης σε δεδομένα εφαρμογών κοινωνικής δικτύωσης, από τη σκοπιά των γράφων κοινωνικής δικτύωσης, όπως εξηγείται στη συνέχεια. (Coull SE, Monrose F, Reiter MK, Bailey M., 2009 : 230–236)

1.4 Κοινωνικά δίκτυα και ιδιωτικότητα

Σήμερα, όλο και περισσότεροι άνθρωποι έχουν ενεργή συμμετοχή σε σελίδες κοινωνικής δικτύωσης όπως είναι για παράδειγμα το Facebook, το LinkedIn, το Instagram κ.ο.κ ώστε να μοιράζονται προσωπικές πληροφορίες με φίλους και γνωστούς τους και να επικοινωνούν εξ' αποστάσεως. Όμως όταν οι πληροφορίες αυτές διαδίδονται μέσω τους διαδικτύου, την ίδια ώρα αυτομάτως αποθηκεύονται σε αυτές τις σελίδες και αυτό τις θέτει σε κίνδυνο από διάφορους χρήστες και αυτό σημαίνει ότι ανά πάσα στιγμή η ιδιωτικότητα μπορεί να παραβιαστεί. Επομένως, για παρόχους όπως το Facebook και το LinkedIn είναι πολύ σημαντικό να προστατεύουν την ιδιωτικότητα των χρηστών και την ίδια ώρα να μπορούν να παρέχουν χρήσιμες για αυτούς πληροφορίες.

Έχουν γίνει πολλές προτάσεις ώστε να προστατευτούν οι πληροφορίες των χρηστών των κοινωνικών δικτύων. Όμως όλες αυτές οι προτάσεις παραβλέπουν κάτι πολύ σημαντικό. Πολλοί διαφορετικοί χρήστες μπορούν να έχουν διαφορετικές ρυθμίσεις και προτιμήσεις για την προστασία ιδιωτικότητας. Ως εκ τούτου η παροχή του ίδιου επιπέδου ασφάλειας σε όλους τους χρήστες εκτός ότι δεν είναι δίκαιο, μπορεί να κάνει τις πληροφορίες μέσω αυτών των δικτύων άχρηστες.

Για να λύσουν τις αδυναμίες της χρήσης ενός και μόνο επιπέδου ασφάλειας, καθορίζουμε διαφορετικά επίπεδα ασφάλειας για τους χρήστες και ενσωματώνουμε αυτά στο ίδιο δίκτυο. Στην πραγματικότητα, τα δίκτυα κοινωνικής δικτύωσης επιτρέπουν σε καθένα χρήστη ξεχωριστά να ρυθμίσει πόσες πληροφορίες θα είναι διαθέσιμες δημόσια σε

άλλους χρήστες. Για παράδειγμα στο Facebook, ένας χρήστης μπορεί να ρυθμίσει ποια κομμάτια από το προσωπικό του προφίλ μπορεί να τα δουν άλλοι χρήστες. Ο χρήστης μπορεί να έχει μια πλήρη εκτίμηση της πληροφορίας που κάποιος μπορεί να έχει για αυτόν.

Για να μπορέσουμε να καλύψουμε το σκοπό να προστατευτεί ο κάθε χρήστης σε ένα επίπεδο ισχυρότερο από αυτό που χρησιμοποιεί, μπορούμε να θέσουμε ένα μοντέλο προστασίας για όλους τους χρήστες σε ένα κοινωνικό δίκτυο με βάση το ισχυρότερο επίπεδο προστασίας που έχει απαιτηθεί από τους χρήστες. Στην απλοϊκή αυτή μέθοδο το χειρότερο σενάριο είναι να γίνει πιο κλειστό το κοινωνικό δίκτυο. Για να δείξουμε ένα παράδειγμα αυτής της μεθόδου, ας υποθέσουμε ότι έχουμε ένα κοινωνικό δίκτυο με 7 άτομα που καθένα έχει 3 στοιχεία: Όνομα, Τοποθεσία και Ηλικία. Υποθέτουμε λοιπόν ότι έχουμε 2 στόχους για την προστασία: η πιθανότητα ένας επιτιθέμενος να βρει το όνομα του χρήστη Χ είναι «Γιώργος» πρέπει να λιγότερο από 50% και η πιθανότητα κάποιος επιτιθέμενος να βρει ότι ο χρήστης Χ και ο χρήστης Ψ έχουν κάποια σύνδεση πρέπει να είναι λιγότερο από 50%. (Hay M, Miklau G, Jensen D, Towsley D., 2008:102-114)

1.4.1 Γράφοι κοινωνικών δικτύων

Ένα σύγχρονο ζήτημα, που επίσης άπτεται της ιδιωτικότητας των χρηστών κοινωνικών δικτύων, είναι η λεγόμενη ανωνυμοποίηση των γράφων κοινωνικής δικτύωσης. Συγκεκριμένα, για την ευχερέστερη επιστημονική ανάλυση των υποκείμενων δομών σε ένα κοινωνικό δίκτυο, χρησιμοποιούνται γράφοι που περιγράφουν τις συνδέσεις μεταξύ χρηστών του κοινωνικού δικτύου και τις επιμέρους συσχετίσεις τους. Αυτοί οι γράφοι κοινωνικής δικτύωσης, που περιγράφονται στο Κεφάλαιο 3 της παρούσας διατριβής, αποτελούν ένα πολύτιμο εργαλείο ανάλυσης συμπεριφορών και επιμέρους σχέσεων των χρηστών κοινωνικών δικτύων. Προφανώς, για την επίτευξη αυτών των επιστημονικών σκοπών δεν είναι αναγκαία η ταυτοποίηση των χρηστών του δικτύου, οπότε αναφερόμαστε κατά κανόνα σε ανώνυμους γράφους κοινωνικής δικτύωσης. Ωστόσο, απλά απαλείφοντας τα στοιχεία ταυτοποίησης των χρηστών σε έναν τέτοιο γράφο δεν είναι αρκετό για τη διασφάλιση της ανωνυμίας τους, αυτό ακριβώς είναι και το αντικείμενο της παρούσας διατριβής.

1.5 Δομή της διατριβής

Η παρούσα διατριβή εστιάζει σε τεχνικές ανωνυμοποίησης των γράφων κοινωνικής δικτύωσης. Ειδικότερα, η δομή της διατριβής είναι η εξής: Στο **δεύτερο κεφάλαιο** γίνεται μία εισαγωγή στην έννοια της σημασίας της ανωνυμοποίησης ή, ακριβέστερα, των δυσμενών συνεπειών που μπορούν να επέλθουν από μία μη αποτελεσματική ανωνυμοποίηση, με περιγραφή κάποιων γνωστών περιστατικών διαρροής δεδομένων από όχι καλή ανωνυμοποίηση όπως π.χ. ο κυβερνήτης της Μασαχουσέτης, το Netflix incident, ενώ εν συνεχεία γίνεται περιγραφή των κλασικών τεχνικών ανωνυμοποίησης (k-anonymity, l-diversity, t-closeness). Στο **τρίτο κεφάλαιο** θα γίνει περιγραφή του τι είναι γράφος κοινωνικής δικτύωσης (social networks graphs), τι εξυπηρετεί, αλλά και πώς μπορεί κάποιος να παραβιάσει την ιδιωτικότητα και να ανακαλύψει κάποιους χρήστες, καταδεικνύοντας με αυτόν τον τρόπο την περιγραφή γνωστών επιθέσεων. Στο **τέταρτο κεφάλαιο** γίνεται περιγραφή γνωστών τεχνικών για την ανωνυμοποίηση σε γράφους κοινωνικής δικτύωσης, πώς δηλαδή εφαρμόζονται οι προαναφερθείσες τεχνικές k-anonymity, l-diversity πάνω σε γράφους, αλλά και άλλες τεχνικές. Τέλος στο **πέμπτο κεφάλαιο** θα γίνει περιγραφή μίας νέας τεχνικής που αναπτύχθηκε στο πλαίσιο της παρούσας διατριβής, η οποία βασίζεται στην κατάλληλη κατηγοριοποίηση των κόμβων ενός γράφου με βάση το βαθμό τους, σε σχέση με το μέσο βαθμό όλων των κόμβων του γράφου. Απώτερος στόχος της τεχνικής είναι η προσθαφαίρεση ακμών στο γράφο κατά το λιγότερο δυνατό πλήθος, έτσι ώστε να επιτευχθεί επαρκής ανωνυμοποίηση με τη λιγότερη δυνατή απώλεια χρήσιμης πληροφορίας. Η ανωτέρω τεχνική έχει υλοποιηθεί σε κώδικα με τη χρήση Python. Τέλος, στο **έκτο κεφάλαιο** θα παραθέσουμε τα συμπεράσματα της παρούσας διατριβής, καθώς και ανοιχτά ερευνητικά ζητήματα.

Κεφάλαιο 2
Τεχνικές ανωνυμοποίησης

Στο παρόν κεφάλαιο θα αναφέρουμε τους τρόπους με τους οποίους μπορεί να επιτευχθεί η ανωνυμοποίηση κατά τη δημοσίευση βάσεων δεδομένων, όπως και κάποια γνωστά περιστατικά διαρροής δεδομένων λόγω μη επαρκούς ανωνυμοποίησης.

2.1 Γνωστά περιστατικά διαρροής δεδομένων από όχι επαρκή ανωνυμοποίηση

Τα τελευταία χρόνια όλο και περισσότερα δεδομένα προσώπων (για παράδειγμα, χρηστών διαφόρων τύπων υπηρεσιών, ασθενών κτλ.) δημοσιεύονται από διάφορους οργανισμούς (π.χ. νοσοκομεία, δημόσιους φορείς) ή τίθενται στη διάθεση ερευνητών, με σκοπό κυρίως να επιτρέψουν την εξαγωγή χρήσιμων ερευνητικών συμπερασμάτων, κατόπιν επιστημονικών (π.χ. στατιστικών) αναλύσεων. Από την άλλη πλευρά, η ασφάλεια των προσωπικών μας πληροφοριών που συνδέεται στενά με την ιδιωτικότητά μας, αρχίζει να απειλείται όταν τα προσωπικά δεδομένα μας τεθούν σε διάθεση τρίτων. Ακριβώς για αυτό το λόγο, και επειδή τα δεδομένα που είναι αξιοποιήσιμα για τους ανωτέρω σκοπούς περιέχουν πολλές προσωπικές πληροφορίες, τα προσωπικά δεδομένα «αποκρύβονται». Δυστυχώς, αυτό δεν σημαίνει ότι κάποιος κακόβουλος χρήστης δεν είναι σε θέση να εκμεταλλευτεί τον μεγάλο όγκο πληροφοριών – δεδομένων που μπορεί να συλλέγει για κάθε άτομο σε διάφορες βάσεις δεδομένων και να ανακαλύψει σε ποιον αναφέρονται τα δεδομένα, ακόμα και αν η ταυτότητά του έχει αποκρυφτεί. Με άλλα λόγια, ο επιτιθέμενος μπορεί να συνδυάσει δεδομένα τα οποία έχουν δημοσιευθεί από διάφορες πηγές, καθώς και τυχόν δική του πληροφόρηση, για να άρει την ανωνυμοποίηση των δεδομένων.

Ως χαρακτηριστικό παράδειγμα, υπάρχει ένα συμβάν με «ανώνυμα» δεδομένα που σχετίζεται με το νοσοκομείο της Μασαχουσέτης. Το νοσοκομείο δημοσίευσε ανώνυμα

δεδομένα, τα οποία παρουσιάζονται στον Πίνακα 1. Ο πίνακας αυτός θεωρήθηκε ανωνυμοποιημένος, καθώς έχουν απομακρυνθεί όλα τα στοιχεία ταυτοποίησης (ονοματεπώνυμο, αριθμός ταυτότητας κτλ.). Ωστόσο, ερευνητές κατάφεραν να αποκαλύψουν την ταυτότητα προσώπων αυτής της λίστας, αξιοποιώντας δημόσια προσβάσιμα δεδομένα από τους εκλογικούς καταλόγους (Πίνακας 2).Όπως μπορεί να διαπιστωθεί, οι δύο λίστες (η ανώνυμη του νοσοκομείου και η επώνυμη εκλογική) έχουν τρία κοινά πεδία: ταχυδρομικός κώδικας, φύλο και ημερομηνία γέννησης (Σχήμα 1). Αν ένας έχει μοναδική τιμή σε αυτήν την τριπλέτα δεδομένων και βρίσκεται και στις δύο λίστες, προφανώς ταυτοποιείται στην «ανώνυμη» λίστα. Αξίζει να σημειωθεί ότι, με αυτόν τον τρόπο, ταυτοποιήθηκε ο κυβερνήτης της Πολιτείας της Μασαχουσέτης στην ανώνυμη λίστα του νοσοκομείου.(https://epic.org/privacy/reidentification/Sweeney_Article.pdf).Οπότε και ανακτήθηκαν ευαίσθητες πληροφορίες υγείας από – φαινομενικά - ανωνυμοποίημενα δεδομένα.

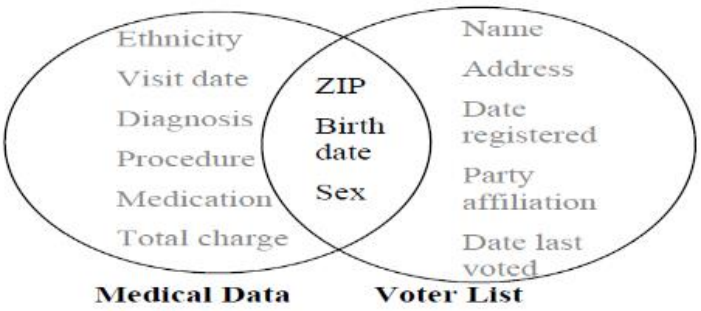
Ουσιαστικά, στο παράδειγμα που περιγράφηκε ανωτέρω, έγινε χρήση της σύνδεσης στοιχείων (linking ή matching) - δηλαδή έγινε αντιστοίχιση ή σύνδεση εγγραφών ανάμεσα σε δεδομένα διαφορετικής προέλευσης, επειδή υπήρχαν κοινά πεδία.

Birthdate	Sex	Zipcode	Disease
1/21/76	Male	53715	Flu
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Sprained Ankle
2/28/76	Female	53706	Hang Nail

Πίνακας 1. Δημοσιευθέντα ιατρικά δεδομένα

Name	Birthdate	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237

Πίνακας 2. Δεδομένα δημοσίων εκλογικών καταλόγων (Voter Registration Data)



Σχήμα 1. Σύνδεση (Linking ή matching) δεδομένων για εκ νέου εντοπισμό δεδομένων.

Συνεπώς, απώτερος στόχος μας είναι τα δεδομένα μας να μας παρέχουν επαρκείς πληροφορίες για τους ερευνητικούς σκοπούς που θέλουμε, αλλά ταυτόχρονα ένας επίδοξος επιτιθέμενος που απειλεί να αποκαλύψει τα προσωπικά δεδομένα να μην έχει αυτή τη δυνατότητα, με άλλα λόγια δηλαδή να μην μπορεί να εξάγει ευαίσθητες προσωπικές πληροφορίες ή να μην είναι σε θέση να βρει σε ποιο φυσικό πρόσωπο ανήκει μια καταχώρηση στον ανωνυμοποιημένο πίνακα. Κάθε τεχνική τροποποίησης των αρχικών δεδομένων με σκοπό τη διατήρηση της ανωνυμίας αποκαλείται τεχνική ανωνυμοποίησης (anonymization technique).

Ήδη από το παραπάνω παράδειγμα γίνεται σαφές ότι, απλά απομακρύνοντας τα μοναδικά στοιχεία ταυτοποίησης σε μία βάση δεδομένων, δεν διασφαλίζεται απαραίτητα η ανωνυμία. Για παράδειγμα, στον Πίνακα 3 που ακολουθεί, αν πρόκειται για υπαλλήλους ενός οργανισμού, στον οποίο οργανισμό εργάζεται μόνο ένας άνδρας δικηγόρος ετών 38, τότε αυτομάτως αποκαλύπτεται η πάθηση του εν λόγω προσώπου (μια που αντιστοιχεί μονοσήμαντα στην τρίτη εγγραφή του πίνακα).

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

Πίνακας 3. Δημοσιευθέντα δεδομένα σχετικά με ασθενείς

Το περιστατικό της Netflix

Το έτος 2006 η Netflix, που είναι μια Αμερικανική πολυεθνική εταιρία με αντικείμενο τη διαδικτυακή ενοικίαση ταινιών, δημοσιοποίησε τις αξιολογήσεις που έκαναν οι

εγγεγραμμένοι σε αυτή χρήστες σε ταινίες. Οποιοδήποτε στοιχείο με το οποίο θα μπορούσε να γίνει ταυτοποίηση είχε εξαλειφθεί. Το έτος 2007 οι ερευνητές Narayanan and Shmatikov ταυτοποίησαν σημαντικό ποσοστό των χρηστών της Netflix, με βάση τις δημόσια προσβάσιμες αξιολογήσεις σε ταινίες που έκαναν οι ίδιοι χρήστες στην πλατφόρμα IMDB. Με αυτή την ταυτοποίηση αποκαλύφθηκαν και ευαίσθητα δεδομένα βάσει συγκεκριμένων αξιολογήσεων ταινιών που έγιναν στη Netflix που είχε υπολογιστεί την προηγούμενη χρονιά ότι δεν υπήρχε περίπτωση να δημοσιοποιηθούν.

Τελικά αντιλαμβανόμαστε ότι χρειάζεται ιδιαίτερη προσοχή στη δημοσιοποίηση των δεδομένων όχι μόνο των δικών μας αλλά και των εταιριών που τυχόν εκπροσωπούμε. Με ένα πιθανό λάθος μπορούν να αποκαλυφθούν πολύ περισσότερες πληροφορίες από εκείνες που υπολογίζαμε.

2.2 Επιθέσεις ως προς την άρση της ανωνυμοποίησης

Αν θέλουμε να αποτυπώσουμε τους κινδύνους που ελλοχεύουν από μία μη ορθή ανωνυμοποίηση – δηλαδή από αποκάλυψη δεδομένων που θεωρούνται ανώνυμα αλλά που όμως κάποιος μπορεί να οδηγηθεί σε ταυτοποίηση κάποιων προσώπων – αυτοί θα μπορούσαν να περιγραφούν ως εξής:

- **Η αποκάλυψη ταυτότητας (identity disclosure)**

  Είναι η κατηγορία της επίθεσης, κατά την οποία ο επιτιθέμενος αποσκοπεί στην αναγνώριση της ταυτότητας ενός ατόμου χρησιμοποιώντας έναν ή περισσότερους δημοσιευμένους πίνακες δεδομένων.

- **Η αποκάλυψη του ''ευαίσθητου'' γνωρίσματος (attribute disclosure)**

  Είναι από τις κατηγορίες επίθεσης κατά την οποία ο επιτιθέμενος αποσκοπεί στον καθορισμό της τιμής ενός ή περισσότερων ευαίσθητων δεδομένων μίας εγγραφής (ή ισοδύναμα ενός ατόμου) με μεγάλη πιθανότητα σε ένα πίνακα δεδομένων. Η αποκάλυψη της ευαίσθητης πληροφορίας μπορεί να είναι είτε «θετική» (δηλαδή εξάγουμε συμπέρασμα ότι κάποιο πρόσωπο έχει μία συγκεκριμένη ιδιότητα, και αυτό ισχύει με μεγάλη πιθανότητα) ή «αρνητική» (δηλαδή εξάγουμε συμπέρασμα ότι κάποιο πρόσωπο δεν έχει μία συγκεκριμένη ιδιότητα, και αυτό ισχύει με μεγάλη πιθανότητα).

Οι συνέπειες μίας τέτοιας άρσης της ανωνυμοποίησης ποικίλουν, ανάλογα με το είδος των δεδομένων: προφανώς, μπορούν να είναι εξαιρετικά δυσμενείς (π.χ. αποκάλυψη κάποιας σοβαρής πάθησης).

2.3 Περιγραφή των κλασικών τεχνικών ανωνυμοποίησης

Για τη διαβίβαση ή δημοσιοποίηση βάσεων (πινάκων) δεδομένων της μορφής που είδαμε στην προηγούμενη ενότητα κατά τρόπο τέτοιο ώστε να είναι ανώνυμες, υπάρχουν κάποιες τεχνικές ανωνυμοποίησης. Πριν τις παρουσιάσουμε, είναι σημαντικό να αναλυθούν κάποιοι όροι οι οποίοι θα βοηθήσουν στην κατανόηση των τεχνικών ανωνυμοποίησης.

- **Πίνακας (Table):** Τα δεδομένα που βρίσκονται αποθηκευμένα σε μία βάση, είναι οργανωμένα σε μορφή πίνακα σχεσιακής βάσης δεδομένων όπου τα A1, A2,..,An είναι οι στήλες -γνωρίσματά του.
- **Στήλη – Γνώρισμα (Attribute):** Γνώρισμα που αντιπροσωπεύει μια κατηγορία πληροφορίας και έχει ένα σύνολο πιθανών τιμών (π.χ. φύλο, ηλικία, κτλ.). Πρόκειται ουσιαστικά για τις κατηγορίες που εμφανίζονται στις στήλες του πίνακα.
- **Εγγραφή ή Πλειάδα (Tuple):** Πρόκειται για μια καταχώρηση στον πίνακα (με απλά λόγια, μία γραμμή του πίνακα)η οποία αφορά ένα άτομο. Κάθε εγγραφή έχει συγκεκριμένες τιμές στα αντίστοιχα γνωρίσματα
- **Αναγνωριστικά (Identifiers):** Είναι γνωρίσματα των οποίων οι τιμές που μπορούν να αποκαλύψουν απ' ευθείας το άτομο που αντιστοιχεί σε μία εγγραφή του πίνακα (π.χ. ο αριθμός ταυτότητας, αριθμός κοινωνικής ασφάλισης, ονοματεπώνυμο).
- **Ψευδό-αναγνωριστικά(Quasi-Identifier):** Γνωρίσματα τα οποία κατ' αρχάς δεν ταυτοποιούν απ' ευθείας το άτομο αλλά αν συνδυαστούν με κάποιες εξωτερικές πληροφορίες μπορούν να προσδιορίσουν μοναδικά το άτομο που αντιστοιχεί σε μια εγγραφή ενός πίνακα (π.χ. ταχυδρομικός κώδικας, ημερομηνία γέννησης, φύλο).

- **Ευαίσθητα Γνωρίσματα (Sensitive Attributes):** Είναι γνωρίσματα των οποίων την πληροφορία θέλουμε να προστατεύσουμε γιατί είναι κρίσιμη για το άτομο στο οποίο αντιστοιχεί (π.χ. ασθένεια, μισθός)
- **Η κλάση ισοδυναμίας (equivalence class):** Ως κλάση ισοδυναμίας, ορίζουμε κάθε σύνολο εγγραφών που έχουν ακριβώς τις ίδιες τιμές στα ψευδο-αναγνωριστικά. Οι κλάσεις ισοδυναμίας είναι εντελώς ξένες μεταξύ τους.

Στο προηγούμενο παράδειγμα (Πίνακας 3), στον «ανωνυμοποιημένο» πίνακα του νοσοκομείου, η τριπλέτα «Τ.Κ. – ημερομηνία γέννησης – φύλο» αποτελεί ψευδό-αναγνωριστικά (που, τελικά, επέτρεψαν την αναγνώριση κάποιου προσώπου), ενώ η ασθένεια αποτελεί το ευαίσθητο γνώρισμα.

- **Γενίκευση δεδομένων (generalization)**

Η γενίκευση δεδομένων είναι μία διαδικασία για την τροποποίηση των τιμών στα ψευδό-αναγνωριστικά κατά τρόπο τέτοιο ώστε να σχηματίζονται κλάσεις ισοδυναμίας που να αποτελούνται από πολλές εγγραφές.

Αυτό επιτυγχάνεται γενικότερα με χρήση μίας ιεραρχίας γενίκευσης (generalization hierarchy) όπου κάθε τιμή του αρχικού πεδίου (γνωρίσματος) αντικαθίσταται από μία «γενικευμένη» μορφή, έτσι ώστε να εξακολουθεί να φέρει σχετική πληροφορία.

Η γενίκευση μπορεί να υφίσταται τόσο σε αριθμητικές τιμές όσο και σε μη αριθμητικές. Οι αριθμητικές τιμές κατά την γενίκευσή τους αντιστοιχίζονται σε ένα διάστημα τιμών. Για παράδειγμα οι τιμές 30, 35 και 38 θα μπορούσαν να γενικευθούν στην τιμή "3*" που σημαίνει οποιαδήποτε τιμή από 30 ως 38 ή στην τιμή [31-40] κλπ. Η τιμή [31-40] θα μπορούσε να γενικευθεί κι αυτή σε μια άλλη τιμή όπως "≤50". Για τις μη αριθμητικές τιμές μπορούμε να θεωρήσουμε ως παράδειγμα μια ιεραρχία όπου ο Δικηγόρος (Lawyer) και ο Μηχανικός (Engineer) γενικεύονται στην τιμή "Επαγγελματίας", η Συγγραφέας (Writer) και η Χορεύτρια (Dancer) γενικεύονται στην τιμή "Καλλιτέχνης" και με την σειρά τους οι δύο αυτές τιμές γενικεύονται στην τιμή "Επαγγέλματα".

Σε ό,τι αφορά την ανωνυμοποίηση, το πόσα επίπεδα πρέπει να "αυξηθεί" μια τιμή κατά την γενίκευσή της εξαρτάται κυρίως από την συχνότητα εμφάνισης των διάφορων επιμέρους τιμών.

- **Η αρχή της ιδιωτικότητας (privacy principle):**

Η επιτυχία του επιτιθέμενου μπορεί να μετρηθεί με την διαφορά της αρχικής του πεποίθησης ότι η ζητούμενη τιμή είναι s και της τελικής του πεποίθησης, η οποία διαμορφώνεται μετά τον εντοπισμό ενός συνόλου εγγραφών που μπορεί να αντιστοιχούν στο αναζητούμενο άτομο στον ανωνυμοποιημένο πίνακα δεδομένων. Η διαφορά αυτή δεν θα πρέπει να είναι μεγάλη.

- **Η κανονικοποιημένη Ποινή Βεβαιότητας (Normalized Certainty Penalty)**

  Έστω κλάση ισοδυναμίας G και γνώρισμα A_N. Για αριθμητικά δεδομένα η Κανονικοποιημένη Ποινή Βεβαιότητας μίας κλάσης ισοδυναμίας ορίζεται ως:

$$NCP_{A_N}(G) = \frac{max_{A_N}^G - min_{A_N}^G}{max_{A_N} - min_{A_N}}$$

  Ο παραπάνω τύπος ισχύει για ένα γνώρισμα A_N (ψευδοαναγνωριστικό). Αν θέλουμε μία μετρική που να αντανακλά την απώλεια πληροφορίας σε μία κλάση ισοδυναμίας από τη γενίκευση όλων των ψευδο-αναγνωριστικών, θα έχουμε:

$$NCP(G) = \sum_{i=1}^{d} w_i NCP_{A_i}(G)$$

  όπου d το πλήθος των ψευδο-αναγνωριστικών και w_i μία τιμή βαρύτητας που έχει κάθε ψευδο-αναγνωριστικό (όπου $\sum_\iota \ w_i = 1$). Φυσικά, είναι προφανές πως ενώ η παραπάνω μετρική προσφέρει χρήσιμη πληροφορία για την απώλεια πληροφορίας μέσα σε μία κλάση ισοδυναμίας, χρειαζόμαστε μία μετρική για να την μετράμε καθολικά κατά μήκος όλων των κλάσεων. Για αυτό το σκοπό εισάγεται η έννοια της Συνολικής Ποινής Βεβαιότητας.

- **Η Συνολική Ποινή Βεβαιότητας (Global Certainty Penalty)**

  Έστω P το σύνολο όλων των κλάσεων ισοδυναμίας στον ανωνυμοποιημένο πίνακα, N ο αριθμός των εγγραφών στον αρχικό πίνακα και d το πλήθος των ψευδο-αναγνωριστικών. Η Συνολική Ποινή Βεβαιότητας ορίζεται ως :

$$GCP(P) = \frac{\sum_{G \in P} |G| NCP(G)}{dN}$$

Όπου |G| είναι το πλήθος των εγγραφών στην κλάση ισοδυναμίας G. Το πλεονέκτημα αυτού το ορισμού είναι κυρίως το εύρος από 0 έως 1 που δίνει, με 0 να είναι η ιδανική περίπτωση μη απώλειας δεδομένων.

2.3.1 Ανωνυμία k τάξης (k-anonymity)

Η τεχνική της ανωνυμίας k τάξης μας βοηθά να προστατεύσουμε την ιδιωτικότητα των δεδομένων μας. Προτάθηκε από την Sweeney και έχει σαν σκοπό να διασφαλιστεί η μη δυνατότητα ταυτοποίησης της κάθε εγγραφής που περιέχεται σε κάποια βάση δεδομένων. Στόχος της συγκεκριμένης τεχνικής είναι να τροποποιηθούν οι διάφορες τιμές των δεδομένων στις εγγραφές ώστε ο επιτιθέμενος να έχει ελάχιστη πιθανότητα να προσδιορίσει μια οντότητα ακόμα και με την διασταύρωση εγγράφων που μπορεί να είναι ανωνυμοποιημένες. Ουσιαστικά με την μέθοδο αυτή εξασφαλίζεται ότι, η πιθανότητα ανακάλυψης της ταυτότητας μιας εγγραφής είναι το πολύ 1/k – και αυτό γιατί κάθε νέα εγγραφή στον ανωνυμοποιημένο πίνακα θα ανήκει σε μία κλάση ισοδυναμίας με πληθικότητα τουλάχιστον k.

Παράδειγμα

Παρακάτω παρουσιάζεται ένας πίνακας με ιατρικά δεδομένα ενός συνόλου ασθενών.

#	Age	Zipcode	Disease
1	29	47677	Heart Disease
2	22	47602	Heart Disease
3	27	47678	Heart Disease
4	43	47905	Flu
5	52	47909	Heart Disease
6	47	47906	Cancer
7	30	47605	Heart Disease
8	36	47673	Cancer
9	32	47607	Cancer

Πίνακας 4. Δημοσιευθέντα δεδομένα Ασθενών

Όπως έχει ήδη αναφερθεί, γνωρίσματα όπως το ΑΦΜ ή ο Αριθμός Ταυτότητας ενός ατόμου αποτελούν μοναδικά αναγνωριστικά. Όταν δημοσιοποιούμε μια βάση δεδομένων και στόχος μας είναι να προστατεύουμε την ιδιωτικότητα των φυσικών προσώπων, αυτού του είδους οι πληροφορίες δεν εμφανίζονται για να μην μπορεί κάποιος να προσδιορίσει μια οντότητα. Τα γνωρίσματα που απαιτείται να δημοσιευθούν και περιέχουν προσωπική πληροφορία, χωρίζονται σε δύο σύνολα αναλόγως με την πληροφορία που αντιπροσωπεύουν, σε ψευδό-αναγνωριστικά και ευαίσθητα γνωρίσματα. Ένα χαρακτηριστικό σύνολο γνωρισμάτων που μπορεί να αποτελέσει ψευδό-αναγνωριστικό (**Quasi-Identifier)** για άτομα που εμφανίζονται σε δημόσιους καταλόγους είναι το σύνολο {Ημερομηνία γεννήσεως, Διεύθυνση, Ταχυδρομικός κωδικός, Φύλο}. Στον Πίνακα 4 ένα σύνολο ψευδό- αναγνωριστικού είναι {ηλικία, τ.κ.}, συνεπώς αν κάποιος το συσχετίσει με άλλους καταλόγους (για παράδειγμα, τους εκλογικούς καταλόγους) μιας περιοχής, μόνο κάποιος περιορισμένος αριθμός ατόμων θα εμφανίζει τις ίδιες τιμές στα κοινά γνωρίσματα των δύο συλλογών. Αυτό έχει σαν συνέπεια ο επιτιθέμενος να μπορεί να εξάγει μια προσωπική πληροφορία που θα πρέπει να μείνει απόρρητη και διασφαλισμένη και συνήθως κρίνεται ως ευαίσθητο γνώρισμα.

Στον Πίνακα 4 το ευαίσθητο γνώρισμα είναι η ασθένεια του ασθενούς. Συνεπώς μπορεί να γίνει εύκολα αντιληπτό ότι με την κατάλληλη επιλογή ενός συνόλου ψευδό-αναγνωριστικών ο επιτιθέμενος μπορεί με το κατάλληλη συσχέτιση/διασύνδεση, όπως περιγράψαμε στην αρχή του Κεφαλαίου, να άρει την ανωνυμία.

Στόχος μας είναι να διαφοροποιήσουμε τα δεδομένα μας στον πίνακα έτσι ώστε να ικανοποιεί να δημιουργούνται κλάσεις ισοδυναμίας όπου η κάθε μια να έχει τουλάχιστον k εγγραφές: με αυτόν τον τρόπο, κάθε προσπάθεια συσχέτισης με εξωτερικά δεδομένα θα οδηγεί σε μία αβεβαιότητα k εγγραφών (δηλαδή θα γνωρίζουμε ότι ένα πρόσωπο είναι ένας εκ των τουλάχιστον k ανωνυμοποιημένων εγγραφών, αλλά δεν θα ξέρουμε ποια ακριβώς εγγραφή). Ουσιαστικά, οι εγγραφές μιας κλάσης ισοδυναμίας είναι αξεχώριστες μεταξύ τους.

#	**Age**	**Zipcode**	**Disease**
1	20-37	476**	Heart Disease
2	20-37	476**	Heart Disease

3	20-37	476**	Heart Disease
4	38-55	4790*	Flu
5	38-55	4790*	Heart Disease
6	38-55	4790*	Cancer
7	20-37	476**	Heart Disease
8	20-37	476**	Cancer
9	20-37	476**	Cancer

Πίνακας 5. 3ης τάξης k-ανωνυμίας

Στόχος μας είναι να έχουμε μεγαλύτερη ανωνυμοποίηση των δεδομένων μας χωρίς να χάνουμε μέρος από τις πληροφορίες μας. Από τον Πίνακα 5 συμπεραίνουμε ότι όσο μεγαλύτερο είναι το k τόσο πιο μεγάλη ανωνυμοποίηση των δεδομένων έχουμε. Σε αυτό σημείο παρουσιάζεται ένα πολύ σοβαρό μειονέκτημα τις τεχνικής αυτής, όσο μεγαλύτερο είναι το k τόσο περισσότερη πληροφορία χάνουμε.

Στον ανωτέρω πίνακα έχει επιτευχθεί ανωνυμία τάξης k=2 – υπάρχουν δύο κλάσεις ισοδυναμίας (η πρώτη αποτελείται από τις εγγραφές 1,2,3,7,8,9 και η δεύτερη από τις εγγραφές 4,5,6), όπου το ελάχιστο πλήθος εγγραφών σε μία κλάση είναι 2. Αυτό έγινε μέσω της διαδικασίας της γενίκευσης γνωρισμάτων, όπως περιγράφηκε ανωτέρω. Άλλος τρόπος επίτευξης της ανωνυμίας k τάξης είναι να απομακρύνουμε κάποιες εγγραφές («γενίκευση» - generalization).

Γενικότερα, θα πρέπει να λαμβάνομε υπόψιν μας τρεις παραμέτρους:

- Κατάργηση (Suppression): πόσες εγγραφές αφαιρούνται από τα δεδομένα στη διαδικασία της ανωνυμοποίησης
- Γενίκευση(Generalization): πόση πληροφορία χάνεται γενικεύοντας τα δεδομένα σε κάποιο επίπεδο γενίκευσης
- Ανωνυμία (Anonymity): ποιο είναι το ελάχιστο ανεκτό μέγεθος k για κάθε κλάση ισοδυναμίας.

<u>Επίθεση ομοιογένειας</u>

Παρότι ένας πίνακας μπορεί να ικανοποιεί την k-ανωνυμία, ο επιτιθέμενος μπορεί να ανακαλύψει και να καταλήξει στην αποκάλυψη ενός ευαίσθητου γνωρίσματος εξαιτίας την ομογένειας που μπορεί να υπάρχει στον πίνακα.

#	**Age**	**Zipcode**	**Disease**
1	2*	476**	Heart Disease

2	2*	476**	Heart Disease
3	2*	476**	Heart Disease
4	≥40	4790*	Flu
5	≥40	4790*	Heart Disease
6	≥40	4790*	Cancer
7	3*	476**	Heart Disease
8	3*	476**	Cancer
9	3*	476**	Cancer

Πίνακας 6. 3ης τάξης k-ανωνυμία

Στον παραπάνω δημοσιευμένο πίνακα με τα ιατρικά δεδομένα ασθενών ικανοποιείται η k-ανωνυμία για k=3. Μάλιστα, σε σχέση με τον Πίνακα 5, ο Πίνακας 6 έχει μικρότερη απώλεια πληροφορίας γιατί υπάρχουν τρεις κλάσεις ισοδυναμίας, ενώ στον Πίνακα 5 είχαμε, όπως ήδη είδαμε, δύο κλάσεις. Αλλά αν ο επιτιθέμενος γνωρίζει ότι το πρόσωπο το οποίο αναζητά είναι 27 χρονών και ζει στην περιοχή με τκ 47678 συνεπώς, ανήκει σε μια εκ των εγγράφων 1-3 (δηλαδή στην πρώτη κλάση ισοδυναμίας), αυτομάτως αποκτά την πληροφορία ότι το πρόσωπο αυτό έχει εγκεφαλικό. Κατά συνέπεια η ομοιογένεια ανάμεσα στα στοιχεία του πίνακα αποκάλυψε το ευαίσθητο γνώρισα για το συγκεκριμένο πρόσωπο. Επίσης, συγκρίνοντας τους Πίνακες 5 και 6 γίνεται κατανοητό ότι η λιγότερη απώλεια πληροφορίας συνεπάγεται άυξηση της πιθανότητας για ανεπαρκή ανωνυμοποίηση.

Επίθεση με προηγούμενη γνώση

Τα μοντέλα επιθέσεων που περιγράφηκαν νωρίτερα βασίζονται στην υπόθεση της προηγούμενης γνώσης που ο επιτιθέμενος κατέχει σχετικά με ένα πρόσωπο: συγκεκριμένα, γνωρίζει ότι το πρόσωπο το οποίο αναζητά βρίσκεται στο δημοσιευμένο πίνακα και επίσης έχει και κάποια πρόσθετη πληροφορία ή γενικότερη γνώση, η οποία του επιτρέπει να αποκλείσει πιθανές ευαίσθητες τιμές παρόλο που ο πίνακας ικανοποίει την k-ανωνυμία. Σύμφωνα με το παράδειγμα μας ο επιτιθέμενος γνωρίζει ότι το άτομο το οποίο αναζητά είναι 32 χρονών και ζει στην περιοχή με τκ 47622 συνεπώς, ανήκει σε μια εκ των εγγράφων 7-9 (δηλαδή στην τρίτη κλάση ισοδυναμίας), αυτομάτως αποκτά την πληροφορία ότι το πρόσωπο αυτό έχει καρκίνο. Κατά συνέπεια η ομοιογένεια ανάμεσα στα στοιχεία του πίνακα αποκάλυψε το ευαίσθητο γνώρισα για το συγκεκριμένο πρόσωπο.

2.3.2 Ποικιλομορφία τάξης λ (l-diversity)

Η δεύτερη τεχνική για την ανωνυμοποίηση δεδομένων που χρησιμοποιείται είναι η ποικιλομορφία τάξης λ (l-diversity). Στόχος και αυτής της τεχνικής είναι ο επιτιθέμενος να μην μπορεί να ανακαλύψει την τιμή των ευαίσθητων γνωρίσματων. Σύμφωνα με τον ορισμό της λ-ποικιλομορφίας (l-diversity) [MGK+06], ο αρχικός πίνακας δεδομένων RT(A1,A2,….An,S) περιέχει ένα ευαίσθητο γνώρισμα S και έχει ανωνυμοποιηθεί με τεχνικές γενίκευσης, από όπου προκύπτει ο γενικευμένος πίνακας RT*(A1,A2,….An,S). Σε αυτόν, οι εγγραφές χωρίζονται σε κλάσεις ισοδυναμίας ως προς τις τιμές των γνωρισμάτων του ψευδό-αναγνωριστικού τους. Μια κλάση ισοδυναμίας ορίζεται ως λ-ποικιλόμορφη (l-diverse) αν περιέχει τουλάχιστον λ «καλώς ορισμένες τιμές» για το ευαίσθητο γνώρισμα. Αντίστοιχα, ένας πίνακας είναι λ-ποικιλόμορφος εάν κάθε κλάση ισοδυναμίας του είναι λ-ποικιλόμορφη.

Age	Zipcode	Salary	Disease
29	47677	3K	gastric ulcer
22	47602	4K	gastritis
27	47678	5K	stomach cancer
43	47905	6K	gastritis
52	47909	11K	Flu
47	47906	8K	bronchitis
30	47605	7K	bronchitis
36	47673	9K	pneumonia
32	47607	10K	stomach cancer

Πίνακας 7. Αρχικός Πίνακας Μισθού(Salary)/ Ασθένειας(Disease)

Age	Zipcode	Salary	Disease
2*	476**	3K	gastric ulcer
2*	476**	4K	gastritis
2*	476**	5K	stomach cancer
≥40	4790*	6K	gastritis
≥40	4790*	11K	Flu
≥40	4790*	8K	bronchitis
3*	476*	7K	bronchitis
3*	476*	9K	pneumonia
3*	476*	10K	stomach cancer

Πίνακας 8. 3ης τάξης λ-ποικιλομορφία

Στον Πίνακα 8 υπάρχει ποικιλομορφία τάξης λ=3, αφού σε κάθε κλάση ισοδυναμίας υπάρχουν τουλάχιστον 3 διαφορετικές τιμές στο ευαίσθητο γνώρισμα (μάλιστα εδώ έχουμε δύο ευαίσθητα γνωρίσματα, το μισθό και την ασθένεια: η ποικιλομορφία τάξης 3 ισχύει και για τα δύο αυτά ευαίσθητα γνωρίσματα). Η λ-ποικιλομορφία, όπως μπορεί να παρατηρήσει κάποιος, αποτελεί σημαντική τεχνική για όποιον αποφασίσει να

δημοσιεύει έναν πίνακα, κυρίως γιατί μπορεί να αποτρέψει επιθέσεις με προηγούμενη γνώση. Οποιαδήποτε γνώση μπορεί να έχει ο επιτιθέμενος θεωρείται απλώς ένας τρόπος να αποκλείσει κάποια τιμή και, θα χρειάζεται να αποκλείσει άλλες λ-1 τιμές για να μπορέσει να εντοπίσει την εγγραφή την οποία αναζητά.

Ένα βασικό μειονέκτημα της τεχνικής αυτής είναι ότι είναι δύσκολο να υλοποιηθεί σε μικρές βάσεις δεδομένων, ιδίως αν υπάρχουν κάποιες τιμές στο ευαίσθητο γνώρισμα πιο σπάνια εμφανιζόμενες από άλλες. Σε αυτήν την περίπτωση, ο επιτιθέμενος μπορεί να εξάγει κάποια συμπεράσματα με αρκετή βεβαιότητα. Όπως ότι μπορεί να συμπεράνει με πολύ μεγάλη πιθανότητα ότι η νόσος του ατόμου που ενδιαφέρεται έχει να κάνει με στομαχικές διαταραχές.

Ένα άλλο αρνητικό της λ-ποικιλομορφίας αποτελεί το γεγονός ότι είναι ευάλωτη σε κάποιου είδους επιθέσεις, όπως για παράδειγμα επιθέσεις ασυμμετρίας (skewness attack) και επιθέσεις ομοιότητας (similarity attack). Οι επιθέσεις ασυμμετρίας σχετίζονται με τη μη ομοιόμορφη κατανομή των τιμών του ευαίσθητου πεδίου: αν μία τιμή εμφανίζεται μόνο στο 1% των εγγραφών συνολικά, αλλά κατά την ανωνυμοποίηση εμφανίζεται σε μία κλάση ισοδυναμίας με π.χ. 10 εγγραφές, τότε για κάποιον που γνωρίζουμε ότι ανήκει σε αυτήν την κλάση ισοδυναμίας μεγαλώνει η πιθανότητα να έχει αυτήν τη σπάνια τιμή στο ευαίσθητο γνώρισμα. Από την άλλη πλευρά, οι επιθέσεις ομοιότητας έχουν να κάνουν με το ότι μπορεί να υπάρχουν λ διαφορετικές τιμές στο ευαίσθητο γνώρισμα αλλά να είναι πανομοιότυπες (π.χ. όλες να είναι δερματικές παθήσεις), οπότε πάλι μπορεί να εξαχθεί ασφαλές συμπέρασμα για κάποιο πρόσωπο που βρίσκεται σε αυτήν την κλάση ισοδυναμίας.

2.3.3 Η τ-Εγγύτητα (t-Closeness)

Μια άλλη βελτίωση της ανωνυμίας k τάξης είναι η t-εγγύτητα (t-closeness): στόχος αυτής της τεχνικής είναι η κατανομή των τιμών ενός ευαίσθητου γνωρίσματος σε κάθε κλάση ισοδυναμίας να είναι κοντά (κατά μία παράμετρο t που εκφράζει την απόσταση των κατανομών) στην κατανομή των τιμών του γνωρίσματος αυτού στον αρχικό πίνακα. Όσο πιο μικρή είναι η τιμή του t, τόσο πιο κοντά βρίσκονται οι δύο κατανομές.

Προκειμένου να ελαχιστοποιηθούν οι τιμές του t στην πρώτη κλάση ισοδυναμίας, ανακατασκευάζεται ο πίνακας 7 όπως φαίνεται πιο κάτω στον Πίνακα 9.

Age	Zipcode	Salary	Disease
≤40	4767*	3K	gastric ulcer
≤40	4767*	5K	stomach cancer
≤40	4767*	9K	pneumonia
≥40	4790*	6K	gastritis
≥40	4790*	11K	Flu
≥40	4790*	8K	bronchitis
≤40	4760*	7K	gastritis
≤40	4760*	7K	bronchitis
≤40	476*	10K	stomach cancer

Πίνακας 9. Ανωνυμία t-εγγύτητας

Η τεχνική αυτή έχει την δυνατότητα να προστατεύσει επιθέσεις που έχουν στόχο να αποκαλύψουν κάποιο ευαίσθητο γνώρισμα αλλά, σε καμία περίπτωση δεν παρέχει προστασία σε επιθέσεις που σχετίζονται με την αποκάλυψη της ταυτότητας μια εγγραφής.

Κεφάλαιο 3

Γράφοι κοινωνικής δικτύωσης και θέματα ιδιωτικότητας

Στο κεφάλαιο αυτό θα μελετήσουμε τους γράφους κοινωνικής δικτύωσης, σε τι μας εξυπηρετούν, τη σχέση του ατόμου με αυτούς και θα παρακολουθήσουμε πόσο εύκολο είναι κάποιος επιτιθέμενος να παραβιάσει την ιδιωτική μας ζωή.

3.1 Τι είναι γράφος κοινωνικής δικτύωσης (social network graph)

Ένας τρόπος για την αναπαράσταση των κοινωνικών δικτύων είναι μέσω των γράφων.

3.1.1. Η έννοια του γράφου

Ένας γράφος G χαρακτηρίζεται από δύο σύνολα V και E. Το σύνολο V είναι ένα πεπερασμένο σύνολο, που περιέχει ως στοιχεία τους κόμβους (vertices). Το σύνολο E περιέχει τα ζεύγη κόμβων του γράφου, τα οποία ορίζουν τις ακμές (edges) ή τόξα (arcs) ή συνδέσμους (links) του. Οι ακμές ενός γράφου χαρακτηρίζονται από ένα μοναδικό όνομα που ονομάζεται ετικέτα (label). Γράφος με βάρη στις ακμές (weighted graph) λέγεται ένας γράφος, όπου με κάθε ακμή του έχει συσχετισθεί ένας αριθμός που ονομάζεται βάρος (weight). Διαδρομή (path) καλείται μια διάταξη κόμβων οι οποίοι ενώνονται με ακμές. Απόσταση (distance) μεταξύ δύο κόμβων καλείται το μήκος της συντομότερης διαδρομής που τους ενώνει.

Ένας γράφος είναι μη κατευθυνόμενος αν τα ζεύγη των κόμβων που ορίζουν τις ακμές του στερούνται διάταξης, π.χ., τα ζεύγη (v, u) και (u, v) αναφέρονται στην ίδια ακμή (όπου u, v κόμβοι που συνδέονται με ακμή). Στους κατευθυνόμενους γράφους κάθε ακμή συμβολίζεται με το κατευθυνόμενο ζεύγος <v, u>, όπου v είναι η ουρά (tail) και u είναι η κεφαλή (head) της ακμής (έτσι, οι ακμές <v, u> και <u, v> είναι δυο διαφορετικές ακμές).

Συνεκτικός (connected) ονομάζεται ο γράφος για τον οποίο υπάρχει διαδρομή από κάθε κόμβο σε κάθε άλλο κόμβο. Βαθμός ενός κόμβου σε έναν μη κατευθυνόμενο γράφο καλείται ο αριθμός των συνδεδεμένων με αυτόν ακμών. Βαθμός εισόδου (in-degree) σε έναν κατευθυνόμενο γράφο καλείται ο αριθμός των ακμών που καταλήγουν σε αυτόν. Βαθμός εξόδου (*out-degree*) σε έναν κατευθυνόμενο γράφο καλείται ο αριθμός των ακμών που ξεκινούν από αυτόν. Υπογράφος (*subgraph*) B ενός γράφου A καλείται ο γράφος του οποίου όλες οι ακμές έχουν ως ένα άκρο τον κόμβο A. Ένας μη κατευθυνόμενος γράφος μπορεί να θεωρηθεί ως ένας συμμετρικός κατευθυνόμενος γράφος.

Αν (v, u) ∈ E(G), τότε οι κόμβοι v και u λέγονται διπλανοί (adjacent) ή γειτονικοί (neighboring) και η ακμή (v, u) ονομάζεται προσκείμενη στους κόμβους v και u. Αν δύο κόμβοι v και u δεν συνδέονται μεταξύ τους με ακμή λέγονται ανεξάρτητες (independent).

Αν (v, u) είναι μια ακμή τότε ο κόμβος v λέγεται διπλανός (adjacent) της u. Επίσης, οι κόμβοι v και u λέγονται γειτονικοί.

Ένας γράφος μπορεί εύκολα να παρασταθεί με δύο διαφορετικούς τρόπους και προϋποθέτουν ένα μονοσήμαντο σύστημα αρίθμησης των κόμβων.

- Πίνακα γειτνίασης (***adjacency matrix***).
  Σε έναν γράφο με ν κόμβους ο πίνακας ν*ν γειτνίασης περιέχει 1 στις θέσεις (κ,λ) όπου υπάρχει ακμή από τον κόμβο κ στον κόμβο λ.

- Λίστα γειτνίασης (***adjacency list***).
  Κάθε κόμβος συσχετίζεται με μια συνδεδεμένη λίστα γειτνίασης η οποία περιέχει τα ονόματα των άλλων κόμβων με τους οποίους συνδέεται ο συγκεκριμένος κόμβος. Οι λίστες για όλους τους κόμβους ν μπορούν να ξεκινούν από έναν μονοδιάστατο πίνακα μήκους ν ή από μια άλλη συνδεδεμένη λίστα.

3.1.2 Αναπαράσταση κοινωνικών δικτύων με γράφους

Υπάρχουν διάφορες τεχνικές ανάλυσης των κοινωνικών δικτύων εκ των οποίων η πιο σημαντική είναι η ανίχνευση της κοινότητας σε κοινωνικά δίκτυα.

Λόγω της τάσης των ανθρώπων για τη δημιουργία καθώς και την ανάγκη τους να ανήκουν σε διαφορετικές ομάδες, όπως οικογενειακές, φιλικές, επαγγελματικές είτε και σε ομάδες που έχουν κοινά ενδιαφέροντα, μπορεί να προκύψουν διαφορετικές αναπαραστάσεις αυτών των κοινοτήτων μέσα σε ένα κοινωνικό δίκτυο.

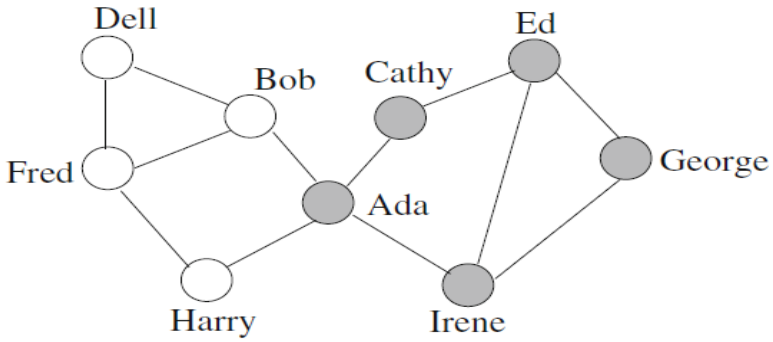
Λόγω των πολύπλοκων κοινωνικών δομών στα κοινωνικά δίκτυα, πολλές ενδιαφέρουσες γνώσεις είναι κρυμμένες στις δομές των γράφων. Ο θεμελιώδης στόχος της ανάλυσης των κοινωνικών δικτύων είναι ότι πρέπει να ανακαλυφθούν τα κρυμμένα κοινωνικά πρότυπα. Πρώιμες μελέτες της ανάλυσης των κοινωνικών δικτύων έχουν επικεντρωθεί στην ανάλυση των χαρακτηριστικών μεμονωμένων κοινωνικών φορέων.

Ωστόσο, πρόσφατη ανάπτυξη της ανάλυσης των κοινωνικών δικτύων έχει δείξει ότι οι πολύπλοκες κοινωνικές σχέσεις μεταξύ των κοινωνικών φορέων είναι συχνά πιο σημαντική και ενημερωτική από τα χαρακτηριστικά των επιμέρους κοινωνικών φορέων.

Πολλοί διαφορετικοί τύποι των τεχνικών ανάλυσης των κοινωνικών δικτύων έχουν αναπτυχθεί για την άντληση ενδιαφερουσών πληροφοριών από τα κοινωνικά δίκτυα. Μεταξύ αυτών, η ανίχνευση της κοινότητας σε κοινωνικά δίκτυα είναι μία από τις πιο σημαντικές τεχνικές ανάλυσης των κοινωνικών δικτύων. Οι άνθρωποι τείνουν να σχηματίζουν ομάδες, μέσα σε εργασιακό περιβάλλον, την οικογένειά τους ή τους φίλους σας. Οι κοινότητες στα κοινωνικά δίκτυα, που μερικές φορές αναφέρονται και ως συστάδες ή συμπλέγματα, είναι ομάδες των κοινωνικών φορέων που ίσως έχουν κάποια κοινές ιδιότητες.

Αναλόγως με τις διαφορετικές μοντελοποιήσεις των ιδιοτήτων, οι κοινότητες στο κοινωνικό δίκτυο μπορεί να έχουν διαφορετικές αναπαραστάσεις. Ένας κλασικός τρόπος αναπαράστασης των κοινωνικών δικτύων, που επιτρέπει αναλύσεις της μορφής που περιγράφηκε ανωτέρω, είναι μέσω γράφων: κάθε κόμβος αντιστοιχεί σε έναν χρήστη του κοινωνικού δικτύου – που μπορεί να φέρει πληροφορία αναγνωριστικών και ψευδό-αναγνωριστικών, κατά την έννοια που αυτά περιγράφηκαν στο Κεφάλαιο 2 – όπου συνδέεται με ακμή με έναν άλλον κόμβο μόνο αν οι δύο αυτοί χρήστες στο δίκτυο είναι «φίλοι». Η ακμή που τους συνδέει μπορεί να φέρει ετικέτα που να περιγράφει το είδος της σύνδεσης (π.χ. «οικογένεια», «συμμαθητές», «συνάδελφοι» κ.ο.κ.). Αυτή η δομή χαρακτηρίζεται και ως γράφος κοινωνικής δικτύωσης (social network graph).

Στο παρακάτω παράδειγμα μπορούμε να παρατηρήσουμε αυτές τις αναπαραστάσεις (όπου, για λόγους απλότητας, δεν υπάρχουν ετικέτες στις ακμές):



Σχήμα 2. Ένα απλό δίκτυο φιλίας με δυο κοινότητες. Η μια κοινότητα περιέχει κόμβους σε γκρι χρώμα και η άλλη κοινότητα σε λευκό χρώμα.

3.2 Τι εξυπηρετεί ο γράφος κοινωνικής δικτύωσης – Ανωνυμοποίηση γράφου

Ο γράφος κοινωνικής δικτύωσης μας βοηθάει να κατηγοριοποιήσουμε τα δεδομένα στα οποία βασίζεται η συμπεριφορά των χρηστών. Γενικότερα μας βοηθά στο να κατανοήσουμε διάφορα θέματα βάσει των διασυνδέσεων των ατόμων στα κοινωνικά δίκτυα, όπως είναι για παράδειγμα τα ενδιαφέροντα, οι προτιμήσεις ή οι τοποθεσίες, καθώς επίσης και η αναγνώριση των πιο σημαντικών ανθρώπων που δημιουργούν τις δικές τους ομάδες.

Για την επίτευξη των ανωτέρω στόχων, δεν είναι απαραίτητο να αντανακλώνται στο γράφο οι ταυτότητες των χρηστών: για αυτόν το λόγο, τα αναγνωριστικά των κόμβων-χρηστών (ονοματεπώνυμα, ηλεκτρονικές διευθύνσεις κτλ.) δεν περιέχονται σε τέτοιες δομές, όταν πρόκειται να γίνει ανάλυση του γράφου για τους σκοπούς που περιγράφηκαν ανωτέρω. Ωστόσο, όπως είδαμε και στο Κεφάλαιο 2, η απαλοιφή των αναγνωριστικών δεν είναι αρκετή για να χαρακτηριστεί ένα σύνολο δεδομένων ως ανώνυμο. Μάλιστα, ειδικά σε γράφους κοινωνικής δικτύωσης, οι επιθέσεις που απειλούν να άρουν την ανωνυμία μπορεί να είναι ακόμα πιο ισχυρές, όπως θα δούμε στη συνέχεια.

3.3 Πως μπορεί κάποιος να παραβιάσει την ιδιωτικότητα των γράφων κοινωνικής δικτύωσης

Το γνωστικό πεδίο του επιτιθέμενου είναι ένας σημαντικός παράγοντας σε πολλές έρευνες για την προστασία της ιδιωτικής ζωής όσον αφορά τη δημοσίευση των δεδομένων στα κοινωνικά δίκτυα (Kleinberg JM., 2007:4-5). Ο επιτιθέμενος χρησιμοποιεί διαφορετικές στρατηγικές για την παραβίαση της ιδιωτική ζωής του χρήστη αναλόγως την προηγούμενη γνώση που διαθέτει (Liu K, Terzi E ., 2008: 93-106).

Λόγω των πολύπλοκων δομών γράφων των κοινωνικών δικτύων, οι βασικές γνώσεις που μπορεί να έχει ο επιτιθέμενος κατηγοριοποιούνται με διαφορετικούς τρόπους – κάποιους εκ των οποίων θα δούμε στη συνέχεια.

1. **Προσδιορισμός των χαρακτηριστικών των κόμβων**. Στα κοινωνικά δίκτυα, ο κόμβος μπορεί να συνδέεται με μοναδικό τρόπο με ένα άτομο με βάση ορισμένα χαρακτηριστικά. Τα χαρακτηριστικά αυτά μπορούμε να τα ταυτίσουμε, όπως αναφέρθηκε και παραπάνω, με τα ψευδό-αναγνωριστικά που υπάρχουν στα σχεσιακά δεδομένα (Liu K., Das K., Grandison T., Kargupta H., 2008). Τα χαρακτηριστικά του κόμβου συχνά αναπαρίστανται ως ετικέτες στα κοινωνικά δίκτυα (Zheleva E, Getoor L., 2011) και (Zhou B., Pei J., Luk W.-S., 2008: 3).

2. Ένας επιτιθέμενος μπορεί να γνωρίζει κάποιες τιμές από αυτά τα χαρακτηριστικά για συγκεκριμένα «θύματα» (Campan A., Marius RT., 2008).

3. **Ο βαθμός ενός κόμβου στο δίκτυο**. Ο βαθμός ενός κόμβου στα κοινωνικά δίκτυα καταγράφει πόσοι άλλοι άνθρωποι είναι συνδεδεμένοι με το αντίστοιχο άτομο στο δίκτυο. Οι πληροφορίες αυτές είναι συχνά εύκολο να συλλεχθούν από τους επιτιθέμενους. Για παράδειγμα, ο γείτονας ενός χρήστη μπορεί εύκολα να εκτιμήσει πόσους φίλους έχει το θύμα. Ένας επιτιθέμενος γνωρίζοντας το βαθμό του θύματος και εξετάζοντας τους βαθμούς των κόμβων μπορεί να επαναπροσδιορίσει το αντικείμενο-στόχο από τα δημοσιευμένα δίκτυα (Hay M., Miklau G., Jensen D., Weis Ph. & Don T., 2008: 102-114).

4. **Η σχέση της γειτονιάς ενός ατόμου στο δίκτυο.** Ένας επιτιθέμενος μπορεί να γνωρίζει τη γειτονιά ορισμένων αντικειμένων-στόχων (Liu K., Terzi E., 2008:93-106). Για παράδειγμα, ένας επιτιθέμενος μπορεί να ξέρει ότι το θύμα έχει τέσσερις στενούς φίλους που γνωρίζουν επίσης ο ένας τον άλλον (Ying X., Wu X., 2008: 739-750). Χρησιμοποιώντας αυτό το είδος της γνώσης υποβάθρου, ο επιτιθέμενος μπορεί να

επαναπροσδιορίσει το θύμα αναζητώντας κόμβους στο δημοσιευμένο γράφημα, των οποίων οι γειτονικές δομές περιέχουν μια κλίκα μεγέθους τουλάχιστον τέσσερα (Hay M., Miklau G., Jensen D., Weis Ph., Srivastava S., 2007:7-19).

5. **Υπογράφος.** Αυτό αναφέρεται σε ένα σύνολο σχέσεων στην οποία ο χρήστης είναι μέλος ενός συγκεκριμένου υπογράφου εντός του συνολικού γράφου. Για παράδειγμα, ένας επιτιθέμενος μπορεί να δημιουργήσει διάφορα ψεύτικα προφίλ και να συνδεθεί κοινωνικά μέσω αυτών χρησιμοποιώντας συγκεκριμένα μοτίβα. Ο επιτιθέμενος στη συνέχεια χρησιμοποιεί αυτά τα ψεύτικα προφίλ για να δημιουργήσει μια σύνδεση με τον χρήστη που θέλει να επιτεθεί. Αυτή η σύνδεση μπορεί να αποδειχθεί τόσο εύκολη όσο και η προσθήκη σύνδεσης στη λίστα φίλων του χρήστη (Backstrom Â. L., Dwork C., Kleinberg J., 2007:181-190).

6. **Η σχέση μεταξύ των ατόμων στο δίκτυο.** Ένας επιτιθέμενος μπορεί να γνωρίζει ότι υπάρχουν κάποιες συγκεκριμένες σχέσεις μεταξύ ορισμένων αντικειμένων-στόχων. Για παράδειγμα, σε ένα κοινωνικό δίκτυο για τη σχέση φιλίας, οι ακμές μπορούν να φέρουν ετικέτες που καταγράφουν το πώς οι άνθρωποι επικοινωνούν μεταξύ τους, όπως τηλέφωνο, ηλεκτρονικό ταχυδρομείο, ή/και σε άμεσο μήνυμα. Ένας επιτιθέμενος μπορεί να προσπαθήσει να χρησιμοποιήσει αυτή τη βασική γνώση όπου το θύμα χρησιμοποιεί μόνο τα μηνύματα για να επικοινωνήσει με τους φίλους του στο δίκτυο και με αυτόν τον τρόπο μπορεί να συνδέσει το θύμα σε συγκεκριμένους κόμβους στο δίκτυο (Campan A., Marius T. T., 2008) και (Ying X., Wu X., 2008:739-750).

7. **Ορισμένοι υπο-γράφοι ενσωματώνονται στο δίκτυο.** Ένας επιτιθέμενος μπορεί να εισβάλλει σε ορισμένους καλά δομημένους υπογράφους σε ένα κοινωνικό δίκτυο, πριν το δίκτυο κυκλοφορήσει δημοσίως. Αφού συλλέξει το δημοσιευμένο δίκτυο, μπορεί να αναγνωρίσει τον ενσωματωμένο υπογράφο εφόσον είναι μοναδικός. Για παράδειγμα, έχει καταδειχτεί ότι ο επιτιθέμενος με τη δημιουργία 7 κόμβων μπορεί να αποκαλύψει ένα μέσο όρο 70 κόμβων χρηστών σε ένα μεγάλο δίκτυο (Zheleva E., Getoor L., 2008:153-171).

3.3.1 Σενάρια παραβίασης της ιδιωτικής ζωής

Πρώτο σενάριο παραβίασης της ιδιωτικής ζωής

Ένα σύνηθες σενάριο της παραβίασης της ιδιωτικής ζωής θεωρείται όταν τα κοινωνικά δίκτυα διανέμουν τα δεδομένα σε τρίτους. Σε αυτό το σενάριο, ο εκδότης των κοινωνικών δικτύων (π.χ. το Twitter) είναι αξιόπιστος και οι χρήστες των κοινωνικών δικτύων είναι πρόθυμοι να παράσχουν τις προσωπικές τους πληροφορίες στον συγκεκριμένο εκδότη. Ωστόσο, η εμπιστοσύνη αυτή δεν είναι μεταβατική και στους αποδέκτες των δεδομένων (π.χ., οι αναλυτές δεδομένων) που θα διεξάγουν διάφορες αναλυτικές εργασίες στα δεδομένα που έχουν δημοσιευθεί.

Ειδικότερα, παρατηρώντας το κοινωνικό δίκτυο στο Σχήμα 2, η προηγούμενη γνώση θα μπορούσε να είναι μερικά τμήματα των πληροφοριών που ένας επιτιθέμενος γνωρίζει εκ των προτέρων για τα άτομα στο δίκτυο.

Ας υποθέσουμε ότι ο επιτιθέμενος γνωρίζει τα ονόματα του κάθε ατόμου στο δίκτυο. Ο επιτιθέμενος γνωρίζει επίσης ότι ο Ed ζει στη Washington, DC. Κάποιες πληροφορίες έχουν αφαιρεθεί για τη διατήρηση της ιδιωτικότητας πριν δημοσιευθούν τα δεδομένα αυτά σε τρίτους. Για παράδειγμα, ο George δε θέλει να γίνει γνωστή η τοποθεσία του και για αυτό το λόγο την αφαιρεί πριν κυκλοφορήσει στο κοινωνικό δίκτυο.

Παρ' όλα αυτά, το να αφαιρέσει την τοποθεσία του δεν είναι αρκετό για να διασφαλιστεί η προστασία της ιδιωτικής του ζωής. Στην πραγματικότητα, εάν ο επιτιθέμενος διεξάγει μια εργασία ανίχνευσης της κοινότητας σχετικά με τα δεδομένα, ο επιτιθέμενος μπορεί να ανακαλύψει ότι ο George και ο Ed είναι στην ίδια κοινότητα. Σύμφωνα με αυτήν την υπόθεση, δηλαδή ότι οι χρήστες ανήκουν στην ίδια κοινότητα, το οποίο με τη σειρά του συνεπάγεται ότι μπορεί να μοιράζονται κάποιες κοινές ιδιότητες (π.χ., ότι ζουν κοντά ο ένας στον άλλο), ο επιτιθέμενος μπορεί εύκολα να συμπεράνει ότι ο George ζει στην Ουάσιγκτον, DC με μεγάλη πιθανότητα.

<u>Δεύτερο σενάριο παραβίασης της ιδιωτικής ζωής</u>

Η πρόσβαση που μας παρέχουν πολλοί διαδικτυακοί τόποι κοινωνικής δικτύωσης σχετικά με το απόρρητο των πληροφοριών μας περιορίζεται μόνο σε απευθείας φίλους.

Για παράδειγμα, παρατηρώντας το απλό κοινωνικό δίκτυο που απεικονίζεται στο Σχήμα 2, ο Ed μπορεί να θέλει να επιτρέψει μόνο στους απευθείας φίλους του Cathy, Irene, και George να βλέπουν τις κοινωνικές δραστηριότητές του. Ωστόσο, για μερικές δημοφιλείς

ιστοσελίδες κοινωνικής δικτύωσης έχει αναφερθεί ότι έχουν τρωτά σημεία στις ρυθμίσεις απορρήτου τους.

Για παράδειγμα, με την εισαγωγή των ανοικτών APIs στις ιστοσελίδες κοινωνικής δικτύωσης δημιουργείται ένας νέος τρόπος παράκαμψης των ρυθμίσεων του ελέγχου πρόσβασης.

Μια περίπτωση πραγματικού γεγονότος στο Facebook αποτελεί ένα αξιόλογο παράδειγμα:

Το Facebook δημοσίευσε το API της κοινωνικής δικτύωσής του σε προγραμματιστές τρίτων φορέων/εταιριών για να σχεδιάσουν διάφορα είδη εφαρμογών. Ωστόσο, όταν ένας χρήστης εγκαταστήσει αυτές τις εφαρμογές στο ιδιωτικό προφίλ του, οι συγκεκριμένοι προγραμματιστές αμέσως λαμβάνουν τα προνόμια των ιδιοκτητών των προφίλ και μπορούν να ρωτήσουν το API για τις προσωπικές πληροφορίες των χρηστών και των μελών των ομάδων των χρηστών. Επίσης, έχει αναφερθεί στο ότι η κρυφή ημερομηνία γέννησης πολλών χρηστών μπορεί να εκτεθεί στο Facebook (Sophos, 2008).

Ορισμένες πρόσφατες μελέτες έδειξαν επίσης ότι τα προσωπικά δεδομένα των χρηστών ακόμα και αν προστατεύονται από τις ρυθμίσεις απορρήτου μπορεί εύκολα να τεθούν σε κίνδυνο λόγω του ότι η πολιτική για τη προστασία της ιδιωτικής ζωής που παρέχονται από αυτές τις ιστοσελίδες κοινωνικής δικτύωσης είναι εγγενώς ευάλωτη (Ruan X., Yue C., Wang H., 2013).

Για παράδειγμα, σε πολλές δικτυακούς τόπους κοινωνικής δικτύωσης όπως το Facebook, το Twitter και το LinkedIn, η λίστα των φίλων μπορεί να ρυθμιστεί ώστε να "προστατεύεται", το οποίο σημαίνει μόνο οι άμεσοι φίλοι μπορούν να έχουν πρόσβαση στην λίστα των φίλων. Ωστόσο, οι χρήστες σε αυτές τις ιστοσελίδες κοινωνικής δικτύωσης μπορούν να ρυθμίσουν αναλόγως με τις δικές τους προτιμήσεις το απόρρητο. Αυτό σημαίνει ότι άλλες ιδιωτικές ρυθμίσεις έχει ένας χρήστης Α και άλλες ένας χρήστης Β.

Όπως φαίνεται στο Σχήμα 2, ο Ed και ο George είναι απευθείας φίλοι. Αν ο George έχει θέσει τη λίστα των φίλων του ως «προστατευμένη», αλλά ο Ed την έχει θέσει ως

"δημόσια", τότε ένας επιτιθέμενος, παρόλο που δεν έχει οποιαδήποτε σύνδεση με τον Ed ή με το George, εξακολουθεί να είναι σε θέση να καταλήξει στο συμπέρασμα ότι ο Ed είναι ένας από τους φίλους του George. Κατά συνέπεια, η λίστα των φίλων του George δεν προστατεύεται όπως ισχυρίζεται.

Οι Liu, Zheleva και Getoor πρότειναν διαφορετικές κατηγοροποιήσεις σχετικά με την προστασία της ιδιωτικής ζωής στα κοινωνικά δίκτυα (Liu K., Das K., Grandison T., Kargupta H., 2008). Η αποκάλυψη της ιδιωτικότητας σε ένα κοινωνικό δίκτυο μπορεί να περιγραφεί με ομαδοποίηση σε τρεις κατηγορίες:

1) Η αποκάλυψη της ταυτότητας: η ταυτότητα του ατόμου, που συνδέεται με έναν κόμβο, αποκαλύπτεται.

2) Η γνωστοποίηση της σύνδεσης: οι ευαίσθητες σχέσεις μεταξύ δύο ατόμων αποκαλύπτονται (Zheleva E., Getoor L., 2007).

3) Η αποκάλυψη των ευαίσθητων γνωρισμάτων: τα ευαίσθητα δεδομένα που σχετίζονται με κάθε κόμβο διακυβεύονται π.χ., τα μήνυματα ηλεκτρονικού ταχυδρομείου που αποστέλλονται/παραλαμβάνονται από το άτομο σε ένα δίκτυο επικοινωνίας μέσω ηλεκτρονικού ταχυδρομείου (Zheleva E., Getoor L., 2011).

3.3.2 Τρόποι προστασίας της ιδιωτικής ζωής

Με βάση τα παραπάνω χρειαζόμαστε μια συστηματική μέθοδο για την προστασία των δεδομένων όταν αυτά πρόκειται να δημοσιευθούν σε μορφή γράφου κοινωνικής δικτύωσης. Η προστασία της ιδιωτικής ζωής μπορεί να παραβιαστεί εάν τα δεδομένα δημοσιευθούν εσφαλμένα (μη λαμβάνοντας δηλαδή τις απαραίτητες ενέργειες) σε ένα κοινωνικό δίκτυο. Στην ουσία, χρειαζόμαστε μια συστηματική μέθοδο για την ανωνυμοποίηση των δεδομένων κοινωνικών δικτύων πριν από την κυκλοφορία του. Σχετικά με την ανωνυμοποίηση των δεδομένων προκύπτουν τα εξής τρία **ζητήματα**. (Machanavajjhala A., Gehrke J., Kifer D., Venkitasubramaniam M., 2006: 17-18)

Το πρώτο ζήτημα είναι ότι η μοντελοποίηση στο υπόβαθρο γνώσεων των αντιπάλων (προηγούμενη γνώση) και στις επιθέσεις σχετικά με τα δεδομένα των κοινωνικών δικτύων, είναι πολύ πιο περίπλοκη και πολύ πιο δύσκολη από ότι στα σχεσιακά δεδομένα. Ο λόγος είναι ότι είναι πολλά τα τμήματα εκείνα των πληροφοριών σε ένα κοινωνικό δίκτυο που μπορούν να χρησιμοποιηθούν για τον εντοπισμό ατόμων, όπως οι ετικέτες

των κόμβων και ακμών, γειτονικών γράφων, η δομή των υπο-γράφων, καθώς και οι συνδυασμοί τους.

Το δεύτερο ζήτημα είναι ότι η μέτρηση της απώλειας πληροφοριών στην ανωνυμοποίηση των δεδομένων των κοινωνικών δικτύων είναι πολύπλοκη λόγω του ότι μπορεί να υπάρχουν πολλοί διαφορετικοί τρόποι για να καθορίσει κάποιος τα μέτρα απώλειας πληροφοριών και την ποιότητα ανωνυμίας. Για παράδειγμα, ένα κοινωνικό δίκτυο αποτελείται από ένα σύνολο κόμβων και ένα σύνολο ακμών. Εν αντιθέσει με την περίπτωση των σχεσιακών δεδομένων, δεν μπορούμε να συγκρίνουμε τα δύο κοινωνικά δίκτυα με την απλή σύγκριση των κόμβων και των ακμών ξεχωριστά, δεδομένου ότι τα δύο κοινωνικά δίκτυα μπορεί να είναι αρκετά διαφορετικά, ακόμη και αν έχουν τον ίδιο αριθμό κόμβων και τον ίδιο αριθμό ακμών. Ως εκ τούτου, πρέπει να εξετάσουμε περισσότερες ιδιότητες σχετικά με το δίκτυο, όπως τη συνδεσιμότητα, τη σχετικότητα, τη διάμετρο, και τη δομή του δικτύου.

Το τρίτο ζήτημα είναι ότι η επινόηση μεθόδων ανωνυμοποίησης για τα δεδομένα των κοινωνικών δικτύων είναι πολύ πιο δύσκολη, λόγω του ότι η μέθοδοι που είδαμε στο Κεφάλαιο 2 μπορεί να μην είναι χρήσιμες. Και αυτό διότι αλλάζοντας/μεταβάλλοντας τις ετικέτες των κόμβων και των ακμών μπορεί να επηρεαστούν οι γειτονιές και των άλλων κόμβων, ενώ επίσης και με την αφαίρεση ή την προσθήκη κόμβων και ακμών μπορεί να επηρεαστούν άλλοι κόμβοι και ακμές, καθώς και οι ιδιότητες του δικτύου.

Συμπερασματικά, θα πρέπει να προσδιοριστούν ποιες πληροφορίες από την ιδιωτική ζωή θεωρούνται σημαντικές και, άρα, πιθανες να δεχθούν επίθεση. Δεύτερον, πρέπει να διαμορφωθεί η προηγούμενη γνώση του αντιπάλου που μπορεί να χρησιμοποιήσει για να επιτεθεί στην προστασία της ιδιωτικής ζωής. Τέλος, θα πρέπει να καθοριστεί η χρησιμότητα των δημοσιευμένων δεδομένων των κοινωνικών δικτύων έτσι ώστε μια μέθοδος ανωνυμίας να προσπαθήσει να διατηρήσει τη χρησιμότητα των στοιχείων όσο το δυνατόν περισσότερο, διατηρώντας ταυτόχρονα και την προστασία της ιδιωτικής ζωής πλήρως.

Κεφάλαιο 4
Τεχνικές ανωνυμοποίησης σε γράφους κοινωνικής δικτύωσης

Παραπάνω μελετήσαμε τις βασικές έννοιες ανωνυμοποίησης και πιθανούς τρόπους εισβολής των επιτιθέμενων. Στο παρόν κεφάλαιο θα εστιάσουμε στις τεχνικές ανωνυμοποίησης σε γράφους κοινωνικής δικτύωσης.

4.1 Περιγραφή γνωστών τεχνικών για την ανωνυμοποίηση σε γράφους κοινωνικής δικτύωσης και η εφαρμογή τους

Αν και η προστασία και ανωνυμοποίηση της ιδιωτικής ζωής στα δεδομένα των κοινωνικών δικτύων είναι μια νέα πρόκληση, παρ' όλα αυτά έχουν αναπτυχθεί διάφορα μοντέλα και μέθοδοι για τη διατήρησή της (Liu K., Das K., Grandison T., Kargupta H., 2008), (Wu X., Ying X., Liu K., Chen L., 2010). Για τη δημοσίευση των δεδομένων στα κοινωνικά δίκτυα έχουν προταθεί διαφορετικές μέθοδοι (Hay M., Miklau G., Jensen D., Towsley D. F., Weis P., 2008) αναλόγως με την προηγούμενη γνώση που έχει ο επιτιθέμενος.

Έχουμε ήδη κατηγοριοποιήσει (βλ. Κεφάλαιο 3) τις υπάρχουσες μεθόδους ανωνυμίας σχετικά με τη δημοσίευση των δεδομένων των κοινωνικών δικτύων σε τρεις κατηγορίες (He. X., Vaidya J., Shafiq. B., Adam N., Atluri V., 2009), όπως:

α) Μέθοδοι για τη διατήρηση της ταυτότητας (Zhou B., Pei J., Luk W., 2008)
β) Μέθοδοι για τη διατήρησης της σύνδεσης (Liu K., Terzi E., 2008)
γ) Μέθοδοι για τη διατήρηση των ευαίσθητων χαρακτηριστικών.

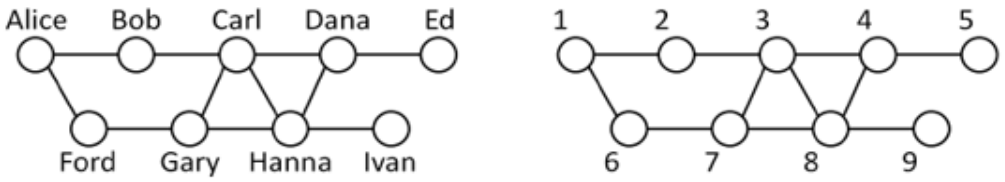
Τα προτεινόμενα μοντέλα σε κάθε μία από αυτές τις κατηγορίες θα αναλυθεί εκτενέστερα παρακάτω.

α) Μέθοδοι για τη διατήρηση της ταυτότητας

Αυτή η μέθοδος ασχολείται με την προστασία της ταυτότητας του ατόμου από το να προσδιοριστεί εκ νέου μέσα σε ένα δίκτυο. Επισήμως, το πρόβλημα μπορεί να οριστεί ως εξής:

Δεδομένου ότι τα στοιχεία δημοσιεύονται στο κοινωνικό δίκτυο, αν ένας επιτιθέμενος είναι σε θέση να προσδιορίσει τον κόμβο ενός χρήστη αναλύοντας τα τοπολογικά χαρακτηριστικά του βασισμένος στην προηγούμενη γνώση που έχει σχετικά με το άτομο από το κοινωνικό δίκτυο, τότε η ταυτότητα του χρήστη αποκαλύπτεται.

Ένας απλός τρόπος για την προστασία του χρήστη από το να προσδιοριστεί εκ νέου από τον επιτιθέμενο είναι να έχουν ανωνυμοποιηθεί τα δεδομένα. Με τη συγκεκριμένη μέθοδο η ανωνυμοποίηση των δεδομένων επιτυγχάνεται με την αφαίρεση των αναγνωριστικών, όπως π.χ. το όνομα, έτσι ώστε το συγκεκριμένο άτομο να μην ταυτοποιείται από τους κόμβους στον πραγματικό κόσμο. Αυτός ο συμβατικός τρόπος είναι γνωστός και ως απλοϊκή Ανωνυμοποίηση όπως φαίνεται στο Σχήμα 3b. Ωστόσο, ακόμη και μετά την αφαίρεση των επιμέρους αναγνωριστικών του χρήστη, δεν αρκεί για να εξασφαλιστεί η προστασία της ιδιωτικής ζωής του, λόγω του ότι η δομή του γράφου παραμένει ίδια.



Σχήμα 3. Παράδειγμα επίθεση φιλίας. α) Αρχικός κοινωνικός γράφος και β) Απλοϊκή ανωνυμοποίηση γράφου

Ένας γράφος είναι k-ανώνυμος αν για κάθε ερώτημα αναγνώρισης κόμβου βάσει της δομής του γράφου υπάρχουν τουλάχιστον k κόμβοι που το ικανοποιούν. Για παράδειγμα, αν ο επιτιθέμενος θέλει να επαναπροσδιορίσει την κόμβο της Hanna και ξέρει επίσης ότι η Hanna έχει τέσσερις φίλους στο γράφο τότε ο επιτιθέμενος προσδιορίζει με 50%

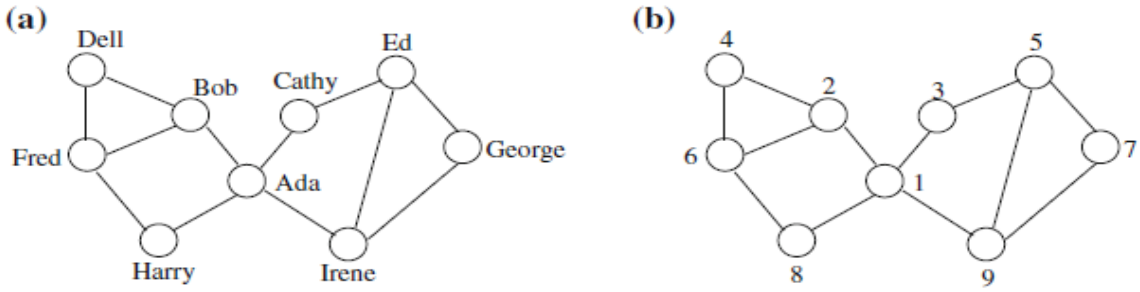
πιθανότητα ποιος είναι ο κόμβος της Hanna, αφού ο κόμβος που θα μπορούσε να αντιστοιχεί στην Hanna είναι είτε αυτός με τον 3 είτε αυτός με τον αριθμό 8 (Σχήμα 3) – γεγονός που σημαίνει ότι, ως προς αυτό το ερώτημα, έχουμε ανωνυμία τάξης 2.

Διάφορες ***μέθοδοι ανωνυμίας*** έχουν αναπτυχθεί για να αντιμετωπίσουν τις επιθέσεις για την εκ νέου ταυτοποίηση κόμβου με βάση τα χαρακτηριστικά του γράφου. Πολλές από αυτές τις μεθόδους είναι επίσης εμπνευσμένες από την έννοια της k-ανωνυμίας που χρησιμοποιείται στα σχεσιακά δεδομένα. Στην παρακάτω υπό-ενότητα, θα αναλυθούν τα διάφοροι πρότυπα σχετικά με την διατήρηση της ταυτότητας από πλευράς των δομικών χαρακτηριστικών του γράφου.

4.1.1 Η k-Ανωνυμία

Το μοντέλο της k-ανωνυμίας (ανωνυμία k τάξης) προτάθηκε έτσι ώστε ο χρήστης να είναι αξεχώριστος από τουλάχιστον k-1 χρήστες, όταν ο επιτιθέμενος χρησιμοποιεί κάποια προηγούμενη γνώση ως προς τα ψευδο-αναγνωριστικά του. Εάν η προηγούμενη γνώση είναι αποκλειστικά συσχετισμένη με τα χαρακτηριστικά ενός χρήστη σε ένα γράφο, τότε μπορούμε να εφαρμόσουμε άμεσα τις τεχνικές ανωνυμίας που είδαμε στο Κεφάλαιο 2. Με άλλα λόγια, θεωρώντας ότι ο κόμβος ενός γράφου περιέχει τα ψευδό-αναγνωριστικά του αντίστοιχου χρήστη, τότε προβαίνουμε στις κλασικές τεχνικές ανωνυμοποίησης (π.χ. γενίκευση) πάνω στα ψευδό-αναγνωριστικά (Sweeney L., 2002:557-570).

Βέβαια το μοντέλο της k-ανωνυμίας αποδείχθηκε ότι δεν είναι επαρκές σε περιπτώσεις που ο επιτιθέμενος έχει προηγούμενη γνώση σχετικά με το βαθμό του κόμβου του, όπως θα δούμε στη συνέχεια.



Σχήμα 4. Μια απλή τεχνική ανωνυμίας αντικαθιστώντας τα ευαίσθητα χαρακτηριστικά, όπως το όνομα, χρησιμοποιώντας ακέραιους. α) Το πρωτότυπο κοινωνικό δίκτυο. β) Ανώνυμο κοινωνικό δίκτυο, αντικαθιστώντας τα ονόματα με ακέραιο αριθμό αναγνώρισης

4.1.2 Η k-βαθμού Ανωνυμία

Ένας γράφος G(V,E) είναι k-βαθμού ανώνυμος εάν κάθε κόμβος στο V μοιράζεται τον ίδιο βαθμό με τουλάχιστον k-1 άλλους κόμβους (Liu K., Terzi E., 2008:71-80). Ο ορισμός αυτός προέκυψε ως αναγκαιότητα όχι μόνο της διατήρησης της ιδιωτικότητας της ταυτότητας του χρήστη αλλά και της διασφάλισης της σχέση του χρήστη με τους άλλους χρήστες. Δηλαδή αν ο επιτιθέμενος γνωρίζει το βαθμό του κόμβου του χρήστη, μπορεί μέσω αυτής της γνώσης να αναγνωρίσει και τον ίδιο χρήστη. Για παράδειγμα, ο κόμβος 1 στο Σχήμα 4(β) είναι επιρρεπής σε επίθεση (δηλαδή μπορεί να οδηγήσει σε μοναδική ταυτοποίηση του αντίστοιχου χρήστη) αφού ο βαθμός του είναι μοναδικός στο δίκτυο.

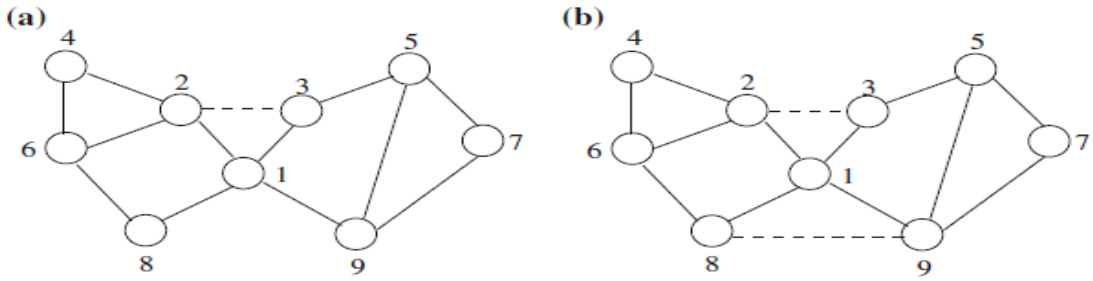
Στη συγκεκριμένη τεχνική εισάγοντας ή αφαιρώντας ακμές σε κάποιον χρήστη έτσι ώστε να τηρούνται οι προϋποθέσεις της k-βαθμού ανωνυμίας, ο επιτιθέμενος γνωρίζοντας τον βαθμού του κόμβου του χρήστη δε θα είναι σε θέση να τον αναγνωρίσει. Για παράδειγμα, το δίκτυο στο Σχήμα 5(α) είναι 2-βαθμού ανώνυμο. Απαιτείται η προσθήκη μια ακμής ανάμεσα στους δυο κόμβους 2 και 3 ώστε να επιτευχθεί η διατήρηση της ανωνυμίας. Η συγκεκριμένη τεχνική χρειάζεται δυο βήματα για να εφαρμοστεί (Zhou B., Pei J., 2008:506-515).

Στο πρώτο βήμα χρησιμοποιείται ένας δυναμικός αλγόριθμος προγραμματισμού ώστε να παραχθεί μια αλληλουχία από τους βαθμούς κόμβων η οποία είναι βασισμένη στους αρχικούς βαθμούς των κόμβων. Αυτή η αλληλουχία υπόκειται σε έναν περιορισμό όπου πρέπει να μειωθεί η απόσταση μεταξύ της ακολουθίας βαθμού που δημιουργήθηκε και της αρχικής. Στόχος είναι η ελαχιστοποίηση των αλλαγών των κόμβων και των ακμών στον αρχικό γράφο ώστε να διατηρήσουμε όσο γίνεται περισσότερο χρήσιμη πληροφορία.

Στο δεύτερο βήμα, ένας νέος γράφος διαμορφώνεται βασισμένος στη νέα ακολουθία των βαθμών. Εν συνέχεια γίνεται αξιολόγηση σχετικά με την απόδοση της τεχνικής, υπολογίζοντας το κόστος ανωνυμοποίησης, συμπεριλαμβάνοντας την απόσταση της ακολουθίας, του συντελεστή συστάδας (έναν μετρητή που δείχνει ποιοι κόμβοι σε έναν γράφο τείνουν να γίνουν συστάδα μαζί) και το μέσο μήκος διαδρομής. Καταλήγουμε στο συμπέρασμα ότι η τεχνική της k-βαθμού ανωνυμίας όχι μόνο προστατεύει τις ταυτότητες των χρηστών από επιθέσεις με την χρήση του βαθμού της κόμβου αλλά διατηρεί και τη χρησιμότητα των δεδομένων.

4.1.3 Η k-γειτνίασης Ανωνυμία

Για να εξηγήσουμε την έννοια της k-γειτνίασης ανωνυμοποίησης, θα περιγράψουμε αρχικά την έννοια του ισομορφισμού μεταξύ δύο γράφων. Συγκεκριμένα, δύο γράφοι G1 και G2 λέγονται ισομορφικοί αν υπάρχει μία-προς-μία αντιστοιχία των κόμβων τους έτσι ώστε ένα ζεύγος κόμβων του G1 να συνδέεται με μια ακμή αν και μόνο αν το αντίστοιχο ζεύγος κόμβων του G2 συνδέεται με μια ακμή (Zhou B., Pei J., 2008: 506-515).



Σχήμα 5. Ανώνυμα κοινωνικά δίκτυα που χρησιμοποιούν k-βαθμού ανωνυμία και η k-γειτνίασης Ανωνυμία. Οι διακεκομμένες ακμές προστίθενται στα δίκτυα για να επιτευχθεί η προστασία της ιδιωτικής ζωής. α) Το 2-βαθμού ανώνυμο δίκτυο. β) 2-γειτνίασης ανώνυμο δίκτυο

Ένας κόμβος v είναι k-γειτνίασης ανώνυμος εάν υπάρχουν τουλάχιστον k-1 άλλοι κόμβοι για κάθε έναν από οποίους υπάρχουν ισομορφικοί γράφοι με τους άμεσους γείτονές τους. Ένας γράφος είναι k-γειτνίασης ανώνυμος αν και μόνο αν όλοι οι κόμβοι είναι ανώνυμοι k-γειτνίασης.

Οι Zhou και Pei μελέτησαν το υπογράφο που δημιουργήθηκε από τους άμεσους γείτονες ενός στόχου/κόμβου (Zhou B., Pei J., 2011:47-77). Υποθέτουμε ότι η μοναδική δομή που έχει ο γειτονικός υπογράφος μπορεί να χρησιμοποιηθεί από τον επιτιθέμενο για να διακρίνει το αντικείμενο-στόχο από τους άλλους στο κοινωνικό δίκτυο. Όπως φαίνεται στο Σχήμα 6, γνωρίζοντας τη δομή της γειτονιάς της Ada, ένας επιτιθέμενος μπορεί να

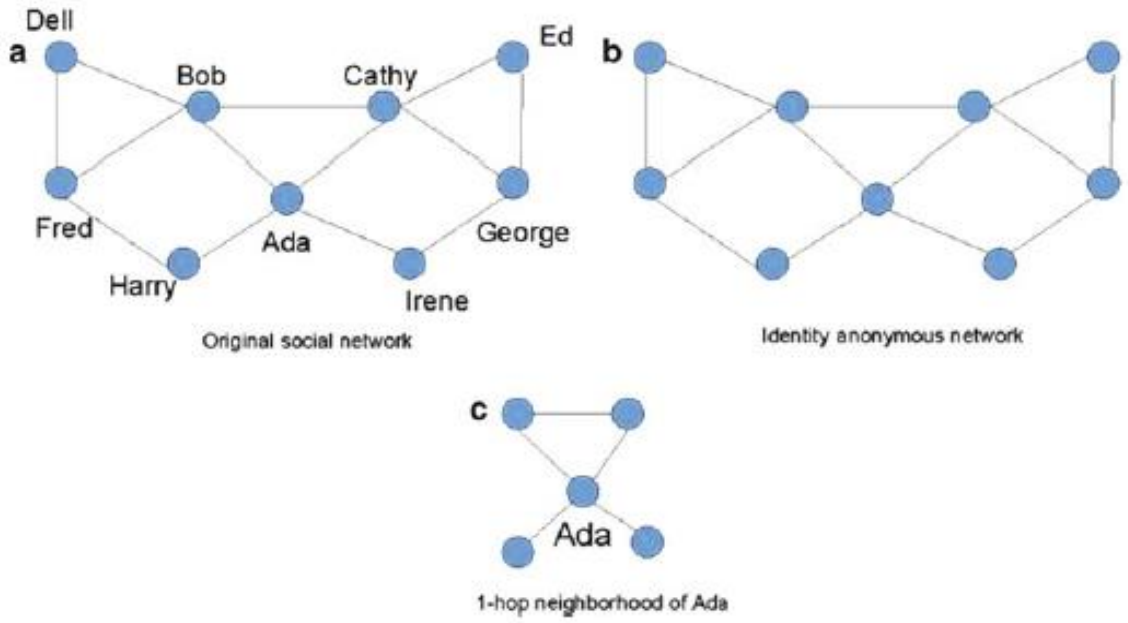
την ταυτοποιήσει από έναν 2-βαθμού ανώνυμο γράφο. Με βάση αυτή την παραδοχή, οι συγγραφείς πρότειναν ένα νέο συμβολισμό της ανωνυμίας στο κοινωνικό δίκτυο που ονομάζεται ως η k-γειτνίασης ανωνυμοποίηση. Η k-γειτνίασης ανωνυμοποίηση ορίζει ότι ένας γράφος είναι k-ανώνυμος, αν για κάθε κόμβος υπάρχει τουλάχιστον k-1 άλλοι κόμβοι που μοιράζονται ισομορφικές γειτονιές (Zhou B., Pei J., 2008:506-515).

Η k-γειτνίασης ανωνυμοποίηση επιτυγχάνεται χρησιμοποιώντας τα ακόλουθα τρία βήματα:

Πρώτο βήμα. Κατ' αρχάς, σημειώνει όλους τους κόμβους ως «μη ανωνυμοποιημένους» και τους ταξινομεί κατά φθίνουσα σειρά ανάλογα με το μέγεθος της γειτονιάς τους. Εδώ το μέγεθος της γειτονιάς ορίζεται ως ο αριθμός των ακμών και κόμβων του υπογράφου που κατασκευάζεται από τους άμεσους γείτονες ενός κόμβου. Η γειτονιά του κάθε κόμβου εξάγεται και κωδικοποιείται με βάση την τοπολογία της. Ο σκοπός αυτής της κωδικοποίησης που γίνεται στη γειτονιά είναι το να μπορεί να γίνει εύκολα σύγκριση μεταξύ των γειτονιών των κόμβων ώστε να βρεθούν οι ισομορφικές γειτονιές.

Δεύτερο βήμα. Δεύτερον, οι κόμβοι ομαδοποιούνται σε μικρές ομάδες με μεγέθη τουλάχιστον k. Οι κόμβοι σε κάθε ομάδα απαιτείται να έχουν παρόμοια κωδικοποίηση, όπως αυτή προέκυψε στο πρώτο βήμα.

Τρίτο βήμα. Οι τεχνικές επεξεργασίας του γράφου, όπως η εισαγωγή/διαγραφή μιας ακμής, εφαρμόζονται στο γράφο για να εξασφαλιστεί η ιδιότητα της k-γειτνίασης ανωνυμίας (Zhou B., Pei J., 2008:506-515).



Σχήμα 6. 2-βαθμού ανώνυμος γράφος με την αποτυχία της προστασίας της ιδιωτικής ζωής κατά την επίθεση γειτονίας με βάση τη γειτονία α) Αρχικός γράφος κοινωνικού δικτύου β)Ανωνυμοποίηση στις ταυτότητες των χρηστών γ) Η γειτονιάς της Ada είναι η μοναδική σε σχέση με όλους τους κόμβους του γράφου (Zou L., Chen L., Ozsu M.T. ,2009:946-957)

4.1.4 Η k-αυτομορφισμού Ανωνυμία

Αυτομορφισμός ενός γραφήματος G(V, E) είναι μια αυτομορφική συνάρτηση f του συνόλου κόμβων V, τέτοια ώστε για κάθε ακμή e = (u, v), η f (e) = (f (u), f (v)) να είναι επίσης μία ακμή στην G. Εάν υπάρχουν k αυτομορφισμοί στο G, αυτό σημαίνει ότι υπάρχουν k-1 διαφορετικές λειτουργίες αυτομορφισμού.

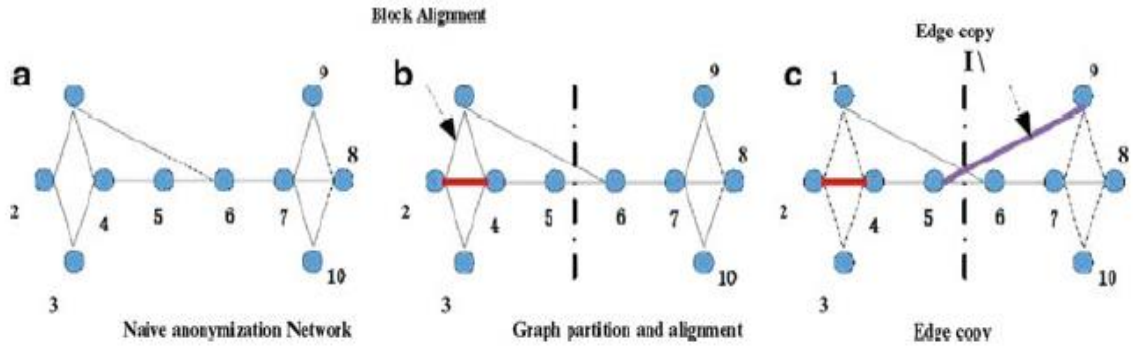
Οι Zou et.al πρότειναν τη μέθοδο του k- αυτομορφισμού βασισμένοι στην παραδοχή ότι ο επιτιθέμενος μπορεί να μελετήσει και να γνωρίζει τους γειτονικούς υπο-γράφους του χρήστη (Zou L., Chen L., Ozsu MT., 2009:946-957). Αν τέτοιοι υπο-γράφοι είναι μοναδικοί στον ανώνυμο γράφο του κοινωνικού δικτύου, τότε ο κόμβος-στόχος v μέσα στον υπο-γράφο εξακολουθεί να είναι ευάλωτος στην αποκάλυψη της ταυτότητάς του. Ο σκοπός αυτής της μεθόδου είναι να κατασκευάσει ένα νέο γράφο, έτσι ώστε κάθε υπο-γράφος γύρω από έναν κόμβο v να έχει τουλάχιστον k παρόμοιους υπο-γράφους που να είναι ισομορφικοί στο v.

Για την επίτευξη του k- αυτομορφισμού οι συγγραφείς πρότειναν τον αλγόριθμο k-αντιστοιχίας που εφαρμόζει ευριστική προσέγγιση στη διαδικασία της ανωνυμίας. Αυτή η διαδικασία ανωνυμίας προσθέτει κατάλληλες ακμές σε κατάλληλους κόμβους για να

πετύχει την ισομορφικότητα. Για να μεγιστοποιηθεί η χρησιμότητα, η προσθήκη ακμών θα πρέπει να είναι η ελάχιστη δυνατή για να πετύχουμε τον επιθυμητό στόχο. Ένα παράδειγμα απεικονίζεται στο Σχήμα 7.

Παράδειγμα

Πρώτα, οι δέκα κόμβοι ομαδοποιούνται σε δύο μπλοκ. Για να γίνουν αυτά τα δύο μπλοκ ισομορφικά, θα πρέπει να εισαχθεί μια ακμή μεταξύ του κόμβου 2 και του 4. Επιπροσθέτως, θα χρειαστεί μια ακμή για να συνδεθεί ο κόμβος 5 με τον κόμβο 9 ώστε να αντισταθμίσει τη διέλευση της ακμής ανάμεσα στους κόμβους 1 και 6. Ο τελικός δημοσιευμένος γράφος είναι ένα 2-αυτομορφικός γράφος.



Σχήμα 7. Κατάτμηση του γράφου και αντιγραφή ακμών a) Απλοϊκή ανωνυμοποίηση κοινωνικού δικτύου b) Κατάτμηση του γράφου και ευθυγράμμιση μπλοκ c) Αντιγραφή ακμής

β) Μέθοδοι Διατήρησης των συνδέσεων

Οι συνδέσεις αποκαλύπτονται όταν διαρρεύσουν ευαίσθητες πληροφορίες σχετικά με τη δομή τους. Η διαρροή των πληροφοριών αυτών είναι είτε αποτέλεσμα της δημοσίευσης των δεδομένων στα κοινωνικών δίκτυα, είτε απόρροια «παραβιασμένων» χρηστών μέσα από τα κοινωνικά δίκτυα. Εφ'οσον μπορούμε να εξάγουμε συμπεράσματα από ανώνυμα δεδομένα μέσω της δομής των συνδέσεων, τότε ο ιδιοκτήτης του κοινωνικού δικτύου θα επιθυμούσε να δημοσιεύσει το κοινωνικό δίκτυο με σκοπό να επιτευχθεί ανάλυση με τέτοιο τρόπο ώστε να μην είναι εφικτό να αποκαλυφθούν οι ευαίσθητες σχέσεις μεταξύ των χρηστών από τα δημοσιευμένα στοιχεία, για παράδειγμα, με τη χρήση τεχνικών εξόρυξης πληροφοριών από γράφους.

γ) Μέθοδοι διατήρησης των ευαίσθητων χαρακτηριστικών

Σε ένα κοινωνικό δίκτυο, ακόμη και αν ο γράφος είναι k-ανώνυμος ή αν χρησιμοποιούμε οποιοδήποτε από τις παραπάνω μεθόδους για την προστασία της ταυτότητας, υπάρχει ακόμα η πιθανότητα για διαρροή προσωπικών δεδομένων. Ο επιτιθέμενος χρησιμοποιώντας το γνωστικό του υπόβαθρο (προηγούμενη γνώση), μπορεί να προσδιορίσει την ευαίσθητη τιμή ενός ατόμου. Αυτό μπορεί να επιτευχθεί είτε εάν ανώνυμες συστάδες κόμβων διαμοιράζονται ίδιες ευαίσθητες πληροφορίες, είτε εάν η πιθανότητα μιας συγκεκριμένης ευαίσθητης τιμής σε μία συστάδα είναι μεγαλύτερη από τις υπόλοιπες ευαίσθητες τιμές, παρόλο που δεν μπορούμε να αναγνωρίσουμε ποιος κόμβος αντιστοιχεί σε ένα συγκεκριμένο άτομο.

4.1.5 Η λ-ποικιλομορφία Ανωνυμία

Οι Zhou και Pei επέκτειναν την μελέτη τους και εισήγαγαν την λ-ποικιλομορφία στην ανωνυμοποίηση του κοινωνικού δικτύου (Zhou B., Pei J., 2008:506-515). Σε αυτήν την περίπτωση, κάθε κόμβος συνδέεται με κάποια χαρακτηριστικά, συμπεριλαμβανομένων των αναγνωριστικών, ψευδο-αναγνωριστικών και ευαίσθητων χαρακτηριστικών. Εάν ένας επιτιθέμενος μπορεί εκ νέου να προσδιορίσει τις ευαίσθητες τιμές των χαρακτηριστικών ενός χρήστη με υψηλή εμπιστοσύνη, τότε η προστασία των προσωπικών δεδομένων του εν λόγω ατόμου παραβιάζεται. Ένας λ-ποικιλόμορφος γράφος εξασφαλίζει ότι ο επιτιθέμενος δεν μπορεί να συμπεράνει την ευαίσθητη τιμή του χαρακτηριστικού με πιθανότητα επιτυχίας μεγαλύτερη πάνω από 1/λ. Οι ερευνητές επεκτείνουν τη μέθοδο k-ανωνυμίας που αναπτύχθηκε κατά της 1-γειτνίασης επίθεσης για να χειριστεί το πρόβλημα λ-ποικιλομορφίας (Zhou B., Pei J., 2011:47-77).

Οι Yu et al πρότειναν ένα λ-ποικιλόμορφο ανώνυμο μοντέλο σε γράφους για τη διατήρηση της προστασίας της ιδιωτικότητας στα κοινωνικά δίκτυα. Αυτό το μοντέλο το οποίο θα μπορούσε να προστατεύσει τον κόμβο καθώς και τα ευαίσθητα χαρακτηριστικά του από την εκ νέου αναγνώριση, είναι βασισμένο στην k-βαθμού ανωνυμοποίηση. Οι ερευνητές πρότειναν έναν αλγόριθμο εύρεσης που θα μπορούσε να μετατρέψει τον αρχικό γράφο σε ένα λ-ποικιλόμορφο μέσω των τριών ανώνυμων στρατηγικών: την Προσαρμογή ομάδας, Ανακατεύθυνση ακμών και Ανάθεση εναπομείναντων αντίστοιχα (Yu L., Zhu J., Wu Z., Yang T., Hu J., Chen Z., 2012).

Παρόλο που η μέθοδος της λ-ποικιλομορφίας προτάθηκε για τα δεδομένα των κοινωνικών δικτύων, εν συνεχεία προέκυψε ότι περιέχει ορισμένες ελλείψεις. Οι Tai et.al.

αναφέρουν ότι αυτή η μέθοδος δεν περιορίζει τη συχνότητα των ευαίσθητων χαρακτηριστικών, έτσι δεν θα μπορούσε να προστατεύσει σε επίθεση συμπερασματικής πιθανότητας. Οι Zhou και Pei από την άλλη πλευρά ανέφεραν ότι δεν προστατεύει τα ευαίσθητα χαρακτηριστικά από άλλες επιθέσεις που βασίζονται σε προηγούμενη γνώση, εκτός από την επίθεση της γειτονιάς (Zhou B., Pei J., 2011:47-77). Οι Yu et.al. αναφέρουν ότι προστατεύει τα ευαίσθητα χαρακτηριστικά που βασίζονται μόνο στην έννοια της ανωνυμίας k-βαθμού (Yu L., Zhu J., Wu Z., Yang T., Hu J., Chen Z., 2012)..

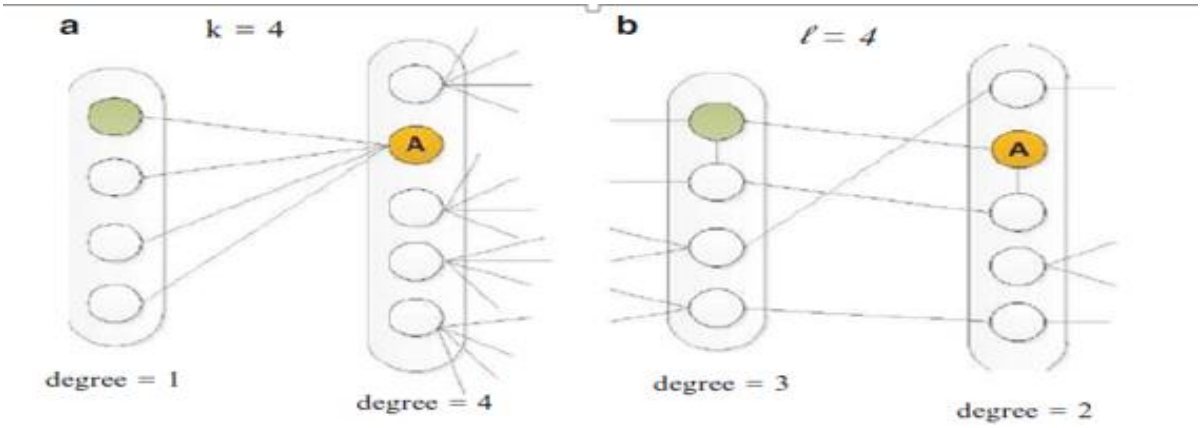
Δυστυχώς, καμία από τις μελέτες δεν μπορεί να λύσει όλα τα προβλήματα που αφορούν την αποκάλυψη των προσωπικών πληροφοριών. Για να προστατευτεί η ιδιωτικότητα από κάθε είδους προηγούμενες γνώσεις ή άλλες επιθέσεις απαιτούνται διαφορετικές μέθοδοι ή συνδυασμοί αυτών. Το παρακάτω παράδειγμα καταδεικνύει την αποκάλυψη του ευαίσθητου χαρακτηριστικού που βασίζεται σε κατάλληλη προηγούμενη γνώση.

Παράδειγμα

Στη συγκεκριμένη περίπτωση, ο επιτιθέμενος επιχειρεί να συμπεράνει μια ευαίσθητη σχέση μεταξύ δύο αντικείμενων-στόχων. Ωστόσο, εξακολουθούμε να επικεντρωνόμαστε σε επιθέσεις που βασίζονται στην τοπολογία. Ως εκ τούτου, θεωρούμε ότι η επιπλέον γνώση που έχει ο επιτιθέμενος είναι ο αριθμός των φίλων του κάθε χρήστη, ο οποίος περιγράφεται στους γράφους των κοινωνικών δικτύων από το βαθμό του κόμβου. Αυτές οι κοινές πληροφορίες είναι διαθέσιμες στο κοινό σε πολλές περιπτώσεις (π.χ. LinkedIn). Επιπλέον, στη χειρότερη περίπτωση, ένας από τους στόχους μπορεί ακόμη να προσδιοριστεί από τον εισβολέα χρησιμοποιώντας την προηγούμενη γνώση που έχει συλλεχθεί από το blog του χρήστη. Σύμφωνα με αυτό το μοντέλο επίθεσης, μια τεχνική ανωνυμίας k-βαθμού είναι ανεπαρκής για τη διασφάλιση της προστασίας της ιδιωτικής ζωής στο βαθμό που επιθυμούμε.

Ένα απλό παράδειγμα απεικονίζεται στο Σχήμα 8(α), όπου οι κόμβοι ομαδοποιούνται με βάση τους βαθμούς τους. Ας υποθέσουμε ότι ο επιτιθέμενος προσπαθεί να συμπεράνει τη σχέση μεταξύ του χρήστη A και του χρήστη B. Σε αυτόν τον 4-βαθμού ανώνυμο γράφο, αν ο επιτιθέμενος γνωρίζει μόνο ότι οι δύο χρήστες έχουν βαθμούς των τεσσάρων και ενός, αντίστοιχα, δεν μπορεί να συμπεράνει τη σχέση τους. Ωστόσο, αν ο επιτιθέμενος μπορεί με κάποιο τρόπο να αναγνωρίσει το χρήστη A, ακόμα κι αν δεν μπορεί να αναγνωρίσει τον χρήστη B (βασισμένος αποκλειστικά και μόνο στη γνώση του βαθμού

του), η ευαίσθητη σχέση μεταξύ του χρήστη Α και χρήστη Β μπορεί να αποκαλυφθεί λόγω του ότι γνωρίζει τις συνδέσεις από όλους τις 1-βαθμού κόμβους του χρήστη Α. Για να προστατευτούμε από αυτή την επίθεση, προτάθηκε η τεχνική ανωνυμοποίησης που είδαμε νωρίτερα ως ***λ-ποικιλομορφία.***



Σχήμα 8. Γράφοι ανώνυμου κοινωνικού δικτύου ενάντια στην αποκάλυψη της σχέσης των χρηστών. α) 4ης τάξης βαθμού ανωνυμία β) 4ης τάξης ποικιλομορφία

4.2 Αλγόριθμοι ανωνυμοποίησης

Όλοι οι γνωστοί αλγόριθμοι ανωνυμοποίησης σε γράφους κοινωνικών δικτύων που υλοποιούν τις ως άνω τεχνικές είναι άπληστοι (greedy), υπό την έννοια ότι δεν παρέχουν απαραίτητα τη βέλτιστη δυνατή συνολικά λύση αλλά επενεργούν σε βήματα, με σκοπό τη βέλτιστη κάθε φορά επιλογή σε κάθε βήμα. Ένας από τους πιο γνωστούς αλγόριθμους είναι ο λεγόμενος SaNGreea.

4.2.1 Ο αλγόριθμος SaNGreeA

Για την άπληστη ανωνυμοποίηση στα κοινωνικά δίκτυα, ο αλγόριθμος SaNGreeA εκτελεί μια επεξεργασία συσταδοποίησης ώστε να δημιουργήσει ένα k-ανώνυμο "μεταμφιεσμένο" κοινωνικό δίκτυο, ξεκινώντας από ένα αρχικό κοινωνικό δίκτυο που είναι δομημένο ως ένας γράφος G = (V, E), όπου V είναι οι κόμβοι που περιγράφονται από τα ψευδο-αναγνωριστικά και τα ευαίσθητα χαρακτηριστικά και E είναι οι μη κατευθυνόμενες (δηλαδή χωρίς ετικέτα) ακμές (Byun, J. W., Kamra A., Bertino E., Li. N., 2007:188-200).

Αρχικά λοιπόν, ο αλγόριθμος δημιουργεί μια «καλή» κατάτμηση όλων των κόμβων σε συστάδες. Στη συνέχεια, όλοι οι κόμβοι μέσα σε κάθε συστάδα γίνονται ομοιόμορφοι σε

σχέση με τα χαρακτηριστικά των ψευδο-αναγνωριστικών και τις σχέσεις τους. Αυτή η ομογενοποίηση επιτυγχάνεται με τη χρήση γενίκευσης (βλ. Κεφάλαιο 2), τόσο για τα χαρακτηριστικά των ψευδο-αναγνωριστικών όσο και για τη σχέση μεταξύ των κόμβων (ακμές) (Ghinita G., Karras P., Kalinis P., Mamoulis N., 2007:758-769).

Είναι σημαντικό να αναφερθεί ότι για να πληρούνται οι απαιτήσεις του μοντέλου k-ανωνυμίας θα πρέπει κάθε ομάδα να περιέχει τουλάχιστον k κόμβους. Κατά συνέπεια, ένα πρώτο κριτήριο για τη διαδικασία ομαδοποίησης είναι να διασφαλιστεί ότι κάθε συστάδα έχει αρκετά στοιχεία. Ως εκ τούτου, με τη γενίκευση των χαρακτηριστικών και τις σχέσεις τους, ως αναπόφευκτο αποτέλεσμα κάποιες πληροφορίες μπορεί να χαθούν.

Ένα δεύτερο κριτήριο που χρησιμοποιείται κατά τη διάρκεια της ομαδοποίησης είναι η ελαχιστοποίηση των πληροφοριών που χάνονται μεταξύ του αρχικού γράφου στο κοινωνικό δίκτυο και στη «μεταμφιεσμένη» εκδοχή του. Η απώλεια πληροφορίας προκαλείται από τη γενίκευση των ψευδο-αναγνωριστικών αλλά και της γενίκευσης των ακμών.

Ο αλγόριθμος ομαδοποίησης χρησιμοποιεί δύο μέτρα απώλειας πληροφοριών που έχουν ως στόχο την απόκτηση καλής ποιότητας "μεταμφιεσμένων" δεδομένων, καθώς και να επιτρέπουν στο χρήστη να ελέγχει τον τύπο και την ποσότητα της απώλειας πληροφοριών που μπορεί να αντέξει να χάσει.

Το πρώτο μέτρο εκφράζει ποσοτικά πόσο περιγραφική λεπτομέρεια δεδομένων χάνονται μέσω της γενίκευσης στα χαρακτηριστικά των ψευδο-αναγνωριστικών και αναφέρεται ως γενίκευση απώλειας πληροφοριών.

Το δεύτερο μέτρο εκφράζει ποσοτικά το πόσο κατασκευαστική λεπτομέρεια χάνεται μέσα από τη γενίκευση των σχέσεων (ακμών) και αναφέρεται ως απώλεια δομικών πληροφοριών. Η δομική πληροφορία χάνεται όταν γίνεται ανωνυμοποίηση ενός γράφου με την διάσπαση των συστάδων σε κόμβους, μαζί με τις γειτονιές τους. Στην περίπτωση αυτή, η απώλεια πληροφοριών εκφράζει ποσοτικά την πιθανότητα λάθους στην προσπάθεια ανακατασκευής της δομής του αρχικού κοινωνικού δικτύου από τη «μεταμφιεσμένη» εκδοχή του.

Όπως αναφέρθηκε και παραπάνω, ο SaNGreeA αλόριθμος ανωνυμοποίησης ομαδοποιεί εκείνους τους κόμβους που είναι όσο το δυνατόν παρόμοιοι, τόσο από την άποψη των τιμών των ψευδο-αναγνωριστικών τους, όσο και από την άποψη της δομής της γειτονιάς τους. Αυτή η άπληστη (greedy) προσέγγιση προσπαθεί ακριβώς να ελαχιστοποιήσει την απώλεια πληροφοριών αλλά και δομικών πληροφοριών μέσω γενίκευσης, δημιουργώντας ένα k-ανώνυμο "μεταμφιεσμένο" κοινωνικό δίκτυο.

Για την εκτίμηση της εγγύτητας μεταξύ των κόμβων σε σχέση με τα χαρακτηριστικά των ψευδο-αναγνωριστικών, ο αλγόριθμος χρησιμοποιεί ειδικούς τύπους για την κανονικοποιημένη απώλεια πληροφορίας.

4.2.2 Ο αλγόριθμος Basic Labeling Algorithm

Ο αλγόριθμος αυτός ασχολείται με την ανωνυμοποίηση σχετικά με τις ετικέτες των γράφων**.** Για να λυθούν οι αδυναμίες της χρήσης ενός και μόνο επιπέδου ασφάλειας, καθορίζουμε διαφορετικά επίπεδα ασφάλειας για τους χρήστες και τα ενσωματώνουμε στο ίδιο δίκτυο. Στην πραγματικότητα, τα δίκτυα κοινωνικής δικτύωσης επιτρέπουν σε καθένα χρήστη ξεχωριστά να θέσει πόση πληροφορία θα μπορεί να είναι δημόσια σε άλλους χρήστες.

Για παράδειγμα στο Facebook, ένας χρήστης μπορεί να θέσει ποιες πληροφορίες από το προσωπικό του προφίλ ή για τις συνδέσεις του μπορεί να τα δουν άλλοι χρήστες. Ο χρήστης μπορεί να έχει μια σαφή εκτίμηση της γνώσης που ένας επιτιθέμενος μπορεί να έχει για αυτόν. Η γνώση που κάποιος θα χρησιμοποιήσει για να βρει τις προσωπικές πληροφορίες για κάποιον χρήστη λέγεται, όπως ήδη έχουμε πει, προηγούμενη γνώση. Η παροχή διαφορετικών επιπέδων ασφάλειας επιτρέπει στους χρήστες να θέσουν εξατομικευμένες απαιτήσεις για την προστασία της ιδιωτικής ζωής με βάση τις δικές τους υποθέσεις σχετικά με το προηγούμενη γνώση του εισβολέα.

Ειδικότερα, για τον κόμβο u σε ένα δημοσιευμένο επισημασμένο γράφο, ξεκινώντας από το ελάχιστη προηγούμενη γνώση ότι ο επιτιθέμενος γνωρίζει μόνο πληροφορίες για την ετικέτα u χωρίς να έχει οποιαδήποτε άλλη πληροφορία σχετικά με τη δομή, μπορούμε να ορίσουμε 3 επίπεδα επιθέσεων στη u με βάση την σταδιακή αύξηση του γνωστικού υποβάθρου που έχει ο επιτιθέμενος:

ΕΠΙΠΕΔΟ 1: Ο επιτιθέμενος γνωρίζει μόνο τις ετικέτες (ψευδο-αναγνωριστικά) του κόμβου u. Για παράδειγμα γνωρίζει ότι ο Bob είναι 21 ετών.

ΕΠΙΠΕΔΟ 2: Ο επιτιθέμενος γνωρίζει τις ετικέτες του κόμβου u και το βαθμό του. Για παράδειγμα γνωρίζει ότι ο Bob είναι 21 ετών και με βαθμό κόμβου 3.

ΕΠΙΠΕΔΟ 3: Ο επιτιθέμενος γνωρίζει τις ετικέτες του κόμβου u, το βαθμό του κόμβου του και πληροφορίες σχετικά με τις ετικέτες των ακμών που είναι πλησίον του. Για παράδειγμα γνωρίζει ότι ο Bob είναι 21 ετών, ότι έχει βαθμό κόμβου 3 και ότι έχει τρεις κατηγορίες συνδέσεων συμμαθητής, συγκάτοικος, συγκάτοικος.

Οι B. Zhou, J. Pei, and W. Luk, μελέτησαν αυτά τα τρία επίπεδα προηγούμενης γνώσης που βασίζονται στους τρεις διαφορετικούς τύπους ασφάλειας του Facebook (Zhou B., Pei J., Luk W., 2008:12-22). Παρατηρήθηκε φυσικά ότι υπάρχουν πιο ισχυροί τύποι επιθέσεων όπως πληροφορίες σχετικά με τη γειτονιά του κόμβου u αλλά και για τις ετικέτες. Όμως δεν επικεντρώθηκαν σε πόσες και ποιες επιθέσεις είναι δυνατό να γίνουν. Σκοπός ήταν να δείξουν τι θα πρέπει μια σελίδα κοινωνικής δικτύωσης να ενεργοποιήσει για να μπορεί να καλύψει τις διαφορετικές απαιτήσεις. Χρησιμοποίησαν "προηγούμενη γνώση επιπέδου Χ" για να παρουσιάσουν την αντίστοιχη προηγούμενη γνώση που χρησιμοποιήθηκε στην επίθεση Επιπέδου Χ.

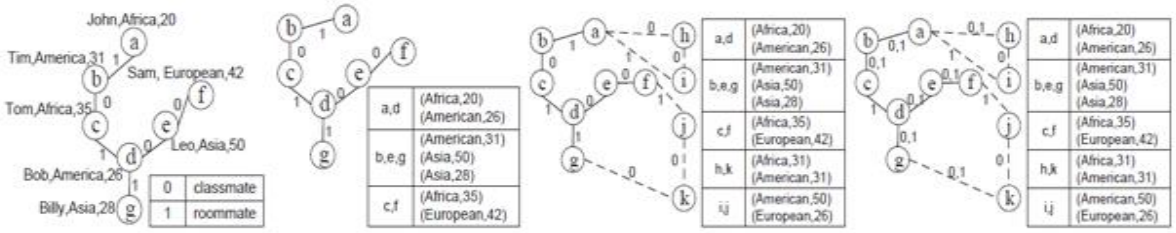
Για την επίτευξη της προστασίας κάθε χρήστη στο επίπεδο που είναι ίση ή μεγαλύτερη από τη δική του ρύθμιση, έθεσαν ένα μοντέλο προστασίας για όλους τους χρήστες σε ένα κοινωνικό δίκτυο με βάση το ισχυρότερο επίπεδο προστασίας που έχει απαιτηθεί από τους χρήστες. Στην απλοϊκή αυτή μέθοδο, το χειρότερο σενάριο είναι να λάβει χώρα αυτή η αλλαγή σε κάθε κόμβο, γεγονός το οποίο μειώνει τη χρησιμότητα του δημοσιευμένου γράφου. Εδώ, η χρησιμότητα αναφέρεται ουσιαστικά στην εγγύτητα μεταξύ του δημοσιευμένου γράφου και του αρχικού (δηλαδή δεν θα πρέπει να διαφέρουν πολύ).

Για παράδειγμα στο Σχήμα 9(α) είναι ένας γράφος με 7 κόμβους που ο καθένας έχει 3 ετικέτες (ψευδο-αναγνωριστικά): όνομα, τοποθεσία και ηλικία. Υποθέτουμε ότι έχουμε δύο στόχους προστασίας για το δημοσιευμένο γράφο οι οποίοι είναι: η πιθανότητα ένας επιτιθέμενος να βρει ότι ένας χρήστης P είναι ο κόμβος u στο δημοσιευμένο γράφο (γνωστή και ως ταυτοποίηση εκ νέου του κόμβου) πρέπει να λιγότερη από 50% και η

πιθανότητα κάποιος επιτιθέμενος να βρει ότι ο χρήστης P1 και ο χρήστης P2 έχουν κάποια σύνδεση (γνωστή και ως ταυτοποίηση εκ νέου της σύνδεσης) πρέπει να είναι μικρότερη από 50%.

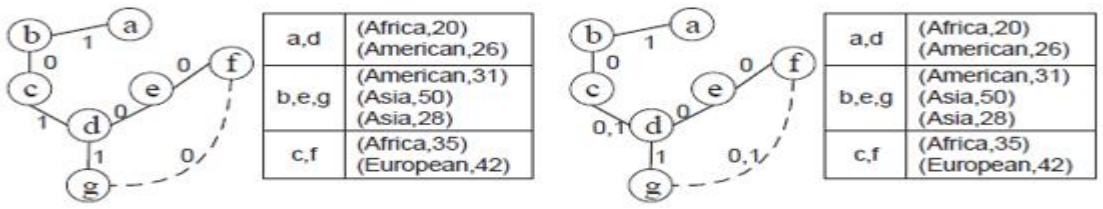
Για το Σχήμα 9(α), θα μπορούσαμε να δημοσιεύσουμε το Σχήμα 9(β) το οποίο παρέχει προστασία Επίπεδου 1 για όλες τους κόμβους σύμφωνα με την απλοϊκή μέθοδο που αναφέραμε παραπάνω. Οι κόμβοι χωρίζονται σε τρεις ομάδες ((a, d), (b, e, g) και (c, f)), όπως φαίνεται στον πίνακα του Σχήματος 9(β). Δεδομένου ότι το μέγεθος της κάθε ομάδας (ο αριθμός των κόμβων σε μια ομάδα) είναι ίσο ή μεγαλύτερο από 2, η πιθανότητα επιτυχίας της εκ νέου ταυτοποίησης μιας κόμβου χρησιμοποιώντας τη γνώση του επιπέδου 1 είναι μικρότερη από 50%.

Δεν υπάρχει καμία ακμή μεταξύ δύο κόμβων στην ίδια ομάδα και ο αριθμός των ακμών μεταξύ των κόμβων σε δύο ομάδες είναι το πολύ 3, η οποία εγγυάται ότι η πιθανότητα να βρεθούν δύο κόμβοι που να έχουν μία σύνδεση είναι μικρότερη από 50%.



Σχήμα 9. Απλοϊκή προστασία α) Αρχικός γράφος β)Προστασία Επιπέδου 1 γ) Προστασία Επιπέδου 2 δ) Προστασία Επιπέδου 3

Οι γραφικές παραστάσεις που παρέχουν Επίπεδο 2 και Επίπεδο 3 προστασίας φαίνεται στο Σχήμα 9(γ) και Σχήμα 9(δ) αντίστοιχα. Στο Σχήμα 9(γ) καμία σύνδεση δε θα μπορούσε να αναγνωριστεί εκ νέου με εμπιστοσύνη μεγαλύτερη από 50%, ακόμη και κάτω από μια επίθεση που χρησιμοποιεί πληροφορίες σχετικά με τις ετικέτες και το βαθμό των κόμβων .



Σχήμα 10. Εξατομικευμένη προστασία α) Προστασία Επιπέδου 2 β) Προστασία Επιπέδου 3

Στο Σχήμα 9(γ) και 9(δ) υποθέτουμε ότι όλοι οι χρήστες θέλουν να έχουν είτε επίπεδο ασφάλειας 2 είτε 3 αντίστοιχα. Ωστόσο, όπως προαναφέρθηκε, μέσα στο ίδιο κοινωνικό δίκτυο διαφορετικοί χρήστες θα θέλουν διαφορετικό επίπεδο ασφάλειας για τις προσωπικές τους πληροφορίες. Για παράδειγμα αν μόνο ο «Leo» και ο «Tom» θέλουν επίπεδο προστασίας 2, το Σχήμα 10(α), μπορεί να ικανοποιήσει όλους τους χρήστες. Στο Σχήμα 10(α) οι κόμβοι c και d έχουν κοινές ετικέτες και βαθμό κόμβων όπως ο «Tom» στο Σχήμα 9(α). Αυτό εξασφαλίζει ότι η πιθανότητα αναγνώρισης του χρήστη «Tom» είναι 50%. Παρομοίως, ο αριθμός των πιθανών κόμβων για τον «Leo» είναι 50%.

Πρέπει να προσθέσουμε 4 κόμβους και 6 ακμές στο Σχήμα 9(α) για να παραχθεί το Σχήμα 9(γ), όμως χρειαζόμαστε μόνο μια σύνδεση για να παραχθεί το Σχήμα 10(α). Αν συγκρίνουμε τα δυο παραγόμενα σχήματα, θα παρατηρήσουμε ότι το Σχήμα 9(α) είναι πιο κοντά στο αρχικό γράφο. Εάν ο «Leo» θέλει επίπεδο προστασίας 2 και ο «Tom» θέλει επίπεδο προστασίας 3 τότε το Σχήμα 10(β) μπορεί να δημοσιευθεί, το οποίο γενικεύει μόνο τις ετικέτες 2 ακμών. Όμως το Σχήμα 9(δ) για να παρέχει την προστασία του επιπέδου 3 πρέπει να γενικεύσει τις ετικέτες 7 ακμών. Έτσι το Σχήμα 10(β) είναι ακόμα πιο κοντά στο Σχήμα 9(α) από το Σχήμα 9(δ). Αυτά τα δυο παραδείγματα δείχνουν ότι η χρησιμότητα των δημοσιευμένων γράφων αυξάνεται όταν οι χρήστες έχουν εξατομικευμένες ρυθμίσεις ασφάλειας.

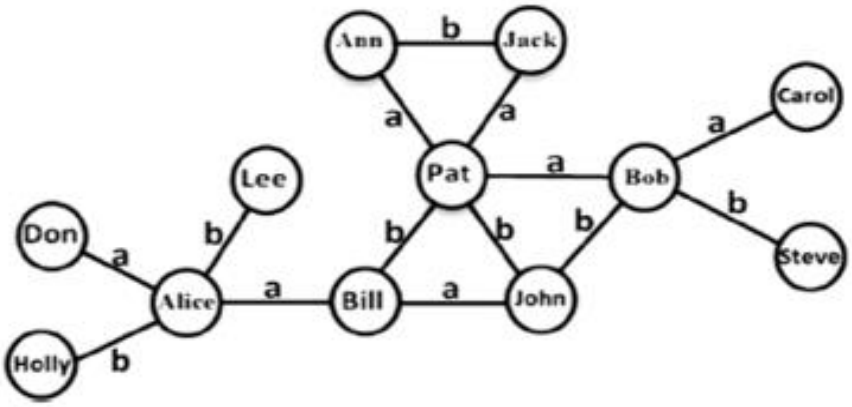
Για την επίτευξη εξατομικευμένης προστασίας της ιδιωτικής ζωής σε κοινωνικά δίκτυα, σχεδιάστηκαν διαφορετικές μέθοδοι αναλόγως των απαιτήσεων. Συγκεκριμένα, για το πρώτο επίπεδο προστασίας χρησιμοποιούμε γενίκευση στην ετικέτα (ψευδο-αναγνωριστικό) του κόμβου, στο επίπεδο 2 συνδυάζουμε μεθόδους όπως προσθήκη θορύβου στον κόμβο και προσθήκη ακμής βασισμένοι στην προστασία επίπεδου 1. Στο τρίτο επίπεδο προστασίας χρησιμοποιούμε γενίκευση στην ετικέτα της ακμής για την επίτευξη των στόχων προστασίας.

Κεφάλαιο 5
Νέα τεχνική ανωνυμοποίησης γράφων κοινωνικής δικτύωσης

Στο κεφάλαιο αυτό θα παραθέσουμε έναν νέο τρόπο με τον οποίο πιθανός επιτιθέμενος μπορεί να παραβιάσει την ανωνυμοποίηση ενός γράφου κοινωνικής δικτύωσης. Το μοντέλο στο οποίο επικεντρωνόμαστε είναι η χρήση γράφου με ετικέτες στις ακμές του, το οποίο σημαίνει ότι η k-βαθμού ανωνυμοποίηση δεν είναι αρκετή (αφού μπορεί δύο κόμβοι να έχουν τον ίδιο βαθμό, αλλά να διαφέρουν οι ετικέτες τους, οπότε και μπορούν να ταυτοποιηθούν από κάποιον επιτιθέμενο που έχει προηγούμενη γνώση επί των ετικετών). Επιθυμούμε να μη δημιουργήσουμε ομάδες κόμβων σε συστάδες ούτε να κάνουμε γενίκευση στις ετικέτες των ακμών, το οποίο όπως είδαμε αποτελεί τις κύριες μέχρι τώρα προσεγγίσεις στο εν λόγω ζήτημα. Πριν προχωρήσουμε στην περιγραφή του βασικού σκεπτικού, παρουσιάζουμε το μοντέλο επίθεσης στο οποίο εστιάζουμε.

5.1 Παρουσίαση τρόπου επίθεσης στα μέσα κοινωνικής δικτύωσης

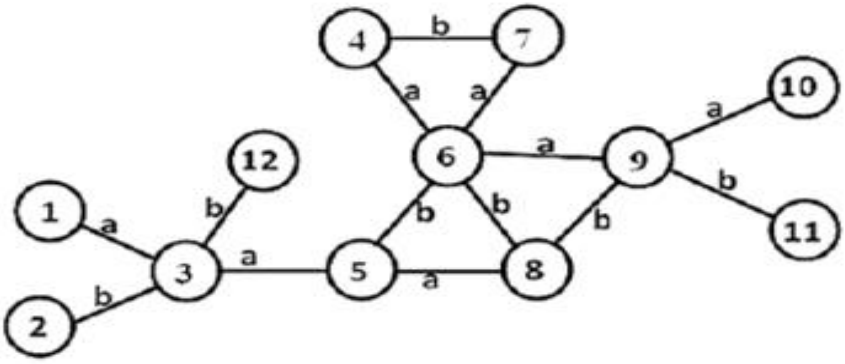
Έστω ότι έχουμε τον εξής γράφο (Sargolzaei E., Khazali M. J., Keikha F. , 2016:1-8):



Σχήμα 11. Μη ανωνυμοποιημένος γράφος κοινωνικού δικτύου με ετικέτες στις ακμές

Όπου οι ετικέτες στις ακμές αντικατοπτρίζουν τη σχέση που συνδέει τους δύο κόμβους («φίλους») – π.χ. «συμμαθητής», «συγγενής», «συνάδελφος» κτλ.

Ο γράφος αυτός, αν απομακρυνθούν τα αναγνωριστικά ονόματα των κόμβων, μετατρέπεται ως εξής:



Σχήμα 12. Αντικατάσταση του ονομάτων των χρηστών του κοινωνικού δικτύου με τυχαίους αριθμούς

Θα περιγράψουμε μία επίθεση βασισμένη σε προηγούμενη γνώση όχι μόνο των φίλων κάποιου χρήστη αλλά των σχέσεων που τον συνδέουν με αυτούς τους φίλους.

Ας υποθέσουμε ότι γνωρίζουμε ότι ο Bill έχει 3 φίλους, με τους οποίους συνδέεται με σχέσεις a, a και b αντίστοιχα. Τότε, από τον ανωτέρω «ανωνυμοποιημένο» γράφο, συμπεραίνουμε ότι ο Bill είναι ο κόμβος 5. Θα επιθυμούσαμε λοιπόν, ακόμα και ο

επιτιθέμενος έχει τέτοια προηγούμενη γνώση, να μην οδηγείται μονοσήμαντα στην αναγνώριση ενός χρήστη αλλά να υπάρχουν k κόμβοι, όπου k>1, με το ίδιο πλήθος «φίλων» και τις ίδιες ετικέτες.

Στόχος μας είναι η προσθαφαίρεση ακμών ή και κόμβων, έτσι ώστε αφενός να είναι η μικρότερη δυνατή – για να περιοριστεί η απώλεια πληροφορίας – και, αφετέρου, να εξασφαλίζει μία ανωνυμία βαθμού k όπως περιγράφεται ανωτέρω (ήτοι k αξεχώριστοι κόμβοι βάσει του εν λόγω κριτηρίου).

5.2 Περιγραφή της νέας τεχνικής

Με τη νέα τεχνική, θέλουμε να έχουμε k αξεχώριστους κόμβους όσον αφορά τις ετικέτες των ακμών που πρόσκεινται σε αυτόν. Για να περιγραφεί η προτεινόμενη αλγοριθμική διαδικασία, ορίζουμε κατ' αρχάς μία διάταξη μεταξύ των δυνατών τιμών που μπορούν να λάβουν οι ετικέτες στις ακμές: για παράδειγμα, αν οι δυνατές τιμές που μπορούν να λάβουν οι ακμές είναι πέντε – έστω $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ – ορίζουμε μία διάταξη
$\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 < \lambda_5$. Στη συνέχεια ορίζουμε ως αλληλουχία ετικετών (sequence label) ενός κόμβου V_i τη διατεταγμένη, βάσει της ορισθείσας διάταξης, ακολουθία των ετικετών των ακμών που πρόσκεινται στον κόμβο V_i. Ως εκ τούτου, συνεχίζοντας το προηγούμενο παράδειγμα, αν οι ετικέτες των ακμών που πρόσκεινται σε έναν κόμβο V_i βαθμού 4 ίναι $\lambda_1, \lambda_1, \lambda_3, \lambda_5$, τότε η αλληλουχία ετικετών του κόμβου είναι $\lambda_1 \lambda_1 \lambda_3 \lambda_5$. Είναι προφανές ότι, λόγω της διάταξης που ορίσαμε στο σύνολο των ετικετών, η αλληλουχία ετικετών ενός οποιουδήποτε κόμβου είναι μοναδική.

Η πρώτη σημαντική παρατήρηση, που αποτελεί βάση για την προτεινόμενη τεχνική, είναι η εξής (η απόδειξη της οποίας είναι προφανής):

Πρόταση 1:

Αν υπάρχουν δύο κόμβοι Α και Β, μη γειτονικοί, με αλληλουχία ετικετών $\lambda_1\lambda_2\ldots\lambda_m$ και ένας κόμβος C με αλληλουχία ετικετών $\lambda_1\lambda_2\ldots\lambda_{m+1}$ τότε ενώνοντας τους κόμβους Α και Β με μία ακμή ετικέτας λ_{m+1} καθιστούμε την τριπλέτα κόμβων Α, Β και C «αξεχώριστη», αφού και οι τρεις κόμβοι Α, Β και C θα έχουν την ίδια αλληλουχία ετικετών $\lambda_1\lambda_2\ldots\lambda_{m+1}$.

Με αντίστροφο σκεπτικό, μπορούμε επίσης να διατυπώσουμε το εξής:

Πρόταση 2:

Αν υπάρχουν δύο κόμβοι Α και Β, μη γειτονικοί, με αλληλουχία ετικετών $\lambda_1\lambda_2\ldots\lambda_m$ και ένας μη γειτονικός τους κόμβος C με αλληλουχία ετικετών $\lambda_1\lambda_2\ldots\lambda_{m+1}$ τότε αφαιρώντας από τον κόμβο C την ακμή ετικέτας λ_{m+1} καθιστούμε την τριπλέτα κόμβων Α, Β και C «αξεχώριστη», αφού και οι τρεις κόμβοι Α, Β και C θα έχουν την ίδια αλληλουχία ετικετών $\lambda_1\lambda_2\ldots\lambda_m$.

Απώτερος στόχος μας λοιπόν είναι η κατά το δυνατόν βέλτιστη προσθήκη ή αφαίρεση ακμών βάσει των ανωτέρω προτάσεων.

Περαιτέρω, επιθυμούμε να επιτύχουμε έναν, κατά το δυνατόν, ισοσκελισμό μεταξύ του πλήθους των ακμών που προσθέτουμε και του πλήθους των ακμών που αφαιρούμε. Ο λόγος που το επιθυμούμε αυτό είναι το να μην αλλοιωθεί σημαντικά το πλήθος των ακμών που συνολικά έχει ο γράφος μας, το οποίο θα συνιστούσε μία σημαντική απώλεια χρήσιμης πληροφορίας – αφού θα ισοδυναμούσε με εισαγωγή «θορύβου» ως προς το σύνολο των συνδέσεων μεταξύ των χρηστών του εν λόγω κοινωνικού δικτύου. Επίσης, δεν επιθυμούμε να τροποποιήσουμε σημαντικά τη μέση τιμή των βαθμών του γράφου – το οποίο συνιστά μία σημαντική πληροφορία ως προς το μέσο πλήθος φίλων που έχουν οι χρηστες του κοινωνικού δικτύου. Η προσέγγιση που ακολουθήθηκε στην παρούσα διατριβή ως προς τα εν λόγω ζητήματα είναι η κατηγοριοποίηση των κόμβων του γράφου σε 2 κατηγορίες βάσει της εξής διαδικασίας:

Κατηγοριοποίηση κόμβων

1) Υπολογίζουμε το μέσο όρο D των βαθμών των κόμβων του γραφήματος
2) Διατρέχουμε όλους τους κόμβους V_i του γραφήματος:

    2a) Αν ο βαθμός του V_i είναι μικρότερος από D, τότε ο V_i εντάσσεται στην Κατηγορία 1 – διαφορετικά, εντάσσεται στην Κατηγορία 2.

Με αυτόν τον τρόπο, κατηγοριοποιούμε τους κόμβους του γράφου σε δύο κατηγορίες: στην Κατηγορία 1 θα είναι εκείνοι οι κόμβοι στους οποίους είναι πιθανό να προσθέσουμε ακμές, ενώ στην Κατηγορία 2 θα είναι εκείνοι οι κόμβοι στους οποίους είναι πιθανό να αφαιρέσουμε ακμές. Με αυτόν τον τρόπο, αποσκοπούμε στο να επιτύχουμε ο τελικός ανωνυμοποιημένος γράφος να έχει ένα πλήθος ακμών που να μη διαφέρει πολύ από τον

αντίστοιχο στον αρχικό γράφο, ενώ ταυτόχρονα να ισχύει το ίδιο και για το μέσο όρο των βαθμών για τους δύο αυτούς γράφους (με τη λογική ότι αποσκοπούμε όσες ακμές προσθέσουμε σε κόμβους της Κατηγορίας 1 να προσεγγίζουν το πλήθος των ακμών που θα αφαιρεθούν από την Κατηγορία 2).

Έχοντας πραγματοποιήσει την ως άνω κατηγοριοποίηση, προχωρούμε ως εξής:
1) Αναζητούμε ζεύγη κόμβων Α και Β της Κατηγορίας 1 και έναν κόμβο C Κατηγορίας 1 ή 2 για τους οποίους θα μπορούσε να εφαρμοστεί η Πρόταση 1. Κάθε τέτοια τριπλέτα κόμβων για την οποία εφαρμόζεται η Πρόταση 1 θα λέγεται «ταξινομημένη».
2) Αναζητούμε ζεύγη κόμβων Α και Β της Κατηγορίας 1 ή 2 και έναν κόμβο C Κατηγορίας 2 για τους οποίους θα μπορούσε να εφαρμοστεί η Πρόταση 2, χωρίς όμως η επιχειρούμενη αφαίρεση ακμής να επηρεάζει κάποιον άλλον ήδη «ταξινομημένο» κόμβο.
3) Κάθε κόμβος στον οποίο δεν προστέθηκε ούτε απομακρύνθηκε ακμή αλλά εμφανίζει την ίδια αλληλουχία ετικετών με κάποιον ταξινομημένο κόμβο θεωρείται επίσης ταξινομημένος.

Η εύρεση κόμβων με τις ανωτέρω ιδιότητες μπορεί να ανακύψει με κάποιον άπληστο αλγόριθμο, κατ' αναλογία με τις άλλες τεχνικές ανωνυμοποίησης – και αποτελεί αναμφίβολα ένα απαραίτητο επόμενο βήμα για την παρούσα έρευνα.

5.3 Παράδειγμα

Στην ενότητα αυτή θα παρουσιάσουμε ένα απλό παράδειγμα εφαρμογής της ως άνω τεχνικής. Για την πρακτική αυτή εφαρμογή θα χρησιμοποιηθεί ο γράφος του Σχήματος 12.

Κατ' αρχάς, δεδομένου ότι οι ακμές λαμβάνουν δύο πιθανές τιμές a και b, ορίζουμε τη διάταξη a<b. Στη συνέχεια, για κάθε κόμβο, καταγράφουμε την αλληλουχία των ετικετών του: προφανώς, με αυτή τη διαδικασία υπολογίζουμε ταυτόχρονα και το βαθμό κάθε κόμβου, αφού ο βαθμός ταυτίζεται πάντα με το μέγεθος (μήκος) της αλληλουχίας ετικετών.

Κόμβος	1	2	3	4	5	6	7	8	9	10	11	12

Άλλη-λουχία ετικετών	a	b	aabb	ab	aab	aaabb	ab	abb	aabb	a	b	b

Πίνακας 7. Αλληλουχία ετικετών των κόμβων του γραφήματος

Οι βαθμοί λοιπόν των κόμβων του γραφήματος είναι 1, 1, 4, 2, 3, 5, 2, 3, 4, 1, 1, 1

Παρατηρούμε ότι στον εν λόγω γράφο υπάρχουν κόμβοι μονοσήμαντα αναγνωρίσιμοι, βάσει της αλληλουχίας των ετικετών τους: είναι οι κόμβοι 5 (με αλληλουχία ετικετών aab) και 8 (abb).

Στη συνέχεια, στο πλαίσιο της κατηγοριοποίησης των κόμβνω, υπολογίζουμε τη μέση τιμή των βαθμών των κόμβων του γραφήματος. Πέντε κόμβοι έχουν βαθμό 1, δύο κόμβοι έχουν βαθμό 2, δύο κόμβοι έχουν βαθμό 3, δύο κόμβοι έχουν βαθμό 4 και ένας κόμβος έχει βαθμό 5. Συνεπώς, ο μέσος βαθμός του γράφου είναι:

(5*1+2*2+2*3+2*4+1*5)/12=(5+4+6+8+5)/12=28/12=2.33

Πλέον μπορούμε να δημιουργήσουμε τις δύο Κατηγορίες 1 και 2: στην Κατηγορία 1 θα είναι εκείνοι οι κόμβοι με βαθμό μικρότερο ή ίσο με 2, και οι εναπομείναντες κόμβοι με βαθμό τουλάχιστον ίσο με 3 θα ενταχθούν στην Κατηγορία 2. Θα έχουμε λοιπόν:

Κατηγορία 1: Κόμβος 1 (a), Κόμβος 2 (b), Κόμβος 4 (ab), Κόμβος 7 (ab), Κόμβος 10 (a), Κόμβος 11 (b), Κόμβος 12 (b)

Κατηγορία 2: Κόμβος 3 (aabb), Κόμβος 5 (aab), Κόμβος 6 (aaabb), Κόμβος8 (abb), Κόμβος 9 (aabb)

όπου, εντός παρενθέσεων, παραθέτουμε την αλληλουχία των ετικετών κάθε κόμβου.

Πλέον, επιδιώκουμε να δημιουργήσουμε αξεχώριστους κόμβους ανά κατηγορία, με βάση το σκεπτικό που περιγράψαμε ανωτέρω.

Στους κόμβους της Κατηγορίας 1 στο παράδειγμά μας, επιδιώκουμε κατ' αρχάς να προσθέσουμε ακμές, βάσει της Πρότασης 1. Προφανώς, ενώνοντας τους κόμβους 1 και

10 με ακμή ετικέτας «b» και τους 2 και 11 με ακμή ετικέτας «a», δημιουργείται η εξής νέα κατάσταση:

Κατηγορία 1:

Κόμβος 1 (ab), Κόμβος 2 (ab), Κόμβος 4 (ab), Κόμβος 7 (ab), Κόμβος 10 (ab), Κόμβος 11 (ab), Κόμβος 12 (b)

Άρα, με την προσθήκη δύο ακμών έχουμε 6 κόμβους αδιαίρετους (μένει μη ταξινομημένος, προς τροποποίηση, ο κόμβος 12).

Με άλλα λόγια, εφαρμόσαμε την Πρόταση 1 για τους κόμβους 1, 10 και 7 (δηλαδή προσθέτοντας μια ακμή κατάλληλης ετικέτας μεταξύ των 1 και 10 γίνονται αξεχώριστοι με τον κόμβο 7), καθώς επίσης και για τους κόμβους 2, 11 και 7 (δηλαδή προσθέτοντας μια ακμή κατάλληλης ετικέτας μεταξύ των 2 και 11 γίνονται αξεχώριστοι επίσης με τον κόμβο 7). Οι κόμβοι 1,10,2,11,7 θεωρούνται ταξινομημένοι και είναι όλοι μεταξύ τους αξεχώριστοι, με αλληλουχία ετικετών ab. Στη συγκεκριμένη περίπτωση, και ο κόμβος 4 είναι επίσης αξεχώριστος με όλους τους ταξινομημένους κόμβους, οπότε και αυτός κρίνεται ως ταξινομημένος.

Στην Κατηγορία 2, εφαρμόζοντας κατάλληλα την Πρόταση 2, αν αφαιρέσουμε μία ακμή με ετικέτα “a” από τον κόμβο 6, αυτός θα γίνει «αξεχώριστος» από τους 3 και 9. Οι δύο από τις τρεις αυτές ακμές όμως με ετικέτα «a» αντιστοιχούν σε κόμβους της Κατηγορίας 1 που είναι ήδη ταξινομημένοι (στους 4 και 7), οπότε - ως μόνη επιλογή - απομακρύνουμε την ακμή που συνδέει τον κόμβο 6 με τον κόμβο 9.

Η νέα κατάσταση λοιπόν είναι η εξής:

Κατηγορία 2:

Κόμβος 3 (aabb), Κόμβος 5 (aab), Κόμβος 6 (aabb), Κόμβος8 (abb), Κόμβος 9 (abb)

Έχουν δημιουργηθεί δύο ζευγάρια «αξεχώριστων» κόμβων, οι 3, 6 και οι 8,9. Αυτό σημαίνει μία ανωνυμία τάξης 2.

Μένει μη ταξινομημένος, προς τροποποίηση, ο κόμβος 5 (aab).

Οι μη ταξινομημένοι κόμβοι λοιπόν είναι οι 12(b) και 5 (aab). Δεν υπάρχει δυνατότητα σύνδεσής τους με κάποια ακμή η οποία να μην αφήνει αναγνωρίσιμους κόμβους. Για την

αντιμετώπιση αυτού, μπορούμε να προσθέσουμε - κατ' αναλογία με άλλες προσεγγίσεις για ανωνυμοποίηση γράφων κοινωνικής δικτύωσης - δύο εικονικούς κόμβους 13, 14, οι οποίοι ενώνονται με τον 12 με ακμές ετικέτας «a» (έτσι ώστε ο κόμβος 12 να γίνει αξεχώριστος με τον κόμβο 5), ενώ ενώνοντας τους εικονικούς κόμβους 13 και 14 με ακμή ετικέτας b, οι δύο αυτοί κόμβοι θα γίνουν αξεχώριστοι με τους 1,2,4,7,10,11.

Τελικά λοιπόν, θα έχουμε:

Κόμβος 1 (ab), Κόμβος 2 (ab), Κόμβος 4 (ab), Κόμβος 7 (ab), Κόμβος 10 (ab), Κόμβος 11 (ab), Κόμβος 12 (aab), Κόμβος 13 (ab), Κόμβος 14 (ab)

Κατηγορία 2:
Κόμβος 3 (aabb), Κόμβος 5 (aab), Κόμβος 6 (aabb), Κόμβος8 (abb), Κόμβος 9 (abb)

Κλάσεις ισοδυναμίας:
1) Κόμβος 1 (ab), Κόμβος 2 (ab), Κόμβος 4 (ab), Κόμβος 7 (ab), Κόμβος 10 (ab), Κόμβος 11 (ab), Κόμβος 13 (ab), Κόμβος 14 (ab)
2) Κόμβος 12 (aab), Κόμβος 5 (aab)

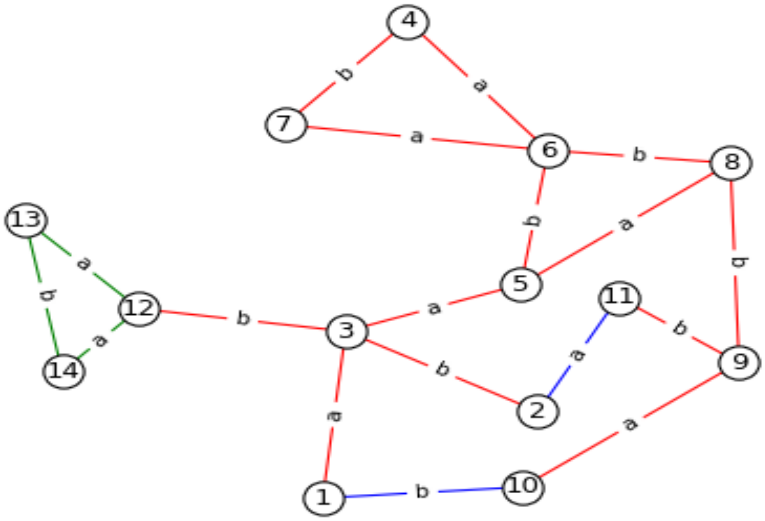
3) Κόμβος 8 (abb), Κόμβος 9 (abb)

4) Κόμβος 3 (aabb), Κόμβος 6 (aabb)

Αφού το ελάχιστο πλήθος κόμβων σε μία κλάση ισοδυναμίας είναι δύο (κλάσεις 2, 3 και 4), έχουμε επιτύχει μία ανωνυμία τάξης 2. Πρέπει ωστόσο να σημειωθεί ότι καταφέραμε να επιτύχουμε και μία κλάση ισοδυναμίας με μεγάλη πληθικότητα – ήτοι η κλάση 1, που περιέχει 8 κόμβους. Αυτό σημαίνει ότι για τους εν λόγω κόμβους καθίσταται εξαρετικά δύσκολη η οποιαδήποτε επίθεση άρσης της ανωνυμίας – αν αναλογιστεί κανείς ότι, στον αρχικό γράφο, το μεγαλύτερο πλήθος αξεχώριστων μεταξύ του κόμβων που υπήρχε ήταν μόλις 3 (οι κόμβοι 2, 11 και 12).

Το νέο μέσο μήκος του ανωνυμοποιημένου γράφου είναι (8*2 + 4*3 + 2*4)/14=(16+12+8)/14=36/14=2,57. Παρατηρούμε ότι είναι μία τιμή που δεν διαφέρει σημαντικά από την αντίστοιχη τιμή στον αρχικό γράφο.

Στο Σχήμα 13 απεικονίζεται ο νέος γράφος που υλοποιήθηκε μέσω κώδικα στη γλώσσα προγραμματισμού Python(βλέπε Παράρτημα Γ).



Σχήμα 13. Κοινωνικό δίκτυο μετά την εφαρμογή της τεχνικης (2ης τάξης ανωνυμία).

Συγκεκριμένα, ο κώδικας αρχικά εισάγει τον αρχικό γράφο από ένα αρχείο .txt (όπου περιέχει την απαραίτητη πληροφορία σχετικά με κόμβους, γείτονες, ετικέτες). Εν συνεχεία, εμφανίζει αυτές τις πληροφορίες και υπολογίζει τον μέσο βαθμό του γράφου. Έπειτα, κατηγοριοποιεί τους κόμβους σε δύο κατηγορίες: σε αυτούς με βαθμό μεγαλύτερο από το μέσο όρο και σε αυτούς με βαθμό μικρότερο από το μέσο όρο. Εμφανίζει το περιεχόμενο της κάθε κατηγορίας (δηλαδή ποιος κόμβος ανήκει στην κάθε κατηγορία καθώς και την ετικέτα του). Τέλος, δημιουργεί αξεχώριστους κόμβους με την εισαγωγή/αφαίρεση ακμών και υπολογίζει το νέο μέσο μήκος του νέου γράφου.

Πρέπει να σημειωθεί ότι οι ακμές με το κόκκινο χρώμα είναι οι αρχικές, οι ακμές με το μπλε χρώμα είναι αυτές που προέκυψαν απο το πρώτο στάδιο και τέλος οι πράσινες ακμές είναι αυτές που συνδέουν τους δυο νέους κόμβους που προστέθηκαν. Πρέπει επίσης να σημειωθεί ότι, σε σχέση με τον αρχικό γράφο, διαγράφηκε η ακμή που συνδέει τους κόμβους 6 και 9.

Κεφάλαιο 6
Συμπεράσματα- Επίλογος

Σύμφωνα με τα παραπάνω και σχετικά με όσα παραδείγματα παραθέσαμε, καταλήγουμε στο συμπέρασμα ότι τελικά είναι πολύ εύκολο κάποιος επιτήδειος που για προσωπικούς λόγους επιθυμεί να εισβάλει στο απόρρητο ενός χρήστη και να συλλέξει υλικό. Αν ο χρήστης λοιπόν δεν έχει εξειδικευμένες γνώσεις στο θέμα αυτό τότε πιθανόν ο επιτιθέμενος με προχωρημένες γνώσεις και με τον κατάλληλο συνδυασμό να μπορέσει να αποσπάσει πληροφορίες. Επομένως είναι απαραίτητο στα μέσα κοινωνικής δικτύωσης να μη χρησιμοποιούμε όλες τις προσωπικές μας πληροφορίες διότι είναι πολύ εύκολο να πάρει κάποιος αυτές τις πληροφορίες και να της χρησιμοποιήσει για προσωπικό σκοπό. Ο επιτιθέμενος αποκτά μεγαλύτερη ισχύ αν διαθέτει πληροφορία σχετικά με το γράφο του κοινωνικού δικτύου, την οποία μπορεί να τη συνδυάσει και με άλλη προηγούμενη γνώση για κάποιον/κάποιους από τους χρήστες. Αξίζει να σημειωθεί ότι οι γράφοι κοινωνικών δικτύων μπορούν να είναι και εύκολα διαθέσιμοι – π.χ. να διατίθενται σε μια μορφή χωρίς τα ονόματα των χρηστών, προκειμένου να μπορούν να αναλυθούν περαιτέρω για άλλους επιστημονικούς σκοπούς.

Αν για παράδειγμα ο επιτιθέμενος γνωρίζει μια ή και περισσότερες πληροφορίες τότε μπορεί να εκπληρώσει το σκοπό του. Όπως αναφέρθηκε υπάρχουν τρία επίπεδα για να συμβεί αυτό. Στο πρώτο ότι επιτιθέμενος γνωρίζει μόνο τις ετικέτες της κόμβου u, στο δεύτερο ότι ο επιτιθέμενος γνωρίζει τις ετικέτες της κόμβου u και το βαθμό της και στο τρίτο ο επιτιθέμενος γνωρίζει τις ετικέτες της κόμβου u , το βαθμό της κόμβου του και πληροφορίες σχετικά με τις ετικέτες των ακμών που είναι κοντά της.

Η νέα τεχνική που περιγράφηκε στο Κεφάλαιο 5 αποτελεί μία πρώτη προσπάθεια μιας καινούριας προσέγγισης: αποσκοπεί στην προσθήκη αλλά και αφαίρεση ακμών σε έναν γράφο, με τρόπο τέτοιο ώστε να επιτυγχάνεται ανωνυμοποίηση αλλά ταυτόχρονα να μη διαφέρει πολύ ο νέος αριθμός των ακμών στον ανωνυμοποιημένο γράφο, ούτε η μέση

τιμή των βαθμών των κόμβων. Προφανώς, πολλά βήματα πρέπει να γίνουν ακόμα ως συνέχεια της ερευνητικής αυτής προσπάθειας: κατ' αρχάς πρέπει να υλοποιηθεί ένας άπληστος (greedy) αλγόριθμος για την επιλογή των κόμβων στους οποίους θα προσθέτουμε/αφαιρούμε ακμές – αφού, κατά κανόνα, αναμένεται να είναι πολύ «υποψήφιοι» τέτοιοι κόμβοι – και να εφαρμοστεί σε μεγάλο σύνολο δεδομένων (μεγάλο γράφο), προκειμένου να ελεγχθεί ακριβέστερα η αποτελεσματικότητά του. Περαιτέρω, πρέπει να υπολογιστούν κατάλληλες μετρικές για την απώλεια της πληροφορίας, προκειμένου να μπορεί να συγκριθεί με άλλες γνωστές τεχνικές. Τέλος, θα πρέπει να ελεγχθεί αν και σε τι βαθμό μπορεί να συνδυαστεί με άλλες τεχνικές, όπως με την τεχνική k-γειτνίασης, καθώς επίσης και να εξεταστεί το ενδεχόμενο, αντί της προσθαφαίρεσης ακμών, να μεταβάλλονται κατάλληλα οι ετικέτες των ακμών.

Παράρτημα Α

Πίνακες

Παράρτημα Β
Σχήματα

Σχήμα 1: Σύνδεση (Linking ή matching) δεδομένων για εκ νέου εντοπισμό δεδομένων.

Σχήμα 2: Ένα απλό δίκτυο φιλίας με δυο κοινότητες. Η μια κοινότητα περιέχει κόμβους σε γκρι χρώμα και η άλλη κοινότητα σε λευκό χρώμα.

Σχήμα 3: Παράδειγμα επίθεση φιλίας. α) Αρχικός κοινωνικός γράφος και β) Απλοϊκή ανωνυμοποίηση γράφου

Σχήμα 4: Μια απλή τεχνική ανωνυμίας αντικαθιστώντας τα ευαίσθητα χαρακτηριστικά, όπως το όνομα, χρησιμοποιώντας ακέραιους. a) Το πρωτότυπο κοινωνικό δίκτυο. b) Ανώνυμο κοινωνικό δίκτυο, αντικαθιστώντας τα ονόματα με ακέραιο αριθμό αναγνώρισης

Σχήμα 5: Ανώνυμα κοινωνικά δίκτυα που χρησιμοποιούν k-βαθμού ανωνυμία και η k-γειτνίασης Ανωνυμία. Οι διακεκομμένες ακμές προστίθενται στα δίκτυα για να επιτευχθεί η προστασία της ιδιωτικής ζωής. Α) Το 2-βαθμού ανώνυμο δίκτυο. b) 2-γειτνίασης ανώνυμο δίκτυο

Σχήμα 6: 2-βαθμού ανώνυμος γράφος με την αποτυχία της προστασίας της ιδιωτικής ζωής κατά την επίθεση γειτονίας με βάση τη γειτονία a) Αρχικός γράφος κοινωνικού δικτύου b)Ανωνυμοποίηση στις ταυτότητες των χρηστών c) Η γειτονιάς της Ada είναι η μοναδική σε σχέση με όλους τους κόμβους του γράφου (Zou L., Chen L., Ozsu M.T., 2009:946-957)

Σχήμα 7: Κατάτμηση του γράφου και αντιγραφή ακμών a) Απλοϊκή ανωνυμοποίηση κοινωνικού δικτύου b) Κατάτμηση του γράφου και ευθυγράμμιση μπλοκ c) Αντιγραφή ακμής

Σχήμα 8: Γράφοι ανώνυμου κοινωνικού δικτύου ενάντια στην αποκάλυψη της σχέσης των χρηστών. α) 4ης τάξης βαθμού ανωνυμία β) 4ης τάξης ποικιλομορφία

Σχήμα 9: Απλοϊκή προστασία α) Αρχικός γράφος β)Προστασία Επιπέδου 1
γ) Προστασία Επιπέδου 2 δ) Προστασία Επιπέδου 3

Παράρτημα Γ
Υλοποίηση κώδικα

```
import networkx as nx
import matplotlib.pyplot as plt
import statistics as s
import sys
from collections import OrderedDict

## Variables definition
labels_sum = ""
group1_dic = {}
group2_dic = {}
file_name = "D:\Users\maggie\Documents\Master\Diatriviv2\diatrivi\Python\data.py"
G = nx.Graph()
## Categorization function
def categories( group_list ):
   group_dic = {}
   counter = 0
   for node in group_list:
      group_dic[node] = ""
      neighbor = G.neighbors(node)
      counter1 = 0
      while counter1 < len(neighbor):
         if group_dic[node] == "":
            group_dic[node] = G[node][neighbor[counter1]]["label"]
         else:
```

```
            val = "".join(sorted(group_dic[node]+G[node][neighbor[counter1]]["label"]))
            group_dic[node] = val
        counter1 += 1
    for counter in group_dic:
        print("{}({})".format(counter, group_dic[counter])),
    print ""
    return group_dic;

## read the file
with open(file_name) as file:
    for txtline in file:
        line = txtline.split()
        if len(line) == 3:
            G.add_edge (int(line[0]),int(line[1]),color = "r", label = line[2])
edge_labels = nx.get_edge_attributes(G,"label")
pos = nx.spring_layout(G)
lst = [] ## create list
count=1

## loop to append adjency list to a new list
while count <= G.order():
    lst.append(G.degree(count))
    count += 1
group1_lst = []
group2_lst = []
count=1
while count <= G.order():
    if G.degree(count) < s.mean(lst) :
        group1_lst.append(count)
    elif G.degree(count) > s.mean(lst) :
        group2_lst.append(count)
```

```
    count += 1
print "Nodes in graph: "
print G.nodes()
print "Edges in graph: "
print G.edges()
print "Labels of edges: "
print edge_labels
print "Average degree of nodes: ",
print s.mean(lst)
print ""
print ("Group 1 nodes: ")
group1_dic = categories(group1_lst)
print "Group 2 nodes: "
group2_dic = categories(group2_lst)

## plus minus
for counter in group1_dic:
    labels_sum += group1_dic[counter]
for counter in group2_dic:
    labels_sum += group2_dic[counter]

## find unique label names
labels_uniq = OrderedDict.fromkeys(labels_sum).keys()

## find the nodes that need adjustment as odd number
max_degree = 0
total_low = len(group1_dic)
for counter in group1_dic:
    if len(group1_dic[counter]) > max_degree:
        max_degree = len(group1_dic[counter])
    elif len(group1_dic[counter]) == max_degree:
```

```
        total_low -= 1
if (total_low %2 != 0):
    total_low -= 1

## add edges to nodes
counter = 0
new_edges=[]
new_group1_dic ={}
new_edges_labels=[]
leftover_nodes=[]
for node in group1_lst:
    if (G.degree(node) < max_degree) and (counter < total_low):
        new_edges.append(node)
        counter1 = 0
        while counter1 < len(labels_uniq):
            if labels_uniq[counter1] not in group1_dic[node]:
                new_edges_labels.append(labels_uniq[counter1])
            counter1 += 1
        counter +=1
    else:
        if G.degree(node) < max_degree:
            leftover_nodes.append(node)
for counter in range(0,len(new_edges)-2):
    G.add_edge(new_edges[counter],new_edges[counter+2], color ="b",
label=new_edges_labels[counter])

## Remove edges to nodes
max_degree = 0
max_prev_degree = 0
for counter in group2_dic:
    if len(group2_dic[counter]) > max_degree:
```

```
        max_prev_degree = max_degree
        max_degree = len(group2_dic[counter])
new_edges=[]
new_group2_dic = {}
for node in group2_lst:
    if (G.degree(node) <= max_degree) and (G.degree(node) > max_prev_degree):
        new_edges.append(node)
        neighbors=G.neighbors(node)
        for nb in neighbors:
            if (G.degree(nb) >= max_prev_degree) and (nb in group2_lst):
                G.remove_edge(node,nb)
    else:
        if G.degree(node) < max_prev_degree:
            leftover_nodes.append(node)

## categorise new nodes
print "New group 1 node degrees (after addition): "
new_group1_dic = categories(group1_lst)
print "New group 2 node degrees (after changes): "
new_group2_dic = categories(group2_lst)
## finding leftovers
result = {}
leftovers = {}
for value in new_group1_dic.values() + new_group2_dic.values():
            if value not in result:
                        result[value]=1
            else:
                        result[value]=result[value]+1
for key,value in result.items():
    if value == 1:
        for key2,value2 in (new_group1_dic.items() + new_group2_dic.items()):
```

```
        if key == value2:
            leftovers[key2]=value2


## Find the needed edges
max_degree = 0
max_node = {}
for counter in leftovers:
    if len(leftovers[counter]) > max_degree:
        max_degree = len(leftovers[counter])
max_node[counter] = leftovers[counter]
del leftovers[counter]
counter = 1
for node in leftovers:
    new_edges_labels = list(str(max_node.values()).replace(leftovers[node],""))
    new_edges_labels=[y for y in new_edges_labels if y != "'"]
    new_edges_labels.remove ("[")
    new_edges_labels.remove ("]")
    for counter in range(len(new_edges_labels)):
        G.add_edge(node,G.nodes()[-1]+1, color = "g", label=new_edges_labels[counter])
G.add_edge(G.nodes()[-1]-1,G.nodes()[-1], color = "g", label=leftovers[node]) # add last
edge between the last nodes


## update the group lists and print a degree view
lst = [] ## create list
count=1
while count <= G.order():
    lst.append(G.degree(count))
    count += 1
group1_lst = []
group2_lst = []
count=1
```

```
while count <= G.order():
   if G.degree(count) < s.mean(lst) :
      group1_lst.append(count)
   elif G.degree(count) > s.mean(lst) :
      group2_lst.append(count)
   count += 1
print "Final view of Group 1 node degrees:"
new_group1_dic = categories(group1_lst)
print "Final view of Group 2 node degrees:"
new_group2_dic = categories(group2_lst)
print "New Average degree of nodes: ",
print s.mean(lst)

## Plot the graph
edges = G.edges()
edge_labels = nx.get_edge_attributes(G,"label") # Getting the latest labels
edge_colors = [G[u][v]["color"] for u,v in edges]
pos= {1: ([ 0.40066374,  0]), 2: ([ 0.69048557,  0.18164301]), 3: ([ 0.43183594,
0.34988748]), 4: ([ 0.51551609,  1.]), 5: ([ 0.66777685,  0.44822789]), 6: ([ 0.70396087,
0.72962364]), 7: ([ 0.34905021,  0.78336586]), 8: ([ 0.94948321,  0.7027138 ]), 9: ([
0.9610698 ,  0.28406085]), 10: ([ 0.6697058 ,  0.02330736]), 11: ([ 0.79878245,
0.42106287]), 12: ([ 0.15403436,  0.39718792]), 13: ([ 0.,  0.58378435]), 14: ([
0.05001318,  0.26655176])}
nx.draw_networkx_edge_labels(G, pos, edge_labels = edge_labels)
nx.draw_networkx(G, pos, edge_color = edge_colors, node_color = "w") ## draw
everything but the edge labels
plt.savefig("simple_path.png") ## save as png
plt.show() ## display the plot
quit()
```

Βιβλιογραφία

Backstrom A. L., Dwork C., Kleinberg J. (2007) Wherefore Art Thou R3579x? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In Proceedings of the 16th international conference on World Wide Web (WWW'07), ACM Press, pages 181–190, New York, NY, USA.

Byun J. W., Kamra A., Bertino E., Li N. (2007) Efficient k-Anonymization using Clustering Techniques. In: Kotagiri, R., Radha, Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, Springer, Heidelberg, pages 188–200, Bangkok, Thailand.

Campan A., Truta T.M. (2008) A Clustering Approach for Data and Structural Anonymity in Social Networks. In Privacy, Security, and Trust in KDD Workshop PinKDD, pages 33-54, Las Vegas, NV, USA.

Ghinita G., Karras P., Kalnis P., Mamoulis N. (2007) Fast Data Anonymization with Low Information Loss. In: Very Large Data Base Conference (VLDB), pages 758–769, University of Vienna, Austria

Hay M., Miklau G., Jensen D., Towsley D., Weis Ph. (2008) Resisting Structural Re-identification Anonymized Social Networks. Proc. VLDB Endow., pages 102-114.

Hay M., Miklau G., Jensen D., Towsley D., Weis Ph., Srivastava S. (2007) Anonymizing social networks. Technical Report, University of Massachusetts Amherst, pages 07-19.

He X., Vaidya J., Shafiq B., Adam N., Atluri V. (2009) Preserving privacy in social networks: A structure-aware approach. In Proc. of WI-IAT, pages 647-654, Milan, Italy.

Kleinberg J. M. (2007) Challenges in mining social network data: processes, privacy, and paradoxes. In: Berkhin P, Caruana R, Wu X (eds) Proceedingsof the 13th ACM SIGKDD international conference on knowledge discoveryand data mining (KDD'07), pages 4-5, San Jose, California, USA.

Liu K., Terzi E. (2008) Towards identity anonymization on graphs. In: Proceedings of the 2008 ACM SIGMOD international conference on management of data (SIGMOD'08). ACM Press, pages 93–106, Vancouver, Canada, USA

Liu K., Das K., Grandison T., Kargupta H. (2008) Privacy-preserving data analysis on graphs and social networks. In: Kargupta H., Han J., Yu P., Motwani R., Kumar V. (eds) Next generation data mining. CRC Press, pages 419–437.

Machanavajjhala A., Gehrke J., Kifer D., Venkitasubramaniam M. (2006) L-diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd IEEE international conference on data engineering (ICDE'06), IEEE Computer Society, Washington, DC pages 24, Atlanta, Georgia, USA.

Ruan X., Yue C., Wang H. (2013) Unveiling privacy setting breaches in online social networks. In: Proceedings of the 10th international conference on security and privacy in communication networks, pages 323-341, Sydney, Australia.

Sargolzaei E., Khazali M. J., Keikha F. (2016) Privacy preserving approach of published social networks data with vertex and edge modification algorithm. Ind. J. of Science and Technology, pages 1–8, Zabol, Iran.

Sophos (2008) Facebook privacy breach exposed user's hidden date of birth. In: Global security mag, July 2008.

Sweeney L. (2002) K-anonymity: a model for protecting privacy. Int J. Uncertain Fuzziness Knowl-Based Systems, pages 557–570, Pittsburgh, Pennsylvania, USA

Wu X., Ying X., Liu K., Chen L. (2010) A Survey of Privacy-Preservation of Graphs and Social Networks, Springer US, pages 421-453.

Ying X., Wu X. (2008) Randomizing Social Networks: A Spectrum Preserving Approach, Proceedings of the 2008 SIAM International Conference on Data Mining, pages 739-750.

Yu L., Zhu J., Wu Z., Yang T., Hu J., Chen Z. (2012) Privacy Protection in SocialNetworks Using l-Diversity. In ICICS, Springer, pages 435-444, Berlin Heidelberg.

Zheleva E., Getoor L. (2008) Preserving the Privacy of Sensitive Relationships in Graph Data. In Proceedings of the 1st ACM SIGKDD International Conference on Privacy, Security, and Trust in KDD, PinKDD'07, Springer-Verlag, pages 153-171, Berlin, Heidelberg.

Zheleva E., Getoor L. (2011) Privacy issues in social networks: a brief survey social network data analytics. Springer US, pages 277-306, Springer, Berlin.

Zhou B., Pei J., Luk W.-S. (2008) A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM SIGKDD Explorations Newsletter, pages 12–22, New York, USA.

Zhou B., Pei J. (2008) Preserving Privacy in Social Networks AgainstNeighborhood Attacks. In Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, ICDE '08, pages 506-515, Washington, DC, USA

Zou L., Chen L., Ozsu M.T. (2009) K-automorphism: a general framework for privacy preserving network publication. In: Proceedings of the VLDB Endowment (VLDB) pages 946–957.

Zhou B., Pei J. (2011) The k-anonymity and l-diversity Approaches for Privacy Preservation in Social Networks Against Neighborhood Attacks. Knowl. Inf. Syst., Springer-Verlag, pages 47-77, London