Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Εξόρυξη γνώσης σε κοινωνικά δίκτυα**
Ονοματεπώνυμο Φοιτητή	**Νίκη Ταράτσα**
Πατρώνυμο	**Χρήστος**
Αριθμός Μητρώου	**ΜΠΣΠ/09012**
Επιβλέπων	**Γιάννης Θεοδωρίδης, Αν. Καθηγητής**

Ημερομηνία Παράδοσης **Δεκέμβριος 2011**

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Τριμελής Εξεταστική Επιτροπή

Γιάννης Θεοδωρίδης Αν. Καθηγητής	Χάρης Κωνσταντόπουλος Λέκτορας	Νίκος Πελέκης Λέκτορας

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΙΝΑΚΑΣ ΕΙΚΟΝΩΝ

ΠΕΡΙΛΗΨΗ

Από το Διαδίκτυο έχουν προκύψει πολλά δίκτυα ανταλλαγής πληροφοριών, το πιο γνωστό από τα οποία είναι το World Wide Web. Πρόσφατα, μια νέα κατηγορία πληροφοριακών δικτύων που ονομάζονται Κοινωνικά Δίκτυα έχουν γίνει ιδιαίτερα δημοφιλή. Οι ιστότοποι κοινωνικών δικτύων όπως το MySpace (πάνω από 125 εκατομμύρια χρήστες), το Facebook (πάνω από 800 εκατομμύρια χρήστες), το Orkut (πάνω από 100 εκατομμύρια χρήστες) και το LinkedIn (πάνω από 110 εκατομμύρια χρήστες) αποτελούν παραδείγματα των εξωφρενικά δημοφιλών δικτύων που χρησιμοποιούνται για την εύρεση και την οργάνωση επαφών. Άλλα κοινωνικά δίκτυα όπως το Flickr, το YouTube και το Google Video χρησιμοποιούνται από τους χρήστες με σκοπό να μοιράζονται περιεχόμενο πολυμέσων και κάποια άλλα όπως το LiveJournal και το BlogSpot χρησιμοποιούνται για να μοιράζονται απόψεις / συζητήσεις (blogs).

Η μελέτη και η ανάλυση αυτών των δικτύων μας προσφέρουν σημαντικές πληροφορίες σε θέματα δομής και χαρακτηριστικών του δικτύου, εμπιστοσύνης και διάδοσης πληροφοριών. Το βασικό πρόβλημα που καλούμαστε να αντιμετωπίσουμε με αυτή την εργασία είναι η ανάλυση των τρόπων με τους οποίους μπορούμε να εξάγουμε χρήσιμα συμπεράσματα από τα δεδομένα κοινωνικών δικτύων χρησιμοποιώντας τεχνικές εξόρυξης γνώσης. Η εργασία αυτή αποτελείται από δύο βασικά μέρη. Το θεωρητικό μέρος το οποίο περιέχει την αναλυτική παρουσίαση/καταγραφή των τεχνικών εξόρυξης γνώσης (αλγόριθμοι ανίχνευσης κοινοτήτων, κατηγοριοποίησης κόμβων, πρόβλεψης συνδέσμου, εύρεσης ειδικών, εμπιστοσύνης κτλ.) σε κοινωνικά δίκτυα και το πρακτικό μέρος στο οποίο γίνεται μια προσπάθεια να εξαχθούν συμπεράσματα από μια βάση δεδομένων κοινωνικού δικτύου χρησιμοποιώντας τις τεχνικές εξόρυξης γνώσης των SQL Server Analysis Services.

ABSTRACT

Since the conception of the Internet, WWW has become the most well-known sharing information network. Recently, a new class of information networks has gained tremendous popularity and now rivals the traditional WWW in terms of usability. Social networks like MySpace (over 125 million users), Facebook (over 800 million users), Orkut (over 100 million users), and LinkedIn (over 110 million "professionals") are examples of wildly popular networks used to find and manage contacts. Other social networks such as Flickr, YouTube, and Google Video provide multimedia content sharing; on the other hand networks such as LiveJournal and BlogSpot are used to share blogs.

The study and analysis of social networks provide us with important information about the network structure, its associated properties and issues related with trust and information distribution. The main problem we face with this study is to survey and analyze various mining techniques that can be used to draw useful conclusions based on the data of the social networks. This dissertation consists of two parts: a) the theoretical one which contains a detailed survey of data mining techniques (community detection, classification nodes, link prediction, finding experts, trust, etc.) and b) the practical part which presents an effort to draw conclusions from a social network database using the data mining techniques provided by SQL Server Analysis Services.

ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ

1. Εισαγωγή

Τα τελευταία χρόνια παρατηρείται ένα αυξανόμενο ακαδημαϊκό ενδιαφέρον για την έρευνα σχετικά με τους ιστότοπους κοινωνικών δικτύων μέσα από ένα ευρύ φάσμα επιστημονικών κλάδων. Η ανάλυση κοινωνικών δικτύων εστιάζει στη δομή των σχέσεων, η οποία περιλαμβάνει από απλές γνωριμίες έως στενούς δεσμούς. Είναι μία μέθοδος με αυξανόμενη εφαρμογή εκτός από τους τομείς της πληροφορικής και σε άλλους τομείς όπως η κοινωνιολογία, η εγκληματολογία, η πολιτική και το μάρκετινγκ. Η παρατήρηση της ροής πληροφοριών σε ένα κοινωνικό δίκτυο συμβάλλει σημαντικά στη βελτίωση των τεχνικών διαφήμισης. Επίσης, ο τρόπος αλληλεπίδρασης των μελών ενός κοινωνικού δικτύου μπορεί να φανεί χρήσιμος ακόμη και στην ανίχνευση τρομοκρατών.

Η ανάλυση κοινωνικών δικτύων βοηθά τους ερευνητές να κατανοήσουν τον τρόπο με τον οποίο επικοινωνούν και συνεργάζονται οι άνθρωποι και να αναγνωρίσουν τη ροή της γνώσης σε επίπεδο τόσο μεταξύ διάφορων οργανισμών όσο και μέσα σε κάθε οργανισμό. Αρκετές επιχειρήσεις προσφέρουν υπηρεσίες βασισμένες στην ανάλυση κοινωνικών δικτύων, υποσχόμενες βελτιστοποίηση της πληροφοριακής ροής ως ένα τρόπο για τη βελτίωση της αποτελεσματικότητας, τη μείωση του κόστους και την αύξηση της παραγωγικότητας. Σε αντίθεση με το παραδοσιακό Web το οποίο σε μεγάλο βαθμό οργανώνεται από το περιεχόμενο, τα κοινωνικά δίκτυα ενσωματώνουν τους χρήστες ως πρώτης κατηγορίας οντότητες. Οι χρήστες συνδέονται σε ένα δίκτυο, δημοσιεύουν το δικό τους περιεχόμενο και δημιουργούν συνδέσεις με άλλους χρήστες του δικτύου που ονομάζεται «φίλοι». Αυτή η βασική δομή σύνδεσης «χρήστη-προς-χρήστη» διευκολύνει την ηλεκτρονική αλληλεπίδραση παρέχοντας έναν μηχανισμό για την οργάνωση τόσο του πραγματικού κόσμου όσο και των εικονικών επαφών με σκοπό την εύρεση άλλων χρηστών με παρόμοια ενδιαφέροντα καθώς και τον εντοπισμό περιεχομένου και γνώσης που έχει προστεθεί από «φίλους» .

Όπως αναφέραμε και παραπάνω η κατανόηση της συμπεριφοράς των χρηστών όταν συνδέονται στα κοινωνικά δίκτυα και η εξαγωγή συμπερασμάτων από τα δεδομένα των κοινωνικών δικτύων είναι σημαντικές για πολλούς λόγους. Οι μελέτες της συμπεριφοράς των χρηστών επιτρέπουν την αξιολόγηση της απόδοση των υπαρχόντων συστημάτων με κύριο στόχο τον καλύτερο σχεδιασμό των ιστοσελίδων και της τοποθέτησης διαφημίσεων. Τα μοντέλα συμπεριφοράς των χρηστών στα κοινωνικά δίκτυα είναι ζωτικής σημασίας στις κοινωνικές επιστήμες όπως το viral marketing. Η εξόρυξη γνώσης από τα δεδομένα κοινωνικών δικτύων μπορεί να αποκαλύψει «κρυφές» κοινωνικές τάσεις και να προκύψουν σημαντικά ευρήματα σχετικά με τη συμπεριφορά των ανθρώπων.

Συνεπώς, όλα τα παραπάνω αποτελούν βασικό κίνητρο για την ύπαρξη της συγκεκριμένης εργασίας η οποία αποτελείται από επτά κεφάλαια: Στο Κεφάλαιο 1 γίνεται μια προσπάθεια εισαγωγής και παρουσίασης του αντικειμένου της εργασίας. Στο Κεφάλαιο 2 παρουσιάζονται τα βασικά χαρακτηριστικά της εξόρυξης γνώσης. Στο Κεφάλαιο 3 δίνεται το θεωρητικό υπόβαθρο των κοινωνικών δικτύων παρουσιάζοντας ορισμούς, ιστορική αναδρομή, χαρακτηριστικά, κίνητρα έρευνας κτλ. Στο Κεφάλαιο 4 αναλύονται οι ιδιότητες των κοινωνικών δικτύων και οι αλγόριθμοι ανάλυσης/εξόρυξης γνώσης σε θέματα ανίχνευσης κοινοτήτων, κατηγοριοποίησης κόμβων, πρόβλεψης συνδέσμου, εμπιστοσύνης, εύρεσης ειδικών. Στο Κεφάλαιο 5 παρουσιάζονται βασικές έννοιες του Microsoft SQL Server Analysis Services που θα χρησιμοποιηθεί στο πρακτικό μέρος της εργασίας με σκοπό την εξόρυξη γνώσης από ένα κοινωνικό δίκτυο. Στο Κεφάλαιο 6 παρουσιάζεται το σύνολο δεδομένων που θα χρησιμοποιηθεί, η σχεδίαση της βάσης δεδομένων καθώς και ορισμένα ερωτήματα που βοηθούν στην εξερεύνηση/κατανόηση των δεδομένων. Στο Κεφάλαιο 7 σχολιάζονται τα αποτελέσματα των αλγορίθμων εξόρυξης γνώσης που προέκυψαν από τη δημιουργία μοντέλων στον SQL Server Analysis Services μέσω του εργαλείου Microsoft Visual Studio. Τέλος, παρουσιάζονται τα συμπεράσματα, η συνεισφορά της εργασίας.

2. Εξόρυξη γνώσης

Η εξόρυξη γνώσης (data mining) έχει εξελιχθεί σε ένα από τα βασικότερα ερευνητικά ζητήματα στον τομέα των βάσεων δεδομένων αποτελώντας αντικείμενο μελέτης από πολλούς ερευνητές ιδιαίτερα τα τελευταία χρόνια λόγω της ραγδαίας αύξησης του όγκου της πληροφορίας. Ένας ευρύς ορισμός για την εξόρυξη γνώσης θα μπορούσε να είναι η διαδικασία ημι-αυτόματης ανάλυσης μεγάλων βάσεων δεδομένων με στόχο την εύρεση χρήσιμης πληροφορίας – «γνώσης». Από την άλλη πλευρά η ανακάλυψη γνώσης από δεδομένα (knowledge discovery in data – KDD) είναι μια διαδικασία που αναπτύσσεται με ραγδαίους ρυθμούς και χαρακτηρίζεται ως η μη τετριμμένη διαδικασία εύρεσης έγκυρων, πρωτότυπων, πιθανώς χρήσιμων και οπωσδήποτε κατανοητών προτύπων μέσα στα δεδομένα. Η ανάγκη για τη χρησιμοποίηση των τεχνικών KDD στην εξόρυξη χρήσιμων πληροφοριών γίνεται κυρίως για δύο λόγους, οικονομικούς και επιστημονικούς.

Ο τομέας της *εξόρυξης γνώσης* (data mining) σχετίζεται με πολλούς άλλους τομείς όπως η στατιστική (statistics), η τεχνητή νοημοσύνη (artificial intelligence), η μηχανική μάθησης (machine learning), οι βάσεις δεδομένων (data bases), οι μηχανές αναζήτησης (search engines), τα συστήματα υποστήριξης αποφάσεων (decision support systems), τα συστήματα άμεσης ανάλυσης δεδομένων (OLAP) και το ταίριασμα προτύπων (pattern matching). Στη συνέχεια, θα αναλύσουμε τη σχέση που έχει η εξόρυξη γνώσης με μερικούς από τους πιο βασικούς τομείς που μόλις αναφέρθηκαν [1]:

- *Στατιστική:* είναι γνωστό πως ένα μεγάλο μέρος της ερευνητικής βάσης της εξόρυξης γνώσης βασίζεται στη στατιστική. Αυτό είναι λογικό μιας και η στατιστική έχει ανάλογους σκοπούς με την εξόρυξη γνώσης αφού αποσκοπούν στην αναγνώριση χρήσιμων πληροφοριών και προτύπων στα δεδομένα. Μέρος των διαδικασιών σε ένα μοντέλο εξόρυξης γνώσης μπορεί να αποτελεί η αναζήτηση των δεδομένων και η εξαγωγή συμπερασμάτων από τα αποτελέσματα μιας αναζήτησης. Μια συχνά χρησιμοποιούμενη τεχνική στην εξόρυξη γνώσης είναι αυτή της δειγματοληψίας. Αυτός ο τρόπος στη στατιστική λέγεται «στατιστική εξαγωγή συμπεράσματος». Ακόμα Όπως και με τις κλασικές τεχνικές στατιστικής έτσι και στην εξόρυξη γνώσης ακολουθούμε ανάλυση παλινδρόμησης (regression analysis), ανάλυση συστάδων (cluster analysis) κτλ.

- *Τεχνητή νοημοσύνη:* ένας άλλος τομέας που σχετίζεται με αυτόν της εξόρυξης γνώσης είναι η τεχνητή νοημοσύνη. Σκοπός της τεχνητής νοημοσύνης είναι να βγάζει λογικά συμπεράσματα από μη επεξεργασμένα δεδομένα, κάτι που κάνει και ο τομέας της εξόρυξης δεδομένων. Επίσης ο τομέας της εξόρυξης γνώσης κάνει εκτεταμένη χρήση εργαλείων τεχνητής νοημοσύνης. Μερικά παραδείγματα είναι τα νευρωνικά δίκτυα, τα δέντρα απόφασης και οι μηχανές διανυσμάτων (vector machines).

- *Μηχανική μάθηση*: είναι μια περιοχή της τεχνητής νοημοσύνης η οποία εξετάζει πως μπορούμε να δημιουργούμε προγράμματα τα οποία μπορούν να μαθαίνουν. Στην εξόρυξη γνώσης, η μηχανική μάθηση χρησιμοποιείται για τεχνικές πρόβλεψης ή κατηγοριοποίησης. Με την μηχανική μάθηση, ο υπολογιστής κάνει κάποιες προβλέψεις και στη συνέχεια βασιζόμενος στην ανατροφοδότηση (feedback) μαθαίνει από αυτές. Όταν συμβεί μελλοντικά ανάλογη περίπτωση, η ανατροφοδότηση χρησιμοποιείται για να κάνει την ίδια πρόβλεψη ή για να κάνει μια εντελώς διαφορετική πρόβλεψη.

- *Βάσεις δεδομένων*: ένα μεγάλο μέρος των σημερινών ερευνητών στην εξόρυξη γνώσης είναι άτομα προερχόμενα από τον τομέα των βάσεων δεδομένων. Η σχέση των δύο αυτών τομέων είναι εμφανής μιας και πριν επεξεργαστούμε τα δεδομένα μας πρέπει πρώτα να μπορούμε να τα διαχειριστούμε ορθά. Έτσι χωρίς καλά συστήματα διαχείρισης δεδομένων δεν μπορούμε να εφαρμόσουμε αλγόριθμους εξόρυξης γνώσης.

2.1 Εφαρμογές εξόρυξης γνώσης

Σε αυτή την ενότητα θα παρουσιάσουμε τις βασικές περιοχές εφαρμογής του τομέα της εξόρυξης γνώσης [1] & [2]:

- *Μάρκετινγκ:* στο μάρκετινγκ, η κύρια εφαρμογή είναι τα συστήματα μάρκετινγκ βάσεων δεδομένων, τα οποία αναλύουν τις βάσεις δεδομένων πελατών με σκοπό τον προσδιορισμό διαφορετικών ομάδων πελατών και την πρόβλεψη της συμπεριφοράς τους. Αναζητούν απαντήσεις σε ερωτήματα όπως, τι είναι αυτό που θέλουν οι πελάτες, ποιες είναι οι ανάγκες τους κτλ. Ο τομέας της εξόρυξης γνώσης έχει συνεισφέρει σημαντικά σε αυτή την κατεύθυνση από την ανάλυση δεδομένων μια επιχείρησης μέχρι και την εξαγωγή χρήσιμων συμπερασμάτων για την συμπεριφορά των πελατών. Ένας αρκετά γνωστός αλγόριθμος εξόρυξης γνώσης είναι ο A-Priori. Ο αλγόριθμος αυτός κάνει ανάλυση δεδομένων αγοράς, όπου υπάρχουν δεδομένα σχετικά με πελάτες ή αγορές σε καταστήματα προσφέροντας συμπεράσματα όπως «ένας πελάτης ο οποίος αγοράζει το x προϊόν θα αγοράσει και το y προϊόν με μεγάλη πιθανότητα».

- *Επένδυση:* πολυάριθμες χρηματιστηριακές εταιρίες βασίζονται σε τεχνικές εξόρυξης με σκοπό να επιτύχουν καλύτερες επενδύσεις. Για το λόγο αυτό πολλές έρευνες στον τομέα εξόρυξης γνώσης έχουν γίνει έχοντας ως αφετηρία χρηματιστηριακές εφαρμογές. Επίσης, μια άλλη χρήση των τεχνικών εξόρυξης γνώσης είναι οι εφαρμογές εξόρυξης γνώσης από κείμενα.

- *Πρόληψη και ασφάλεια:* η εξόρυξη γνώσης έχει με επιτυχία εφαρμοστεί τόσο στην πρόληψη όσο και στην αποφυγή διάφορων τύπων απάτης. Τα συστήματα HNC Falcon και Nestor PRISM χρησιμοποιούνται για την παρακολούθηση της απάτης πιστωτικών καρτών. Από την αναγνώριση κακόβουλων ενεργειών σε συναλλαγές κάποιος μπορεί να αντιληφθεί συναλλαγές που μπορεί να σχετίζονται με οικονομικές παρανομίες ή άλλου είδους απάτες. Ένα παράδειγμα συστήματος είναι το FAIS.

- *Επιστήμη:* οι αλγόριθμοι εξόρυξης γνώσης χρησιμοποιούνται ευρέως σε εφαρμογές από διάφορους άλλους επιστημονικούς τομείς. Ένα αξιοσημείωτο παράδειγμα είναι το SKYCAT, ένα σύστημα εξόρυξης γνώσης που αναλαμβάνει ανάλυση και κατηγοριοποίηση χωρικών αντικειμένων. Αυτό που είναι αξιοσημείωτο, είναι πως το SKYCAT εκτελεί αλγόριθμους για την ανίχνευση αντικειμένων από εικόνες.

- *Βιομηχανία:* το σύστημα αντιμετώπισης προβλημάτων CASSIOPEE εφαρμόστηκε από τρεις μεγάλες ευρωπαϊκές αεροπορικές εταιρείες για τη διάγνωση και την πρόβλεψη των προβλημάτων στον τύπο Boeing737. Χρησιμοποιήθηκαν τεχνικές συσταδοποίησης με σκοπό να δημιουργηθούν οικογένειες των καταγεγραμμένων βλαβών. Το CASSIOPEE έλαβε το πρώτο ευρωπαϊκό βραβείο ως προς την καινοτομία της εφαρμογής.

- *Τηλεπικοινωνίες:* το σύστημα ανάλυσης τηλεπικοινωνιών και συναγερμών TASA χρησιμοποιεί ένα νέο πλαίσιο για τον εντοπισμό συχνών επεισοδίων συναγερμού και τα παρουσιάζει ως κανόνες. Με τον τρόπο αυτό προσφέρει μια ποικιλία επιλογής και ιεράρχησης των κριτηρίων σχετικά με τα επεισόδια και την υποστήριξη επαναληπτικής ανάκτησης πληροφοριών.

- *Καθαρισμός δεδομένων:* το σύστημα MERGE-PURGE εφαρμόστηκε για τον εντοπισμό των διπλών αιτήσεων πρόνοιας από το τμήμα κοινωνικής ευημερίας της Ουάσιγκτον. Σε άλλους τομείς, το σύστημα ADVANCED SCOUT της IBM είναι ένα εξειδικευμένο σύστημα εξόρυξης γνώσης που βοηθά τους προπονητές του NBA να οργανώνουν και να ερμηνεύουν στοιχεία και αποτελέσματα αγώνων. Χρησιμοποιήθηκε από αρκετές ομάδες του NBA το 1996, συμπεριλαμβανομένων και των Seattle Supersonics οι οποίοι έφτασαν στον τελικό.

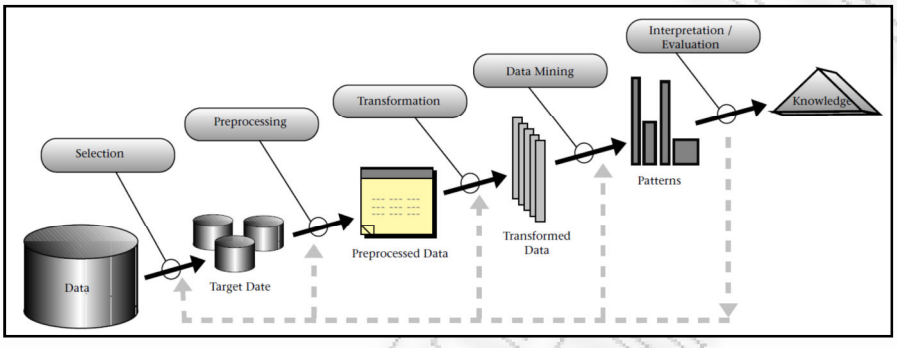
- *Παγκόσμιος ιστός:* ο τομέας της εξόρυξης γνώσης είχε άμεση εφαρμογή με επιτυχία και στο Διαδίκτυο. Το πιο δημοφιλές παράδειγμα εξόρυξης γνώσης στο διαδίκτυο είναι η Google. Για

να γίνει πιο κατανοητή η σημαντικότητα της συνεισφοράς αυτής θα πρέπει να αντιληφθούμε πως ο όγκος της πληροφορίας που υπάρχει μέχρι τώρα στο διαδίκτυο είναι αδύνατο να μετρηθεί με ακρίβεια.

2.2 Διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων

Η ανακάλυψη γνώσης σε βάσεις δεδομένων είναι μια σύνθετη διαδικασία για τον προσδιορισμό έγκυρων, νέων, χρήσιμων και κατανοητών προτύπων σε δεδομένα. Χρησιμοποιεί τεχνικές από πολλούς τομείς όπως στατιστική, μηχανική μάθηση, βάσεις δεδομένων, αναγνώριση προτύπων, πράκτορες, επεξεργασία φυσικής γλώσσας, κτλ. Η διαδικασία ανακάλυψης γνώσης είναι μια ολοκληρωμένη διαδικασία που περιλαμβάνει την επεξεργασία των δεδομένων, την εφαρμογή των αλγορίθμων ανακάλυψης γνώσης και τέλος την ερμηνεία των αποτελεσμάτων.

Τα στάδια που αποτελούν τη διαδικασία ανακάλυψης γνώσης απεικονίζονται στην εικόνα 2.1 και αναλύονται στη συνέχεια [2] & [3]:



Εικόνα 2-1 Στάδια διαδικασίας ανακάλυψης γνώσεις σε βάσεις δεδομένων

- *Κατανόηση του τομέα της εφαρμογής*: της σχετικά προγενέστερης γνώσης και προσδιορισμός του στόχου του τελικού χρήστη.
- *Επιλογή του συνόλου δεδομένων*: δημιουργείται το σύνολο δεδομένων (μεταβλητές, δείγματα δεδομένων) στο οποίο θα εφαρμοστούν οι αλγόριθμοι ανακάλυψης.
- *Καθορισμός και προ-επεξεργασία δεδομένων*: στο στάδιο αυτό αντιμετωπίζονται περιπτώσεις ελλιπών δεδομένων, αφαίρεση του θορύβου, συλλογή των απαραίτητων πληροφοριών για τη διαμόρφωση ή τη μέτρηση του θορύβου κτλ.
- *Μετασχηματισμός των δεδομένων*: τα δεδομένα μετασχηματίζονται ώστε να διευκολύνουν την ανακάλυψη γνώσης. Τέτοιοι μετασχηματισμοί μπορεί να περιλαμβάνουν τη μείωση του αριθμού των υπό εξέταση χαρακτηριστικών με επιλογή ορισμένων από αυτών, τη διακριτοποίηση δηλαδή τη μετατροπή συνεχόμενων αριθμητικών τιμών σε διακριτές τιμές καθώς και την εύρεση κατάλληλης αντιπροσώπευσης των δεδομένων χωρίς μεταβλητές.
- *Επιλογή στόχων και μεθόδων εξόρυξης δεδομένων*: σε αυτό το στάδιο αποφασίζουμε τους στόχους εξόρυξης γνώσης που θέλουμε να επιτύχουμε. Επίσης, επιλέγονται οι μέθοδοι που θα χρησιμοποιηθούν. Αυτό περιλαμβάνει την επιλογή του κατάλληλου αλγορίθμου.
- *Εξόρυξη γνώσης*: στο στάδιο αυτό εφαρμόζονται οι επιλεγμένοι αλγόριθμοι αναζητώντας ενδιαφέροντα πρότυπα γνώσης. Τα πρότυπα θα μπορούσαν να είναι μιας συγκεκριμένης

αντιπροσωπευτικής μορφής ή ενός συνόλου όπως κανόνες συσχέτισης, δένδρα απόφασης, παλινδρόμηση, συσταδοποίηση κλπ.

- *Αξιολόγηση των προτύπων*: τα πρότυπα που προκύπτουν αξιολογούνται βάσει κάποιων κανόνων ή μέτρων, προκειμένου να προσδιοριστούν τα πρότυπα τα οποία αντιπροσωπεύουν τη γνώση.
- *Ερμηνεία και παρουσίαση της γνώσης*: στο τελικό στάδιο γίνεται η ερμηνεία και η παρουσίαση της γνώσης που προκύπτει στον τελικό χρήστη. Η γνώση αυτή μπορεί να χρησιμοποιηθεί επίσης σε ένα σύστημα γνώσης, περίπτωση κατά την οποία ίσως να απαιτείται και η επίλυση πιθανών συγκρούσεων με προγενέστερη γνώση.

2.3 Κατηγορίες αλγορίθμων εξόρυξης δεδομένων

Οι βασικές κατηγορίες αλγορίθμων εξόρυξης γνώσης είναι: η κατηγοριοποίηση, η συσταδοποίηση, οι κανόνες συσχέτισης, τα πρότυπα ακολουθιών, η παλινδρόμηση και τα δέντρα απόφασης. Οι κατηγορίες αυτές αναπαριστούν σχεδόν όλη την περιοχή των αλγορίθμων που χρησιμοποιούνται στον τομέα αυτό και αναλύονται στη συνέχεια [1] & [4]:

- *Κατηγοριοποίηση:* η κατηγοριοποίηση (classification) αποτελεί μια από τις βασικές κατηγορίες στην εξόρυξη γνώσης. Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από τον προκαθορισμένο ορισμό κατηγοριών και από το σύνολο που θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου το οποίο αποτελείται από προ-κατηγοριοποιημένα παραδείγματα. Ο στόχος είναι η δημιουργία ενός μοντέλου το οποίο θα είναι σε θέση να αναθέσει σε κατηγορίες δεδομένα τα οποία δεν έχουν ακόμα κατηγοριοποιηθεί βάσει επιλεγμένων κατηγοριών.
- *Συσταδοποίηση:* η συσταδοποίηση (clustering) είναι η εργασία του καταμερισμού ενός πληθυσμού σε ένα σύνολο περισσότερων διαφορετικών συστάδων (clusters). Η βασική διαφορά μεταξύ συσταδοποίησης και κατηγοριοποίησης είναι ότι η συσταδοποίηση δε βασίζεται σε προκαθορισμένες κατηγορίες. Οι τρεις βασικές κατηγορίες μεθόδων της συσταδοποίησης είναι:
    i. *Μέθοδοι διαχωρισμού (partitioning methods)*: δημιουργούν k ομάδες από ένα δεδομένο αρχικό σύνολο n αντικειμένων με κάθε ομάδα να αντιπροσωπεύει μια συστάδα και να ικανοποιούνται οι εξής δύο συνθήκες: (α) κάθε συστάδα περιέχει τουλάχιστον ένα αντικείμενο και (β) κάθε αντικείμενο ανήκει σε μια μόνο συστάδα.
    ii. *Ιεραρχικές μέθοδοι (hierarchical methods)*: το αρχικό σύνολο δεδομένων διασπάται δημιουργώντας μια ιεραρχική δομή από συστάδες και διακρίνονται σε συζευκτικοί (agglomerative bottom-up) ή διαιρετικοί (divisive top-down) ανάλογα με τον τρόπο που γίνεται η διάσπαση.
    iii. *Μέθοδοι βασισμένες σε μοντέλα (model-based methods)*: υποθέτουν ότι κάθε μια από τις συστάδες περιγράφεται από ένα μαθηματικό μοντέλο και εντοπίζουν τα αντικείμενα που ανήκουν σε κάθε συστάδα, με στόχο να ικανοποιείται το αντίστοιχο μοντέλο.
- *Κανόνες συσχέτισης:* η εξαγωγή κανόνων συσχέτισης (association rules) θεωρείται μια από τις σημαντικότερες διεργασίες εξόρυξης δεδομένων. Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Μια συσχέτιση έχει συνήθως τη μορφή ενός κανόνα συσχέτισης (association rule) $A \rightarrow B$, ο οποίος υποδηλώνει μια σχέση εξάρτισης ανάμεσα στα δύο (ξένα) σύνολα στοιχείων (itemsets) A και B. Ο πλέον δημοφιλής αλγόριθμος για την ανακάλυψη κανόνων συσχέτισης είναι ο Apriori.
- *Δέντρα απόφασης:* τα δέντρα απόφασης (decision trees) χρησιμοποιούνται για να προβλέψουν, με κάποιο βαθμό ακρίβειας, την τιμή της μεταβλητής που μοντελοποιούν με βάση τις τιμές των

ανεξάρτητων μεταβλητών (χαρακτηριστικών). Ένα δέντρο απόφασης στο σύνολο των εγγραφών είναι ένα δέντρο όπου σε κάθε κόμβο του (που δεν είναι φύλλο) υπάρχει ένα ερώτημα που αναφέρεται στα χαρακτηριστικά των εγγραφών και κάθε ερώτημα καταλήγει σε ένα συγκεκριμένο παιδί ενός κόμβου. Τα φύλλα του δηλώνουν τις κλάσεις. Έτσι ένα δέντρο απόφασης εκτελεί κατηγοριοποίηση χρησιμοποιώντας ερωτήματα σχετικά με τα χαρακτηριστικά των εγγραφών.

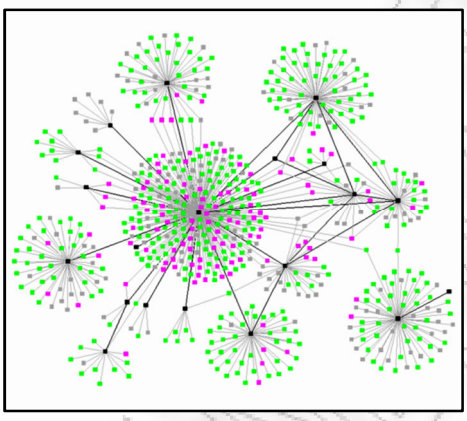
- *Πρότυπα ακολουθιών:* η εξόρυξη πρότυπων ακολουθιών (sequential patterns) είναι η εξόρυξη των συχνά εμφανιζόμενων προτύπων τα οποία σχετίζονται με το χρόνο ή άλλες ακολουθίες. Παραδείγματα προτύπων ακολουθιών έχουμε στην καθημερινή μας ζωή όπως τα κείμενα, οι μουσικές νότες, τα δεδομένα του καιρού και οι ακολουθίες του DNA.
- *Παλινδρόμηση:* η ανάλυση παλινδρόμησης χρησιμοποιείται για τη μοντελοποίηση και την ανάλυση αριθμητικών δεδομένων, μιας εξαρτημένης μεταβλητής και κάποιων ανεξάρτητων μεταβλητών. Κύριος σκοπός στην εξόρυξη γνώσης είναι η πρόβλεψη της τιμής μιας μεταβλητής μελετώντας τις τιμές που είχε στο παρελθόν.

3 Κοινωνικά δίκτυα

Το Διαδίκτυο έχει προκαλέσει τη γέννηση διάφορων τύπων συστημάτων επιμερισμού πληροφοριών, συμπεριλαμβανομένου και του Ιστού. Πρόσφατα, κοινωνικά δίκτυα που βρίσκονται στον Ιστό έχουν κερδίσει σπουδαία δημοτικότητα και αυτή την στιγμή βρίσκονται ανάμεσα στις πιο δημοφιλείς ιστοσελίδες του παγκόσμιου ιστού. Για παράδειγμα, το MySpace (πάνω από 125 εκατομμύρια χρήστες), το Facebook (πάνω από 800 εκατομμύρια χρήστες), το Orkut (πάνω από 100 εκατομμύρια χρήστες), το LinkedIn (πάνω από 110 εκατομμύρια) και το LiveJournal (πάνω από 5,5 εκατομμύρια) είναι δημοφιλείς ιστοσελίδες που βασίζονται στα κοινωνικά δίκτυα. Αντίθετα από τον παγκόσμιο ιστό ο οποίος οργανώνεται σε μεγάλο βαθμό γύρω από το περιεχόμενο, τα κοινωνικά δίκτυα οργανώνονται γύρω από τους χρήστες. Οι συμμετέχοντες χρήστες συνδέονται σε ένα δίκτυο, δημοσιοποιούν το προφίλ τους και οποιοδήποτε περιεχόμενο και δημιουργούν συνδέσμους με οποιουσδήποτε άλλους χρήστες θέλουν να συναναστραφούν. Η μελέτη των αποτελεσμάτων των κοινωνικών δικτύων παρέχει μία βάση για τη διατήρηση των κοινωνικών σχέσεων, για την εύρεση χρηστών με παρόμοια ενδιαφέροντα και για τον εντοπισμό περιεχομένων και γνώσης την οποία έχουν συνεισφέρει ή εγκρίνει άλλοι χρήστες.

3.1 Ορισμοί

Ένα *κοινωνικό δίκτυο* (social network) σύμφωνα με το [5] είναι μία κοινωνική δομή η οποία περιλαμβάνει κόμβους (που γενικά είναι φυσικά πρόσωπα ή οργανισμοί) οι οποίοι συνδέονται με μια ή περισσότερες σχέσεις, όπως αξίες, οράματα, στόχοι, ιδέες, οικονομικές συναλλαγές, εμπορικές συναλλαγές, φιλία. Στην απλούστερη του μορφή, η αποτύπωση ενός κοινωνικού δικτύου, περιλαμβάνει όλους τους κόμβους και τις σχέσεις μεταξύ τους. Ο ορισμός αυτός αποτυπώνεται σε διαγράμματα όπου οι κόμβοι είναι σημεία και οι σχέσεις μεταξύ των κόμβων είναι οι γραμμές που συνδέουν τα σημεία. Ένα τέτοιο διάγραμμα εμφανίζεται στην εικόνα 3.1.



Εικόνα 3-1 Απεικόνιση ενός κοινωνικού δικτύου

Τα κοινωνικά δίκτυα βασίζονται στην άποψη ότι υπάρχει μία καθορισμένη δομή για τον τρόπο με τον οποίο ένα άτομο συνδέεται με ένα άλλο, είτε άμεσα είτε έμμεσα, η οποία έχει γίνει γνωστή ως *"six degrees of separation"*. Ο όρος κοινωνικό δίκτυο, χρησιμοποιήθηκε πρώτη φορά από τον J. A. Barnes το 1951 και μπορεί να ερμηνευτεί ως η κοινωνική δομή μεταξύ φορέων, κυρίως μεμονωμένων ατόμων, οργανισμών αλλά και ολόκληρων κοινωνιών. Ένα κοινωνικό δίκτυο μπορεί να θεωρηθεί ως: *ένα συγκεκριμένο σύνολο συνδέσμων ανάμεσα σε ένα ορισμένο σύνολο φορέων, με την επιπρόσθετη ιδιότητα, ότι τα χαρακτηριστικά αυτών των συνδέσμων ως σύνολο μπορεί να χρησιμοποιηθούν για την ερμηνεία της κοινωνικής συμπεριφοράς των εμπλεκόμενων φορέων* [5].

Η *ανάλυση κοινωνικών δικτύων* (SNA - social network analysis) είναι μία τεχνική κλειδί σε πολλά σύγχρονα πεδία έρευνας όπως η κοινωνιολογία, η ανθρωπολογία, η κοινωνική ψυχολογία, οι οργανωτικές μελέτες κτλ. Στο σύγχρονο οικονομικό περιβάλλον, τα κοινωνικά δίκτυα χρησιμοποιούνται για να εξεταστούν οι σχέσεις και οι συνδέσεις μεταξύ μεμονωμένων υπαλλήλων σε διαφορετικές εταιρείες, καθώς επίσης και ο τρόπος με τον οποίο οι εταιρείες αλληλεπιδρούν μεταξύ τους.

Η ανάλυση κοινωνικών δικτύων βοηθά τους ερευνητές να κατανοήσουν τον τρόπο με τον οποίο επικοινωνούν και συνεργάζονται οι άνθρωποι και να αναγνωρίσουν τη ροή της γνώσης σε επίπεδο τόσο μεταξύ διάφορων οργανισμών όσο και μέσα σε κάθε οργανισμό. Αρκετές επιχειρήσεις προσφέρουν υπηρεσίες βασισμένες στην ανάλυση κοινωνικών δικτύων, υποσχόμενες βελτιστοποίηση της πληροφοριακής ροής ως έναν τρόπο για τη βελτίωση της αποτελεσματικότητας, τη μείωση του κόστους και την αύξηση της παραγωγικότητας.

Τα κοινωνικά δίκτυα εξυπηρετούν αρκετούς σκοπούς αλλά τρεις ξεχωρίζουν ως κοινοί σε όλες τις ιστοσελίδες [5]:

i. η διατήρηση και η ενδυνάμωση των ήδη υπαρχόντων κοινωνικών δεσμών ή η δημιουργία νέων κοινωνικών δεσμών

ii. η δυνατότητα κοινοποίησης προσωπικού περιεχομένου κάθε χρήστη

iii. η εύρεση νέων περιεχομένων μέσω της ανακάλυψης, της σύστασης και της οργάνωσης περιεχομένων που έχουν ήδη κοινοποιηθεί από τους χρήστες.

3.2 Ιστορική αναδρομή

Στη συνέχεια, παρουσιάζεται μια σύντομη ιστορική αναδρομή των κοινωνικών δικτύων σύμφωνα με το [6]. Η ιστοσελίδα Classmates.com, θεωρείται ως η πρώτη ιστοσελίδα που επέτρεψε στους χρήστες να συνδέονται με άλλους χρήστες. Ξεκίνησε το 1995 ως μία ιστοσελίδα η οποία έδινε τη δυνατότητα επανασύνδεσης σε χρήστες με παλαιούς συμμαθητές και σήμερα μετρά πάνω από 40 εκατομμύρια εγγεγραμμένους χρήστες. Ωστόσο, η ιστοσελίδα Classmates.com δεν επέτρεπε στους χρήστες να δημιουργούν γενικότερα συνδέσμους με άλλους χρήστες, αλλά επέτρεπε στους χρήστες να συνδέονται με άλλους χρήστες μόνο μέσω των σχολείων στα οποία είχαν παρακολουθήσει. Το 1997, δημιουργήθηκε η ιστοσελίδα SixDegrees.com, η οποία ήταν η πρώτη ιστοσελίδα κοινωνικής δικτύωσης, η οποία επέτρεπε στους χρήστες της να συνδέονται απευθείας με άλλους χρήστες. Ως εκ τούτου, η SixDegrees.com είναι η πρώτη ιστοσελίδα που εκφράζει τον ορισμό του online κοινωνικού δικτύου.

Τα online κοινωνικά δίκτυα, ξεκίνησαν να αναπτύσσονται και να γίνονται δημοφιλή καθώς όλο και περισσότεροι χρήστες συνδέονταν στο Διαδίκτυο. Στις αρχές της δεκαετίας του 2000, καθιερώθηκε ένα σύνολο ιστοσελίδων γενικού σκοπού για την εύρεση φίλων, η πιο αξιοσημείωτη από τις οποίες είναι η Friendster. Άλλες παρόμοιες ιστοσελίδες δημιουργήθηκαν στο ίδιο χρονικό διάστημα, συμπεριλαμβανομένων των CyWorld, Ryze, και LinkedIn. Το 2003 δημιουργήθηκε η ιστοσελίδα MySpace ως εναλλακτική της Friendster. Η MySpace επέτρεπε στους χρήστες της να τροποποιούν την εμφάνιση του προφίλ τους, γεγονός το οποίο αποδείχθηκε πολύ δημοφιλές στους χρήστες, με αποτέλεσμα το MySpace να γίνει γρήγορα το μεγαλύτερο online κοινωνικό δίκτυο. Μέχρι αυτή τη στιγμή το MySpace έχει 125 εκατομμύρια λογαριασμούς.

Με την άνοδο της δημοτικότητας των κοινωνικών δικτύων, πολλές ιστοσελίδες διαφορετικού είδους άρχισαν να ενσωματώνουν χαρακτηριστικά κοινωνικής δικτύωσης. Τα παραδείγματα περιλαμβάνουν τις ιστοσελίδες δημοσίευσης περιεχομένου πολυμέσων (Flickr, YouTube και Zoomr), τις ιστοσελίδες ιστολογιών (Live-Journal και BlogSpot), τις ιστοσελίδες επαγγελματικής δικτύωσης (LinkedIn and Ryze) και τις ιστοσελίδες συλλογής ειδήσεων (Digg, Reddit, και del.icio.us). Όλες αυτές οι ιστοσελίδες έχουν διαφορετικούς στόχους αλλά ακολουθούν την ίδια στρατηγική εκμετάλλευσης των κοινωνικών δικτύων

για τη βελτίωση τους. Η παραπάνω λίστα δεν είναι εφικτό να είναι εξαντλητική καθώς συνεχώς καινούργιες σελίδες δημιουργούνται. Μια πιο ολοκληρωμένη και ενημερωμένη λίστα των αξιοσημείωτων ιστοσελίδων κοινωνικών δικτύων μπορούμε να βρούμε στην Wikipedia (www.wikipedia.com).

3.3 Χαρακτηριστικά κοινωνικών δικτύων

Τα κύρια χαρακτηριστικά των κοινωνικών δικτύων σχετίζονται με τη φύση των σχέσεων που δημιουργούνται μέσα σε αυτά, το σχήμα του δικτύου, τους μηχανισμούς και τις δυνάμεις που επηρεάζουν την ανάπτυξη τους [7]:

- Οι σχέσεις μεταξύ των φορέων του δικτύου βασίζονται στην οργάνωση του προσωπικού ή σε τεχνικό-οργανωτικές διασυνδέσεις μακροπρόθεσμης βάσης και μπορεί να θεωρηθούν ότι απορρέουν από την αυτονομία και την ανεξαρτησία τους, τη συνύπαρξη της συνεργασίας και του ανταγωνισμού καθώς επίσης την αμοιβαιότητα και τη σταθερότητα. Οι σχέσεις που αναπτύσσονται μεταξύ των φορέων του κοινωνικού δικτύου μπορούν να κατηγοριοποιηθούν σύμφωνα με τα περιεχόμενα (προϊόντα ή υπηρεσίες, πληροφορίες και συναισθήματα), το είδος (διάρκεια και εγγύτητα της σχέσης) και την ένταση (συχνότητα επικοινωνίας). Το είδος και η ένταση των σχέσεων εδραιώνει τη δικτυακή δομή.
- Το σχήμα του κοινωνικού δικτύου είναι ένας καθοριστικός παράγοντας για τη χρησιμότητα του δικτύου. Τα κοινωνικά δίκτυα που έχουν πολλές συνδέσεις με άτομα εκτός του κύριου δικτύου είναι πιο χρήσιμα για τα μέλη τους από ότι τα μικρότερα και «αυστηρότερα» δίκτυα, επειδή οι συμμετέχοντες των «ανοιχτών» δικτύων έχουν πρόσβαση σε μία ευρεία γκάμα πληροφοριών και ως εκ τούτου είναι πιο πιθανό να έχουν πρόσβαση σε καινούργιες ιδέες και ευκαιρίες.
- Τα κοινωνικά δίκτυα μπορούν να έχουν αποτέλεσμα μέσω της εσωτερίκευσης ή της εξωτερίκευσης. Εσωτερίκευση σημαίνει μία εντατικοποίηση της συνεργασίας, ενώ εξωτερίκευση είναι μία περιορισμένη λειτουργική ανάθεση σε τρίτους, η οποία επιτυγχάνεται με τη χαλάρωση ιεραρχικών συντονιστικών μηχανισμών. Τόσο η εσωτερίκευση όσο και η εξωτερίκευση μπορεί να λαμβάνουν χώρα ταυτόχρονα στην ίδια εταιρεία.

3.4 Μηχανισμοί και πολιτικές

Στη συνέχεια, παρουσιάζεται μια σύνοψη των μηχανισμών και των πολιτικών που παρέχουν τα κοινωνικά δίκτυα [5] & [6]:

- *Χρήστες:* για να συμμετέχουν πλήρως σε ένα κοινωνικό δίκτυο, οι χρήστες πρέπει να εγγραφούν σε μια ιστοσελίδα πιθανώς με κάποιο ψευδώνυμο. Κάποιες ιστοσελίδες επιτρέπουν την περιήγηση σε δημόσια δεδομένα χωρίς απαραίτητη εγγραφή. Οι χρήστες μπορεί να προσφέρουν εθελοντικά πληροφορίες για τους εαυτούς τους (π.χ. γενέθλια, τόπο κατοικίας, ενδιαφέροντα), οι οποίες προστίθενται στο προφίλ τους.
- *Σύνδεσμοι:* ένα κοινωνικό δίκτυο αποτελείται από λογαριασμούς χρηστών και συνδέσμους ανάμεσα στους χρήστες. Κάποιες ιστοσελίδες (όπως το Flickr, το LiveJournal) επιτρέπουν στο χρήστη να συνδεθεί με οποιονδήποτε άλλο χρήστη χωρίς να απαιτείται η αποδοχή από τον σύνδεσμο που αποτελεί στόχο. Άλλες ιστοσελίδες (π.χ. Orkut, LinkedIn) απαιτούν τη συγκατάθεση και του δημιουργού αλλά και του στόχου για να δημιουργηθεί ένας σύνδεσμος. Οι χρήστες διαμορφώνουν συνδέσμους για διάφορους λόγους. Οι κόμβοι που συνδέονται από ένα σύνδεσμο μπορεί να αποτελούν πραγματικές γνωριμίες, γνωριμίες μέσω του ιστού ή

επαγγελματικές επαφές, μπορεί να μοιράζονται κάποιο κοινό ενδιαφέρον ή μπορεί να ενδιαφέρονται για κάποιο κοινό περιεχόμενο.

- *Ομάδες:* οι περισσότερες ιστοσελίδες δίνουν στους χρήστες τη δυνατότητα να σχηματίσουν και να γίνουν μέλη σε ομάδες ειδικού ενδιαφέροντος. Οι χρήστες μπορούν να αναρτούν μηνύματα και να κοινοποιούν περιεχόμενα στην ομάδα. Η κύρια χρήση των ομάδων στα κοινωνικά δίκτυα είναι είτε για να εφαρμόσουν πολιτικές ελέγχου πρόσβασης ή για να παρέχουν μία πλατφόρμα (forum) για κοινόχρηστα περιεχόμενα.
- *Περιεχόμενο:* με τη δημιουργία ενός νέου λογαριασμού δίνεται η δυνατότητα κοινοποίησης περιεχομένου. Πολλές ιστοσελίδες δίνουν τη δυνατότητα στο χρήστη να χαρακτηρίσει το περιεχόμενο ως δημόσιο (ορατό από τον καθένα) ή ιδιωτικό (ορατό μόνο στους άμεσους «φίλους» του). Το περιεχόμενο που έχει κοινοποιηθεί από έναν συγκεκριμένο χρήστη, απεικονίζεται στο προφίλ του, επιτρέποντας σε άλλους χρήστες οι οποίοι περιηγούνται στο κοινωνικό δίκτυο να το ανακαλύψουν. Τυπικά, το περιεχόμενο ενσωματώνεται αυτόματα σε ευρετήρια με αποτέλεσμα εφόσον είναι δημόσια διαθέσιμο να γίνεται προσβάσιμο μέσω μιας μηχανής αναζήτησης. Ένα παράδειγμα είναι η αναζήτηση φωτογραφιών του Flickr, η οποία επιτρέπει στους χρήστες να εντοπίζουν φωτογραφίες μέσω αναζήτησης που βασίζεται σε επισημάνσεις και σχόλια.

3.5 Κίνητρα έρευνας

Τα online κοινωνικά δίκτυα ανήκουν ήδη σε μερικές από τις δημοφιλέστερες ιστοσελίδες. Καθώς η τεχνολογία ωριμάζει, είναι πιο πιθανό να βγουν στην επιφάνεια περισσότερες εφαρμογές. Είναι επίσης πολύ πιθανόν ότι η κοινωνική δικτύωση θα παίξει σημαντικό ρόλο στη μελλοντική προσωπική και εμπορική διαδικτυακή αλληλεπίδραση καθώς επίσης και στον εντοπισμό και την οργάνωση της πληροφορίας και της γνώσης. Σύμφωνα με τις έρευνες [6] και [8] τα βασικά κίνητρα έρευνας των κοινωνικών δικτύων παρουσιάζονται παρακάτω:

- *Κοινά ενδιαφέροντα και εμπιστοσύνη:* γειτονικοί χρήστες σε ένα κοινωνικό δίκτυο έχουν την τάση να εμπιστεύονται ο ένας τον άλλον και αρκετές φορές να μοιράζονται κοινά ενδιαφέροντα. Οι χρήστες περιηγούνται σε γειτονικές περιοχές στο δικό τους κοινωνικό δίκτυο επειδή είναι πολύ πιθανό να βρουν ένα περιεχόμενο του δικού τους ενδιαφέροντος. Διάφορα συστήματα χρησιμοποιούν τα κοινωνικά δίκτυα για να κατατάξουν τα αποτελέσματα αναζήτησης στο διαδίκτυο σχετικά με τα ενδιαφέροντα του γείτονα ενός χρήστη στο κοινωνικό δίκτυο.
- *Μέσα επικοινωνίας:* οι ιστοσελίδες κοινωνικών δικτύων είναι ένα καινούργιο εργαλείο επικοινωνίας και έχουν την ικανότητα να παρέχουν ασύγχρονη επικοινωνία μέσω των διαδικτυακών μηνυμάτων, την κοινοποίηση δημόσιων ανακοινώσεων, κτλ. Αντίθετα με τα εργαλεία επικοινωνίας όπως είναι το ηλεκτρονικό ταχυδρομείο και τα άμεσα μηνύματα, οι χρήστες των κοινωνικών δικτύων χρειάζεται να αναρτήσουν την πληροφορία τους μόνο μία φορά και οι φίλοι τους θα ειδοποιηθούν άμεσα. Οι χρήστες μπορούν να δηλώνουν το ενδιαφέρον τους για ανθρώπους και γεγονότα έτσι ώστε να λαμβάνουν τις τελευταίες ενημερώσεις.
- *Επιπτώσεις στο μελλοντικό Διαδίκτυο:* εάν τα μελλοντικά κατανεμημένα κοινωνικά δίκτυα είναι δημοφιλή και με τάσεις μεγάλου εύρους ζώνης, μπορεί να έχουν σημαντικό αντίκτυπο στην κυκλοφορία του Διαδικτύου, όπως ακριβώς κάνουν και τα τρέχοντα δίκτυα διανομής περιεχομένου peer-to-peer. Η κατανόηση της δομής των κοινωνικών δικτύων δεν είναι κρίσιμη μόνο στην κατανόηση της δύναμης και της ασφάλειας των κατανεμημένων κοινωνικών δικτύων αλλά επίσης στην κατανόηση των επιπτώσεών τους στο Διαδίκτυο του μέλλοντος.

- *Κοινωνικό κεφάλαιο:* οι ιστότοποι κοινωνικής δικτύωσης είναι ηλεκτρονικές «τράπεζες» του κοινωνικού κεφαλαίου στον πραγματικό κόσμο. Οι χρήστες μπορούν να «διασχίσουν» τις συνδέσεις των φίλων τους και να επιτύχουν καινούργιες επαφές μέσω κοινών συνδέσεων. Η μέθοδος αυτή μειώνει δραματικά το χρόνο και την προσπάθεια που απαιτείται για τη δημιουργία σχέσεων εμπιστοσύνης. Βοηθά τους χρήστες να χτίσουν και να διατηρήσουν το κοινωνικό τους κεφάλαιο, το οποίο έχει ισχυρή επίδραση στον τομέα της εργασίας, των οικονομικών, της οργανωτικής συμπεριφοράς, της πολιτικής επιστήμης, της δημόσιας υγείας και της κοινωνιολογίας.
- *Κοινωνικός ιστός:* από την άλλη μεριά, ένα αξιόπιστο κοινωνικό δίκτυο διευκολύνει τη διάδοση της γνώσης και της πληροφορίας, χάρη στη δύναμη της διάδοσης «από στόμα σε στόμα». Πολυάριθμες μελέτες έχουν δείξει ότι ένα από τα πιο αποτελεσματικά κανάλια διάδοσης της πληροφορίας και της γνώσης μέσα σε έναν οργανισμό είναι το ανεπίσημο δίκτυο των συνεργατών, συναδέλφων και φίλων. Τα κοινωνικά δίκτυα μπορούν να εξαπλώσουν/διαδώσουν πληροφορίες πολύ πιο γρήγορα από την υποδομή του παγκόσμιου ιστού, όπου μία ιστοσελίδα μπορεί να βρεθεί μόνο εάν διατηρεί υψηλή κατάταξη και επισκεψιμότητα στις μηχανές αναζήτησης.
- *Κοινωνική νοημοσύνη:* οι σελίδες κοινωνικής δικτύωσης μπορεί να θεωρηθούν μεγάλης κλίμακας συστήματα αλληλεπίδρασης χρηστών. Η ικανότητα κατανόησης των ανθρώπινων όντων και η σοφή δράση στις ανθρώπινες σχέσεις ονομάζεται κοινωνική νοημοσύνη και είναι ισάξια με τη διαπροσωπική νοημοσύνη. Η κοινωνική νοημοσύνη είναι ζωτικής σημασίας στην ανάπτυξη της ανθρώπινης νοημοσύνης διότι βελτιώνει την εμπιστοσύνη, την επικοινωνία και το συντονισμό μεταξύ των πρακτόρων σε ένα πολυπρακτορικό σύστημα.
- *Επιπτώσεις σε άλλους κλάδους:* στις κοινωνικές επιστήμες τα κοινωνικά δίκτυα προσφέρουν μια πρωτοφανή ευκαιρία για μελέτη των κοινωνικών δικτύων σε μεγάλη κλίμακα. Οι πολιτικές εκστρατείες έχουν αντιληφθεί τη σημασίας των ιστολογιών στις εκλογές. Ομοίως, οι ειδικοί του μάρκετινγκ πειραματίζονται με το μάρκετινγκ με τη μέθοδο viral marketing με σκοπό την καλύτερη προώθηση προϊόντων και εταιρειών. Ανεξάρτητα από τη στάση κάποιου σε αυτά τα φαινόμενα, μια καλύτερη κατανόηση της δομής των κοινωνικών δικτύων είναι πολύ πιθανόν να βελτιώσει την κατανόησή για τις ευκαιρίες, τους περιορισμούς και τις απειλές που σχετίζονται με αυτές τις ιδέες.

3.6 Θέματα έρευνας

Στο πεδίο των κοινωνικών δικτύων τα τελευταία χρόνια παρατηρείται ένα αυξανόμενο ενδιαφέρον. Μια βασική πτυχή των κοινωνικών δικτύων είναι ότι είναι *πλούσια σε δεδομένα* και παρέχουν πρωτοφανείς *προκλήσεις και ευκαιρίες* σε ότι αφορά την ανακάλυψη γνώσης και την εξόρυξη γνώσης. Υπάρχουν δύο κύρια είδη δεδομένων, τα οποία συχνά αναλύονται στο πλαίσιο των κοινωνικών δικτύων [9]:

- **Δομική ανάλυση η οποία βασίζεται στους συνδέσμους** (Linkage-based and Structural Analysis): σε αυτό το είδος κατασκευάζουμε μία ανάλυση της συμπεριφοράς συνδέσμων του δικτύου, με σκοπό τον καθορισμό των σημαντικών κόμβων, των κοινοτήτων, των συνδέσμων και των εξελισσόμενων περιοχών του.
- **Επιπρόσθετη ανάλυση η οποία βασίζεται στο περιεχόμενο** (Adding Content-based Analysis): πολλά κοινωνικά δίκτυα όπως το Flickr, το Youtube περιέχουν μία τεράστια ποσότητα περιεχομένου η οποία μπορεί να χρησιμοποιηθεί ως μοχλός με στόχο τη βελτίωση της ποιότητας της ανάλυσης.

Έχει παρατηρηθεί ότι ο συνδυασμός της ανάλυσης που βασίζεται στο περιεχόμενο και της ανάλυσης που βασίζεται στους συνδέσμους παρέχει πιο αποδοτικά αποτελέσματα σε μία μεγάλη ποικιλία εφαρμογών. Στη συνέχεια, παρουσιάζονται οι βασικές περιοχές που συγκεντρώνουν το μεγαλύτερο ενδιαφέρον στο πλαίσιο της έρευνας των κοινωνικών δικτύων σύμφωνα με το [9] οι οποίες αναλύονται στο επόμενο κεφάλαιο μέσω των σχετικών αλγορίθμων:

- *Ανίχνευση κοινοτήτων:* ένα από τα πιο σημαντικά προβλήματα στο πλαίσιο της ανάλυσης των κοινωνικών δικτύων, είναι αυτό της ανίχνευσης κοινοτήτων. Αυτό το πρόβλημα είναι στενά συνδεδεμένο με αυτό της συσταδοποίησης και επιχειρεί να καθορίσει περιοχές του δικτύου, στις οποίες εμφανίζεται έντονη συμπεριφορά σύνδεσης. Τα θέμα αυτό σχετίζεται με το γενικότερο πρόβλημα του διαμερισμού γραφήματος, το οποίο διαχωρίζει το δίκτυο σε πυκνές περιοχές βασισμένο στη συμπεριφορά της σύνδεσης.

- *Κατηγοριοποίηση κόμβων:* σε πολλές εφαρμογές, σε κάποιους από τους κόμβους του κοινωνικού δικτύου μπορεί να τοποθετηθεί ετικέτα και μπορεί να είναι επιθυμητή η χρήση χαρακτηριστικών και δομικών πληροφοριών στα κοινωνικά δίκτυα με σκοπό τη διάδοση αυτών των ετικετών. Τα κοινωνικά δίκτυα περιέχουν πλούσιες πληροφορίες για το περιεχόμενο και τη δομή του δικτύου, οι οποίες μπορεί να χρησιμοποιηθούν για αυτό τον σκοπό.

- *Ανάλυση κοινωνικής επιρροής:* από τη στιγμή που τα κοινωνικά δίκτυα σχεδιάστηκαν κυρίως πάνω στις αλληλεπιδράσεις των διαφορετικών συμμετεχόντων, είναι φυσικό ότι τέτοιες αλληλεπιδράσεις μπορεί να οδηγήσουν στην επιρροή των φορέων σε ότι αφορά τη συμπεριφορά. Ένα παράδειγμα είναι το viral marketing στο οποίο χρησιμοποιούνται μηνύματα μεταξύ συνδεδεμένων συμμετεχόντων με σκοπό τη διάδοση της πληροφορίας ανάμεσα στα διαφορετικά μέλη του δικτύου.

- *Εντοπισμός ειδικών:* τα κοινωνικά δίκτυα μπορούν να αποτελέσουν ένα εργαλείο για την εύρεση ειδικών για τη διεκπεραίωση μιας συγκεκριμένης εργασίας. Συχνά, οι ειδικοί οργανώνονται σε ομάδες που αντιστοιχούν σε κοινωνικά δίκτυα ή σε οργανωτικές δομές εταιρειών. Πολλές πολύπλοκες εργασίες απαιτούν τη συλλογική εξειδίκευση περισσότερων από έναν ειδικό. Σε αυτές τις περιπτώσεις είναι πιο ρεαλιστικό να απαιτήσουμε μια ομάδα ειδικών που μπορούν να συνεργαστούν για ένα κοινό στόχο.

- *Πρόβλεψη συνδέσμου:* στις περισσότερες εφαρμογές κοινωνικής δικτύωσης, οι σύνδεσμοι είναι δυναμικοί και μπορεί να αλλάξουν με το χρόνο. Για παράδειγμα, σε ένα κοινωνικό δίκτυο οι σύνδεσμοι φιλίας δημιουργούνται συνεχώς με το πέρασμα του χρόνου. Συνεπώς, ένα φυσικό ερώτημα είναι ο καθορισμός ή η πρόβλεψη μελλοντικών συνδέσμων στο κοινωνικό δίκτυο. Η διαδικασία πρόβλεψης μπορεί να χρησιμοποιήσει είτε τη δομή του δικτύου είτε τις χαρακτηριστικές πληροφορίες στους διαφορετικούς κόμβους.

- *Εμπιστοσύνη:* είναι μια σημαντική πτυχή της σχέσης μεταξύ δύο οντοτήτων. Στο πλαίσιο ενός κοινωνικού δικτύου το ερώτημα «ποιος εμπιστεύεται ποιον» διαδραματίζει σημαντικό ρόλο στον τομέα πληροφοριών και ασφαλείας. Η εμπιστοσύνη αποτελεί τη βάση για σχηματισμό συνασπισμών (ισχυρές κοινότητες που αποτελούνται από φορείς, οι οποίοι «εμπιστεύονται» ο ένας τον άλλο) γεγονός το οποίο είναι πιθανό να φανεί χρήσιμο στον εντοπισμό επιρροής μεταξύ των κόμβων ενός δικτύου και στον καθορισμό της ροής των πληροφοριών σε ένα κοινωνικό δίκτυο.

3.7 Ανάλυση κοινωνικών δικτύων

Το κύριο χαρακτηριστικό της *ανάλυσης κοινωνικών δικτύων* (social network analysis) είναι η εστίαση στη δομή των σχέσεων, η οποία περιλαμβάνει από απλές γνωριμίες έως στενούς δεσμούς. Η ανάλυση

των κοινωνικών δικτύων υποθέτει ότι οι σχέσεις είναι σημαντικές. Είναι μία μέθοδος με αυξανόμενη εφαρμογή στις κοινωνικές επιστήμες και έχει εφαρμοστεί σε πεδία αρκετά διαφορετικά όπως η ψυχολογία, η υγεία, η επιχειρησιακή οργάνωση και οι ηλεκτρονικές επικοινωνίες. Πιο πρόσφατα, το ενδιαφέρον έχει αυξηθεί για την ανάλυση ηγετικών δικτύων με σκοπό τη διατήρηση και την ενίσχυση των σχέσεων μεταξύ ομάδων, οργανισμών και σχετικών συστημάτων.

Τυπικά, η ανάλυση των κοινωνικών δικτύων βασίζεται σε ερωτηματολόγια και συνεντεύξεις με σκοπό τη συλλογή πληροφοριών για τις σχέσεις μέσα σε μία ορισμένη ομάδα. Οι απαντήσεις που συλλέγονται στη συνέχεια χαρτογραφούνται. Η συλλογή δεδομένων και η διαδικασία της ανάλυσης παρέχουν βασικές πληροφορίες, τις οποίες μπορούμε να θέσουμε κατά προτεραιότητα και να σχεδιάσουμε παρεμβάσεις για να βελτιώσουμε τη ροή της γνώσης η οποία μπορεί να συνεπάγεται αναδιατύπωση των κοινωνικών συνδέσμων. Η ανάλυση ενθαρρύνει τις συναισθηματικά συμμετοχικές και ερμηνευτικές προσεγγίσεις της περιγραφής και ανάλυσης των κοινωνικών δικτύων, εστιάζοντας στα πιο απλά και πιο χρήσιμα βασικά στοιχεία. Τα στάδια κλειδιά της βασικής διαδικασίας σύμφωνα με το [10] απαιτούν από τους εμπλεκόμενους να:

- αναγνωρίσουν το δίκτυο των ατόμων, των ομάδων και των φορέων με σκοπό την ανάλυσή του
- καθορίσουν το σκοπό και να ξεκαθαρίσουν το πεδίο εφαρμογής της ανάλυσης με στόχο τον καθορισμό των αναφορών που απαιτούνται
- διαμορφώσουν υποθέσεις και ερωτήματα
- αναπτύξουν την ερευνητική μεθοδολογία
- σχεδιάσουν τα ερωτηματολόγια, διατηρώντας τις ερωτήσεις σύντομες και ενδιαφέρουσες
- ερευνήσουν τα άτομα, τις ομάδες και τους φορείς στο δίκτυο για να αναγνωρίσουν τις σχέσεις και τη ροή γνώσης ανάμεσά τους
- χρησιμοποιήσουν ένα εργαλείο ανάλυσης κοινωνικών δικτύων έτσι ώστε να δημιουργηθεί ένας οπτικός χάρτης του δικτύου
- επανεξετάσουν τον χάρτη, τα προβλήματα και τις ευκαιρίες που έχουν επισημανθεί χρησιμοποιώντας τις συνεντεύξεις
- σχεδιάσουν τις ενέργειες εφαρμογής έτσι ώστε να επιτευχθούν οι επιθυμητές αλλαγές

4 Αλγόριθμοι εξόρυξης γνώσης σε κοινωνικά δίκτυα

Στο συγκεκριμένο κεφάλαιο ακολουθεί η ανάλυση και η παρουσίαση αλγορίθμων εξόρυξης γνώσης σε κοινωνικά δίκτυα. Οι βασικές περιοχές που καλύπτονται είναι: η ανίχνευση κοινοτήτων, η κατηγοριοποίηση κόμβου, η πρόβλεψη συνδέσμου, η κοινωνική επιρροή, η εμπιστοσύνη και η εύρεση ειδικών. Για την κατανόηση αυτών των αλγορίθμων είναι απαραίτητη η κατανόηση των βασικών ιδιοτήτων των κοινωνικών δικτύων οι οποίες παρουσιάζονται παρακάτω.

4.1 Ιδιότητες κοινωνικών δικτύων

Ένα κοινωνικό δίκτυο με μαθηματικούς όρους είναι απλώς ένα γράφημα $G(V, E)$ όπου οι κόμβοι αναπαριστούν άτομα και μία ακμή $(u, v)\ \epsilon\ E$ υποδεικνύει ένα είδος σχέσης ανάμεσα στα άτομα u και v. Οι σύνδεσμοι σε ένα γράφημα μπορεί να είναι κατευθυνόμενοι (directed) που σημαίνει ότι κάθε σύνδεσμος ξεκινά από έναν κόμβο (πηγή) και καταλήγει σε κάποιον άλλο ή μη-κατευθυνόμενοι (undirected) το οποίο σημαίνει ότι σε κάθε σύνδεσμο μεταξύ δύο κόμβων δεν υπάρχει πηγή και προορισμός.

4.1.1 Βαθμός

Ο βαθμός (degree) ενός κόμβου i, ορίζεται ως d_i και είναι ο αριθμός των συνδέσμων του κόμβου με άλλους κόμβους. Για κατευθυνόμενα δίκτυα (directed networks) γίνεται διαχωρισμός σε indegree (αριθμός των εισερχόμενων συνδέσμων) και outdegree (αριθμός εξερχόμενων συνδέσμων). Επίσης σε κατευθυνόμενα δίκτυα υφίσταται και η έννοια της συμμετρίας (symmetry) ως το κλάσμα των συνδέσμων οι οποίοι έχουν μια αντίστοιχη αντίστροφη σύνδεση [5].

4.1.2 Ακτίνα και διάμετρος

Η ακτίνα (radius) και η διάμετρος (diameter) ενός γραφήματος εκφράζουν σε γενικές γραμμές πόσο μακριά είναι οι κόμβοι στο δίκτυο. Η ακτίνα αντιπροσωπεύει τη μέγιστη απόσταση από τον πιο «κεντρικό» κόμβο του γραφήματος για όλους τους άλλους κόμβους, και η διάμετρος αντιπροσωπεύει τη μέγιστη απόσταση από το λιγότερο «κεντρικό» κόμβου του γραφήματος για όλους τους άλλους κόμβους [5].

4.1.3 Κατανομή βαθμού

Η κατανομή βαθμού (degree distribution) ενός γραφήματος είναι η συνάρτηση $P(k)$ η οποία περιγράφει το κλάσμα των κόμβων του δικτύου οι οποίοι έχουν βαθμό k. Η κατανομή βαθμού περιγράφει ουσιαστικά τον τρόπο με τον οποίο οι σύνδεσμοι του γραφήματος κατανέμονται μεταξύ των κόμβων. Για παράδειγμα, η κατανομή βαθμού ενός γραφήματος με τυχαίες ακμές μεταξύ n κόμβων δίνεται με βάση τη διωνυμική κατανομή ως εξής [5]:

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

όπου το p αναπαριστά την πιθανότητα ότι δύο κόμβοι είναι συνδεδεμένοι.

4.1.4 Συντελεστής σχετικότητας

Η μετρική η οποία σχετίζεται με το συντελεστή σχετικότητας (assortativity coefficient) r, μετρά την πιθανότητα οι κόμβοι να συνδέονται με άλλους κόμβους με παρόμοιους βαθμούς. Η σχετικότητα ορίζεται ως συντελεστής συσχέτισης Pearson μεταξύ των βαθμών όλων των ζευγών κόμβων οι οποίοι συνδέονται με μια ακμή. Για το λόγο αυτό, ο συντελεστής κυμαίνεται μεταξύ -1 και 1, όπου υψηλός συντελεστής σχετικότητας σημαίνει ότι οι κόμβοι έχουν την τάση να συνδέονται με κόμβους του ίδιου βαθμού, ενώ ένας αρνητικός συντελεστής σημαίνει ότι οι κόμβοι τείνουν να συνδέονται με κόμβους διαφορετικού βαθμού [5].

4.1.5 Συντελεστής συσταδοποίησης

Ο συντελεστής συσταδοποίησης (clustering coefficient) ενός κόμβου i, συμβολίζεται ως $c(i)$ και υποδηλώνει τον αριθμό των κατευθυνόμενων συνδέσμων μεταξύ των γειτόνων του κόμβου διαιρούμενος από τον αριθμό των πιθανών κατευθυνόμενων συνδέσμων που θα μπορούσαν να υπάρχουν μεταξύ των γειτόνων του κόμβου. Συνεπώς, εάν ο γείτονας ενός κόμβου i έχει n κατευθυνόμενους συνδέσμους τότε ο συντελεστής συσταδοποίησης του i ορίζεται ως εξής [5]:

$$c(i) = \frac{n}{d_i(d_i - 1)}$$

Ο συντελεστής συσταδοποίησης ολόκληρου του γραφήματος είναι ο μέσος όρος των συντελεστών συσταδοποίησης των κόμβων του και συμβολίζεται ως $C(G)$ [5]:

$$C(G) = \frac{\sum_{v \in V} c(v)}{|V|}$$

Ο συντελεστής συσταδοποίησης του γραφήματος κυμαίνεται μεταξύ 0 και 1, με τις υψηλές τιμές να υποδεικνύουν έναν μεγάλο βαθμό «cliquishness» μεταξύ των κόμβων. Συγκεκριμένα, ένα γράφημα με συντελεστή συσταδοποίησης 0 δεν περιέχει καθόλου «τρίγωνα» συνδεδεμένων κόμβων ενώ ένα γράφημα με συντελεστή συσταδοποίησης 1 είναι μια τέλεια «κλίκα» (clique).

4.1.6 Κεντρικότητα διαμεσότητας

Η κεντρικότητα διαμεσότητας (betweenness centrality) B μιας ακμής ορίζεται ως ο αριθμός των συντομότερων διαδρομών μεταξύ όλων των ζευγών κορυφών του γραφήματος διαμέσου αυτής της ακμής. Εάν ένα ζευγάρι κορυφών έχει πολλές σύντομες διαδρομές τότε σε κάθε διαδρομή ανατίθεται ένα βάρος τέτοιο ώστε το άθροισμα όλων των διαδρομών να είναι 1. Η κεντρικότητα διαμεσότητας για μια ακμή e εκφράζεται ως [5]:

$$B(e) = \sum_{u \in V, v \in V} \frac{\sigma_e(u, v)}{\sigma(u, v)}$$

όπου το $\sigma(u, v)$ αναπαριστά τον αριθμό των σύντομων διαδρομών μεταξύ των u και v και το $\sigma_e(u, v)$ αναπαριστά τον αριθμό των σύντομων διαδρομών μεταξύ των u και v οι οποίες περιλαμβάνουν την e. Η κεντρικότητα διαμεσότητας μιας ακμής είναι μια μετρική η οποία εκφράζει τη σημαντικότητα της ακμής στο γράφημα όπου υψηλές τιμές υποδηλώνουν σημαντικές ακμές για τη δομή του γραφήματος.

4.1.7 Τμηματικότητα

Όταν εξετάζονται κοινότητες σε ένα δίκτυο, συχνά απαιτείται μια μετρική αξιολόγησης της ποιότητας του διαμερισμού του γραφήματος σε κοινότητες. Ένα τέτοιο μέτρο είναι η τμηματικότητα (modularity). Θεωρούμε μια δομή k κοινοτήτων, το e να είναι ένας συμμετρικός πίνακας $k \times k$, του οποίου το στοιχείο e_{ij} είναι το κλάσμα των ακμών του δικτύου οι οποίες ενώνουν τις κορυφές της κοινότητας i με την κοινότητα j. Ορίζουμε επίσης ως $a_i = \sum_j e_{ij}$ το κλάσμα των ακμών που εφάπτονται με τις κορυφές της κοινότητας i. Έτσι, ο πίνακας $\mathrm{Tr\,e} = \sum_i e_{ii}$ δίνει το κλάσμα των ακμών του δικτύου της ίδιας κοινότητας. Η τμηματικότητα ορίζεται ως [5]:

$$Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr\,e} - ||\mathrm{e}^2||$$

με τιμές μεταξύ -1 και 1, με το 0 να εκφράζει ότι δεν υπάρχουν δομές κοινοτήτων ενώ αντίθετα οι θετικές τιμές εκφράζουν την ύπαρξη κοινοτικών δομών. Στην πράξη, τμηματικότητα ίση ή μεγαλύτερη του 0.3 παρατηρείται σε πραγματικά δίκτυα παρουσιάζοντας σημαντικό ποσοστό δομής κοινοτήτων.

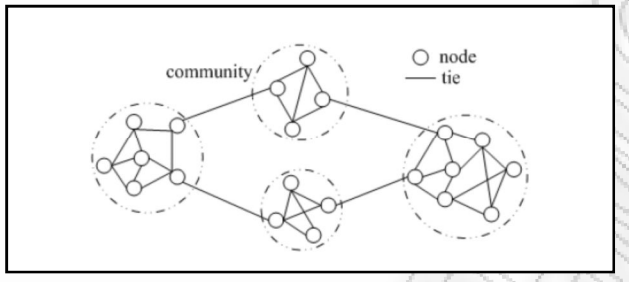
4.1.8 Συνδεδεμένα συστατικά

Για ένα μη-κατευθυνόμενο γράφημα ένα συνδεδεμένο συστατικό (connected component) ορίζεται ως ένα υποσύνολο κόμβων τέτοιο ώστε να υπάρχει ένα μονοπάτι του δικτύου μεταξύ όλων των ζευγών αυτών των κόμβων. Για ένα κατευθυνόμενο γράφημα, γίνεται διάκριση μεταξύ ενός έντονα συνδεδεμένου συστατικού και ενός λιγότερο έντονα συνδεδεμένου συστατικού [5].

4.2 Ανίχνευση κοινοτήτων

Το πιο γνωστό πρόβλημα ανάλυσης δομής στο πλαίσιο των κοινωνικών δικτύων είναι αυτό της *ανίχνευσης κοινοτήτων* (community detection). Το πρόβλημα της ανίχνευσης κοινοτήτων συνδέεται με το πρόβλημα της εύρεσης στενά συνδεδεμένων ομάδων σε ένα δίκτυο. Ενώ το πρόβλημα αυτό έχει μελετηθεί παραδοσιακά στο πλαίσιο του διαμερισμού ενός γραφήματος (graph partitioning) τα κοινωνικά δίκτυα διαφέρουν σημαντικά ως προς το μέγεθος. Βέβαια, όταν έχουμε διαθέσιμο ένα σημαντικό ποσό περιεχομένου έχει ως συνέπεια την αύξηση της αποτελεσματικότητας στο πρόβλημα της ανίχνευσης κοινοτήτων.

Μια κοινότητα σε ένα δίκτυο είναι μια ομάδα κόμβων με εσωτερικά ισχυρότερους δεσμούς σε σχέση με το υπόλοιπο δίκτυο όπως απεικονίζεται και στην εικόνα 4.1. Αυτός ο ορισμός έχει τυποποιηθεί με μια σειρά *«ανταγωνιστικών»* τρόπων, συνήθως μέσω μιας ποιοτικής συνάρτησης, η οποία ποσοτικοποιεί την ποιότητα ενός συγκεκριμένου διαμερισμού του δικτύου σε κοινότητες. Ορισμένες από τις μετρικές ποιότητας όπως οι *κανονικοποιημένες τομές* (Normalized Cuts) [11] και η *τμηματικότητα* (Modularity) [12] είναι πιο δημοφιλής σε σχέση με άλλες, αλλά καμία δεν έχει κερδίσει την καθολική αποδοχή, δεδομένου ότι καμία μετρική δεν εφαρμόζεται σε όλες τις περιπτώσεις.



Εικόνα 4-1 Κοινωνική δομή ενός δικτύου

Οι αλγόριθμοι για την ανίχνευση κοινοτήτων διαφέρουν σημαντικά σε σχέση με την προσέγγισή τους στο πρόβλημα καθώς και στα χαρακτηριστικά των επιδόσεών τους. Μια σημαντική διαφοροποίηση στις προσεγγίσεις τους είναι αν βελτιστοποιούν ξεκάθαρα ή όχι μια συγκεκριμένη μετρική ποιότητας. Οι *φασματικές μέθοδοι* (spectral methods) είναι παραδείγματα αλγορίθμων που προσπαθούν ρητά να βελτιστοποιήσουν μια συγκεκριμένη μετρική ποιότητας ενώ άλλοι αλγόριθμοι όπως ο *αλγόριθμος Markov Clustering* (MCL) και οι αλγόριθμοι *συσταδοποίησης shingling* δεν το πράττουν.

Μια άλλη διάσταση στην οποία οι αλγόριθμοι διαφέρουν είναι στο πώς (ή ακόμα και στο αν) θα επιτρέψουν στον χρήστη τον έλεγχο του βαθμού ανάλυσης της διαίρεσης του δικτύου σε κοινότητες. Μερικοί αλγόριθμοι (όπως οι φασματικές μέθοδοι) εστιάζουν κυρίως στον διμερή διαμερισμό του δικτύου, αλλά αυτό μπορεί να χρησιμοποιηθεί επαναληπτικά με σκοπό την υποδιαίρεση του δικτύου σε όσες κοινότητες είναι επιθυμητό.

Άλλοι αλγόριθμοι όπως της *συζευκτικής συσταδοποίησης* (agglomerative clustering) επιτρέπουν στον χρήστη να ελέγχει έμμεσα τον βαθμό ανάλυσης των κοινοτήτων μέσω ορισμένων παραμέτρων. Επιπλέον, αυτοί που βελτιστοποιούν τη μετρική της τμηματικότητας δεν επιτρέπουν ο χρήστης να ελέγχει τον αριθμό των κοινοτήτων που πρόκειται να δημιουργηθούν.

Ένα άλλο σημαντικό χαρακτηριστικό διαφοροποίησης των αλγορίθμων ανίχνευσης κοινοτήτων είναι η σημασία που αποδίδουν σε μια ισορροπημένη κατανομή του δικτύου. Τέλος, σε ότι αφορά τα χαρακτηριστικά απόδοσης, οι αλγόριθμοι διαφέρουν επίσης στη δυνατότητα επεκτασιμότητας τους σε μεγάλα δίκτυα.

4.2.1 Συναρτήσεις ποιότητας

Θεωρούμε ως A τον πίνακα γειτνίασης του δικτύου ή του γραφήματος, το $A(i,j)$ αντιπροσωπεύει το βάρος της ακμής ή τη συνάφεια μεταξύ των κόμβων i,j και V υποδηλώνει το σύνολο των κορυφών ή κόμβων του δικτύου.

Η *κανονικοποιημένη τομή* (normalized cut) μιας ομάδας κορυφών $S \subset V$ ορίζεται ως εξής [11]:

$$N_{cut}(S) = \frac{\sum_{i\in S, j\in \bar{S}} A(i,j)}{\sum_{i\in S} degree(i)} + \frac{\sum_{i\in S, j\in \bar{S}} A(i,j)}{\sum_{j\in S} degree(j)}$$

Πιο αναλυτικά, η κανονικοποιημένη τομή της ομάδας κορυφών S είναι το άθροισμα των βαρών των ακμών που συνδέουν το S με το υπόλοιπο του γραφήματος το οποίο κανονικοποιείται από το συνολικό βάρος των ακμών του S και το βάρος των ακμών του υπόλοιπου γραφήματος. Διαισθητικά, ομάδες με χαμηλή κανονικοποιημένη τομή είναι κατάλληλες για ανίχνευση καλών κοινοτήτων δεδομένου ότι είναι ισχυρά συνδεδεμένες μεταξύ τους αλλά ταυτόχρονα αραιά συνδεδεμένες με το υπόλοιπο του γραφήματος.

Μια άλλη μετρική, η *αγωγιμότητα* (conductance) της ομάδας κορυφών $S \subset V$ είναι στενά συνδεδεμένη και ορίζεται ως εξής:

$$Conductance(S) = \frac{\sum_{i\in S, j\in \bar{S}} A(i,j)}{\min(\sum_{i\in S} degree(i), \sum_{i\in \bar{S}} degree(i))}$$

Η αγωγιμότητα ενός διαμερισμού γραφήματος σε k συστάδες $\{V_1, \dots, V_k\}$ είναι το άθροισμα των κανονικοποιημένων τομών κάθε συστάδας $V_i = \{1, \dots, k\}$.

Η τμηματικότητα έχει γίνει πρόσφατα αρκετά δημοφιλής ως ένας τρόπος μέτρησης της ποιότητας συσταδοποίησης ενός γραφήματος. Ένα από τα πλεονεκτήματα της τμηματικότητας είναι ότι είναι ανεξάρτητη από τον αριθμό των συστάδων στις οποίες διαμερίζεται ένα γράφημα. Η διαίσθηση πίσω από τον ορισμό της τμηματικότητας είναι όσο πιο πολύ ένας υπογράφος αντιστοιχεί σε κάθε κοινότητα τόσο καλύτερη ή πιο σημαντική είναι η ανίχνευση κοινοτήτων.

Η τμηματικότητα Q για το διαμερισμό ενός γραφήματος σε k συστάδες $\{V_1, \dots, V_k\}$ προκύπτει ως εξής [12]:

$$Q = \sum_{c=1}^{k} \left[\frac{A(V_i, V_i)}{m} - \left(\frac{degree(V_i)}{2m} \right)^2 \right]$$

Στην παραπάνω εξίσωση το V είναι οι συστάδες, το m είναι ο αριθμός των ακμών στο γράφημα και το $degree(V_i)$ είναι ο συνολικός βαθμός της συστάδας V_i. Για κάθε συστάδα, υπολογίζουμε τη διαφορά μεταξύ του κλάσματος των ακμών στο εσωτερικό της συστάδας και του κλάσματος των ακμών το οποίο αναμένεται να ανήκει στο εσωτερικό μιας τυχαίας συστάδας με τον ίδιο συνολικό βαθμό.

4.2.2 Φασματικοί αλγόριθμοι

Οι *φασματικοί αλγόριθμοι* (spectral algorithms) χαρακτηρίζονται ως κλασικές μεθόδους στη συσταδοποίηση και στην ανίχνευση κοινοτήτων. Αναφέρονται εν γένει σε αλγόριθμους που αναθέτουν τους κόμβους σε κοινότητες με βάση τα ιδιοδιανύσματα πινάκων, όπως είναι ο πίνακας γειτνίασης του ίδιου του δικτύου ή άλλων σχετικών πινάκων. Τα πρώτα k ιδιοδιανύσματα καθορίζουν την ενσωμάτωση των κόμβων στο δίκτυο ως σημεία ενός k-διάστατου χώρου με αποτέλεσμα να υπάρχει στη συνέχεια η δυνατότητα εφαρμογής κλασσικών τεχνικών συσταδοποίησης όπως ο αλγόριθμος K-means με στόχο να

προκύψουν οι τελικές αναθέσεις των κόμβων σε συστάδες. Η κύρια ιδέα πίσω από τη φασματική συσταδοποίηση είναι ότι χαμηλών διαστάσεων απεικόνιση η οποία προκαλείται από τα πρώτα k ιδιοδιανύσματα, εκθέτει τη δομή συστάδας στο αρχικό γράφημα με μεγαλύτερη σαφήνεια.

Ο κύριος πίνακας ο οποίος χρησιμοποιείται στις εφαρμογές φασματικής συσταδοποίησης είναι ο πίνακα Laplace L. Θεωρώντας ως A τον πίνακα γειτνίασης του δικτύου και ως D τον διαγώνιο πίνακα των βαθμών των κόμβων κατά μήκος της διαγωνίου έχουμε τον μη κανονικοποιημένο πίνακα Laplace L ο οποίος προκύπτει ως $L = D - A$. Ο πίνακας Laplace (ή ο κανονικοποιημένος πίνακας Laplace) συμβολίζεται με $\mathcal{L}$ και προκύπτει ως $\mathcal{L} = D^{-1/2}(D - A)D^{-1/2} = I - D^{-1/2}AD^{-1/2}$. Μπορεί να αποδειχθεί ότι οι L και $\mathcal{L}$ είναι συμμετρικοί και έχουν πραγματικές και θετικές ιδιοτιμές.

Στην εργασία [13] παρουσιάζονται οι πιο κοινοί φασματικοί αλγόριθμοι συσταδοποίησης και αναλύονται συνοπτικά παρακάτω. Θεωρούμε ότι έχουμε ένα σύνολο σημείων $x_1, \dots x_n$ τα οποία μπορεί να είναι αυθαίρετα αντικείμενα και οι ομοιότητές τους $s_{ij} = s(x_i, x_j)$, να προκύπτουν σύμφωνα με κάποια συνάρτηση ομοιότητας η οποία είναι συμμετρική και μη-αρνητική. Ο αντίστοιχος πίνακας ομοιότητας δίνεται από $S = (s_{ij})_{ij=1\dots n}$

Οι τρεις αλγόριθμοι που ακολουθούν φαίνονται αρκετά όμοιοι, εκτός από το γεγονός ότι χρησιμοποιούν τρία διαφορετικά γραφήματα Laplace. Ο *Αλγόριθμός 1* χρησιμοποιεί ως δεδομένα εισόδου τον πίνακα ομοιότητας και τον αριθμό των συστάδων που πρέπει να δημιουργηθούν. Στη συνέχεια, υπολογίζεται ο μη κανονικοποιημένος πίνακας Laplace και τα πρώτα k ιδιοδιανύσματα του L. Εφαρμόζεται σε επόμενη φάση ο αλγόριθμος k-means για το τελικό στάδιο της συσταδοποίησης.

Unnormalized spectral clustering

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct

- Construct a similarity graph by one of the ways described in Section 2. Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L.
- **Compute the first k eigenvectors $v_1, \dots, v_k$ of L.**
- Let $V \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $v_1, \dots, v_k$ as columns.
- For $i = 1, \dots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of V.
- Cluster the points $(y_i)_{i=1,\dots,n}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \dots, C_k$.

Output: Clusters $A_1, \dots, A_k$ with $A_i = \{j | y_j \in C_i\}$.

Εικόνα 4-2 Αλγόριθμος 1-Μη κανονικοποιημένη φασματική συσταδοποίηση

Στη συνέχεια, ακολουθούν δύο διαφορετικές εκδοχές κανονικοποιημένης φασματικής συσταδοποίησης οι οποίες διαφοροποιούνται από το κανονικοποιημένο γράφημα Laplace. Ο *Αλγόριθμος 2* χρησιμοποιεί τα γενικευμένα ιδιοδιανύσματα του L, τα οποία αντιστοιχούν στα ιδιοδιανύσματα του κανονικοποιημένου πίνακα L_{rw}.

Normalized spectral clustering according to Shi and Malik (2000)

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct

- Construct a similarity graph by one of the ways described in Section 2. Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L.
- **Compute the first k eigenvectors $v_1, \ldots, v_k$ of the generalized eigenproblem $Lv = \lambda Dv$.**
- Let $V \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $v_1, \ldots, v_k$ as columns.
- For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of V.
- Cluster the points $(y_i)_{i=1,\ldots,n}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \ldots, C_k$.

Output: Clusters $A_1, \ldots, A_k$ with $A_i = \{j | y_j \in C_i\}$.

Εικόνα 4-3 Αλγόριθμος 2-Κανονικοποιημένη φασματική συσταδοποίηση

Ο *Αλγόριθμος 3* χρησιμοποιεί επίσης έναν κανονικοποιημένο πίνακα Laplace αλλά αυτή τη φορά ο πίνακας L_{sym} αντί του L_{rw}. Όπως μπορούμε να δούμε αυτός ο αλγόριθμος χρειάζεται να εισάγει ένα επιπρόσθετο βήμα κανονικοποίησης το οποίο δεν απαιτείται στους άλλους αλγορίθμους.

Normalized spectral clustering according to Ng, Jordan, and Weiss (2002)

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct

- Construct a similarity graph by one of the ways described in Section 2. Let W be its weighted adjacency matrix.
- Compute the normalized Laplacian L_{sym}.
- **Compute the first k eigenvectors $v_1, \ldots, v_k$ of L_{sym}.**
- Let $V \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $v_1, \ldots, v_k$ as columns.
- **Form the matrix $U \in \mathbb{R}^{n \times k}$ from V by normalizing the row sums to have norm 1,** that is $u_{ij} = v_{ij} / (\sum_k v_{ik}^2)^{1/2}$.
- For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of U.
- Cluster the points $(y_i)_{i=1,\ldots,n}$ with the k-means algorithm into clusters $C_1, \ldots, C_k$.

Output: Clusters $A_1, \ldots, A_k$ with $A_i = \{j | y_j \in C_i\}$.

Εικόνα 4-4 Αλγόριθμος 3-Κανονικοποιημένη φασματική συσταδοποίηση

Στην εργασία [14] οι J. Ruan και W. Zhang θεωρούν το $G = (V : E)$ ως ένα δίκτυο n κορυφών στο V και m ακμών στο E και $A = (A_{ij})$ είναι ο πίνακας γειτνίασης του G. Ένα πρόβλημα που προκύπτει κατά το διαμερισμό γραφήματος είναι η εύρεση δύο ή περισσοτέρων υποσυνόλων κορυφής σχεδόν ίσου μεγέθους, καθώς και η ελαχιστοποίηση του αριθμού των ακμών που προκύπτουν από τον διαμερισμό. Ανάμεσα σε πολλές ευριστικές μεθόδους που έχουν αναπτυχθεί για το παραπάνω πρόβλημα, οι φασματικές μέθοδοι έχουν λάβει αξιόλογη προσοχή ενώ είναι και οι πιο διαδεδομένες.

Με δεδομένο τον Γ^k διαμερισμό του δικτύου, ο οποίος χωρίζει τις κορυφές σε k κοινότητες, η τμηματικότητα (modularity) στο [14] ορίζεται ως εξής:

$$Q(\Gamma^k) = \sum_{i=1}^{k} (e_{ii}/c) - (a_i/c)^2)$$

όπου το e_{ii} ισούται με τον αριθμό των ακμών των οποίων οι κορυφές τους είναι στην κοινότητα i, το a_i ισούται με τον αριθμό των ακμών με τη μία ή και τις δύο κορυφές μέσα στην κοινότητα i και το c είναι ο συνολικός αριθμός των ακμών. Συνεπώς, η συνάρτηση Q μετρά το κλάσμα των ακμών που περιέχονται μέσα σε κοινότητες από το οποίο αφαιρείται το αναμενόμενο αποτέλεσμα μιας τυχαίας κατανομής.

Η μεγαλύτερη αξία Q συνεπάγεται ισχυρότερες δομές κοινοτήτων. Εάν ένας διαμερισμός δεν δίνει μεγαλύτερο αριθμό ακμών μέσα σε κοινότητες από τον αριθμό που θα προέκυπτε από μια τυχαία κατανομή, τότε η τμηματικότητα είναι $Q \leq 0$. Για έναν ασήμαντο διαμερισμό με μία μόνο κοινότητα, η τμηματικότητα είναι $Q = 0$. Έχει παρατηρηθεί πως σε όλα τα πραγματικά εφαρμοσμένα δίκτυα η τμηματικότητα είναι $Q > 0.3$. Όπως έχουμε αναφέρει η συνάρτηση Q αποτελεί μία πολύ καλή μετρική ποιότητας στη σύγκριση διαφορετικών κοινοτικών δομών.

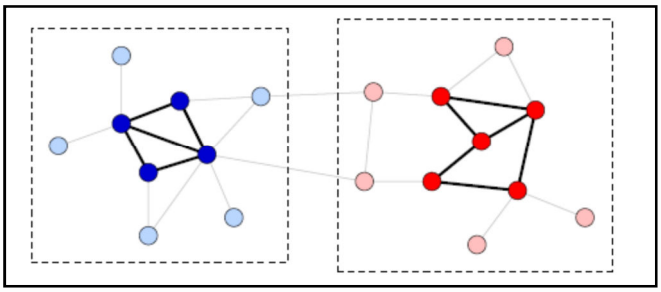
Οι J. Ruan και W. Zhang στο [14] στην προσπάθειά τους να αναπτύξουν μία μέθοδο η οποία θα μπορεί να κλιμακώνεται με επιτυχία σε τεράστια δίκτυα, ενώ παράλληλα θα διατηρεί την αποτελεσματικότητά της στην εύρεση καλών κοινοτήτων, δανείζονται τη στρατηγική που χρησιμοποιείται στον αλγόριθμο ο οποίος παρουσιάζεται στο [11], και αφορά την επαναλαμβανόμενη διαίρεση ενός δικτύου σε μικρότερα τμήματα. Ωστόσο, δύο προβλήματα παραμένουν. Αρχικά, ποια είναι η χρονική στιγμή που θα πρέπει να σταματήσει ο αλγόριθμος ή με άλλα λόγια με ποιο τρόπο γνωρίζουμε εάν ένα υποδίκτυο πρέπει να διαμεριστεί; Η απάντηση σε αυτή την ερώτηση είναι ότι από την στιγμή που ο στόχος είναι να βρεθεί ένας διαμερισμός με υψηλή τμηματικότητα, εξετάζεται εάν η αξία Q αυξάνεται μετά τον διαμερισμό. Σε περίπτωση που κανένας διαμερισμός δεν μπορεί να αυξήσει την τμηματικότητα, τότε το υποδίκτυο δεν πρέπει να διαχωριστεί. Δεύτερον, έχει παρατηρηθεί εμπειρικά, πως εάν υπάρχουν πολλαπλές κοινότητες, η χρήση πολλαπλών ιδιοδιανυσμάτων για τον απευθείας υπολογισμό ενός k-way διαμερισμού, είναι η καλύτερη επιλογή από τις μεθόδους επαναλαμβανόμενης διχοτόμησης. Για τους λόγους αυτούς προτείνεται ο αλγόριθμος K-cut στο [14] που είναι ένας μοναδικός συνδυασμός επαναλαμβανόμενου διαχωρισμού και απευθείας k-way μεθόδου, με τον οποίο επιτυγχάνεται η επάρκεια μιας επαναλαμβανόμενης προσέγγισης αλλά και η ακρίβεια μιας απευθείας k-way μεθόδου διαμερισμού.

Στον αλγόριθμο αυτόν ακολουθείται μια άπληστη στρατηγική για τον κατά επανάληψη διαμερισμό ενός δικτύου με σκοπό τη βελτιστοποίηση της τμηματικότητας Q. Αντίθετα με τους υπάρχοντες αλγόριθμους που αναζητούν μονίμως μία διχοτόμηση προτείνεται μια στρατηγική απευθείας k-way διαμερισμού όπως στον αλγόριθμο ο οποίος παρουσιάζεται στο [15]. Εν συντομία, υπολογίζεται ο καλύτερος k-way διαμερισμός με $k = 2, 3, \ldots l$ και επιλέγεται το k το οποίο δίνει την υψηλότερη αξία Q. Στη συνέχεια, για κάθε υποδίκτυο ο αλγόριθμος εφαρμόζεται επαναληπτικά. Για να μειωθεί το κόστος υπολογισμού, περιορίζεται το l σε μικρούς ακέραιους αριθμούς. Μέσω των πειραμάτων, ο αλγόριθμος όπου το l είναι τόσο μικρό όσο το 3 ή το 4, μπορεί να βελτιώσει σημαντικά την τμηματικότητα πάνω από την καθιερωμένη στρατηγική διχοτόμησης. Επιπλέον, το κόστος υπολογισμού μπορεί επίσης να μειωθεί από μία ελαφρώς αυξημένη αξία του l σε σύγκριση με την διχοτόμηση.

4.2.3 Συζευκτικοί αλγόριθμοι

Η *συζευκτική συσταδοποίηση* (agglomerative clustering) αποτελεί παράδειγμα συσταδοποίησης ιεραρχικής προσέγγισης. Το αποτέλεσμα του αλγόριθμου είναι μια ιεραρχία συστάδων. Ο αλγόριθμος ξεκινά με ένα σύνολο n συστάδων όπου κάθε συστάδα αντιπροσωπεύει ένα αντικείμενο του συνόλου δεδομένων το οποίο σκοπός είναι να τοποθετηθεί σε μια συστάδα. Σε κάθε βήμα δύο συστάδες ενώνονται βάσει ενός κριτηρίου. Το πιο συχνά χρησιμοποιούμενο κριτήριο είναι η απόσταση μεταξύ δύο συστάδων. Ο αλγόριθμος χρειάζεται επίσης ένα κριτήριο τερματισμού με σκοπό η συζευκτική διαδικασία συσταδοποίησης να συνεχισθεί μέχρι το σημείο που όλες οι περιπτώσεις θα ανήκουν σε μία συστάδα.

Οι μέθοδοι συζευκτικής συσταδοποίησης είναι συνήθως καλοί στην ανίχνευση στενά συνδεδεμένων πυρήνων των κοινοτήτων (έντονες κορυφές και ακμές) αλλά έχουν την τάση να αφήνουν εκτός απομακρυσμένες κορυφές, ακόμη και όταν, οι περισσότερες από αυτές σαφώς ανήκουν σε κάποια κοινότητα. Στην εικόνα 4.5 που ακολουθεί παρουσιάζεται αυτό το χαρακτηριστικό της συζευκτικής συσταδοποίησης.



Εικόνα 4-5 Συζευκτική συσταδοποίηση

Η συζευκτική συσταδοποίηση μπορεί να θεωρηθεί ως μια «άπληστη» μέθοδος αναζήτησης κατά την οποία σε κάθε βήμα αναζητούμε την τρέχουσα βέλτιστη διάταξη των συστάδων με δεδομένο ότι δύο συστάδες πρέπει να ενωθούν. Ας υποθέσουμε ότι αναζητούμε δύο συστάδες με ελάχιστη μέση απόσταση μεταξύ των μελών τους. Για να βρούμε την τρέχουσα βέλτιστη σύζευξη ο αλγόριθμος εξετάζει όλα τα πιθανά ζεύγη συστάδων. Μια τέτοια προσέγγιση όμως είναι ανέφικτη για το πρόβλημα της ανίχνευσης κοινοτήτων, λόγω του μεγέθους των κοινωνικών δικτύων και του αριθμού των αναμενόμενων κοινοτήτων.

Οι J. Newman et. al. στο [16] προτείνουν έναν άπληστο συζευκτικό αλγόριθμο συσταδοποίησης ο οποίος βελτιστοποιεί τη μετρική της τμηματικότητας. Η βασική ιδέα του αλγορίθμου είναι ότι σε κάθε στάδιο, οι ομάδες των κορυφών διαδοχικά συγχωνεύονται σε μεγαλύτερες κοινότητες έτσι ώστε η τμηματικότητα του αποτελέσματος από τον διαμερισμό του δικτύου να αυξάνεται σε κάθε συγχώνευση.

Πιο αναλυτικά, θεωρούμε ότι A_{vw} είναι ένα στοιχείο του πίνακα γειτνίασης του δικτύου οπότε από την εξίσωση (1) έχουμε [16]:

$$A_{vw} = \begin{cases} 1 & \text{εάν οι κορυφές } v \text{ και } w \text{ συνδέονται} \\ 0 & \text{σε όποια άλλη περίπτωση} \end{cases} \quad (1)$$

και υποθέτουμε ότι οι κορυφές διαιρούνται σε κοινότητες όταν η κορυφή v ανήκει στην κοινότητα c_v.

Τότε το κλάσμα των ακμών που συμπεριλαμβάνονται σε κοινότητες προκύπτει από την εξίσωση (2) [16]:

$$\frac{\sum_{vw} A_{vw}\delta(c_v,c_w)}{\sum_{vw} A_{vw}} = \frac{1}{2m}\sum_{vw} A_{vw}\delta(c_v,c_w) \quad (2)$$

όπου η δ-συνάρτηση $\delta(i,j)$ είναι 1 εάν $i = j$ και 0 σε κάθε άλλη περίπτωση και $m = \frac{1}{2}\sum_{vw} A_{vw}$ είναι ο αριθμός των ακμών στο γράφημα. Όσο μεγαλύτερη είναι η τιμή της συνάρτησης δ τόσο καλύτερο διαμερισμό έχουμε στο γράφημα με τη λογική ότι έχουμε πολλές ακμές οι οποίες συμμετέχουν σε κοινότητες. Όμως γενικότερα δεν είναι ένα καλό μέτρο της δομής της κοινότητας από τη στιγμή που η μεγαλύτερη τιμή που παίρνει είναι 1 και αντιστοιχεί στην απλή περίπτωση όπου όλες οι κορυφές ανήκουν σε μία μόνο κοινότητα.

Ο βαθμός k_v της κορυφής v ορίζεται ως το βάρος των ακμών που επιδρούν πάνω σε αυτή και δίνεται από την εξίσωση (3) [16]:

$$k_v = \sum_w A_{vw} \quad (3)$$

Η πιθανότητα μια ακμή να υπάρχει μεταξύ των κορυφών v και w εάν οι συνδέσεις έχουν δημιουργηθεί τυχαία αλλά με σεβασμό στους βαθμούς των κορυφών είναι $k_v k_w/2m$.

Ορίζουμε την τμηματικότητα Q μέσω της εξίσωσης (4) [16]:

$$Q = \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta(c_v,c_w) \quad (4)$$

Εάν οι υψηλές τιμές της τμηματικότητας αντιστοιχούν σε καλό διαμερισμό του δικτύου τότε πρέπει να είμαστε σε θέση να βρούμε τόσο καλούς διαμερισμούς μέσω της αναζήτησης πιθανών υποψηφίων για τους οποίους έχουμε υψηλή τμηματικότητα. Ενώ η εύρεση της μέγιστης τμηματικότητας όλων των δυνατών διαμερισμών φαίνεται αρκετά δύσκολη σε γενικές γραμμές, αρκετά καλές λύσεις μπορούν να βρεθούν κατά προσέγγιση με τις τεχνικές βελτιστοποίησης.

Για να απλοποιήσουμε την περιγραφή του αλγορίθμου ορίζουμε τις παρακάτω δύο ποσότητες. Η πρώτη μέσω της εξίσωσης (5) [16]:

$$e_{ij} = \frac{1}{2m}\sum_{vw} A_{vw}\delta(c_v, i)\delta(c_w, j) \quad (5)$$

η οποία είναι το κλάσμα των ακμών που ενώνουν τις κορυφές στην κοινότητα i με τις κορυφές της κοινότητας j και η εξίσωση (6) [16]:

$$a_i = \frac{1}{2m}\sum_v k_v\delta(c_v, i) \quad (6)$$

η οποία είναι το κλάσμα των ακμών οι οποίες συνδέονται στις κορυφές της κοινότητας i.

Στη συνέχεια, γράφοντας $\delta(c_v, c_w) = \sum_i \delta(c_v, i)\delta(c_w, i)$, έχουμε από την εξίσωση (4) την παρακάτω εξίσωση (7) της τμηματικότητας [16]:

$$Q = \sum_i (e_{ii} - a_i^2) \quad (7)$$

Στον αλγόριθμό αυτό αντί να διατηρείται ο πίνακας γειτνίασης και να υπολογίζεται η ΔQ_{ij}, διατηρείται και ενημερώνεται ένας πίνακας την τιμής Q_{ij}. Από την ένωση δύο κοινοτήτων χωρίς καμία ακμή μεταξύ τους δεν πρόκειται ποτέ να δημιουργηθεί μια αύξηση στην Q, για το λόγο αυτό απαιτείται η αποθήκευση της ΔQ_{ij} για εκείνα τα ζεύγη $\{i, j\}$ τα οποία ενώνονται με μια ή περισσότερες ακμές.

Συνολικά, διατηρούνται τρείς δομές δεδομένων οι οποίες είναι [16]:

- Ένας αραιός πίνακας (sparse matrix) ο οποίος περιέχει το ΔQ_{ij} για κάθε ζευγάρι $\{i, j\}$ των κοινοτήτων με μια τουλάχιστον ακμή μεταξύ τους.
- Μια μέγιστη «στοίβα» (max heap) H η οποία περιέχει το μεγαλύτερο στοιχείο της κάθε γραμμής του πίνακα ΔQ_{ij} των $\{i, j\}$ του αντίστοιχου ζευγαριού κοινοτήτων.
- Ένα κοινό διάνυσμα των στοιχείων a_i.

Όπως αναφέραμε και παραπάνω ο αλγόριθμος ξεκινά θεωρώντας κάθε κορυφή ως το μοναδικό μέλος κάθε κοινότητας όπου σε κάθε περίπτωση ισχύει $e_{ij} = {}^{1}/_{2m}$ εάν τα i και j συνδέονται και μηδέν σε οποιαδήποτε άλλη περίπτωση.

Συνεπώς, έχουμε τις εξισώσεις (8) και (9) αντίστοιχα [16]:

$$\Delta Q_{ij} = \begin{cases} {}^{1}/_{2m} - {}^{k_i k_j}/_{(2m)^2} & \text{εάν τα } i, j \text{ συνδέονται} \\ 0 & \text{σε κάθε άλλη περίπτωση} \end{cases} \quad (8)$$

$$a_\iota = \frac{k_i}{2m} \text{ για κάθε } i \quad (9)$$

Ο αλγόριθμος μπορεί να οριστεί μέσω των παρακάτω βημάτων [16]:

Βήμα 1: Υπολόγισε τις αρχικές τιμές των ΔQ_{ij} και a_i σύμφωνα με τις εξισώσεις (8) και (9) και στη συνέχεια συμπλήρωσε τη μέγιστη στοίβα με το μεγαλύτερο στοιχείο της κάθε γραμμής του πίνακα ΔQ.

Βήμα 2: Επέλεξε το μεγαλύτερο ΔQ_{ij} από το H, σύνδεσε τις αντίστοιχες κοινότητες, ενημέρωσε τον πίνακα ΔQ, τη στοίβα H, το a_i και αύξησε το Q σε ΔQ_{ij} .

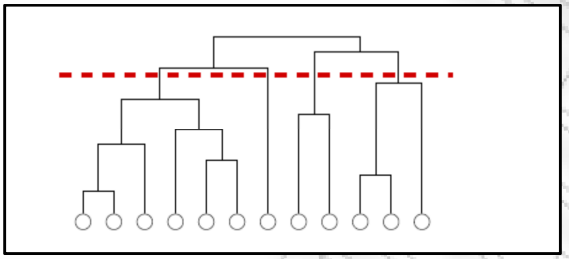
Βήμα 3: Επανέλαβε το βήμα 2 έως ότου να παραμείνει μια κοινότητα.

Ο αλγόριθμος που αναλύθηκε παραπάνω και παρουσιάζεται στο [16] κατάφερε να ανιχνεύσει ξεκάθαρα τις κοινότητες στο δίκτυο του Amazon οι οποίες αντιστοιχούν σε συγκεκριμένα θέματα ή είδη βιβλίων/μουσικής, υποδηλώνοντας ότι οι αγοραστικές τάσεις των πελατών του Amazon είναι στενά συνδεδεμένες με το θέμα. Τέλος, από θέμα πολυπλοκότητας ο αλγόριθμος απαιτεί $O|V|^2$ χρόνο.

4.2.4 Διαιρετικοί αλγόριθμοι

Η *διαιρετική συσταδοποίηση* (divisive clustering) αποτελεί επίσης παράδειγμα συσταδοποίησης ιεραρχικής προσέγγισης. Το αποτέλεσμα του αλγόριθμου είναι και σε αυτή την περίπτωση μια ιεραρχία συστάδων. Εντούτοις, οι αλγόριθμοι αυτοί λειτουργούν με αντίστροφο τρόπο σε σχέση με τους συζευκτικούς αλγορίθμους που αναλύσαμε παραπάνω. Συγκεκριμένα, σε μια διαιρετική προσέγγιση ξεκινάμε με το σύνολο του δικτύου ως μία κοινότητα και σε κάθε βήμα, επιλέγεται μια συγκεκριμένη κοινότητα η οποία διαχωρίζεται σε δύο μέρη. Με τον τρόπο αυτό επαναληπτικά διαιρούμε το δίκτυο σε μικρότερα τμήματα και μπορούμε να σταματήσουμε σε οποιοδήποτε στάδιο και να πάρουμε ως αποτέλεσμα το δίκτυο των κοινοτήτων.

Τόσο οι συζευκτικοί όσο και οι διαιρετικοί αλγόριθμοι συσταδοποίησης παρουσιάζουν ως αποτέλεσμα ένα δενδρόγραμμα το οποίο είναι ένα δυαδικό δέντρο, όπου τα φύλλα είναι κόμβοι του δικτύου και κάθε εσωτερικός κόμβος είναι μια κοινότητα όπως φαίνεται και στην εικόνα 4.6.



Εικόνα 4-6 Απεικόνιση δενδρογράμματος

Στην περίπτωση των διαιρετικών αλγορίθμων, μια σχέση γονέα-παιδιού υποδηλώνει ότι η κοινότητα η οποία αναπαρίσταται από τον κόμβο-γονέα χωρίστηκε με στόχο να προκύψουν οι κοινότητες οι οποίες αναπαρίστανται τους κόμβους-παιδιά. Στην περίπτωση όμως των συζευκτικών αλγορίθμων, μια σχέση γονέα-παιδιού υποδηλώνει ότι οι κοινότητες οι οποίες αναπαρίστανται από τους κόμβους-παιδιά συγχωνεύτηκαν με στόχο τη δημιουργία της κοινότητας η οποία αναπαρίσταται από τον κόμβο-γονέα.

Οι J. Newman και M. Girvan στο [12] προτείνουν ένα διαιρετικό αλγόριθμο ανίχνευσης κοινοτήτων χρησιμοποιώντας τη μετρική της *διαμεσότητας ακμών* (edge betweenness). Η μετρική αυτή εκφράζει ότι οι ακμές με υψηλούς βαθμούς διαμεσότητας είναι πιο πιθανό να είναι οι ακμές οι οποίες θα ενώνουν διαφορετικές κοινότητες. Οι ακμές που βρίσκονται μεταξύ κοινοτήτων (inter-community edges) είναι σχεδιασμένες έτσι ώστε να έχουν υψηλούς βαθμούς διαμεσότητας ακμών σε σχέση με τις ακμές που βρίσκονται στο εσωτερικό των κοινοτήτων (intra-community edges). Συνεπώς, με τον εντοπισμό και την απόρριψη των ακμών με υψηλό βαθμό διαμεσότητας μπορεί κανείς να αποσυνθέσει το κοινωνικό δίκτυο στα συστατικά των κοινοτήτων του.

Το *shortest path betweenness* είναι ένα παράδειγμα ενός μέτρου διαμεσότητας ακμών όπου θεωρεί ότι από τη στιγμή που υπάρχουν λίγες ακμές μεταξύ κοινοτήτων, οι σύντομες διαδρομές μεταξύ των κόμβων που ανήκουν σε διαφορετικές κοινότητες θα περνούν υποχρεωτικά από αυτές τις λίγες μεταξύ των κοινοτήτων ακμές. Επίσης, υπάρχουν άλλα δύο παραδείγματα μέτρων διαμεσότητας ακμών τα *random-walk betweenness* όπου η επιλογή της διαδρομής που συνδέει δύο οποιουσδήποτε κόμβους είναι το αποτέλεσμα ενός τυχαίου περιπάτου και το *current-flow betweenness* το οποίο υποκινείται από τη θεωρία των κυκλωμάτων όπου αρχικά το δίκτυο μετατρέπεται σε ένα δίκτυο αντίστασης όπου κάθε πλευρά έχει αντικατασταθεί από μια μονάδα αντίστασης και δύο κόμβοι επιλέγονται ως πηγή και δέκτης.

Η γενική μορφή του αλγορίθμου είναι [12]:

Βήμα 1: Υπολόγισε το βαθμό διαμεσότητας για όλες τις ακμές του δικτύου χρησιμοποιώντας οποιοδήποτε μέτρο.

Βήμα 2: Βρες την ακμή με τον υψηλότερο βαθμό και απομάκρυνε την από το δίκτυο.

Βήμα 3: Επαναϋπολόγισε τη διαμεσότητα των υπολοίπων ακμών.

Βήμα 4: Επανέλαβε το βήμα 2.

Η παραπάνω διαδικασία συνεχίζεται μέχρι να επιτευχθεί ένας αρκετά μικρός αριθμός κοινοτήτων. Σε αντίθεση προς τις εικασίες ότι τα διάφορα μέτρα της διαμεσότητας ακμών μπορεί να οδηγήσουν σε αποκλίσεις στην εύρεση κοινοτικών δομών, τα πειράματα έδειξαν ότι το ακριβές μέτρο της διαμεσότητας που χρησιμοποιείται δεν είναι τόσο ζωτικής σημασίας.

Το μόνο μειονέκτημα αυτής της προσέγγισης είναι το υπολογιστικό κόστος. Συγκεκριμένα, ο υπολογισμός της διαμεσότητας όλων των ακμών απαιτεί $O(|V||E|)$ χρόνο ενώ η εκτέλεση ολόκληρου του αλγορίθμου απαιτεί $O(|V|^3)$ χρόνο.

4.2.5 Τοπική συσταδοποίηση γραφήματος

Ένας *τοπικός αλγόριθμος συσταδοποίησης* (local clustering algorithm) είναι αυτός ο οποίος βρίσκει μια λύση η οποία περιέχεται ή είναι κοντά σε μια συγκεκριμένη κορυφή ή κορυφές χωρίς να ενδιαφέρεται για ολόκληρο το γράφημα. Οι τοπικοί αλγόριθμοι έχουν τραβήξει το ενδιαφέρον στη διαχείριση μεγάλων γραφημάτων από τη στιγμή που ο χρόνος πολυπλοκότητάς τους εξαρτάται από το μέγεθος της λύσης παρά από το μέγεθος του γραφήματος. Οι D. Spielman and N. Teng στο [17] περιέγραψαν από τους πρώτους τοπικούς αλγορίθμους συσταδοποίησης χρησιμοποιώντας «τυχαίους περιπάτους» (random walks).

Πολλές μέθοδοι ανίχνευσης κοινοτήτων χρησιμοποιούν ιεραρχική συσταδοποίηση είτε μέσω συζευκτικών είτε μέσω διαιρετικών αλγορίθμων. Και οι δύο κατηγορίες συμπεριλαμβανομένων και εκείνων που χρησιμοποιούν τη διαμεσότητα (betweenness) ή άλλες μεθόδους χαρακτηρίζονται ως γενικοί αλγόριθμοι και δεν αναπαριστούν πραγματικές ενέργειες τις οποίες ένα μέλος του δικτύου θα μπορούσε να πραγματοποιήσει με στόχο την αναγνώριση της κοινοτικής δομής του δικτύου. Η προτεινόμενη μέθοδος στο [18] μπορεί ίσως να αναπαραστήσει με καλύτερο τρόπο τις ενέργειες των μελών του δικτύου με στόχο την αναγνώριση των δικών τους κοινοτήτων.

Μια σύντομη περιγραφή του αλγορίθμου δίνεται στη συνέχεια. Για μια κορυφή j, τα βήματα είναι [18]:

Βήμα 1: Ξεκίνα με ένα κέλυφος l, όπου $l = 0$, πρόσθεσε την κορυφή j στη λίστα των μελών μιας κοινότητας και υπολόγισε το K_j^0.

Βήμα 2: Επέκτεινε το κέλυφος l, σε $l = 1$, πρόσθεσε τους γείτονες της j και υπολόγισε το K_j^1.

Βήμα 3: Υπολόγισε το ΔK_j^1 . Εάν το $\Delta K_j^1 < \alpha$ τότε η κοινότητα έχει εντοπιστεί και ο αλγόριθμος σταματά.

Βήμα 4: Διαφορετικά επανέλαβε το βήμα 2 για τα επόμενα l κελύφη, μέχρι να γίνει διασταύρωση με το α ή ολόκληρο το συνδεδεμένο συστατικό να προστεθεί στη λίστα κοινοτήτων.

Όταν το l κέλυφος φτάσει στο όριο της κοινότητας τότε ο αριθμός των αναδυόμενων ακμών μειώνεται απότομα. Αυτό συμβαίνει επειδή, σε αυτό το σημείο, οι μόνες αναδυόμενες ακμές είναι εκείνες που συνδέουν την κοινότητα με το υπόλοιπο του γραφήματος οι οποίες εξ ορισμού είναι λιγότερες σε αριθμό από εκείνες εντός της κοινότητας.

Εισάγοντας μια παράμετρο α και παρακολουθώντας το ΔK_j^l, η ανάπτυξη του l κελύφους μπορεί να σταματήσει όταν έχει καλύψει την κοινότητα. Αυτό ακριβώς το γεγονός επιτρέπει στην κορυφή εκκίνησης να ανιχνεύσει τις κοινότητες της σε τοπικό επίπεδο. Η μέθοδός αυτή δεν είναι τέλεια, όμως, είναι δυνατό το κέλυφος l να «εξαπλωθεί» κατά την ανίχνευση κοινοτήτων. Αυτό εξαρτάται από το πώς η αρχική κορυφή βρίσκεται στο γράφημα: αν είναι πιο κοντά (ή εξίσου κοντά) σε κάποια κορυφή εκτός κοινότητας από ότι σε κάποια κορυφή της κοινότητας, το κέλυφος l μπορεί να εξαπλωθεί κατά μήκος δύο ή περισσότερων κοινοτήτων την ίδια στιγμή. Για να αντιμετωπιστεί αυτή η επίδραση, μπορεί κανείς να τρέξει τον αλγόριθμο N φορές, χρησιμοποιώντας κάθε κορυφή ως κορυφή-αφετηρίας και στη συνέχεια να επιτευχθεί ομοφωνία ομάδας ως προς το ποιες κορυφές ανήκουν σε ποιες κοινότητες.

4.2.6 Διαμερισμός γραφήματος πολλαπλών επιπέδων

Οι μέθοδοι *διαμερισμού γραφήματος πολλαπλών επιπέδων* (multi-level graph partitioning) παρέχουν ένα ισχυρό πλαίσιο για γρήγορο και υψηλής ποιότητας διαμερισμό γραφήματος και στην πραγματικότητα έχουν χρησιμοποιηθεί για την επίλυση διαφόρων άλλων προβλημάτων. Η βασική ιδέα είναι η συρρίκνωση του γραφήματος εισόδου διαδοχικά έτσι ώστε να επιτευχθεί ένα μικρό γράφημα, στη συνέχεια να διαμεριστεί αυτό το μικρό γράφημα και τότε διαδοχικά να προβληθεί αυτός ο διαμερισμός και πάλι στο αρχικό γράφημα, καθορίζοντας τον διαμερισμό σε κάθε βήμα με αυτόν τον τρόπο. Οι μέθοδοι διαμερισμού πολλαπλών επιπέδων περιλαμβάνουν τη φασματική συσταδοποίηση πολλαπλών επιπέδων, την προσέγγιση Metis (η οποία βελτιστοποιεί την αντικειμενική συνάρτηση KL), την προσέγγιση Graclus (η οποία βελτιστοποιεί τις κανονικοποιημένες τομές και άλλες σταθμισμένες τομές) και την προσέγγιση MLR-MCL.

Τα βασικά συστατικά μια στρατηγικής διαμερισμού γραφήματος πολλαπλών επιπέδων σύμφωνα με το [19] είναι:

- *Coarsening:* ο στόχος είναι να δημιουργηθεί ένα μικρότερο γράφημα το οποίο είναι παρόμοιο με το πρωτότυπο γράφημα. Το βήμα αυτό μπορεί να εφαρμοστεί επαναληπτικά μέχρι τη στιγμή που θα προκύψει ένα γράφημα τόσο μικρό έτσι ώστε να διαμερίζεται γρήγορα και με υψηλή ποιότητα.
- *Initial partitioning:* σε αυτό το βήμα εκτελείται ο διαμερισμός του coarsest γραφήματος. Από τη στιγμή που το γράφημα σε αυτό το στάδιο είναι αρκετά μικρό είναι εφικτή η χρήση στρατηγικών όπως ο φασματικός διαμερισμός ο οποίος ναι μεν είναι αργός αλλά δίνει αποτελέσματα υψηλής ποιότητας.
- *Uncoarsening:* στη φάση αυτή, ο διαμερισμός του συγκεκριμένου γραφήματος χρησιμοποιείται για την αρχικοποίηση ενός διαμερισμού στο μεγαλύτερο γράφημα. Σε αυτό το βήμα στην προσέγγιση Metis χρησιμοποιείται μια εκδοχή του KL αλγορίθμου με στόχο να τελειοποιήσει το διαμερισμό που προέρχεται από τα προηγούμενα βήματα. Ο Graclus, από την άλλη πλευρά χρησιμοποιεί έναν σταθμισμένο k-means για την επίτευξη του βέλτιστου διαμερισμού.

4.2.7 Σχετικές εργασίες

Όπως έχουμε αναφέρει η συσταδοποίηση στην ανάλυση κοινωνικών δικτύων είναι διαφορετική από την παραδοσιακή συσταδοποίηση. Για το λόγο αυτό πολλές φορές απαιτεί ομαδοποίηση αντικειμένων σε κλάσεις με βάση τις συνδέσεις τους και τις ιδιότητές τους. Οι παραδοσιακοί αλγόριθμοι συσταδοποίησης βασίζονται στην ομοιότητα των αντικειμένων με αποτέλεσμα να μην μπορούν να εφαρμοστούν στην ανάλυση κοινωνικών δικτύων. Στην εργασία [20], με βάση τον BSP (Business System Planning) αλγόριθμο συσταδοποίησης, προτείνεται ένας αλγόριθμος ανάλυσης κοινωνικών δικτύων. Ο προτεινόμενος αλγόριθμος, διαφέρει από τους παραδοσιακούς αλγόριθμους διότι μπορεί να ομαδοποιήσει αντικείμενα σε ένα κοινωνικό δίκτυο σε διαφορετικές κλάσεις με βάση τις συνδέσεις τους και να εντοπίσει τυχόν σχέσεις μεταξύ των κλάσεων.

Στην εργασία [21] προτείνεται ένας νέος αλγόριθμος με σκοπό την εύρεση κοινοτικών δομών σε πολύπλοκα δίκτυα με βάση το συνδυασμό της φασματικής ανάλυσης και της δομικής βελτιστοποίησης. Το κύριο πλεονέκτημα του αλγορίθμου είναι η αποτελεσματικότητά του. Η καλύτερη αντιστοιχία για τον συγκεκριμένο αλγόριθμό είναι ο γρήγορος αλγόριθμος του Newman ο οποίος είναι αλγόριθμος αναφοράς σε θέματα συσταδοποίησης που αφορούν μεγάλα δίκτυα λόγω της απόδοσης του. Τα αποτελέσματα δείχνουν ότι ο προτεινόμενος αλγόριθμος είναι μια καλή επιλογή για την ανάλυση της δομής κοινοτήτων των μεσαίων και μεγάλων δικτύων σε ένα εύρος δεκάδων έως και εκατοντάδων χιλιάδες κορυφών.

Στην εργασία [22] προτείνεται ένας συζευκτικός γενετικός αλγόριθμος συσταδοποίηση (agglomerative clustering genetic algorithm - ACGA) ο οποίος σε κάθε βήμα θεωρεί δύο τοπικές συστάδες και με βάση τη συνάρτηση καταλληλότητας (fitness function) προσπαθεί να τις ενώσει ή να τις αναδιατάξει έτσι ώστε μια νεοσυσταθείσα συστάδα να βελτιώσει το αποτέλεσμα της συσταδοποίησης. Κάθε βήμα της εξελικτικής διαδικασίας πραγματοποιείται σε τοπικό επίπεδο, συνεπώς επηρεάζει μόνο ένα μικρό μέρος του κοινωνικού δικτύου το οποίο περιορίζεται σε δύο συστάδες και στην άμεση γειτονιά του. Αυτό κάνει τον αλγόριθμο πρακτικά χρήσιμο, ανεξάρτητα από το μέγεθος του δικτύου. Η αξιολόγηση του σε δύο μοντέλα κοινωνικού δικτύου δείχνει ότι είναι σε θέση να ανιχνεύσει κοινότητες με ακρίβεια συγκρίσιμη ή καλύτερη από δύο τυπικούς αλγορίθμους συσταδοποίησης.

Στην εργασία [23] προτείνεται ένα νέο μοντέλο για κοινωνικά δίκτυα, το οποίο αναφέρεται ως μοντέλο λανθάνουσας θέσης συστάδας (latent position cluster model). Με αυτό τον τρόπο μπορούμε να συλλάβουμε τρεις σημαντικές ιδιότητες των δικτύων: τη μεταβατικότητα (transitivity), την ομοφιλία (homophily) και τη συσταδοποίηση. Αναπτύσσονται δύο μέθοδοι για τον υπολογισμό των λανθανουσών θέσεων των παραμέτρων του μοντέλου: μια απλή δύο σταδίων διαδικασία μεγιστοποίησης της πιθανοφάνειας και μια πλήρως μπεϋζιανή προσέγγιση για την εύρεση του αριθμού των συστάδων στα δεδομένα. Η προσέγγιση των δύο σταδίων λειτουργεί αρκετά καλά και είναι απλή στην εφαρμογή ενώ η πλήρως μπεϋζιανή προσέγγιση αποδίδει καλύτερα αλλά είναι πιο πολύπλοκη.

Στην εργασία [24] προτείνεται ένας αλγόριθμος για την ανίχνευση κοινοτήτων χρησιμοποιώντας γενετικούς αλγορίθμους. Η προσέγγιση ορίζει μία μετρική ποιότητας η οποία βασίζεται στον αριθμό και την τοπολογία των υπαρχόντων συνδέσεων ανάμεσα στους κόμβους που αποτελούν μία κοινότητα και στη συνέχεια προσπαθεί να βελτιστοποιήσει αυτή τη μετρική με την εφαρμογή του γενετικού αλγορίθμου. Ο αλγόριθμος χρησιμοποιεί μία αναπαράσταση η οποία βασίζεται σε ένα γράφημα πληθυσμού στα οποία ένα χρωμόσωμα αποτελείται από N γονίδια, καθένα από τα οποία μπορεί να λάβει διάφορες αξίες j, στο εύρος $\{1, \dots, N\}$. Τα γονίδια συμβολίζουν τους κόμβους του γραφήματος $G = (V, E)$. Μοντελοποιώντας ένα κοινωνικό δίκτυο, μία αξία j που αντιστοιχεί στο i_{th} γονίδιο, ερμηνεύεται ως ένας σύνδεσμος ανάμεσα στους κόμβους i και j του V. Αυτό σημαίνει ότι στη λύση συσταδοποίησης τα i και j θα είναι στην ίδια συστάδα. Ειδικοί τελεστές επιτρέπουν τη μείωση του χώρου των πιθανών λύσεων και έτσι βελτιώνουν τη σύγκλιση της μεθόδου. Πειράματα σε πραγματικά δίκτυα δείχνουν την ικανότητα της γενετικής προσέγγισης να ανιχνεύει σωστά κοινότητες με αποτελέσματα συγκρινόμενα με τις προσεγγίσεις της τελευταίας τεχνολογίας.

Στην εργασία [25] παρουσιάζεται η μοντελοποίηση κοινωνικών δικτύων ως μη κατευθυνόμενα γραφήματα και διαμορφώνονται πρότυπα προστασίας της ιδιωτικής ζωής, μοντέλα επιθέσεων για το πρόβλημα της ανωνυμίας και συγκεκριμένα για το πρόβλημα ανωνυμίας που βασίζεται στον βαθμό i-βήματος (i-hope degree) δηλαδή η προγενέστερη γνώση του αντιπάλου περιλαμβάνει το βαθμό του στόχο και τους βαθμούς των γειτόνων εντός i βημάτων από το στόχο. Για το λόγο αυτό παρουσιάζονται δύο νέες και αποτελεσματικές τεχνικές συσταδοποίησης μη κατευθυνόμενων γραφημάτων: οριοθετούνται οι αλγόριθμοι συσταδοποίησης t-means και union-split οι οποίοι ομαδοποιούν όμοιους κόμβους γραφημάτων σε συστάδες με ένα ελάχιστο περιορισμό μεγέθους. Αυτοί οι αλγόριθμοι συσταδοποίησης συνεισφέρουν τόσο στη μελέτη προβλημάτων κοινωνικών δικτύων όσο και στη συσταδοποίηση γενικών τύπων δεδομένων.

Τέλος, στην εργασία [26] παρουσιάζεται μια νέα τεχνική για την ποσοτικοποίηση της ακρίβειας μιας τοπικής μεθόδου με σκοπό να μπορεί κάποιος να προσδιορίσει πως διαφορετικοί αλγόριθμοι εκτελούνται ο ένας σε σχέση με τον άλλο. Λόγω της μοναδικής εξάρτησης μιας τοπικής μεθόδου κατά την εκκίνηση του κόμβου, αναπτύχθηκε επίσης ένα απλό σύνολο adhoc δικτύων αναφοράς, με τη γενικευμένη κατανομή βαθμού, επιτρέποντας έτσι τον έλεγχο της ακριβείας, όταν ο αρχικός κόμβος είναι ένα κομβικό σημείο.

4.3 Κατηγοριοποίηση κόμβου

Με δεδομένο ένα κοινωνικό δίκτυο (ή γενικότερα, οποιαδήποτε δομή δικτύου), με ετικέτες (labels) σε κάποιους κόμβους, τίθεται το ερώτημα πώς μπορούμε να παρέχουμε υψηλής ποιότητας σήμανση (labeling) για κάθε κόμβο. Αναφερόμαστε σε αυτό το ερώτημα ως πρόβλημα *κατηγοριοποίησης κόμβου* (node classification).

Όπως συνηθίζεται στη μηχανική μάθηση, αρχικά πρέπει να εντοπιστούν ορισμένα «χαρακτηριστικά» των κόμβων που μπορούν να χρησιμοποιηθούν για την καθοδήγηση της κατηγοριοποίησης. Τα προφανή χαρακτηριστικά είναι οι ίδιες οι ιδιότητες του κόμβο: πληροφορίες που μπορεί να είναι γνωστές για όλους (ή τους περισσότερους) κόμβους, όπως η ηλικία, η θέση κτλ. Όμως, η παρουσία μιας δομής συνδέσμου καθιστά το πρόβλημα κατηγοριοποίησης κόμβου αρκετά διαφορετικό από τις παραδοσιακές μεθόδους κατηγοριοποίησης της μηχανικής μάθησης όπου τα αντικείμενα τα οποία έχουν κατηγοριοποιηθεί θεωρούνται ανεξάρτητα. Ένα σύνολο χαρακτηριστικών μπορεί να βασίζεται σε απλές ιδιότητες του γραφήματος όπως είναι: ο βαθμός (αριθμός των γειτόνων) του κόμβου, το μέγεθος της γειτονιάς το οποίο είναι προσβάσιμο μέσα από δύο ή τρία βήματα, ο αριθμός των συντομότερων διαδρομών που διασχίζουν τον κόμβο κτλ. Αλλά ίσως τα πιο ενδιαφέροντα χαρακτηριστικά είναι αυτά που προέρχονται από τις ιδιότητες των κοντινών κόμβων.

Ένα ερώτημα που μπορεί να τεθεί είναι για ποιο λόγο οι ετικέτες των γειτονικών κόμβων μπορεί να είναι χρήσιμες για την πρόβλεψη της ετικέτας ενός κόμβου. Ο λόγος είναι ότι κυρίως στα κοινωνικά δίκτυα, οι σύνδεσμοί μεταξύ των κόμβων δεν είναι αυθαίρετοι και συνήθως υποδηλώνουν κάποιας μορφής σχέση μεταξύ των ατόμων που εκπροσωπούν τους κόμβους. Ειδικότερα, ένας σύνδεσμος μπορεί να υποδηλώνει κάποια ομοιότητα μεταξύ των ατόμων που συνδέονται. Οι ετικέτες στους κόμβους μπορεί να είναι διαφόρων τύπων όπως προτείνεται στο [27]:

- *δυαδικές (binary):* μόνο δύο τιμές επιτρέπονται (για παράδειγμα το γένος έχει αυστηρά τις τιμές αρσενικό ή θηλυκό) ή μια ετικέτα μπορεί να εμφανίζει μια θετική τιμή (για παράδειγμα «καπνιστής» και μία αρνητική τιμή («μη καπνιστής»).
- *αριθμητικές (numeric):* οι ετικέτες μπορεί να πάρουν μια αριθμητική τιμή (πχ. ηλικία). Είναι πιθανό να επιτρέπονται μόνο τιμές οι οποίες ανήκουν σε ένα συγκεκριμένο εύρος (για παράδειγμα επιτρεπόμενες τιμές ηλικίας από το εύρος 0-120 ή από τις ηλικιακές ομάδες 0-17; 18-35; 36-50, 50+).
- *κατηγορικές (categorical):* η ετικέτα πρέπει αυστηρά να προέρχεται από ένα σύνολο καθορισμένων κατηγοριών (πχ. ενδιαφέροντα, απασχόληση κτλ.).
- *ελεύθερου κειμένου (free-text):* οι χρήστες μπορούν να εισάγουν αυθαίρετα κείμενο για τον προσδιορισμό των ετικετών που εφαρμόζονται σε έναν κόμβο (σχόλιο / επισήμανση σε μια φωτογραφία).

Μια μεγάλη σειρά προσεγγίσεων στο πρόβλημα της κατηγοριοποίησης δικτύου μπορεί να θεωρηθεί ως προσέγγιση «κεντρικού κόμβου» (node centric) υπό την έννοια ότι σε κάθε χρονική στιγμή εστιάζουμε σε έναν μόνο κόμβο. Για διάφορους λόγους είναι χρήσιμη η ταξινόμηση αυτών των προσεγγίσεων βάσει τριών συστατικών τα οποία παρουσιάζονται παρακάτω [28]:

i. *Μη σχεσιακό μοντέλο* (non-relational-local model): αυτό το συστατικό αποτελείται από ένα μοντέλο εκμάθησης το οποίο χρησιμοποιεί μόνο τοπικές πληροφορίες.
ii. *Σχεσιακό μοντέλο* (relational model): σε αντίθεση με το πρώτο συστατικό, το σχεσιακό μοντέλο χρησιμοποιεί τις σχέσεις του δικτύου καθώς και τις τιμές των χαρακτηριστικών των σχετικών οντοτήτων. Τα σχεσιακά μοντέλα επίσης χρησιμοποιούν τοπικά χαρακτηριστικά των οντοτήτων.
iii. *Συλλογικό συμπέρασμα* (collective inferencing): το μοντέλο συλλογικού συμπεράσματος καθορίζει πως οι άγνωστες τιμές οι οποίες εκτιμώνται μαζί επηρεάζουν η μία την άλλη.

4.3.1 Συλλογική κατηγοριοποίηση

Συχνά, ενδιαφερόμαστε για τον τρόπο με τον οποίο τα αντικείμενα του δικτύου επηρεάζονται μεταξύ τους (πχ., ποιος επηρεάζει ποιον σε ένα δίκτυο), ή είναι πιθανό να θέλουμε να προβλέψουμε ένα χαρακτηριστικό με βάση παρατηρηθέντα χαρακτηριστικά αντικειμένων του δικτύου ή ίσως ενδιαφερόμαστε για την αναγνώριση σημαντικών συνδέσμων του δικτύου. Στα περισσότερα από αυτά τα σενάρια, ένα σημαντικό βήμα στην επίτευξη του τελικού στόχου είναι η κατηγοριοποίηση των αντικειμένων του δικτύου.

Με δεδομένο ένα δίκτυο και ένα αντικείμενο o του δικτύου, υπάρχουν τρεις διακριτοί τύποι συσχετισμών οι οποίοι μπορούν να καθορίσουν την κατηγοριοποίηση ή την ετικέτα του o:

- Οι συσχετισμοί της ετικέτας του o και των χαρακτηριστικών του o που έχουν παρατηρηθεί.
- Οι συσχετισμοί μεταξύ της ετικέτας του o και των χαρακτηριστικών των αντικειμένων στη γειτονιά του o που έχουν παρατηρηθεί.
- Οι συσχετισμοί μεταξύ της ετικέτας του o και των χαρακτηριστικών των αντικειμένων στη γειτονιά του o που δεν έχουν παρατηρηθεί.

Η *συλλογική κατηγοριοποίηση* (collective classification) αναφέρεται σε μια συνδυασμένη κατηγοριοποίηση ενός συνόλου διασυνδεδεμένων αντικειμένων χρησιμοποιώντας τους τρεις τύπους συσχετισμών που περιγράφηκαν στην προηγούμενη ενότητα. Να σημειωθεί, ότι ορισμένες φορές η φράση *σχεσιακή κατηγοριοποίηση* (relational classification) χρησιμοποιείται για να υποδηλώσει μια προσέγγιση η οποία εστιάζει στην κατηγοριοποίηση των δεδομένων του δικτύου χρησιμοποιώντας μόνο τους δύο πρώτους τύπους συσχετισμού.

Η συλλογική κατηγοριοποίηση είναι ένα συνδυαστικό πρόβλημα βελτιστοποίησης στο οποίο δίνονται ένα σύνολο κόμβων $V = \{V_1, \dots, V_n\}$ και μια συνάρτηση «γειτονιάς» N, όπου ισχύει $N_i \subseteq V \backslash \{V_i\}$, περιγράφοντας με αυτό τον τρόπο τη δομή του δικτύου. Κάθε κόμβος στο V είναι μια τυχαία μεταβλητή η οποία μπορεί να πάρει τιμή από μια κατάλληλη περιοχή. Στη συνέχεια, το V διαιρείται σε δύο σύνολα κόμβων: X είναι οι κόμβοι για τους οποίους γνωρίζουμε τις σωστές τιμές και Y είναι οι κόμβοι των οποίων οι τιμές είναι απαραίτητο να καθοριστούν. Στόχος της εργασίας [29] είναι να χαρακτηρισθούν οι κόμβοι $Y_i \in Y$ με μια από τον μικρό αριθμό ετικετών $L = \{L_1, \dots, L_q\}$ χρησιμοποιώντας το συμβολισμό y_i ο οποίος υποδηλώνει την ετικέτα του κόμβου Y_i.

Όπως αναφέρθηκε και παραπάνω, υπάρχει ένας αριθμός προσεγγίσεων στη συλλογική κατηγοριοποίηση. Με μια λογική αφαιρετικότητας, οι προσεγγίσεις αυτές μπορεί να διαιρεθούν σε δύο διακριτούς τύπους, στον πρώτο χρησιμοποιούμε μια συλλογή μη κανονικοποιημένων ταξινομητών και στον δεύτερο ορίζουμε το πρόβλημα της συλλογικής κατηγοριοποίησης σαν μια συνάρτηση αντικειμένου προς βελτιστοποίηση. Στη συνέχεια, παρουσιάζουμε δύο αλγορίθμους αυτών των προσεγγίσεων οι οποίοι χρησιμοποιούνται συχνά. Οι αλγόριθμοι αυτοί είναι ο αλγόριθμος επαναληπτικής κατηγοριοποίησης (*iterative classification algorithm* - ICA) και ο αλγόριθμος *gibbs sampling* (GS).

4.3.2 Επαναληπτική κατηγοριοποίηση

Η βασική ιδέα πίσω από τον αλγόριθμο *επαναληπτικής κατηγοριοποίησης* (iterative classification) ο οποίος αναλύεται στο [29] είναι ιδιαίτερα απλή. Θεωρούμε έναν κόμβο $Y_i \in Y$ την τιμή του οποίου χρειάζεται να καθορίσουμε υποθέτοντας ότι γνωρίζουμε τις τιμές των άλλων κόμβων της γειτονιάς N_i του κόμβου Y_i. Στη συνέχεια, ο αλγόριθμος υποθέτει ότι δίνεται ένας τοπικός ταξινομητής f ο οποίος παίρνει τιμές από το N_i ως ορίσματα και επιστρέφει τις βέλτιστες τιμές του Y_i από το σύνολο ετικετών L. Για τους τοπικούς ταξινομητές f οι οποίοι δεν επιστρέφουν μια ετικέτα κλάσης αλλά μια τιμή

πιθανοφάνειας για ένα δεδομένο σύνολο τιμών χαρακτηριστικών και για μια ετικέτα, απλά επιλέγεται η ετικέτα η οποία αντιστοιχεί στη μέγιστη τιμή πιθανοφάνειας ή με άλλα λόγια αντικαθίσταται η f με $argmax_{l \in L} f$.

Αυτό καθιστά τον τοπικό ταξινομητή f μια ιδιαίτερα ευέλικτη συνάρτηση όπου στη συνέχεια μπορεί να χρησιμοποιηθεί οτιδήποτε στη θέση της από ένα δέντρο απόφασης μέχρι ένας ταξινομητής SVM. Δυστυχώς, στην πραγματικότητα είναι σπάνιο να γνωρίζουμε όλες τις τιμές στο N_i και αυτός είναι ο λόγος για τον οποίο η διαδικασία συμβαίνει επαναληπτικά όπου σε κάθε επανάληψη αναθέτουμε ετικέτες σε κάθε Y_i χρησιμοποιώντας τις τρέχουσες βέλτιστες εκτιμήσεις του N_i και τον τοπικό ταξινομητή f έως ότου οι αναθέσεις των ετικετών να σταθεροποιηθούν.

Algorithm 1 Iterative Classification Algorithm (ICA)

for each node $Y_i \in \mathcal{Y}$ **do** // bootstrapping
  // compute label using only observed nodes in $\mathcal{N}_i$
  compute $\vec{a}_i$ using only $\mathcal{X} \cap \mathcal{N}_i$
  $y_i \leftarrow f(\vec{a}_i)$
end for
repeat // iterative classification
  generate ordering $\mathcal{O}$ over nodes in $\mathcal{Y}$
  for each node $Y_i \in \mathcal{O}$ **do**
    // compute new estimate of y_i
    compute $\vec{a}_i$ using current assignments to $\mathcal{N}_i$
    $y_i \leftarrow f(\vec{a}_i)$
  end for
until all class labels have stabilized or a threshold number of iterations have elapsed

Εικόνα 4-7 Αλγόριθμος (1) επαναληπτικής κατηγοριοποίησης

Οι περισσότεροι τοπικοί ταξινομητές ορίζονται ως συναρτήσεις των οποίων τα ορίσματα αποτελούνται από ορισμένου μήκους διανύσματα τιμών των χαρακτηριστικών. Στην παραπάνω εικόνα ο Αλγόριθμος 1 απεικονίζει τον ψευδοκώδικα του αλγορίθμου επαναληπτικής κατηγοριοποίησης όπου χρησιμοποιεί το $\vec{a_i}$ για να υποδηλώσει το διάνυσμα κωδικοποίησης των τιμών του N_i το οποίο επιτυγχάνεται μετά την συγκέντρωση (aggregation). Να τονίσουμε ότι στην πρώτη επανάληψη του αλγορίθμου όλες οι ετικέτες του y_i είναι απροσδιόριστες και για την αρχικοποίηση τους απλά χρησιμοποιείται ο τοπικός ταξινομητής των χαρακτηριστικών που έχουν παρατηρηθεί στη γειτονιά του Y_i, το οποίο στον αλγόριθμο αναφέρεται ως «bootstrapping».

Μια άλλη πιο γενική προσέγγιση επαναληπτικής κατηγοριοποίησης παρουσιάζεται στο [30]. Οι J. Neville και D. Jensen σε αυτή την εργασία διερευνούν πώς οι συμβατικές τεχνικές για την κατασκευή και τη χρήση μοντέλων κατηγοριοποίησης μπορεί να χρησιμοποιηθούν σε νέους τρόπους για να αξιοποιηθεί η γνώση ενός αντικειμένου με στόχο να προκύψουν συμπεράσματα για άλλα σχετικά αντικείμενα. Συγκεκριμένα, διερευνάται η χρήση απλών μπεϋζιανών ταξινομητών σε μια επαναληπτική διαδικασία με στόχο να βελτιωθεί η ακρίβεια της κατηγοριοποίησης αξιοποιώντας τις σχεσιακές πληροφορίες. Η υπόθεση πίσω από αυτή την προσέγγιση είναι ότι εάν δύο αντικείμενα σχετίζονται τότε συμπεραίνοντας κάτι για ένα αντικείμενο ίσως είμαστε σε θέση να καταλήξουμε σε συμπέρασμα και για το άλλο.

Αυτό σημαίνει ότι υπάρχουν διακριτά χαρακτηριστικά των σχεσιακών δεδομένων που μπορούν να χρησιμοποιηθούν για τη βελτίωση της ακρίβειας της κατηγοριοποίησης. Οι απλοί μπεϋζιανοί ταξινομητές δέχονται ως δεδομένα εισόδου παραδοσιακά δεδομένα τύπου *ιδιότητα-τιμή*. Στην επαναληπτική κατηγοριοποίηση, ένα μοντέλο δημιουργείται χρησιμοποιώντας μια ποικιλία από στατικά και δυναμικά χαρακτηριστικά. Κατά την εκπαίδευση του μοντέλου, οι κατηγορικές ετικέτες όλων των αντικειμένων

και κατά συνέπεια, οι τιμές όλων των δυναμικών χαρακτηριστικών είναι γνωστές. Στην παρακάτω εικόνα παρουσιάζονται τα βασικά βήματα του αλγορίθμου επαναληπτικής κατηγοριοποίησης:

Iterative Classification Algorithm
1. Build model on fully labeled training set
2. Apply trained model to test set of *N* instances. For each iteration *i* : 1 to *m*
   a. Calculate values for dynamic relational attributes
   b. Use model to predict class labels
   c. Sort inferences by probability
   d. Accept *k* class labels, where *k* = *N* (*i* / *m*)
3. Output final inferences made by model on test set

Εικόνα 4-8 Αλγόριθμος επαναληπτικής κατηγοριοποίησης

Ορισμένα συμπεράσματα μπορούν να προκύψουν σχετικά με τις δυνατότητες της επαναληπτικής κατηγοριοποίησης. Έχει αποδειχθεί ότι υπάρχει μια ευκαιρία να χρησιμοποιηθούν οι σχέσεις των δεδομένων με στόχο να αυξηθεί η ακρίβεια της κατηγοριοποίησης με αποτέλεσμα μια επαναληπτική προσέγγιση να είναι σε θέση να εκμεταλλευτεί την ευκαιρία αυτή παράγοντας μια σημαντική βελτίωση στην ακρίβεια της δυαδικής κατηγοριοποίησης σε ένα σύνολο δεδομένων.

4.3.3 Μέδοδος Gibbs sampling

Η μέθοδος gibbs sampling (GS) όπως παρουσιάζεται στο [29], θεωρείται ευρέως ως μία προσέγγιση που παρέχει μεγάλη ακρίβεια στη διεξαγωγή συμπερασμάτων. Αρχικά, προτάθηκε στο πλαίσιο της αποκατάστασης εικόνας. Παρόλο κάποια μειονεκτήματα της μεθόδου χρησιμοποιείται στη συλλογική κατηγοριοποίηση μέσω μιας απλοποιημένης εκδοχής θεωρώντας ότι έχουμε πρόσβαση σε έναν τοπικό ταξινομητή f ο οποίος χρησιμοποιείται για την εκτίμηση της πιθανότητας των ετικετών του Y_i για όλες τιμές των κόμβων στο N_i. Η βασική ιδέα είναι να κάνουμε δειγματοληψία για την καλύτερη εκτίμηση της ετικέτας του Y_i με δεδομένο όλες τις τιμές για τους κόμβους στο N_i χρησιμοποιώντας έναν τοπικό ταξινομητή f για έναν συγκεκριμένο αριθμό επαναλήψεων (αυτή η περίοδος αναφέρεται ως “burn-in”). Μετά από αυτό, όχι μόνο έχουμε δείγμα για τις ετικέτες για κάθε $Y_i \in Y$, αλλά διατηρούμε επίσης στατιστικά στοιχεία για το πόσες φορές έχουμε κάνει δειγματοληψία την ετικέτα l για τον κόμβο Y_i. Μετά τη συλλογή ενός προκαθορισμένου αριθμού δειγμάτων, έχουμε την καλύτερη ανάθεση ετικέτας για τον κόμβο Y_i επιλέγοντας την ετικέτα εκείνη η οποία ανατέθηκε τις περισσότερες φορές στο Y_i, κατά τη διάρκεια της συλλογής των δειγμάτων.

4.3.4 Κατηγοριοποίηση βάσει τύπων συνδέσμου

Τα κοινωνικά δίκτυα επιτρέπουν στους χρήστες να κοινοποιούν προσωπικές πληροφορίες αλλά επίσης τους επιτρέπουν και να μην κοινοποιούν. Εντούτοις, οι «κρυμμένες» αυτές λεπτομέρειες μπορεί να φανούν ιδιαίτερα σημαντικές στους διαχειριστές ενός κοινωνικού δικτύου. Οι περισσότερες από αυτές τις ιστοσελίδες είναι ελεύθερες για τον τελικό χρήστη με αποτέλεσμα να αποτελούν χώρο προβολής διαφημίσεων. Εάν υποθέσουμε ότι οι διαφημιστές θέλουν να εντοπίσουν άτομα τα οποία είναι πιθανό να ενδιαφέρονται για τα προϊόντα τους, τότε ο εντοπισμός αυτών των ατόμων γίνεται προτεραιότητα για τη διατήρηση περισσότερων διαφημιστών στην ιστοσελίδα.

Για την αντιμετώπιση του παραπάνω προβλήματος, στην εργασία [31] επιλέγεται ο network-only Naive Bayes ταξινομητής ως σημείο αφετηρίας, εφόσον είναι αποδεδειγμένο ότι αυτός ο ταξινομητής σε συνδυασμό με τις συλλογικές τεχνικές συμπερασμάτων παρέχει μια αποτελεσματική λύση με ικανοποιητική ακρίβεια στην πράξη. Στη συνέχεια, τροποποιείται αυτός ο ταξινομητής έτσι ώστε να συμπεριλάβει τον «τύπο» του συνδέσμου μεταξύ των κόμβων για τους υπολογισμούς της πιθανότητας. Οι μέθοδοι που παρουσιάζονται παρακάτω βασίζονται όλες στον παραδοσιακό naive Bayes ταξινομητή.

1η μέθοδος: Local Classification

Η χρήση ενός απλού μπεϋζιανού ταξινομητή είναι μια αποτελεσματική μέθοδος διορατικότητας προσωπικών πληροφοριών για ένα κοινωνικό δίκτυο. Εξαιτίας αυτού του χαρακτηριστικού και των σχετικά χαμηλών υπολογιστικών απαιτήσεων χρησιμοποιείται ο βασικός Naive Bayes ταξινομητής με στόχο τον καθορισμό της πιθανότητας ότι ένας συγκεκριμένος χρήστης x_i, είναι της συγκεκριμένης κατηγορίας c_i με δεδομένο ολόκληρο το σύνολο των χαρακτηριστικών T. Η πιθανότητα αυτή δίνεται από τον παρακάτω τύπο [31]:

$$\Pr(x_i = c_i|T) =$$

$$\frac{\Pr(T|x_i = c_i)\Pr(x_i = c_i)}{\Pr(T)} =$$

$$\Pr(x_i = c_i) \times \prod_{t_i \in T} \frac{\Pr(t_i|x_i = c_i)}{\Pr(t_i)}$$

2η μέθοδος: Network-only Bayes Classification

Αυτός είναι ο πιο βασικός από τους τρεις σχεσιακούς ταξινομητές που εξετάζονται στην εργασία [31]. Ο γενικός Network-only Bayes (nBC) ταξινομητής θεωρεί ότι όλοι οι τύποι συνδέσμων είναι οι ίδιοι συνεπώς η πιθανότητα της κατηγορίας ενός συγκεκριμένου κόμβου επηρεάζεται από τις κατηγορίες των γειτόνων του. Χρησιμοποιείται ένας τοπικός ταξινομητής για την ανάθεση κάθε κόμβου σε ένα σύνολο δεδομένων ελέγχου. Για τον υπολογισμό της πιθανότητας χρησιμοποιείται ο παρακάτω τύπος [31]:

$$\Pr(x_i = c_i|N) =$$

$$\frac{\Pr(N|x_i = c_i)\Pr(x_i = c_i)}{\Pr(N)} =$$

$$\Pr(x_i = c_i) \times \prod_{n_i \in N} \frac{\Pr(n_i|x_i = c_i)}{\Pr(n_i)}$$

3η μέθοδος: Link Type nBC

Ο επόμενος σχεσιακός ταξινομητής που εξετάζεται σε αυτή την εργασία είναι ο πρώτος που συμπεριλαμβάνει τους τύπους συνδέσμων σαν μια επιπρόσθετη παράμετρο στην κατηγοριοποίηση κατά Naive Bayes. Συνεπώς, ορίζεται η πιθανότητα οποιουδήποτε κόμβου x_i να είναι σε μια συγκεκριμένη κατηγορία c_i, ως $\Pr(x_i = c_i|N)$. Η πιθανότητα οποιουδήποτε κόμβου να ανήκει σε μια κατηγορία η οποία καθορίζεται από τους γείτονές του και από το σύνολο των συνδέσμων που καθορίζουν αυτοί οι γείτονες είναι $\Pr(x_i = c_i|N, L)$ και υπολογίζεται ως εξής [31]:

$$\Pr(x_i = c_i|N, L) =$$

$$\frac{\Pr(N, L|x_i = c_i)\Pr(x_i = c_i)}{\Pr(N, L)} =$$

$$\prod_{\substack{n_i \in N \\ l_i \in L}} \frac{\Pr(n_i, l_i | x_i = c_i)\Pr(x_i = c_i)}{\Pr(n_i, l_i)}$$

4η μέθοδος: Weighted Link Type rBC

Ο τελευταίος σχεσιακός ταξινομητής θεωρεί ότι ορισμένοι τύποι συνδέσμων είναι ενδεικτικοί μιας δυνατής σχέσης. Για παράδειγμα το σύνολο δεδομένων που χρησιμοποιείται στην εργασία [31] περιλαμβάνει δεδομένα για όλο το προσωπικό που εργάζεται σε μια τηλεοπτική σειρά. Αυτό το σύνολο δεδομένων περιλαμβάνει σκηνοθέτες, παραγωγούς, ηθοποιούς, ενδυματολόγους, ειδικούς τεχνικών εφέ κτλ. θεωρώντας ότι όλες αυτές οι ειδικότητες δεν είναι εξίσου σημαντικές. Συνεπώς, στο προηγούμενο μοντέλο ταξινομητή το Link Type rBC προστίθενται βάρη με αποτέλεσμα η πιθανότητα του κόμβου να ανήκει σε μια κατηγορία και να δίνεται από τον τύπο [31]:

$$\Pr(x_i = c_i | N, L) =$$

$$\prod_{\substack{n_i \in N \\ l_i \in L}} [\frac{w_i}{W} \times \frac{\Pr(n_i, l_i | x_i = c_i)\Pr(x_i = c_i)}{\Pr(n_i, l_i)}]$$

όπου w_i είναι το βάρος που σχετίζεται με τον τύπο συνδέσμου l_i και W είναι το άθροισμα όλων των βαρών του δικτύου.

5η μέθοδος: Relaxation Labeling

Στη συνέχεια, παρουσιάζεται η relaxation labeling ως μια μέθοδος η οποία διατηρεί την αβεβαιότητα των ετικετών που έχουν κατηγοριοποιηθεί. Η μέθοδος αυτή είναι μια επαναληπτική διαδικασία, όπου σε κάθε βήμα $i + 1$ ο αλγόριθμος χρησιμοποιεί τις εκτιμήσεις πιθανότητας και όχι μια μοναδική ετικέτα, από το βήμα i για τον υπολογισμό των εκτιμήσεων πιθανότητας.

Ανακεφαλαιώνοντας, στην εργασία [31] παρουσιάστηκαν δύο βασικές βελτιώσεις του παραδοσιακού σχεσιακού μπεϋζιανού ταξινομητή οι οποίες ονομάζονται Link Type rBC και Weighted Link Type rBC και χαρακτηρίζονται ως δύο επιτυχημένες επεκτάσεις επαρκώς τεκμηριωμένες στον τομέα της κατηγοριοποίησης σε κοινωνικά δίκτυα.

4.3.5 Κατηγοριοποίηση βάσει σχέσεων

Μια υπόθεση η οποία έχει γίνει σε πολλές εργασίες που αφορούν εξόρυξη γνώσης σε κοινωνικά δίκτυα είναι ότι το κοινωνικό δίκτυο G τυπικά περιέχει μια μόνο σχέση. Ουσιαστικά, υποστηρίζεται ότι οι σχέσεις μεταξύ διαφορετικών χρηστών είναι όμοιες από τη στιγμή που έχουν ένα σύνδεσμο μεταξύ τους. Για παράδειγμα, όλοι οι σύνδεσμοι μπορεί να θεωρηθούν ως «γνωριμία». Με άλλα λόγια, όλοι οι σύνδεσμοι έχουν μια κοινή ετικέτα «γνωριμία». Εάν βέβαια αναζητήσουμε στις σχέσεις ανάμεσα στους χρήστες θα έχουμε περισσότερες τιμές όπως «συμφοιτητές», «συμμαθητές», «συνάδελφοι» κτλ.

Θεωρούμε ως δεδομένο ένα κοινωνικό δίκτυο $G = (V, E)$, μια ετικέτα $l()$ για ένα μικρό υποσύνολο $K \subset E$ των ακμών και το σύνολο όλων των πιθανών ετικετών C το οποίο μπορεί να χρησιμοποιηθεί. Συνεπώς, έχουμε μια ετικέτα $l(e)$ της ακμής $e = (u, v)$ η οποία υποδηλώνει ότι η σχέση μεταξύ του u και του v είναι τύπου $l(e)$.

Ο στόχος της εργασίας [32] είναι η ανάθεση σε κάθε σύνδεσμο e στο E μιας ετικέτας από το C έτσι ώστε να μεγιστοποιείται η ακρίβεια. Ορίζουμε την ακρίβεια ως την αναλογία του αριθμού των συνδέσμων που

έχουν σωστές ετικέτες προς το συνολικό αριθμό των συνδέσμων χωρίς ετικέτες. Συνεπώς, τυποποιούμε ένα κοινωνικό δίκτυο ως ένα γράφημα $G = (V, E)$, στο οποίο το V είναι ένα σύνολο χρηστών του κοινωνικού δικτύου και E είναι ένα σύνολο ακμών μεταξύ των χρηστών. Θεωρούμε ότι οι σύνδεσμοι του κοινωνικού δικτύου μπορούν να ταξινομηθούν σε κατηγορίες $\mathcal{L} = \{l_0, l_1, l_2, \dots, l_k\}$ τέτοιες ώστε *{friends, classmates, officemates, family, others}*. Με άλλα λόγια, κάθε σύνδεσμος $e \in E$ έχει μια από τις πολλές ετικέτες $l(e) \in \mathcal{L}$.

Χρησιμοποιούμε το $G(V')$ για να υποδηλώσουμε έναν υπογράφο ο οποίος προκύπτει από το σύνολο των κορυφών του $V' \subseteq V$, με το $N(v)$ να υποδηλώνει το σύνολο των γειτόνων του v στο G. Στην εργασία αυτή το $| \mathcal{S} |$ υποδηλώνει το μέγεθος του συνόλου $\mathcal{S}$. Στη συνέχεια, παρουσιάζεται η βασική ιδέα που κρύβεται πίσω από τον αλγόριθμο *κατηγοριοποίησης βάσει σχέσεων* (relationship classification algorithm - RCA).

Για έναν χρήστη $v \in V$, ο υπογράφος $G(N(v))$ τείνει να διαμορφώνει ασυνεχείς κοινότητες ή συστάδες όπου ισχύουν βάσει του [32] τα παρακάτω:

- οι χρήστες οι οποίοι έχουν την ίδια σχέση l_i με τον v τείνουν να ανήκουν στην ίδια συστάδα του $G(N(v))$ και επίσης τείνουν να έχουν την ίδια σχέση l_i μεταξύ τους και
- οι χρήστες οι οποίοι έχουν διαφορετικές σχέσεις με τον v είναι πιο πιθανό να ανήκουν σε διαφορετικές συστάδες του $G(N(v))$.

Στη συνέχεια, αναζητούμε έναν τρόπο εύρεσης ετικετών για όλους τους συνδέσμους τέτοιες ώστε να διατηρηθούν οι δύο ιδιότητες που αναφέρθηκαν παραπάνω. Για ένα χρήστη v, έχουμε $V_i \subseteq (v)$ να υποδηλώνει ένα σύνολο χρηστών οι οποίοι έχουν την ίδια σχέση l_i με τον v να αναφέρεται στους συνδέσμους πριν την ανάθεση των συνδέσμων. Να υπενθυμίσουμε ότι οι χρήστες οι οποίοι έχουν τις ίδιες σχέσεις l_i, με τον v τείνουν να ανήκουν στην συστάδα $N(v)$. Συνεπώς, αν είμαστε σε θέση να ανακαλύψουμε το σύνολο των χρηστών που ανήκουν στην ίδια συστάδα V_i, μπορούμε να ισχυριστούμε ότι όλοι οι χρήστες έχουν την ίδια σχέση l_i με τον v.

Για να εντοπίσουμε εκείνους τους χρήστες που ανήκουν στην ίδια συστάδα V_i, εκτελούμε τυχαίους «περιπάτους» αρχίζοντας από κάθε χρήστη στο V_i με στόχο να αξιολογηθεί η περιοχή που έχει επηρεαστεί (affecting range) από το V_i. Έτσι, χρησιμοποιούμε την πιθανότητα ότι τουλάχιστον ένας από τους τυχαίους περιπάτους επισκέπτεται κάποιον χρήστη μέσω κάποιων σταθερών βημάτων με στόχο τη μέτρηση της επίδρασης από το V_i στον χρήστη. Όσο περισσότερο ένας χρήστης $w \in (v)$ επηρεάζεται από το V_i, τόσο πιο πιθανό είναι να ανήκει στην ίδια συστάδα που δημιουργήθηκε από τον V_i. Στη συνέχεια, μπορούμε να αναθέσουμε την ετικέτα l_i στον σύνδεσμο wv, $l(wv) = l_i$, με μεγάλη αυτοπεποίθηση.

4.3.6 Κατηγοριοποίηση βάσει συνδέσμων

Στην εργασία [33] παρουσιάζεται ένα σύνολο δεδομένων D το οποίο αποτελείται από D^l δεδομένα με ετικέτες και D^u δεδομένα χωρίς ετικέτες όπου καθορίζεται μια εκ των υστέρων πιθανότητα του D^u ως:

$$P(c(X): X \in D^u | D) =$$

$$\prod_{X \in D^u} P(c(X) | OA(X), LD(X))$$

Χρησιμοποιείται επίσης ένας τύπου EM (Expecation-Maximization) επαναληπτικός αλγόριθμος για να χρησιμοποιηθούν τόσο τα δεδομένα $D^l = \{(x_i, c(x_i): i = 1, \dots, n\}$ όσο και τα δεδομένα $D^u = \{(x_j^*, c(x_j^*): j = 1, \dots, m\}$ για την εκμάθηση του μοντέλου. Αρχικά, ένα δομημένο λογικό μοντέλο

παλινδρόμησης δημιουργείται χρησιμοποιώντας τα δεδομένα D^l. Στη συνέχεια, κατηγοριοποιούμε τα δεδομένα στο D^u ως εξής [33]:

$$c(x_j^*) = argmax_{c \in O} \frac{P\left(c \middle| OA(x_j^*)\right) P(c|LD(x_j^*))}{P(c)}$$

όπου $j = 1, \ldots, m$. Χρησιμοποιώντας τα κατηγοριοποιημένα δεδομένα D^u και τα δεδομένα D^l δημιουργούμε ένα νέο μοντέλο μέσω των παρακάτω βημάτων:

Βήμα 1: (Αρχικοποίηση) Δημιουργούμε ένα δομημένο ταξινομητή παλινδρόμησης χρησιμοποιώντας το περιεχόμενο και τα χαρακτηριστικά μόνο των δεδομένων εκπαίδευσης με ετικέτες.

Βήμα 2: (Επανάληψη) Επαναλαμβάνουμε τη διαδικασία μέχρι η πιθανότητα των δεδομένων δοκιμής χωρίς ετικέτες να αυξηθεί ως εξής:

i. Ταξινομούμε τα δεδομένα χωρίς ετικέτες με τη χρήση του τρέχοντος μοντέλου

ii. Επαναϋπολογίζουμε τα χαρακτηριστικά συνδέσμου κάθε αντικειμένου.

iii. Επανεκτιμούμε τις παραμέτρους του μοντέλου παλινδρόμησης.

Στην κατηγοριοποίηση που βασίζεται στους συνδέσμους, τα δεδομένα χωρίς ετικέτες παρέχουν χρήσιμες πληροφορίες με τρεις σημαντικούς τρόπους: αρχικά μας δίνουν επιπρόσθετες πληροφορίες σχετικά με την κατανομή τιμών των χαρακτηριστικών του αντικειμένου, δεύτερον οι σύνδεσμοι μεταξύ των δεδομένων χωρίς ετικέτες στο σύνολο δοκιμής παρέχουν χρήσιμες πληροφορίες κατηγοριοποίησης και τρίτον οι σύνδεσμοι μεταξύ των δεδομένων εκπαίδευσης με ετικέτες και των δεδομένων δοκιμής χωρίς ετικέτες παρέχουν επίσης χρήσιμες πληροφορίες που δεν πρέπει να αγνοηθούν. Όταν το πρόβλημα κατηγοριοποίησης μοντελοποιείται κατάλληλα, δεν αλλοιώνονται τα δεδομένα μέσω της κατάργησης συνδέσμων μεταξύ των δεδομένων δοκιμής και των δεδομένων εκπαίδευσης τα οποία χρησιμοποιούνται στη συλλογική κατηγοριοποίηση διότι είμαστε σε θέση να χρησιμοποιήσουμε όλες τις διαθέσιμες πληροφορίες που παρέχουν τα δεδομένα χωρίς ετικέτες.

4.3.7 Σχετικές εργασίες

Τα κοινωνικά δίκτυα έχουν δημιουργήσει μεγάλες προσδοκίες που συνδέονται με τη δυνητική επιχειρηματική τους αξία. Ο σκοπός της εργασίας [34] είναι να παρουσιάσει ότι ακόμη και μια υποτυπώδης εφαρμογή τεχνικών εξόρυξης γνώσης μπορεί να αποφέρει στατιστικά σημαντική βελτίωση στην απόκριση μέσω marketing ενεργειών. Η προσέγγιση αφορά τη μέθοδο CART (classification and regression tree) η οποία χρησιμοποιείται για να προκύψει ένα δέντρο κατηγοριοποίησης το οποίο θα επιτρέπει τη διατύπωση ορισμένων ειδικών κανόνων σε ότι αφορά τον προσδιορισμό της κατάλληλης ομάδας στόχου (target group).

Στην εργασία [35] εξάγονται δομές κοινωνικού δικτύου από δεδομένα συζητήσεων (chat data) χρησιμοποιώντας μερικές βασικές ευριστικές. Στη συνέχεια, παρουσιάζονται ορισμένα προκαταρκτικά αποτελέσματα τα οποία δείχνουν ότι το προκύπτον κοινωνικό γράφημα μπορεί να χρησιμοποιηθεί για τον εντοπισμό της αναγνώρισης θεμάτων σε ένα chat room όταν συνδυαστεί με μοντέλα κατηγοριοποίησης.

Στην εργασία [36] παρουσιάζεται μια διαδικασία μεταφοράς από την εξαγωγή απόψεων στον εντοπισμό συναισθημάτων χρησιμοποιώντας μια μελέτη περίπτωσης των σχολίων του κοινωνικού δικτύου MySpace. Ένας κατηγοριοποιητής δημιουργείται για την ποσοτικοποίηση του βαθμού στον οποίο θετικά και αρνητικά συναισθήματα εκφράζονται σε κάθε σχόλιο χωρίς να λαμβάνεται υπόψη το περιεχόμενο των προηγούμενων σχολίων. Η μελέτη αυτή έδειξε ότι το συναίσθημα είναι προφανώς ο κανόνας σε ιστοσελίδες κοινωνικής δικτύωσης και κατά συνέπεια μελλοντικές έρευνες θα πρέπει να δώσουν ιδιαίτερη προσοχή στη θετικά συναισθηματική έκφραση και το ρόλο των δύο φύλων σε αυτό.

Στην εργασία [37] εισάγεται μια προσέγγιση εκμάθησης με ημι-επίβλεψη η οποία βασίζεται σε ένα τυχαίο Gaussian μοντέλο το οποίο ορίζεται με σεβασμό σε ένα σταθμισμένο γράφημα παρουσιάζοντας δεδομένα με ετικέτες και δεδομένα χωρίς ετικέτες. Τα αποτελέσματα των πειραμάτων ήταν ιδιαίτερα θετικά για την κατηγοριοποίηση κειμένου και ψηφίων, αποδεικνύοντας ότι αυτή η προσέγγιση έχει τη δυνατότητα να εκμεταλλευθεί αποτελεσματικά τη δομή των δεδομένων χωρίς ετικέτες για τη βελτίωση της ακρίβειας της κατηγοριοποίησης.

Το ηλεκτρονικό ταχυδρομείο είναι ένα από τα πιο διαδεδομένα σήμερα επικοινωνιακά εργαλεία με αποτέλεσμα η επίλυση του προβλήματος υπερφόρτωσης των ηλεκτρονικών μηνυμάτων να είναι ιδιαίτερα σημαντική. Ένας καλός τρόπος για την επίλυση αυτού του προβλήματος είναι να δοθεί αυτόματα προτεραιότητα στα μηνύματα σύμφωνα με τις προτεραιότητες του κάθε χρήστη. Η εργασία [38] παρουσιάζει μια μελέτη κάτω από μια τέτοια υπόθεση. Ειδικότερα, επικεντρώνεται στην ανάλυση των προσωπικών κοινωνικών δικτύων για να συλλάβει ομάδες χρηστών και να συγκεντρώσει χαρακτηριστικά που εκπροσωπούν τους κοινωνικούς ρόλους από τη σκοπιά ενός συγκεκριμένου χρήστη. Παράλληλα, αναπτύχθηκε ένας αλγόριθμος εκμάθησης με ημι-επίβλεψη που διαδίδει χαρακτηρισμούς από τα παραδείγματα εκπαίδευσης στα παραδείγματα ελέγχου μέσω μηνυμάτων και κόμβων χρηστών σε ένα προσωπικό δίκτυο ηλεκτρονικού ταχυδρομείου. Αυτές οι μέθοδοι δίνουν τη δυνατότητα να αποκτηθεί μια εμπλουτισμένη εικόνα του διανύσματος κάθε νέου μηνύματος ηλεκτρονικού ταχυδρομείου, το οποίο αποτελείται από τα βασικά χαρακτηριστικά ενός μηνύματος ηλεκτρονικού ταχυδρομείου (όπως οι λέξεις του τίτλου ή το σώμα του μηνύματος, οι ταυτότητες του αποστολέα και του παραλήπτη κ.λπ.) καθώς και από τα κοινωνικά χαρακτηριστικά του αποστολέα και των παραληπτών του μηνύματος. Τέλος, χρησιμοποιήθηκε το εμπλουτισμένο αυτό διάνυσμα ως είσοδος σε έναν SVM ταξινομητή με σκοπό να προβλεφθεί το επίπεδο σημαντικότητας κάθε μηνύματος.

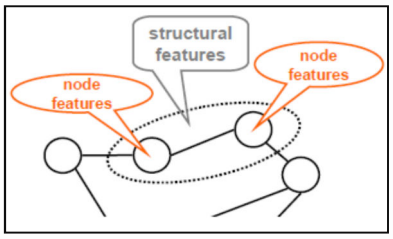
Στην εργασία [28] παρουσιάζεται το NetKit ένα toolkit το οποίο σχεδιάστηκε για κατηγοριοποίηση σε δεδομένα δικτύου. Το κόμβο-κεντρικό του πλαίσιο αποτελείται από τρία βασικά συστατικά: τον σχεσιακό ταξινομητή, τον τοπικό ταξινομητή και το συλλογικό συμπέρασμα (collective inference). Κάθε τοπικός ταξινομητής μπορεί να συνδυαστεί με κάθε σχεσιακό ταξινομητή όπου στη συνέχεια συνδυάζονται με οποιαδήποτε μέθοδο συλλογικού συμπεράσματος.

4.4 Πρόβλεψη συνδέσμου

Τα κοινωνικά δίκτυα είναι αντικείμενα υψηλής δυναμικής, αναπτύσσονται και αλλάζουν με γρήγορους ρυθμούς μέσω της πρόσθεσης ακμών, επισημαίνοντας την εμφάνιση νέων αλληλεπιδράσεων στη βαθύτερη κοινωνική δομή. Ένα θεμελιώδες ερώτημα που ακόμα δεν έχει κατανοηθεί πλήρως είναι η αντίληψη των μηχανισμών μέσω των οποίων εξελίσσονται τα δίκτυα αυτά και αποτελεί το κίνητρο διαφόρων ερευνών.

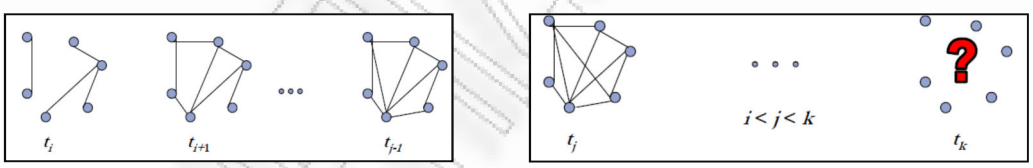
Ορίζουμε και μελετάμε σε αυτήν την ενότητα ένα βασικό υπολογιστικό πρόβλημα της εξέλιξης των κοινωνικών δικτύων, το πρόβλημα *πρόβλεψης συνδέσμου* (link prediction): δεδομένου ενός στιγμιότυπου ενός κοινωνικού δικτύου σε μια χρονική στιγμή t προσπαθούμε να προβλέψουμε επακριβώς τις ακμές που θα προστεθούν στο δίκτυο κατά τη διάρκεια του μεσοδιαστήματος από τη στιγμή t έως μια δεδομένη μελλοντική στιγμή t_0. Κατά την εφαρμογή το πρόβλημα πρόβλεψης συνδέσμου θέτει το εξής ερώτημα: *σε ποιά έκταση μπορεί η εξέλιξη ενός κοινωνικού δικτύου να μοντελοποιηθεί με τη χρήση εσωτερικών στοιχείων του ίδιου του δικτύου;*

Υπάρχουν διάφοροι τύποι γνωστών μεθόδων πρόβλεψης συνδέσμου οι οποίες βασίζονται είτε στα χαρακτηριστικά κόμβων είτε σε δομικά χαρακτηριστικά οι οποίες αναλύονται στη συνέχεια.



Εικόνα 4-9 Χαρακτηριστικά κόμβων vs. Δομικών χαρακτηριστικών

Δίνοντας ένα στιγμιότυπο ενός κοινωνικού δικτύου τη χρονική στιγμή t, το πρόβλημα της πρόβλεψης συνδέσμου προσπαθεί να προβλέψει με ακρίβεια τις ακμές που θα προστεθούν στο δίκτυο κατά τη διάρκεια από τη χρονική στιγμή t στη χρονική στιγμή t' όπως φαίνεται γραφικά στην εικόνα που ακολουθεί.



Εικόνα 4-10 Οπτική αναπαράσταση της πρόβλεψης συνδέσμου

Οι L. Nowell και J. Kleinberg στο [40] πρότειναν ένα από τα πρώτα μοντέλα πρόβλεψης συνδέσμου το οποίο έχει εφαρμογή σε ένα κοινωνικό δίκτυο και αναλύεται παρακάτω. Κάθε κορυφή στο γράφημα αντιπροσωπεύει ένα άτομο και μια ακμή μεταξύ δυο κορυφών αντιπροσωπεύει την αλληλεπίδραση μεταξύ δύο ατόμων. Η διαδικασία της εκμάθησης σε αυτό το μοντέλο τυπικά εξάγει την ομοιότητα μεταξύ ενός ζεύγους κορυφών βάσει διαφόρων μετρικών ομοιότητας και κατατάσσει τις βαθμολογίες ομοιότητας με στόχο την πρόβλεψη της σχέση ανάμεσα σε δύο κορυφές.

Αργότερα, οι Hasan et. al. στο [41] επεκτείνουν την παραπάνω εργασία με δύο τρόπους. Αρχικά, έδειξαν ότι η χρήση εξωτερικών δεδομένων εκτός του πεδίου εφαρμογής της τοπολογίας του γράφου μπορεί να βελτιώσουν σημαντικά το αποτέλεσμα της πρόβλεψης. Δεύτερον, χρησιμοποίησαν διάφορες μετρικές ομοιότητας ως χαρακτηριστικά σε μια διαδικασία εκμάθησης με επίβλεψη, όπου το πρόβλημα

πρόβλεψης συνδέσμου αντιμετωπίζεται ως δυαδική κατηγοριοποίηση. Η εργασία αυτή υπήρξε αφετηρία ώστε να θεωρείται η κατηγοριοποίηση με επίβλεψη δημοφιλής προσέγγιση στο πρόβλημα της πρόβλεψης συνδέσμου.

4.4.1 Μέθοδοι πρόβλεψης συνδέσμου

Οι L. Nowell και J. Kleinberg στο [40] ερευνούν μια σειρά μεθόδων οι οποίες σχετίζονται με την πρόβλεψη συνδέσμου και αναλύονται στη συνέχεια. Όλες οι μέθοδοι προσδιορίζουν μία σύνδεση βάρους $score(x; y)$ σε ζευγάρια κόμβων (x, y) βασισμένες στο γράφημα εισόδους G_{collab} και στη συνέχεια παράγουν μια λίστα κατάταξης $score(x, y)$ σε φθίνουσα σειρά.

***Μέθοδοι οι οποίοι βασίζονται σε γειτονικούς κόμβους*:** για έναν κόμβο x, έστω ότι $\Gamma(x)$ δείχνει το σύνολο των γειτόνων του x στο G_{collab}. Διάφορες προσεγγίσεις βασίζονται στο ότι δύο κόμβοι x και y είναι πιο πιθανό να σχηματίσουν έναν σύνδεσμο στο μέλλον εάν τα γειτονικά σύνολά τους $\Gamma(x)$ και $\Gamma(y)$ παρουσιάζουν μεγάλη επικάλυψη [40]:

- *Κοινοί γείτονες*: η πιο άμεση υλοποίηση της ιδέας που αφορά την πρόβλεψη συνδέσμου είναι ο ορισμός του $score(x, y) = |\Gamma(x) \cap \Gamma(y)|$, ο αριθμός των κοινών γειτόνων των x και y.
- *Συντελεστής Jaccard*: ο συντελεστής Jaccard είναι μια ευρέως χρησιμοποιούμενη μετρική ομοιότητας στην ανάκτηση πληροφοριών ο οποίος μετρά την πιθανότητα τόσο ο x όσο και ο y να έχουν ένα κοινό χαρακτηριστικό γνώρισμα f, όπου f ένα τυχαία επιλεγμένο χαρακτηριστικό είτε του x είτε του y.

  $score(x, y) \coloneqq |\Gamma(x) \cap \Gamma(y)|/|\Gamma(x) \cup \Gamma(y)|$.
- *Adamic και Adar*: εξετάζει ένα σχετικό μέτρο, έτσι ώστε οι κοινοί γείτονες (common neighbors) να θεωρούνται ως χαρακτηριστικά:

  $$score(x, y) \coloneqq \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

Μέθοδοι οι οποίοι βασίζονται στο σύνολο των μονοπατιών: ένας αριθμός μεθόδων οι οποίοι στηρίζονται στη θεωρία του μικρότερου μονοπατιού λαμβάνοντας υπόψη το σύνολο όλων των μονοπατιών μεταξύ των δύο κόμβων [40]:

- *Katz:* ορίζεται ως το μέτρο το οποίο αθροίζει τα βάρη όλων των μονοπατιών μεταξύ δύο κόμβων μειώνοντας εκθετικά από το μήκος:

  $score(x, y) \coloneqq \sum_{l=1}^{\infty} b^l \cdot |paths_{x,y}^{<l>}|$, όπου $paths_{x,y}^{<l>}$ είναι το σύνολο όλων των μονοπατιών μήκους l από το x στο y.
- *Hitting time, Commute time:* το μέτρο hitting time $H_{x,y}$ από το x στο y είναι ο αναμενόμενος αριθμός των βημάτων που απαιτούνται για ένα τυχαίο πέρασμα ξεκινώντας από το x και φτάνοντας στο y. Από τη στιγμή που ο hitting time δεν είναι σε γενικές γραμμές συμμετρικός, είναι φυσικό να απαιτείται ο υπολογισμός του commute time ως $C_{x,y} \coloneqq H_{x,y} + H_{y,x}$. Τα δυο αυτά μέτρα αποτελούν φυσικό μέτρο ομοιότητας και μπορούν να χρησιμοποιηθούν σαν $score(x, y)$.
- *SimRank*: είναι ένα σταθερό σημείο για το οποίο ισχύει ότι δύο κόμβοι είναι όμοιοι ανάλογα το βαθμό με τον οποίο είναι συνδεδεμένοι σε όμοιους γείτονες. Μαθηματικά αυτό ορίζεται ως:

  $similarity(x, x). \coloneqq 1$ και

$$similarity(x,y) \coloneqq \gamma \cdot \frac{\sum_{a \epsilon \Gamma(x)} \sum_{b \epsilon \Gamma(y)} similarity\ (a,b)}{|\Gamma(x)| \cdot |\Gamma(y)|}$$

για κάποιο γ ∈ [0,1]. Ορίζουμε επίσης το $score(x,y) \coloneqq similarity(x,y)$.

Υψηλού επιπέδου προσεγγίσεις: αναλύουμε στη συνέχεια τρεις «μετά-προσεγγίσεις» οι οποίες μπορούν να χρησιμοποιηθούν σε συνδυασμό με οποιαδήποτε από τις παραπάνω μεθόδους σύμφωνα με το [40]:

- *Low-rank προσέγγιση*: δεδομένου ότι ο πίνακας γειτνίασης M μπορεί να χρησιμοποιηθεί για να αντιπροσωπεύσει το γράφημα G_{collab}, το σύνολο των μεθόδων πρόβλεψης συνδέσμου παρουσιάζουν ισοδύναμη διατύπωση στον πίνακα M. Για παράδειγμα, η μέθοδος κοινών γειτόνων συνίσταται απλώς στην καταγραφή κάθε κόμβου x στη γραμμή του $r(x)$ στον πίνακα M, και στη συνέχεια ορίζεται το $score(x,y)$ να είναι το εσωτερικό γινόμενο των γραμμών $r(x)$ και $r(y)$.
- *Unseen bigrams*: μπορούμε να αυξήσουμε τις εκτιμήσεις μας για το $score(x,y)$, χρησιμοποιώντας τις τιμές του $score(x,y)$ για τους κόμβους z που είναι «παρόμοιοι» με το x. Αυτό επιτυγχάνεται στο πρόβλημα πρόβλεψης συνδέσμου ως ακολούθως: ας υποθέσουμε ότι έχουμε τιμές $score(x,y)$ οι οποίες έχουν υπολογιστεί με ένα από τα μέτρα που αναλύθηκαν παραπάνω:

  Θεωρούμε ότι $S_x^{<\delta>}$ δηλώνει τους δ κόμβους οι οποίοι σχετίζονται με το x κάτω από το $score(x,\cdot)$, για μια παράμετρο $\delta \epsilon Z^+$. Συνεπώς, προκύπτουν οι παρακάτω βαθμολογίες:

  $$score^*_{unweighted}(x,y) \coloneqq |\{z : z \in \Gamma(y) \cap S_x^{\delta}\}|$$

  $$score^*_{weighted}(x,y) \coloneqq \sum_{z \in \Gamma(y) \cap S_x^{\delta}} score(x,z)$$

- *Clustering:* θα μπορούσε κανείς να προσπαθήσει να βελτιώσει την ποιότητα μιας πρόβλεψης με τη διαγραφή των πιο «αδύναμων» ακμών του γραφήματος G_{collab} μέσω της διαδικασίας της συσταδοποίησης.

4.4.2 Εφαρμογή αλγορίθμων εκμάθησης με επίβλεψη

Οι Hasan et. al. στο [41] εφαρμόζουν διαφορετικούς τύπους αλγορίθμων εκμάθησης με επίβλεψη (supervised learning algorithms) με σκοπό τη δημιουργία μοντέλων δυαδικών κατηγοριοποιητών τα οποία θα διαχωρίζουν για παράδειγμα το σύνολο των συγγραφέων που θα συνεργαστούν σε μια επιλεγμένη χρονιά από τους υπόλοιπους που δεν θα συνεργαστούν. Η πρόβλεψη αναμενόμενων συνδέσμων στο γράφημα των συγγραφέων είναι μια ενδιαφέρουσα κατεύθυνση έρευνας διότι είναι πανομοιότυπη εννοιολογικά αλλά και δομικά με πολλά πρακτικά προβλήματα των κοινωνικών δικτύων. Ο βασικός λόγος είναι ότι το συγγραφικό δίκτυο είναι ένα πραγματικό παράδειγμα κοινωνικού δικτύου, όπου οι επιστήμονες της κοινότητας συνεργάζονται για την επίτευξη ενός κοινού στόχου.

Θεωρούμε ένα κοινωνικό δίκτυο $G = (V,E)$ στο οποίο κάθε ακμή $e = (u,v) \in E$ αναπαριστά την αλληλεπίδραση μεταξύ u και v σε μία συγκεκριμένη χρονική στιγμή t. Η αλληλεπίδραση αυτή ορίζεται ως η συνεργασία συγγραφής σε ένα ερευνητικό άρθρο. Για κάθε άρθρο είναι γνωστές οι πληροφορίες σχετικά με τον συγγραφέα και το έτος έκδοσης. Με σκοπό την πρόβλεψη ενός συνδέσμου διαμερίζεται η σειρά των ετών έκδοσης σε δύο μη επικαλυπτόμενα υπό-διαστήματα. Η πρώτη κατηγορία επιλέγεται ως έτη εκπαίδευσης (training data) και η επόμενη ως έτη ελέγχου (testing data). Στη συνέχεια, προετοιμάζεται το σύνολο δεδομένων κατηγοριοποίησης με την επιλογή συγγραφέων ανά ζεύγη οι οποίοι

εμφανίστηκαν στα έτη εκπαίδευσης, χωρίς όμως να έχουν δημοσιεύσει κάποια εργασία μαζί στη διάρκεια αυτών των ετών.

Κάθε ένα τέτοιο ζευγάρι δηλώνει ένα θετικό ή ένα αρνητικό παράδειγμα, το οποίο εξαρτάται από το αν αυτό το ζευγάρι έχει εκδώσει έστω και μία εργασία στα έτη ελέγχου ή όχι. Η συνεργασία σε μία εργασία κατά τη διάρκεια των ετών ελέγχου από ένα ζευγάρι συγγραφέων καθιερώνει έναν σύνδεσμο μεταξύ τους ο οποίος δεν υπήρχε στα έτη εκπαίδευσης. Το μοντέλο κατηγοριοποίησης του προβλήματος πρόβλεψης συνδέσμου είναι απαραίτητο με σκοπό την πρόβλεψη του συνδέσμου διακρίνοντας τις θετικές κλάσεις από το σύνολο δεδομένων.

Η επιλογή ενός κατάλληλου συνόλου στοιχείων είναι το πιο σημαντικό κομμάτι κάθε αλγορίθμου μηχανικής εκμάθησης. Για το πρόβλημα της πρόβλεψης συνδέσμου, θα πρέπει να διαλέγουμε στοιχεία που δηλώνουν κάποιας μορφής εγγύτητα μεταξύ των ζευγαριών των κορυφών που εκπροσωπούν ένα δεδομένο στοιχείο. Τα στοιχεία συνήθως σχετίζονται με έναν και μόνο κόμβο. Στην εργασία αυτή χρησιμοποιείται η απλή συνάρτηση συγκέντρωσης *sum* για τη μετατροπή του στοιχείου σε έναν πιθανό υποψήφιο για πρόβλεψη συνδέσμου. Αναλυτικά σύμφωνα με το [41] ορίζουμε ως:

- **sum of papers:** η αξία αυτού του χαρακτηριστικού υπολογίζεται προσθέτοντας τον αριθμό των δημοσιεύσεων που το ζευγάρι των συγγραφέων εξέδωσε κατά τα χρόνια εκπαίδευσης.
- **sum of neighbors:** αυτό το χαρακτηριστικό δηλώνει την κοινωνική συνδεσιμότητα του ζευγαριού των συγγραφέων προσθέτοντας τον αριθμό των γειτόνων που έχουν. Η έννοια της «γειτονιάς» ορίζεται από τις πληροφορίες συνεργασίας.
- **sum of keyword counts:** στις επιστημονικές δημοσιεύσεις, οι λέξεις κλειδιά παίζουν ζωτικό ρόλο στην παρουσίαση του επιστημονικού τομέα της εργασίας των ερευνητών. Οι ερευνητές που έχουν μια ευρεία γκάμα ενδιαφερόντων ή εκείνοι που εργάζονται σε διεπιστημονική έρευνα συνήθως χρησιμοποιούν περισσότερες λέξεις κλειδιά.
- **sum of classification code:** συνήθως οι ερευνητικές εκδόσεις κατηγοριοποιούνται σε συμβολοσειρές κωδικών (code strings) για την οργάνωση σχετικών περιοχών. Ομοίως με τη θεωρία των λέξεων κλειδιών, μια δημοσίευση που περιέχει πολλαπλούς κωδικούς είναι πιο πιθανό να αποτελεί διεπιστημονική δουλειά με αποτέλεσμα οι ερευνητές σε αυτόν τον τομέα να έχουν συνήθως περισσότερους συνεργάτες.
- **sum of log:** ενώ το πλήθος των πρωταρχικών γειτόνων είναι πολύ σημαντικό, ο αριθμός των δευτερευόντων γειτόνων μερικές φορές παίζει έναν σημαντικό ρόλο, ειδικά σε μια συνεργασία επιστημονικής έρευνας. Εάν ένας συγγραφέας συνδέεται απευθείας με έναν άλλον συγγραφέα, ο πρώτος έχει μια μεγάλη πιθανότητα να συνεργαστεί με έναν μακρινό κόμβο (συγγραφέα) μέσω του δεύτερου.

Επίσης χρησιμοποιούνται τα ακόλουθα τρία τοπολογικά χαρακτηριστικά βάσει του [41]:

- **shortest distance:** αυτό το στοιχείο είναι από τα πιο σημαντικά στο πρόβλημα της πρόβλεψης συνδέσμου. Στην εργασία [40] παρατηρήθηκε επίσης ότι στα κοινωνικά δίκτυα οι περισσότεροι από τους κόμβους συνδέονται μεταξύ τους με πολύ μικρή απόσταση. Αυτό το αξιοσημείωτο χαρακτηριστικό κάνει τη μικρότερη απόσταση ένα πολύ καλό στοιχείο για την πρόβλεψη συνδέσμου.
- **clustering index:** πολλές προσεγγίσεις στο πλαίσιο της έρευνας των κοινωνικών δικτύων υποδεικνύουν ότι ο δείκτης συσταδοποίησης είναι ένα σημαντικό στοιχείο των κόμβων του κοινωνικού δικτύου. Έχει παρατηρηθεί ότι ένας κόμβος που βρίσκεται σε πυκνή περιοχή είναι πιο πιθανό να αναπτύξει περισσότερες ακμές σε σχέση με έναν που βρίσκεται σε μια πιο αραιή γειτονιά. Ουσιαστικά, ο δείκτης συσταδοποίησης μετρά την τοπική πυκνότητα.

Μαθηματικά σύμφωνα με το [41], εάν το u είναι ο κόμβος ενός γραφήματος, ο δείκτης συσταδοποίησης του u είναι:

$$\frac{x\ number\ of\ triangles\ with\ u\ as\ one\ node}{number\ of\ connected\ triples\ with\ u\ as\ one\ node}$$

- **shortest distance in author-kw graph:** για τον υπολογισμό του συγκεκριμένου τοπολογικού χαρακτηριστικού επεκτείνεται το κοινωνικό δίκτυο προσθέτοντας κόμβους keyword (kw). Κάθε κόμβος kw συνδέεται σε έναν συγγραφικό κόμβο μέσω μιας ακμής εφόσον η λέξη κλειδί (κόμβος kw) χρησιμοποιείται από τον συγγραφέα σε κάποια από τις εργασίες του. Επιπλέον, δύο λέξεις κλειδιά οι οποίες εμφανίζονται μαζί σε μια εργασία συνδέονται με μία ακμή. Η ελάχιστη απόσταση μεταξύ δύο κόμβων σε αυτό το εκτεταμένο γράφημα υπολογίζεται με σκοπό το χαρακτηριστικό αυτό να αποκτήσει αξία.

Εκτός των παραπάνω χαρακτηριστικών που αναλύθηκαν οι Hasan et al. στο [41] δοκίμασαν και άλλα χαρακτηριστικά, όπως οι συντελεστές Jaccard's, Adamic/Adar, οι οποίοι σχετίζονται κυρίως με την τοπολογία του δικτύου. Δυστυχώς, όμως δεν παρείχαν κάποια σημαντική βελτίωση στην απόδοση του κατηγοριοποιητή. Υπάρχει μια πληθώρα αλγορίθμων κατηγοριοποίησης για εκμάθηση με επίβλεψη. Αν και οι αποδόσεις τους μπορούν να συγκριθούν, συνήθως μερικοί αποδίδουν καλύτερα από άλλους για ένα συγκεκριμένο σύνολο δεδομένων. Στη συγκεκριμένη έρευνα πραγματοποιήθηκαν πειράματα με επτά διαφορετικούς αλγόριθμους: SVM, Decision Tree, Multilayer Perceptron, *K*-Nearest Neighbors, Naive Bayes, RBF Network and Bagging.

Στη συνέχεια, συγκρίθηκε η απόδοση των παραπάνω κατηγοριοποιητών με τη χρήση διαφορετικών μετρικών απόδοσης όπως είναι η ακρίβεια (accuracy), η ανάκληση ακριβείας (precision-recall), η F-value, το τετραγωνικό σφάλμα (squared-error) κτλ. Για όλους τους αλγόριθμους χρησιμοποιήθηκε η μέθοδος 5-fold cross validation. Για αλγόριθμους που περιέχουν ρυθμιζόμενες παραμέτρους όπως οι SVM, K-Nearest Neighbors χρησιμοποιήθηκε ένα ξεχωριστό σύνολο αξιολόγησης με σκοπό την εύρεση των μέγιστων παραμετρικών αξιών. Ανακεφαλαιώνοντας με την εργασία [41] αποδείχθηκε ότι το πρόβλημα της πρόβλεψης συνδέσμου μπορεί να αντιμετωπιστεί αποτελεσματικά μοντελοποιώντας το ως ένα πρόβλημα κατηγοριοποίησης. Επιπλέον τα πειράματα απέδειξαν ότι τα πιο δημοφιλή μοντέλα κατηγοριοποίησης μπορούν να ανταπεξέλθουν με αποδεκτή ακρίβεια στο συγκεκριμένο πρόβλημα με τη μέθοδο SVM να εμφανίζει τα καλύτερα αποτελέσματα.

4.4.3 Εφαρμογές της πρόβλεψης συνδέσμου

Η πρόβλεψη συνδέσμου ισχύει για ένα ευρύ φάσμα εφαρμογών. Στον τομέα του διαδικτύου και του παγκόσμιου ιστού, επίσης μπορεί να χρησιμοποιηθεί σε εργασίες όπως η αυτόματη δημιουργία πρόβλεψης υπερ-συνδέσμου ιστοσελίδας. Στο ηλεκτρονικό εμπόριο, μια από τις βασικές χρήσεις της πρόβλεψης συνδέσμου είναι η δημιουργία συστημάτων συστάσεων (recommendation systems). Επίσης, αποτελεσματικές μέθοδοι της πρόβλεψης συνδέσμου θα μπορούσαν να χρησιμοποιηθούν για την ανάλυση κοινωνικών δικτύων με σκοπό να προτείνουν υποσχόμενες αλληλεπιδράσεις ή συνεργασίες που δεν έχουν ακόμα χρησιμοποιηθεί μέσα σε έναν οργανισμό. Από μία διαφορετική σκοπιά, η έρευνα στον τομέα της ασφάλειας έχει αρχίσει να δίνει έμφαση στη σημασία της ανάλυσης των κοινωνικών δικτύων και αυτό συμβαίνει λόγω του προβλήματος της παρακολούθησης των δικτύων τρομοκρατών.

Επιπλέον, η πρόβλεψη συνδέσμου θα μπορούσε να οδηγήσει στο συμπέρασμα ότι συγκεκριμένα άτομα συνεργάζονται ακόμα και αν η αλληλεπίδραση τους δεν είναι ευθέως ορατή. Το πρόβλημα πρόβλεψης συνδέσμου σχετίζεται επίσης με το πρόβλημα της εύρεσης εκλιπόντων συνδέσεων ενός δικτύου: σε έναν αριθμό γνωστικών πεδίων, κάποιος κατασκευάζει ένα δίκτυο αλληλεπιδράσεων, βασιζόμενος στα δεδομένα που έχουν παρατηρηθεί και προσπαθεί να συμπεράνει πρόσθετους συνδέσμους, οι οποίοι ενώ δεν είναι απευθείας ορατοί είναι πολύ πιθανόν να υπάρχουν. Τέλος, παρουσιάζει διάφορες εφαρμογές και

σε άλλους επιστημονικούς κλάδους. Για παράδειγμα, στην επιστήμη της βιβλιοθηκονομίας μπορεί να χρησιμοποιηθεί για την επικάλυψη και τη σύνδεση εγγραφών, στην Βιοπληροφορική, έχει χρησιμοποιηθεί για πρόβλεψη της αλληλεπίδραση μεταξύ πρωτεϊνών (PPI).

4.4.4 Σχετικές εργασίες

Οι Kashima et al. στο [42] εισάγουν ένα πλαίσιο διάδοσης συνδέσμων (link propagation) με στόχο την αξιοποίηση πολλαπλών τύπων δεσμών μεταξύ των κορυφών. Επίσης, προβλέπουν άγνωστα σημεία της δομής του δικτύου με τη χρήση βοηθητικών πληροφοριών, όπως οι ομοιότητες των κόμβων. Ωστόσο, η εργασία αυτή είναι στο μεγαλύτερο βαθμό χωρίς επίβλεψη και λειτουργεί μόνο για τους απόντες (missing) συνδέσμους στατικών δικτύων.

Μία από τις κύριες προκλήσεις της πρόβλεψης συνδέσμου αφορά το ρυθμό εξέλιξη των κοινωνικών δικτύων σε διαδικτυακή κλίμακα, όπως το Facebook, το MySpace, το Flickr κτλ. Τα δίκτυα αυτά είναι τεράστια σε μέγεθος με αποτέλεσμα πολλοί αλγόριθμοι να μην είναι σε θέση να προσαρμοστούν κατάλληλα. Για παράδειγμα οι Tylenda et. al. στο [43] δείχνουν ότι χρησιμοποιώντας χρονικές σημάνσεις (time stamps) παρελθοντικών αλληλεπιδράσεων μπορούν να βελτιώσουν σημαντικά την απόδοση της πρόβλεψης συνδέσμου. Επίσης, οι Song et. al. στο [44] χρησιμοποίησαν παραγοντικό πίνακα για την εκτίμηση της ομοιότητας μεταξύ των κόμβων σε ένα πραγματικό κοινωνικό δίκτυο με περίπου 2 εκατομμύρια κόμβους και 90 εκατομμύρια ακμές. Συνεπώς, κάθε παραδοσιακός αλγόριθμος ο οποίος έχει ως στόχο να υπολογίζει ζεύγη ομοιότητας μεταξύ κορυφών ενός τόσο μεγάλου γραφήματος είναι καταδικασμένος να αποτύχει.

Σε αυτή την εργασία [45] αντιμετωπίζεται το πρόβλημα της πρόβλεψης συνδέσμου χρησιμοποιώντας το δίκτυο relational Markov (RMN). Μέσα από αυτό το πλαίσιο καθορίζεται ένα ενιαίο πιθανολογικό μοντέλο για τους συνδέσμους ολόκληρου του γραφήματος συμπεριλαμβανομένου των ετικετών αντικειμένων και των συνδέσμων μεταξύ των αντικειμένων. Οι παράμετροι του μοντέλου είναι εκπαιδευμένοι με τέτοιο τρόπο ώστε να μεγιστοποιηθεί η πιθανότητα των ετικετών συνδέσμου για γνωστά χαρακτηριστικά (π.χ. λέξεις, hyperlinks κτλ.). Η εκμάθηση του μοντέλου εφαρμόζεται στη συνέχεια, χρησιμοποιώντας πιθανολογικά συμπεράσματα με σκοπό την πρόβλεψη και την ταξινόμηση των συνδέσμων βάσει των συνδέσμων και των χαρακτηριστικών που έχουν ήδη καταγραφεί.

Στην εργασία [46] αναλύεται το σύνολο των σχέσεων χρηστών της ιστοσελίδας ειδήσεων τεχνολογίας Slashdot. Τα δεδομένα συλλέχθηκαν από το εργαλείο Slashdot Zoo, όπου οι χρήστες της ιστοσελίδας μπορούν να χαρακτηρίσουν άλλους χρήστες ως φίλους και εχθρούς καθώς και να παρέχουν θετικά και αρνητικά σχόλια. Ορισμένες σχέσεις όπως η δυσπιστία και η απέχθεια είναι από τη φύση τους αρνητικές. Σε τέτοιες περιπτώσεις, το κοινωνικό δίκτυο περιέχει αρνητικά βάρη ακμών. Συνεπώς, σε επίπεδο ακμών αναλύονται τα μέτρα ομοιότητας που εφαρμόζονται στα ζεύγη των κόμβων και αξιολογείται η χρήση τους στο πρόβλημα της πρόβλεψης συνδέσμου. Πραγματοποιείται προσαρμογή των τεχνικών ανάλυσης του κοινωνικού δικτύου για το πρόβλημα που αφορά τα αρνητικά βάρη των ακμών. Οι μέθοδοι για την ανάλυση ενός δικτύου με αρνητικά βάρη ακμών εφαρμόζονται σε μεγάλα κοινωνικά δίκτυα και αποκαλύπτουν γεγονότα που δεν μπορούν να αποκαλυφθούν με τη χρήση κοινών τεχνικών**.**

Στην εργασία [47] προτείνεται μια διαφορετική προσέγγιση. Ακολουθώντας το παράδειγμα των κανόνων συσχέτισης και των συχνά εμφανιζόμενων προτύπων εξόρυξης γνώσης αναπτύσσεται το λογισμικό GERM (Graph Evolution Rule Miner) για την εξαγωγή κανόνων οι οποίοι θα εφαρμοστούν για να προβλέψουν τη μελλοντική εξέλιξη του δικτύου. Το μοντέλο αυτό μπορεί να βοηθήσει στην πρόβλεψη ακμών μεταξύ παλαιών και νέων κόμβων και στην πρόβλεψη πότε μια νέα ακμή αναμένεται να εμφανιστεί. Ακόμα και αν η μέθοδος αυτή συγκριθεί απευθείας με το κλασικό πρόβλημα πρόβλεψης συνδέσμου φαίνεται ότι η απλή βαθμολογία (scores) η οποία θα βασιστεί στους κανόνες που θα προκύψουν αποτελεί σημαντικό χαρακτηριστικό το οποίο παρέχει καλές επιδόσεις πρόβλεψης.

4.5 Κοινωνική επιρροή

Η *κοινωνική επιρροή* (social influence) αναφέρεται στην αλλαγή της συμπεριφοράς των ατόμων οι οποίοι επηρεάζονται από άλλα άτομα εντός δικτύου. Αποτελεί ένα διαισθητικό και ευρέως αποδεκτό φαινόμενο κοινωνικών δικτύων. Η δύναμη της κοινωνικής επιρροής εξαρτάται από πολλούς παράγοντες, όπως η δύναμη των σχέσεων μεταξύ των ανθρώπων, η απόσταση μεταξύ των χρηστών του δικτύου, οι προσωρινές επιδράσεις καθώς και τα χαρακτηριστικά των δικτύων και των ατόμων που ανήκουν σε αυτό. Σε αυτή την ενότητα, θα επικεντρωθούμε στην ανάλυση της κοινωνικής επιρροής περιγράφοντας τα μέτρα και τους αλγορίθμους που σχετίζονται με αυτήν. Συγκεκριμένα σκοπός μας είναι η ποιοτική ή η ποσοτική μέτρηση των επιπέδων επιρροής των κόμβων και των ακμών στο δίκτυο.

Όπως έχουμε αναφέρει ένα κοινωνικό δίκτυο μοντελοποιείται ως ένα γράφημα $G = \{V, E\}$, όπου V είναι το σύνολο των κόμβων και E είναι το σύνολο των ακμών. Σε τοπικό επίπεδο, η κοινωνική επιρροή είναι μια επίδραση με κατεύθυνση από τον κόμβο A στον κόμβο B και σχετίζεται με τη δύναμη της ακμής από τον A στον B. Σε γενικότερο επίπεδο, κάποιοι κόμβοι μπορούν να έχουν ουσιαστικά υψηλότερη επιρροή από άλλους, λόγω της δομής του δικτύου.

Στην εργασία [48] δημιουργούνται μοντέλα κοινωνικών δικτύων χρησιμοποιώντας δεδομένα από τη σελίδα δημοσιοποίησης απόψεων Epinions και το συνεργατικό σύστημα κριτικής ταινιών EachMovie. Αυτά τα μοντέλα επιτρέπουν τη δημιουργία σχεδίων για ηλεκτρονική προώθηση πληροφοριών (viral marketing) τα οποία βελτιστοποιούν τη θετική διαφήμιση από στόμα σε στόμα μεταξύ των πελατών. Οι δοκιμές έδειξαν ότι αυτό επέφερε επίτευξη πολύ μεγαλύτερων κερδών από ότι εάν είχαν αγνοηθεί οι αλληλεπιδράσεις ανάμεσα στους πελάτες και τα αντίστοιχα διαδικτυακά αποτελέσματα, κάτι το οποίο υιοθετεί το παραδοσιακό μάρκετινγκ.

Η *αξία πελάτη* (customer value) συνήθως ορίζεται ως το αναμενόμενο κέρδος από τις πωλήσεις σε έναν πελάτη κατά τη διάρκεια της σχέσης του με την εταιρεία. Η αξία πελάτη είναι κρίσιμου ενδιαφέροντος για τις εταιρείες επειδή καθορίζει το ποσό που αξίζει να επενδυθεί για την απόκτηση ενός συγκεκριμένου πελάτη. Ωστόσο, τα παραδοσιακά μέτρα της αξίας του πελάτη αγνοούν το γεγονός ότι εκτός από την αγορά προϊόντων, ένας πελάτης μπορεί να επηρεάσει άλλους να αγοράσουν.

Οι παράγοντες που επηρεάζουν την αξία πελάτη είναι [48]:

- **Γνώμη του πελάτη:** είναι σημαντικό το προϊόν να αρέσει στον πελάτη και κατά προτίμηση πολύ. Οι πελάτες που έχουν υψηλή συνδεσιμότητα (πολλές συνδέσεις σε ένα κοινωνικό δίκτυο) αλλά αντιπαθούν ένα προϊόν μπορεί να έχουν αρνητική διαδικτυακή αξία και θα έπρεπε να αποφύγουμε την προώθηση σε αυτούς.
- **Ασύμμετρη επιρροή:** άλλη μια άποψη κλειδί είναι ότι για να έχουμε υψηλή διαδικτυακή αξία, ο πελάτης θα πρέπει να επηρεάζει τους ανθρώπους με τους οποίους συναναστρέφεται περισσότερο (και ιδανικά πολύ περισσότερο) από ότι τον επηρεάζουν εκείνοι. Εάν η επιρροή είναι συμμετρική, η αναζήτηση των πελατών με την μεγαλύτερη επιρροή δεν έχει καθόλου πλεονεκτήματα.
- **Αλυσίδα επιρροής:** το πιο σημαντικό είναι ίσως το γεγονός ότι η διαδικτυακή αξία ενός πελάτη δεν τελειώνει με τις άμεσες γνωριμίες του. Αυτές οι γνωριμίες με την σειρά τους επηρεάζουν άλλους ανθρώπους και πάει λέγοντας μέχρι να φτάσουμε την πιθανότητα να επηρεαστεί ολόκληρο το δίκτυο. Ένας πελάτης που δεν έχει πολλές διασυνδέσεις, μπορεί στην πραγματικότητα να έχει υψηλή διαδικτυακή αξία εάν κάποιος από τις γνωριμίες του έχει αρκετά μεγάλο κύκλο γνωριμιών.

4.5.1 Επιρροή και δράσεις

Η επιρροή συχνά αντικατοπτρίζεται στις μεταβολές κοινωνικών προτύπων δράσης (συμπεριφορά χρηστών) σε ένα κοινωνικό δίκτυο. Οι εργασίες [49] & [50] έχουν μελετήσει το πρόβλημα της εκμάθησης του βαθμού επιρροής από τις ιστορικές ενέργειες ενός χρήστη, ενώ άλλες εργασίες [51] & [52] διερευνούν με ποιο τρόπο οι κοινωνικές δράσεις εξελίσσονται στο πλαίσιο του δικτύου και πως αυτές επηρεάζονται από παράγοντες κοινωνικής επιρροής.

Οι Goyal et al. στο [49] μελετούν το πρόβλημα της εκμάθησης των βαθμών επιρροής (αποκαλούνται πιθανότητες) από ένα ιστορικό αρχείο καταγραφής των ενεργειών ενός χρήστη. Παρουσιάζουν την έννοια της πιθανότητας ο χρήστης να ασκεί επιρροή (user influenceability) και την έννοια της πιθανότητας η ενέργεια να ασκεί επιρροή (action influenceability). Με την υπόθεση ότι ο χρήστης v_i εκτελεί μια ενέργεια y τη χρονική στιγμή t και αργότερα ($t' > t$) ο φίλος του v_j επίσης εκτελεί την ίδια ενέργεια συμπεραίνουμε ότι υπάρχει μια επιρροή από τον χρήστη v_i στον χρήστη v_j.

Ο στόχος της εκμάθησης των πιθανοτήτων επιρροής είναι η εύρεση (στατικών ή δυναμικών) μοντέλων με σκοπό την καλύτερη σύλληψη της επιρροής των χρηστών και της επιρροής των ενεργειών του δικτύου. Οι χρήστες οι οποίοι είναι εμπνευστές των διαφόρων ενεργειών και οι οποίοι επηρεάζονται περισσότερο από εξωτερικούς παράγοντες είναι σχετικά απρόβλεπτοι ή επηρεάζονται λιγότερο εύκολα. Για το λόγο αυτό ορίζεται ένας βαθμός επιρροής ως ο λόγος μεταξύ του αριθμού των ενεργειών για τις οποίες υπάρχουν στοιχεία που αποδεικνύουν ότι ο χρήστης έχει επηρεαστεί προς το συνολικό αριθμό των ενεργειών που εκτελούνται από το χρήστη. Πιο συγκεκριμένα ορίζεται σύμφωνα με το [49]:

$$infl(v_i) = \frac{|\{y|\exists v, \Delta t : prop(a, v_i, v_j, \Delta t) \wedge 0 \ \leq \Delta t|}{Y_{vi}}$$

Επίσης, ορίζουμε το πηλίκο επιρροής για μια ενέργεια διακρίνοντας μεταξύ των ενεργειών εκείνες για τις οποίες υπάρχει μεγαλύτερη απόδειξη της διάδοσης επιρροής από τις υπόλοιπες ως εξής [49]:

$$infl(y) = \frac{|\{v_i|\exists v_j, \Delta t : prop(a, v_j, v_i, \Delta t) \wedge 0 \ \leq \Delta t|}{number\ of\ users\ performing\ y}$$

όπου $\Delta t = t_j - t_i$ να αντιπροσωπεύει τη διαφορά μεταξύ του χρόνου όπου ο χρήστης v_j εκτελεί μια ενέργεια και του χρόνου όπου ο χρήστης v_i εκτελεί την ίδια ενέργεια και $prop(a, v_i, v_j, \Delta t)$ αναπαριστά το βαθμό διάδοσης της ενέργειας.

Οι Goyal et al. στο [49] προτείνουν τρεις μεθόδους για την προσέγγιση της διάδοσης ενέργειας $prop(a, v_i, v_j, \Delta t)$ οι οποίες παρουσιάζονται στη συνέχεια:

Στατικά μοντέλα: τα μοντέλα αυτά είναι ανεξάρτητα από το χρόνο και είναι αρκετά απλά στην εκμάθηση και τη δοκιμή. Ακολουθούν τρεις περιπτώσεις των στατικών μοντέλων:

- *Κατανομή Bernoulli:* κάθε φορά που ένας χρήστης v προσπαθεί να επηρεάσει τον ανενεργό γείτονά του u, έχει μια σταθερή πιθανότητα ενεργοποίησης του u. Αν ο u ενεργοποιηθεί τότε είναι μια επιτυχημένη προσπάθεια. Η μέγιστη εκτίμηση πιθανοφάνειας της πιθανότητας επιτυχίας είναι ο λόγος του αριθμού των επιτυχημένων προσπαθειών επί του συνολικού αριθμού των δοκιμών. Ως εκ τούτου, η πιθανότητα επιρροής του χρήστη v στον u υπολογίζεται ως εξής [49]:

  $$p_{v,u} = \frac{A_{v2u}}{A_v}$$

- *Δείκτης Jaccard:* ο δείκτης Jaccard χρησιμοποιείται συχνά για τη μέτρηση της ομοιότητας μεταξύ των συνόλων δείγματος και ορίζεται ως το μέγεθος της τομής διαιρούμενο με το μέγεθος

της ένωσης των συνόλων δείγματος. Ο δείκτης αυτός έχει προσαρμοστεί με σκοπό την εκτίμηση $p_{v,u}$ ως εξής [49]:

$$p_{v,u} = \frac{A_{v2u}}{A_{v|u}}$$

- *Μοντέλο μερικών βαθμών (Partial Credits):* όταν ένας χρήστης u σε ένα δίκτυο επηρεάζεται για την εκτέλεση μιας ενέργειας είναι πιθανό να επηρεαστεί από τους γείτονες του οι οποίοι έχουν εκτελέσει την ενέργεια αυτή σε προηγούμενο χρόνο. Ο βαθμός που δίνεται στον χρήστη $v \in S$ ο οποίος πραγματοποιεί την ενέργεια a πριν από τον χρήστη u ορίζεται ως εξής [49]:

  $credit_{v,u}(a) = \frac{1}{\sum_{w \in S} I(t_w(a) < t_u(a))}$ όπου το I είναι μία συνάρτηση δείκτη (indicator function)

Μοντέλα συνεχούς χρόνου: στην πραγματικότητα, η πιθανότητα επιρροής δεν μπορεί να παραμείνει σταθερή ανεξάρτητα από τον χρόνο. Είναι φυσικό να αναμένουμε ότι όταν ένας χρήστης αντιλαμβάνεται για πρώτη φορά ότι ο «γείτονας» του πραγματοποιεί μια ενέργεια είναι πιθανό να νιώσει την επιθυμία να την εξερευνήσει με αποτέλεσμα η αναμενόμενη επιρροή να καθυστερήσει.

Συνεπώς, ορίζουμε $p_{v,u}^t$ την πιθανότητα του v να επηρεάσει τον γείτονα του u τη χρονική στιγμή t ως εξής [49]:

$$p_{v,u}^t = p_{v,u}^0 e^{-(t-t_v)/\tau_{v,u}}$$

όπου $p_{v,u}^0$ είναι η μέγιστη δύναμη του v να επηρεάσει τον u.

Από τη στιγμή που ορίσαμε το $p_{v,u}^t$, μπορούμε να αντλήσουμε την πιθανότητα επιρροής $p_{v,u}^0(S)$, όπου ο u επηρεάζεται τη χρονική στιγμή t από το συνδυασμό των ενεργών γειτόνων του με τον ίδιο τρόπο όπως στα στατικά μοντέλα. Πιο συγκεκριμένα έχουμε [49]:

$$p_u^t(S) = 1 - \prod_{v \in S}(1 - p_{v,u}^t)$$

Η παράμετρος $\tau_{v,u}$ μπορεί να υπολογιστεί ως ο μέσος χρόνος καθυστέρησης στη διάδοση μιας ενέργειας από τον v στον γείτονά του u.

Μοντέλα διακριτού χρόνου: όπως αναφέραμε παραπάνω, τα μοντέλα συνεχούς χρόνου δεν είναι βαθμιαία με αποτέλεσμα να είναι ακριβά στον απαιτούμενο χρόνο εκτέλεσης για δοκιμές. Συνεπώς, προτείνονται κατά προσέγγιση μοντέλα συνεχούς χρόνου τα οποία αποκαλούνται διακριτά μοντέλα χρόνου (discrete time models). Στην περίπτωση αυτή μπορούμε να πούμε ότι η επιρροή ενός ενεργού χρήστη v στον γείτονά του u παραμένει σταθερή στο $p_{v,u}$ για το χρονικό διάστημα $\tau_{v,u}$ αφού ο v πραγματοποιήσει μια ενέργεια. Επιπλέον, ο ορισμός του S πρέπει να τροποποιηθεί κατά τρόπο ώστε να περιέχει μόνο τους «μεταδοτικούς» γείτονες του u. Ως εκ τούτου, όταν ένας μεταδοτικός γείτονας w γίνεται μη μεταδοτικός, πρέπει να ενημερωθεί το $p_u(S)$ ως εξής [49]:

$$p_u(S \backslash w) = \frac{p_u(S) - p_{w,u}}{1 - p_{w,u}}$$

4.5.2 Αλγόριθμοι εκμάθησης βαθμών επιρροής

Σε αυτή την ενότητα παρουσιάζουμε αλγορίθμους για την εκμάθηση των παραμέτρων των διαφόρων μοντέλων που αναλύσαμε παραπάνω. Οι παρακάτω αλγόριθμοι παρουσιάζονται στην εργασία [49]. Όπως σε κάθε προσέγγιση μηχανικής μάθησης, το πρώτο βήμα είναι η εκμάθηση των παραμέτρων του μοντέλου. Ο Αλγόριθμος 1 ο οποίος ακολουθεί, για να μάθει τις παραμέτρους για τα στατικά και τα συνεχή μοντέλα απαιτεί μόνο μία σάρωση του αρχείου καταγραφής ενεργειών. Καθώς διαβάζει μια νέα πλειάδα της μορφής $(u, \mathrm{a}, \mathrm{t_u})$ λέει στον χρήστη u να εκτελέσει την ενέργεια a τη χρονική στιγμή $\mathrm{t_u}$.

Algorithm 1 Learning - Phase1

```
1: for each action a in training set do
2:   current_table = φ
3:   for each user tuple < u, a, t_u > in chronological order do
4:     increment A_u
5:     parents = φ
6:     for each user v : (v, a, t_v) ∈ current_table && (v, u) ∈
       E^{t_v} do
7:       if t_u > t_v then
8:         increment A_{v2u}
9:         update τ_{v,u}
10:        insert v in parents
11:      increment A_{v&u}
12:    for each parent v ∈ parents do
13:      update credit_{v,u}
14:    add (u, a, t_u) to current_table
```

Εικόνα 4-11 Αλγόριθμος (1) Φάση εκμάθησης

Για την εκμάθηση της $infl(u), \tau$ απαιτείται μια δεύτερη σάρωση του αρχείου καταγραφής ενεργειών. Ομοίως, η εκμάθηση των παραμέτρων των διακριτών μοντέλων χρόνου απαιτούν επίσης τ εκ των προτέρων. Ομοίως με τον Αλγόριθμό 1 στην εργασία [49] προτείνεται ο Αλγόριθμος 2 (δεύτερη φάση εκμάθησης) ο οποίος διαφέρει με τον Αλγόριθμο 1, στο Βήμα 6 όπου απαιτεί $t_u - t_v \leq \tau_{v,u}$.

Στον Αλγόριθμο 3 [49] δίνεται μια βασική αξιολόγηση των στατικών μοντέλων. Διατηρούμε έναν πίνακα αποτελεσμάτων (results table) με καταχωρήσεις της μορφής $< u, p_u, perform_u >$, όπου το $perform_u$ αντιπροσωπεύει εάν ο χρήστης u έχει πραγματοποιήσει την εν λόγω ενέργεια ή όχι: η αξία της είναι 0 εάν ο u δεν έχει εκτελέσει ποτέ την ενέργεια αλλά το έχει κάνει τουλάχιστον ένας από τους γείτονες του, 1 είναι εάν ο u εκτελεί την ενέργεια και τουλάχιστον ένας από τους γείτονες του την έχει εκτελέσει πριν από αυτόν και 2 εάν ο u είναι ο δημιουργός της ενέργειας στη γειτονιά του. Το p_u αντιπροσωπεύει την πιθανότητα ο χρήστης u να εκτελεί την ενέργεια με δεδομένο τους γείτονες που έχουν εκτελέσει ήδη την ενέργεια αυτή. Σε κάθε στιγμή ο πίνακας αποτελεσμάτων περιέχει όλους τους ενεργοποιημένους χρήστες για την τρέχουσα ενέργεια και τους γείτονες τους.

Algorithm 3 Evaluate-Basic

```
1: for each action a in test set do
2:   results_table = φ
3:   for each user tuple < v, a, t_v > in chronological order do
4:     if v ∈ results_table then
5:       set perform_v flag to 1
6:     else
7:       add v to results_table with p_v=0 and perform_v=2
8:     for each user u : (v, u) ∈ E^{t_v} do
9:       if u ∈ results_table then
10:        update p_u incrementally as in Theorem 1
11:      else
12:        add u to results_table with appropriate p_u and
           perform_u=0
13:  for each entry < u, p_u, perform_u > in results_table do
14:    if (perform_u == 1 && p_u ≥ θ_u) it is TP
15:    if (perform_u == 1 && p_u < θ_u) it is FN
16:    if (perform_u == 0 && p_u ≥ θ_u) it is FP
17:    if (perform_u == 0 && p_u < θ_u) it is TN
```

Εικόνα 4-12 Αλγόριθμος (3) Φάση αξιολόγησης

Αντίστοιχα, με τον Αλγόριθμο 3 στην εργασία [49] παρουσιάζεται ο Αλγόριθμος 4 για την αξιολόγηση μοντέλων συνεχούς χρόνου όπου οι δοκιμές γίνονται αρκετά περίπλοκες. Στην περίπτωση αυτή αποθηκεύουμε τον πίνακα αποτελεσμάτων τη χρονική στιγμή t_u στην οποία ο χρήστης u πραγματοποιεί την ενέργεια a.

4.5.3 Μεγιστοποίηση επιρροής

Η ανάλυση της κοινωνικής επιρροής έχει πολλές εφαρμογές στον πραγματικό κόσμο. Η μεγιστοποίηση της επιρροής στο viral marketing είναι ένα παράδειγμα σημαντικής εφαρμογής και συχνά υποκινείται από την ανάγκη καθορισμού πιθανών πελατών ή υποψήφιων πελατών για τις ανάγκες του μάρκετινγκ. Ο βασικός στόχος είναι η ελαχιστοποίηση του κόστους προώθησης με την ταυτόχρονη μεγιστοποίηση του κέρδους. Για παράδειγμα, μια εταιρεία επιθυμεί να προωθήσει ένα νέο προϊόν μέσα από τη φυσική επίδραση της διάδοσης «από στόμα σε στόμα» η οποία προκύπτει από τις αλληλεπιδράσεις ενός κοινωνικού δικτύου. Ο στόχος είναι η εύρεση ενός μικρού αριθμού χρηστών οι οποίοι θα ασκούν επιρροή και θα υιοθετήσουν το προϊόν και συνεπώς θα προκαλέσουν μεγάλη κλιμάκωση στην αποδοχή του προϊόντος από άλλους.

Με σκοπό την επίτευξη αυτού του στόχου είναι απαραίτητο ένα μέτρο ποσοτικοποίησης των έμφυτων χαρακτηριστικών (πχ. το αναμενόμενο κέρδος από αυτό τον χρήστη) και της διαδικτυακής αξίας του χρήστη (πχ. το αναμενόμενο κέρδος από τους χρήστες που μπορεί να επηρεαστούν από τον χρήστη). Οι Kempe et al. στο [53] έκαναν το πρώτο βήμα για τον ορισμό αυτής της διαδικασίας μέσω δύο μοντέλων διάδοσης (diffusion models) και θεωρητικά απέδειξαν ότι το πρόβλημα βελτιστοποίησης της επιλογής των κόμβων που ασκούν τη μεγαλύτερη επιρροή σε δύο μοντέλα είναι ένα NP-hard πρόβλημα.

Η ποσοτικοποίηση της επιρροής κάθε κόμβου στην εργασία [53] αναλύεται με τα παρακάτω μοντέλα:

i. *Linear threshold model:* σε αυτή την κατηγορία μοντέλων, εάν ένας κόμβος v ενεργοποιηθεί μπορεί να βασιστεί σε μια αυθαίρετα μονότονη συνάρτηση του συνόλου των γειτόνων του v οι οποίοι είναι ήδη ενεργοποιημένοι. Συνδέουμε μια μονότονη συνάρτηση κατωφλιού f_v η οποία αντιστοιχεί υποσύνολα των γειτόνων του v σε πραγματικούς αριθμούς στο $[0,1]$. Συνεπώς, για κάθε κόμβο v δίνεται ένα κατώφλι θ_v και ενεργοποιείται ο v στο βήμα t εάν $f_v(S) > \theta_v$, όπου S είναι το σύνολο των γειτόνων του v οι οποίοι είναι ενεργοί στο βήμα $t-1$. Συγκεκριμένα, στο [53] η συνάρτηση κατωφλιού $f_v(S)$ ορίζεται ως $f_v(S) = \sum_{u \in S} b_{v,u}$ όπου το $b_{v,u}$ μπορεί να θεωρηθεί ως ένα σταθερό βάρος, υπό τον ακόλουθο περιορισμό:

$$\sum_{u \, neightbors \, of \, u} b_{v,u} \leq 1$$

ii. *General cascade model:* αρχικά ορίζουμε μια στοιχειώδη συνάρτηση $p_v(u, S) \in [0, 1]$ ως την πιθανότητα επιτυχίας ο χρήστης u να ενεργοποιήσει τον χρήστη v, για παράδειγμα ο χρήστης u προσπαθεί να ενεργοποιήσει τον v και τελικά το επιτυγχάνει όπου S είναι εκείνοι οι γείτονες του v οι οποίοι έχουν ήδη προσπαθήσει αλλά απέτυχαν να ενεργοποιήσουν τον v. Μια ειδική έκδοση του μοντέλου χρησιμοποιείται στο [53] και αποκαλείται *independent cascade model* στο οποίο το $p_v(u, S)$ είναι μια σταθερά, το οποίο σημαίνει ότι αν ο κόμβος v πρόκειται να ενεργοποιηθεί δεν εξαρτάται από τη σειρά με την οποία οι γείτονες του v θα προσπαθήσουνε να το κάνουν. Μια περίπτωση του independent cascade model είναι το *weighted cascade model*, όπου κάθε ακμή από τον κόμβο u στον κόμβο v έχει πιθανότητα ενεργοποίησης του v η οποία είναι $1/d_v$.

Δοσμένης μιας συνάρτησης f η οποία είναι υπό-τμηματική και δέχεται μόνο μη αρνητικές τιμές έχουμε:

$$f(S \cup \{u\}) \geq f(S)$$

για όλα τα στοιχεία v και τα σύνολα S. Έτσι, το πρόβλημα μπορεί να μετασχηματιστεί στην εύρεση ενός k-στοιχείου του συνόλου S για το οποίο η $f(S)$ μεγιστοποιείται. Το πρόβλημα, μπορεί να λυθεί χρησιμοποιώντας έναν άπληστο hill-climbing αλγόριθμο ο οποίος κατά προσέγγιση υπολογίζει τη βέλτιστη λύση μέσω ενός παράγοντα $(1 - 1/e)$.

Στην εργασία [54] προτείνεται ο Αλγόριθμος 2 έτσι ώστε να βρεθεί το σύνολο S το οποίο μεγιστοποιεί την f.

Algorithm 2 Hill-Climbing Algorithm

1: $S_{[1]} := \arg\max_{v \in V; c(v) \leq B}\{w(\{v\})\}$;
2: $S_{[2]} := \emptyset$;
3: **for** $v \in V \setminus S_{[2]}$ **do**
4: $\quad v \leftarrow \arg\max_{v \in V; c(S_{[2]})+c(v) \leq B}\{\frac{w(S_{[2]} \cup \{v\}) - w(S_{[2]})}{c(v)}\}$;
5: $\quad S_{[2]} = S_{[2]} \cup v$;
6: $S := \arg\max_{S \in \{S_{[1]}, S_{[2]}\}}\{w(S)\}$;

Εικόνα 4-13 Αλγόριθμος (2) Hill-Climbing

Το μοντέλο μπορεί να επεκταθεί περαιτέρω με την υπόθεση ότι κάθε κόμβος v έχει ένα μη αρνητικό βάρος w_v, το οποίο μπορεί να χρησιμοποιηθεί για τη σύλληψη της προηγούμενης ανθρώπινης γνώσης για μια συγκεκριμένη εργασία πχ. πόσο σημαντικό είναι να ενεργοποιηθεί ο κόμβος v στο τελικό αποτέλεσμα.

Για να υιοθετήσουμε το μοντέλο σε ένα πιο ρεαλιστικό σενάριο, θα χρειαστεί να έχουμε έναν διαθέσιμο αριθμό m διαφορετικών ενεργειών προώθησης M_i, κάθε μία εκ των οποίων μπορεί να επηρεάσει κάποιο υποσύνολο των κόμβων αυξάνοντας τις πιθανότητες να γίνουν ενεργοί. Εντούτοις διαφορετικοί κόμβοι είναι πιθανό να αντιδρούν σε ενέργειες προώθησης με διαφορετικό τρόπο. Για το λόγο αυτό θεωρούμε ένα πιο γενικό μοντέλο εισάγοντας την επένδυση t_i για κάθε ενέργεια προώθησης M_i. Με αυτό τον τρόπο ο στόχος είναι η εύρεση του μέγιστου κέρδους χωρίς οι συνολικές επενδύσεις να υπερβαίνουν το διαθέσιμο προϋπολογισμό. Μια στρατηγική μάρκετινγκ είναι ένα m-διαστάσεων διάνυσμα t επενδύσεων. Η πιθανότητα ο κόμβος v να γίνει ενεργός καθορίζεται από το $h_v(t)$. Θεωρώντας ότι η συνάρτηση είναι μη φθίνουσα και ικανοποιεί την ιδιότητα της φθίνουσας απόδοσης για όλα τα $t \geq t'$ και $a \geq 0$ έχουμε:

$$h_v(t + a) - h_v(t) \leq h_v(t' + a) - h_v(t')$$

Η ικανοποίηση της παραπάνω ανισότητας οδηγεί σε μια ενδιαφέρουσα διαίσθηση μάρκετινγκ η οποία υποστηρίζει ότι: *η ενέργεια μάρκετινγκ θα είναι περισσότερο αποτελεσματική όταν τα άτομα στόχοι είναι λιγότερο κορεσμένα σε ενέργειες μάρκετινγκ.* Τέλος, ο στόχος του μοντέλου είναι η μεγιστοποίηση του αναμενόμενου μεγέθους του τελικού συνόλου ενεργών κόμβων. Δίνοντας ένα αρχικό σύνολο A και θεωρώντας το αναμενόμενο μέγεθος του τελικού ενεργού συνόλου $\sigma(A)$, τότε τα αναμενόμενα κέρδη από την στρατηγική μάρκετινγκ t μπορούν να οριστούν ως εξής:

$$g(t) = \sum_{A \subset V} \sigma(A) \cdot \prod_{u \in A} h_v(t) \cdot \prod_{u \in V - A} 1 - h_u(t))$$

Ένα δύσκολο πρόβλημα στο μοντέλο διάδοσης και μεγιστοποίησης της επιρροής είναι η αξιολόγηση της αποτελεσματικότητας και της αποδοτικότητας του. Οι αλγόριθμοι που προτείνουν στο [53] μπορεί επίσης σε θεωρητικό επίπεδο να εγγυηθούν ότι η εξάπλωση της επιρροή είναι μεταξύ του $(1 - 1/e)$ της βέλτιστης εξάπλωσης επιρροής. Από εμπειρικής σκοπιάς, οι Kempe et al. αποδεικνύουν ότι τα μοντέλα τους είναι σε θέση να ξεπεράσουν τις παραδοσιακές ευριστικές από την άποψη της μεγιστοποίησης της

κοινωνικής επιρροής. Ένα άλλο πρόβλημα είναι η αξιολόγηση της αποτελεσματικότητας των μοντέλων στη μεγιστοποίηση της επιρροής το οποίο είναι αντικείμενο των εργασιών [55] & [56].

4.5.4 Προβλέποντας τους πελάτες

Το viral marketing αποσκοπεί στην αύξηση της αναγνωρισιμότητας μιας ετικέτας και των εσόδων με τη βοήθεια των κοινωνικών δικτύων και της κοινωνικής επιρροής. Το άμεσο μάρκετινγκ είναι μια σημαντική εφαρμογή, η οποία επιχειρεί να προσεγγίσει μόνο ένα επιλεγμένο σύνολο εν δυνάμει πελατών.

Στην εργασία [55] προτείνεται ένα μοντέλο που προσπαθεί να συνδυάσει την αξία του δικτύου με την αξία του πελάτη. Η αξία του πελάτη αντιπροσωπεύει ιδιότητες (π.χ. το ιστορικό της συμπεριφοράς) που συνδέονται άμεσα με αυτόν οι οποίες μπορεί να επηρεάσουν την πιθανότητα ο πελάτης να αγοράσει το προϊόν, ενώ η αξία του δικτύου αντιπροσωπεύει το κοινωνικό δίκτυο (π.χ. τους φίλους του) η οποία μπορεί να επηρεάσει την απόφαση αγοράς του πελάτη. Η βασική ιδέα της εργασίας είναι η τυποποίηση του κοινωνικού δικτύου ως τυχαία πεδία Markov, όπου η πιθανότητα κάθε πελάτη να αγοράσει μοντελοποιείται ως η συνάρτηση τόσο της επιθυμίας του πελάτη για το προϊόν όσο και της επιρροής των άλλων πελατών. Τα δεδομένα εισόδου μπορεί να οριστούν ως εξής: θεωρούμε ένα κοινωνικό δίκτυο $G = (V, E)$ με n πιθανούς πελάτες, τις σχέσεις τους να καταγράφονται στο E, και το x_i να εκφράζει τα χαρακτηριστικά που σχετίζονται με κάθε πελάτη v_i.

Στη συνέχεια, αναθέτουμε μια y_i μεταβλητή boole σε κάθε πελάτη η οποία παίρνει την τιμή 1 αν ο v_i πελάτης αγοράζει το προϊόν που διατίθεται στο εμπόριο ή την τιμή 0 στην αντίθετη περίπτωση. Επιπρόσθετα, ορίζουμε ως NB_i τους γείτονες του v_i στο κοινωνικό δίκτυο και z_i είναι μια μεταβλητή που αντιπροσωπεύει την ενέργεια μάρκετινγκ που πραγματοποιείται για τον v_i πελάτη. Η z_i είναι μια μεταβλητή boole, με z_i = 1 να σημαίνει ότι ο πελάτης έχει επιλεχθεί για το σύνολο προώθησης (π.χ. του προσφέρεται ένα προϊόν δωρεάν) και με z_i = 0 στην αντίθετη περίπτωση. Εναλλακτικά, η z_i θα μπορούσε να είναι μια συνεχής μεταβλητή που δείχνει μια έκπτωση που προσφέρεται στον πελάτη. Με δεδομένα τα παραπάνω μπορούμε να ορίσουμε τη διαδικασία προώθησης για τον πελάτη v_i σε ένα τυχαίο πεδίο Markov ως εξής [55]:

$$P\left(y_i \middle| y_{N\,B^i}, x_i, z\right) = \sum_{C(N\,B^i)} P\left(y_i, y_{N\,B^i} \middle| x_i, z\right)$$

$$= \sum_{C(N\,B^i)} P\left(y_i \middle| y_{N\,B^i}, x_i, z\right) P\left(y_{N\,B^i} \middle| X, z\right)$$

όπου $C(NBi)$ είναι το σύνολο όλων των πιθανών γειτόνων του v_i και το X αναπαριστά όλα τα χαρακτηριστικά των πελατών. Για την εκτίμηση του $P(y_{NB^i} |X,z)$, χρησιμοποιείται η εκτίμηση της μέγιστης εντροπίας για την προσέγγιση της πιθανότητας η οποία βασίζεται σε μια ανεξάρτητη υπόθεση [55]:

$$P\left(y_{N\,B^i} \middle| X, z\right) = \prod_{u_j \epsilon N\,B^i} P(y_j | X, z)$$

Η ενέργεια προώθησης z μοντελοποιείται με μια μεταβλητή Boole. Το κόστος προώθησης σε έναν πελάτη αναφέρεται στο μοντέλο Markov. Ας θεωρήσουμε ότι το r_0είναι τα έσοδα από την πώληση του προϊόντος στον πελάτη εάν δεν πραγματοποιηθεί καμία ενέργεια προώθησης και r_1είναι τα έσοδα αν πραγματοποιηθεί κάποια ενέργεια. Το κόστος μπορεί να θεωρηθεί ως μια προσφερόμενη έκπτωση στον πελάτη που συμμετέχει στην ενέργεια προώθησης. Έτσι, το αναμενόμενο κέρδος στον πελάτη v_i (χωρίς επιρροή) μπορεί να οριστεί ως εξής [55]:

$$ELP_i^1(Y,z) = r_1 P\left(y_i = 1 \middle| Y, f_i^1(M)\right) - r_o P\left(y_i = 1 \middle| Y, f_i^o(z)\right) - C$$

Το συνολικό κέρδος για μια συγκεκριμένη ενέργεια z είναι:

$$ELP_i^1(Y,z) = \sum_{i=1}^{n} [r_i P(X_i = 1|Y,z) - r_0 P(X_{i=1}|Y,z_0) - c_i]$$

Η συνολική αξία ενός πελάτη είναι το γενικό κέρδος από την προώθηση σε αυτόν:

$$ELP\left(Y, f_i^1(z_i)\right) - ELP(Y, f_i^0(z_i))$$

4.5.5 Σχετικές εργασίες

Στην εργασία [57] παρουσιάζεται μια επισκόπηση του αντίκτυπου της κοινωνικής επιρροής στο ηλεκτρονικό εμπόριο. Τα βασικά στοιχεία στα οποία εστιάζει είναι πως μπορούμε να συλλάβουμε τις κοινωνικές αλληλεπιδράσεις στις ιστοσελίδες ηλεκτρονικού εμπορίου, πώς μπορούμε να συνδυάσουμε τα κοινωνικά στοιχεία επιρροής στις προτιμήσεις των χρηστών και με ποιο τρόπο μπορούμε να ασκήσουμε κοινωνική επιρροή στις αποφάσεις αγορών των πελατών με στόχο τη μεγιστοποίηση της κοινωνικής επιρροής στο ηλεκτρονικό εμπόριο και στη διαδικασία λήψης αποφάσεων.

Στην εργασία [58] αναλύεται η χρήση των κοινωνικών δικτύων στην υλοποίηση στρατηγικών viral marketing. Στο προτεινόμενο μοντέλο, η απόφαση ενός αγοραστή να αγοράσει ένα αντικείμενο επηρεάζεται από το σύνολο άλλων αγοραστών οι οποίοι κατέχουν το αντικείμενο αυτό στην τιμή που προσφέρεται. εστιάζοντας σε ένα αλγοριθμικό ερώτημα εύρεσης του μέγιστου εισοδήματος στρατηγικών προώθησης. Όταν οι αγοραστές είναι απολύτως συμμετρικοί, μπορούμε να βρούμε τη βέλτιστη στρατηγική προώθησης σε πολυωνυμικό χρόνο. Προσδιορίζουμε μια οικογένεια στρατηγικών οι οποίες αποκαλούνται στρατηγικές επιρροής και αξιοποίησης και βασίζονται στην ακόλουθη ιδέα: αρχικά επηρεάζουμε τον πληθυσμό μέσω της προσφοράς του αντικειμένου δωρεάν σε ένα επιλεγμένο σύνολο αγοραστών. Στη συνέχεια, αυξάνουμε το εισόδημα από τους υπόλοιπους αγοραστές χρησιμοποιώντας μια άπληστη στρατηγική τιμολόγησης.

Τα συστήματα συστάσεων συνεργατικού φιλτραρίσματος (collaborative filtering recommender systems) βασίζουν τις αποφάσεις τους στη γνώμη των χρηστών. Σε αντίθεση με άλλα κοινωνικά δίκτυα, τα συστήματα συστάσεων συλλαμβάνουν αλληλεπιδράσεις οι οποίες είναι επίσημες και ποιοτικές. Στην εργασία [59], αποδεικνύεται ότι όσες απόψεις εκφράζει ένας χρήστης αποτελεί μια σημαντική συνιστώσα επιρροής. Προτείνεται επίσης ο αλγόριθμος NUPD ο οποίος είναι ανεξάρτητος και μπορεί να εφαρμοστεί σε κάθε σύστημα συστάσεων στο πλαίσιο των προβλέψεων.

4.6 Εμπιστοσύνη

Η *εμπιστοσύνη* (trust) είναι μια σημαντική πτυχή της σχέσης μεταξύ δύο οντοτήτων. Στο πλαίσιο ενός κοινωνικού δικτύου (ποιος εμπιστεύεται ποιον) διαδραματίζει σημαντικό ρόλο στον τομέα πληροφοριών και ασφαλείας. Η εμπιστοσύνη αποτελεί τη βάση για σχηματισμό συνασπισμών (ισχυρές κοινότητες που αποτελούνται από φορείς, οι οποίοι «εμπιστεύονται» ο ένας τον άλλο) γεγονός το οποίο είναι πιθανό να φανεί χρήσιμο στον εντοπισμό επιρροής μεταξύ των κόμβων ενός δικτύου, στον καθορισμό της ροής των πληροφοριών σε ένα κοινωνικό δίκτυο και στο κατά πόσο οι κόμβοι θα υιοθετούν τις πληροφορίες που λαμβάνουν ή κατά πόσο θα επιλέγουν να τις διαβιβάσουν σε κάποιον άλλο κόμβο. Σε γενικές γραμμές, όταν κάποιος αποφασίζει αν θα εμπιστεύεται ή αν δεν θα εμπιστεύεται ένα άτομο, επηρεάζεται από μια σειρά παραγόντων, όπως [60]:

i. η δική του προδιάθεση σε θέματα εμπιστοσύνης, η οποία συνδέεται κυρίως με την ψυχολογία του,

ii. η σχέση του και οι εμπειρίες του παρελθόντος οι οποίες σχετίζονται με το συγκεκριμένο πρόσωπο και τους φίλους του και

iii. οι απόψεις του για τις ενέργειες και τις αποφάσεις που το άτομο έχει κάνει στο παρελθόν.

Σε ένα κοινωνικό δίκτυο, οι πληροφορίες δημιουργούνται και «καταναλώνονται» από τους χρήστες του. Οι χρήστες ανταλλάσσουν πληροφορίες με βάση το επίπεδο της εμπιστοσύνης το οποίο ξεκάθαρα αναθέτουν σε άλλους χρήστες. Η δυνατότητα να καθοριστεί πόσο ένας χρήστης εμπιστεύεται την πηγή των πληροφοριών, όταν ο χρήστης δεν γνωρίζει άμεσα την πηγή μπορεί να χρησιμοποιηθεί για τη συγκέντρωση, το φιλτράρισμα και την ταξινόμηση των πληροφοριών. Επιπλέον, εάν η εμπιστοσύνη μπορεί να εκτιμηθεί με ακρίβεια, ο χρήστης μπορεί να χρησιμοποιήσει στη συνέχεια, αυτή την εκτίμηση εμπιστοσύνης για τη λήψη αποφάσεων σχετικά με τις πληροφορίες.

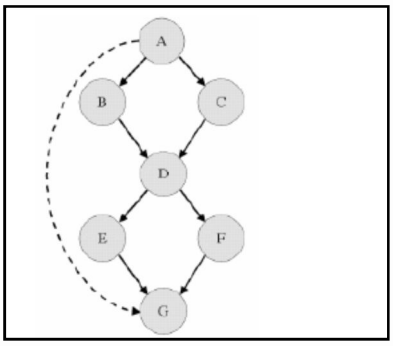
Η εκτίμηση εμπιστοσύνης για να είναι χρήσιμη στα κοινωνικά δίκτυα εκφράζεται συχνά ως βαθμοί εμπιστοσύνης (trust ratings) ή τιμές που ένας χρήστης μπορεί να αναθέσει ρητά σε άλλο χρήστη. Στη συνέχεια, μπορούμε να χρησιμοποιήσουμε αυτές τις τιμές εμπιστοσύνης οι οποίες μπορεί να υπάρχουν ανάμεσα σε δύο ανθρώπους που δεν είναι άμεσα συνδεδεμένοι. Με άλλα λόγια, όταν η εμπιστοσύνη εκφράζεται ρητά μέσω βαθμολογίας σε αριθμητική κλίμακα, μπορεί να προκύψουν πληροφορίες σχετικά με την εμπιστοσύνη μεταξύ των δύο ατόμων οι οποίοι δεν έχουν άμεση σύνδεση.

4.6.1 Ορισμοί και χαρακτηριστικά

Ορισμένες ιστοσελίδες κοινωνικής δικτύωσης έχουν υιοθετήσει την εμπιστοσύνη στη δομή του δικτύου τους. Για παράδειγμα, το LinkedIn συνδέει κάποιον με τις αξιόπιστες επαφές του και τον βοηθά να ανταλλάξει γνώσεις, ιδέες και ευκαιρίες με ένα ευρύτερο δίκτυο επαγγελματιών. Στην εργασία [61], ορίζουμε την εμπιστοσύνη στο πλαίσιο των κοινωνικών δικτύων, ως *την πεποίθηση ενός ατόμου ότι η ενέργεια που πραγματοποιεί θα είναι ένα επιθυμητό αποτέλεσμα αν στηριχθεί στα πιστεύω άλλων ατόμων.*

Για την προσομοίωση της εμπιστοσύνης στο περιβάλλον των κοινωνικών δικτύων οι J. Golbeck και J. Hendler στην εργασία [62], προτείνουν τρεις βασικές ιδιότητες της εμπιστοσύνης οι οποίες είναι η *μεταβατικότητα* (transivity), η *ασυμμετρία* (asymmetry) και η *εξατομίκευση* (personalisation).

Στην εικόνα 4.14, μπορούμε να αντιληφθούμε τι είναι εμπιστοσύνη. Θεωρούμε τον κόμβο A (ή την «πηγή») ότι είναι απευθείας συνδεδεμένος με τους κόμβους B και C αλλά δεν συνδέεται απευθείας με τους κόμβους D, E, F και G. Επιπλέον, μπορούμε να πούμε ότι ο κόμβος A συνδέεται έμμεσα με τον G μέσω των εξής τεσσάρων διαδρομών A→B→D→E→G, A→C→D→F→G, A→B→D→F→G, και A→C→D→E→G, γεγονός το οποίο δημιουργεί τέσσερις προοπτικές εμπιστοσύνης του G όταν καθορίζουμε την αξία εμπιστοσύνης του A στον G.



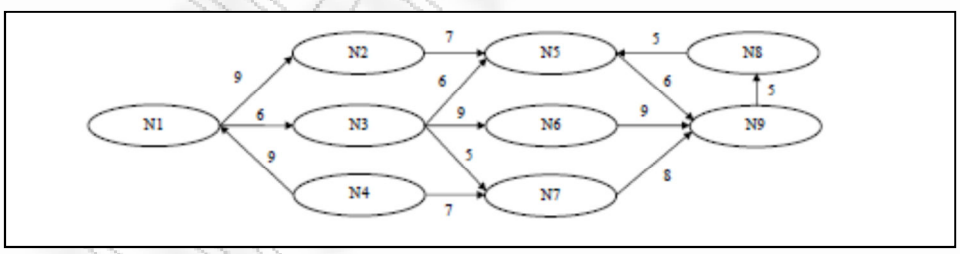
Εικόνα 4-14 Εξαγωγή συμπεράσματος εμπιστοσύνης από τον κόμβο A στον G

Ένας μηχανισμός εξαγωγής συμπεράσματος εμπιστοσύνης μπορεί να περιγραφεί ως ένας αλγόριθμος ο οποίος καθορίζει τις αξίες εμπιστοσύνης οι οποίες αποτελούν συστάσεις σε ένα χρήστη για το πόσο να εμπιστεύεται κάποιον άλλο χρήστη. Αυτός ο αλγόριθμος μπορεί να δημιουργήσει μια προτεινόμενη βαθμολογία εμπιστοσύνης για ένα άγνωστο άτομο στο κοινωνικό δίκτυο, με βάση τις πληροφορίες από άλλους χρήστες οι οποίοι συνδέονται με τον άγνωστο χρήστη. Η μοντελοποίηση ενός τέτοιου μηχανισμού εμπιστοσύνης επιτυγχάνεται με τη θεώρηση των ατόμων του κοινωνικού δικτύου ως κόμβους, τις φιλίες ή τις σχέσεις ως ακμές και τις τιμές εμπιστοσύνη ως ετικέτες ακμών [61].

4.6.2 Ο αλγόριθμος TidalTrust

Ο αλγόριθμος TidalTrust προτείνεται από την J. Golbeck στο [63] και θεωρεί ότι οι αξίες εμπιστοσύνης είναι αριθμοί σε ένα συνεχές εύρος [0…10]. Είναι ένας απλός αλγόριθμος χαμηλής πολυπλοκότητας ο οποίος επιτρέπει υψηλή επεκτασιμότητα στις εφαρμογές του. Στην εργασία [63] γίνεται η υπόθεση ότι οι τιμές εμπιστοσύνης μέσω των συντομότερων διαδρομών μπορεί να είναι πιο ακριβείς, για το λόγο αυτό λαμβάνονται υπόψη μόνο οι συντομότερες διαδρομές από την πηγή στον δέκτη. Ο αλγόριθμος TidalTrust είναι ένας δημοφιλής αλγόριθμος σε θέματα εμπιστοσύνης και χρησιμοποιείται συχνά ως αλγόριθμος αναφοράς.

Ο αλγόριθμος λειτουργεί σε ένα δίκτυο αξιολογήσεων στο οποίο κάθε κόμβος έχει βαθμολογήσει αρκετούς άλλους κόμβους. Η έννοια της αξιολόγησης μπορεί να διαφέρει μεταξύ σεναρίων εφαρμογής. Στην εικόνα 4.15 απεικονίζεται ένα παράδειγμα δικτύου βαθμολόγησης. Οι κόμβοι αντιπροσωπεύουν παρόχους πληροφοριών και οι ακμές αντιπροσωπεύουν τις αξιολογήσεις τους σε κλίμακα από το 1 (χαμηλή ποιότητα) έως το 10 (υψηλή ποιότητα).



Εικόνα 4-15 Παράδειγμα δικτύου βαθμολόγησης

Ο αλγόριθμος TidalTrust [64] καθορίζει την κατάταξη ενός κόμβου στο δίκτυο (που ονομάζεται δέκτης) από την οπτική γωνία ενός άλλου κόμβου (που ονομάζεται πηγή). Εάν το δίκτυο αξιολόγησης περιέχει μια άμεση αξιολόγηση από την πηγή στον δέκτη, τότε ο αλγόριθμος επιστρέφει αυτή τη βαθμολογία ως

αποτέλεσμα. Στην περίπτωση που το δίκτυο δεν περιέχει μια τέτοια εκτίμηση, ο αλγόριθμος συνάγει μία εκτίμηση κατά προσέγγιση από τις διαδρομές που συνδέουν τους δύο κόμβους. Το συμπέρασμα αυτό βασίζεται σε δύο υποθέσεις: αναμένεται ότι τα άτομα τα οποία ο χρήστης βαθμολογεί με υψηλά ποσοστά τείνουν να συμφωνούν με τον χρήστη περισσότερο σχετικά με τη βαθμολογία άλλων σε σχέση με τα άτομα στα οποία ο χρήστης δίνει χαμηλή βαθμολογία και δεύτερον η ακρίβεια των βαθμολογιών που προκύπτουν αναμένεται να μειωθεί με το μήκος των διαδρομών που συνδέουν δύο άτομα.

Με βάση αυτές τις υποθέσεις ο αλγόριθμος TidalTrust συνάγει τη βαθμολογία που αναζητούμε, διενεργώντας τα παρακάτω βήματα σύμφωνα με το [64]:

Βήμα 1ο: Αναζήτα όλες τις διαδρομές ελάχιστου μήκους του δικτύου που συνδέουν την πηγή με το δέκτη. Το μήκος της διαδρομής είναι κατανοητό ως ο αριθμός των ακμών που σχηματίζουν ένα μονοπάτι. Ας υποθέσουμε για παράδειγμα, ότι ο κόμβος N1 θέλει να συμπεράνει μια εκτίμηση για τον κόμβο N9 από το παράδειγμα του δικτύου στην εικόνα 4.15. Υπάρχουν τέσσερα ελάχιστα μονοπάτια μήκους από τον κόμβο N1 στον κόμβο N9: Η πρώτη διαδρομή συνδέει την πηγή με τον δέκτη μέσω των κόμβων N2 και N5, το δεύτερο μονοπάτι μέσω των κόμβων N3 και N5, το τρίτο μονοπάτι μέσω των κόμβων N3 και N6 και το τέταρτο μονοπάτι μέσω των κόμβων N3 και N7.

Βήμα 2ο: Ο αλγόριθμος καθορίζει την ισχύ του κάθε μονοπατιού. Η ισχύς ενός μονοπατιού ισούται με τη χαμηλότερη βαθμολογία στο μονοπάτι. Στο παράδειγμά που εξετάζουμε, η ισχύς των τριών πρώτων μονοπατιών είναι 6 ενώ η ισχύς του τέταρτου μονοπατιού είναι 5.

Βήμα 3ο: Στη συνέχεια, ο αλγόριθμος καθορίζει το μέγιστο όριο. Το όριο αυτό χρησιμοποιείται για να καθορίσει τις αξιολογήσεις που λαμβάνονται υπόψη στον τελικό υπολογισμό. Το μέγιστο όριο καθορίζεται από τη μέγιστη ισχύ όλων των ελάχιστων μονοπατιών μήκους που οδηγούν στον δέκτη. Στο παράδειγμά η μέγιστη τιμή είναι 6.

Βήμα 4ο: Με τον καθορισμό της μέγιστης τιμής, κάθε κόμβος σε ένα μονοπάτι ο οποίος δεν έχει εκφράσει εκτίμηση για τον δέκτη μπορεί να υπολογίσει τη βαθμολογία ως τον σταθμισμένο μέσο όρο των αξιολογήσεων χρησιμοποιώντας τον παρακάτω μαθηματικό τύπο.

$$t_{is} = \frac{\sum_{j \in suc(i) | t_{ij} \geq max} t_{ij} t_{js}}{\sum_{j \in suc(i) | t_{ij} \geq max} t_{ij}}$$

Η μεταβλητή t_{ij} αντιπροσωπεύει την αξιολόγηση ενός κόμβου i για τον κόμβο j. Η συνάρτηση $suc(i)$ επιστρέφει όλους τους κόμβους που είναι διάδοχοι του κόμβου i. Ο τύπος λαμβάνει υπόψη μόνο βαθμολογίες από κόμβους οι οποίες αξιολογούνται είτε στο μέγιστο όριο είτε πάνω από αυτό του προκατόχου τους. Κάθε αξιολόγηση έχει σταθμιστεί με την αξιολόγηση που λαμβάνει ένας διάδοχος από τον προκάτοχό του. Εντός του δικτύου στο παράδειγμά μας, οι κόμβοι N5, N6 και N7 έχουν ήδη βαθμολογήσει τον δέκτη N9. Ο κόμβος N3 συμπεράνει τη βαθμολογία του για τον δέκτη από τις αξιολογήσεις των κόμβων N5 και N6. Δεν λαμβάνει υπόψη την αξιολόγηση από τον κόμβο N7 διότι η βαθμολογία αυτού του κόμβου είναι κάτω από το μέγιστο όριο.

4.6.3 Ο αλγόριθμος SUNNY

Οι U. Kuter και J. Golbeck στο [65] ανέπτυξαν έναν αλγόριθμο εμπιστοσύνης ο οποίος χρησιμοποιεί μια πιθανολογική τεχνική δειγματοληψίας για την εκτίμηση της αυτοπεποίθησης του χρήστη σχετικά με τις πληροφορίες εμπιστοσύνης από καθορισμένες πηγές. Υπολογίζεται μια εκτίμηση εμπιστοσύνης η οποία βασίζεται μόνο στις πηγές πληροφόρησης με υψηλά επίπεδα αυτοπεποίθησης, ανεξάρτητα από το μήκος διαδρομής. Αυτό έρχεται σε αντίθεση με τον αλγόριθμο TidalTrust [64] που εξετάσαμε παραπάνω ο οποίος λαμβάνει υπόψη μόνο τις συντομότερες διαδρομές. Επίσης, διαφέρει από άλλους αλγορίθμους διότι είναι ο πρώτος αλγόριθμος ο οποίος περιλαμβάνει ένα μέτρο αυτοπεποίθησης στον υπολογισμό του.

Είναι πιο ακριβής από τον αλγόριθμο TidalTrust της Golbeck στο πλαίσιο των δοκιμών σε δεδομένα του κοινωνικού δικτύου FilmTrust.

Procedure SUNNY(T, n_0, n_∞)
$B_T \leftarrow$ GENERATEBN(T)
for every leaf node n in B_T do
$decision[n] \leftarrow$ UNKNOWN
$\langle P_\perp(n_0), P_\top(n_0)\rangle \leftarrow$ SAMPLE-BOUNDS(B_T)
for every leaf node n in B_T do
set the lower and upper probability bounds such that
$P_\perp(n) = P_\top(n) = 1.0$
$\langle P'_\perp(n_0), P'_\top(n_0)\rangle \leftarrow$ SAMPLE-BOUNDS(B_T)
if $|P'_\top(n_0) - P_\top(n_0)| < \epsilon$ and $|P'_\perp(n_0) - P_\perp(n_0)| < \epsilon$
then $decision[n] \leftarrow$ TRUE
else $decision[n] \leftarrow$ FALSE
return (COMPUTE-TRUST$(B_T, decision)$)

Εικόνα 4-16 Βήματα αλγορίθμου SUNNY

Στην παραπάνω εικόνα απεικονίζεται ο ψευδοκώδικας του αλγορίθμου Sunny για τον υπολογισμό των τιμών αυτοπεποίθησης και εμπιστοσύνης σε ένα δίκτυο. Τα δεδομένα εισόδου αποτελούνται από ένα πρόβλημα εμπιστοσύνης (T, n_0, n_1). Με αυτά τα δεδομένα εισόδου, ο αλγόριθμος Sunny παράγει ένα μπεϋζιανό δίκτυο BT. Στη συνέχεια, παράγει εκτιμήσεις για τα κατώτερα και τα ανώτερα όρια των τιμών αυτοπεποίθησης κάθε κόμβου στο BT δίκτυο για τον κόμβο δέκτη. Ο Sunny χρησιμοποιεί αυτές τις εκτιμήσεις με σκοπό να λάβει μια απόφαση για κάθε κόμβο στο BT. Η απόφαση αυτή αφορά αν θα πρέπει να συμπεριλάβει αυτόν τον κόμβο στον τελικό υπολογισμό εμπιστοσύνης ή όχι. Αυτό επιτυγχάνεται με τη χρήση μιας πιθανοκρατικής τεχνικής δειγματοληψίας προκειμένου να υπολογιστεί η τιμή αυτοπεποίθησης της πηγής n_0 στον δέκτη n_1 στο δίκτυο BT. Η πιθανοκρατική λογική δειγματοληψίας του αλγορίθμου εκτελεί διαδοχικές προσομοιώσεις σε ένα μπεϋζιανό δίκτυο ελέγχοντας αν κάθε μεταβλητή παίρνει μια από τις τιμές True ή False.

Η υπορουτίνα Sample-Bounds εκτελεί την πιθανοκρατική διαδικασία δειγματοληψίας, η οποία παρέχει έναν τρόπο για την εκτίμηση του άνω και κάτω φράγματος, σχετικά με την αξία της εμπιστοσύνης ενός κόμβου n. Σε κάθε προσομοίωση, η Sample-Bounds επιλέγει όλους τους «γονείς» του n. Στη συνέχεια, ο αλγόριθμος Sunny χρησιμοποιεί αυτά τα όρια σε μια hill-climbing αναζήτηση προκειμένου να οριστικοποιηθεί πραγματικά μια απόφαση για το αν θα συμπεριλάβει ή θα αποκλείσει έναν κόμβο στον υπολογισμό εμπιστοσύνης. Όταν ολοκληρωθούν όλες οι αποφάσεις επί όλων των κόμβων του δικτύου BT ο Sunny υπολογίζει την τιμή εμπιστοσύνης για την πηγή εκτελώντας μια προς τα πίσω αναζήτηση από τους κόμβους του δικτύου BT προς την πηγή.

Η υπορουτίνα Compute-Trust εκτελεί την προς τα πίσω αναζήτηση. Σε κάθε επανάληψη κατά τη διάρκεια αυτής της αναζήτησης, η αξία εμπιστοσύνης ενός κόμβου υπολογίζεται με βάση τις αξίες της εμπιστοσύνης μεταξύ αυτού του κόμβου και των άμεσων γονέων του και των αξιών εμπιστοσύνης των γονέων στον κόμβο δέκτη που έχουν ήδη υπολογιστεί στην αναζήτηση. Αν και υπάρχουν αρκετοί τρόποι για τον συνδυασμό αυτών των τμημάτων πληροφορίας εμπιστοσύνης, ο Sunny χρησιμοποιεί την ίδια μέθοδο του σταθμισμένου μέσου όρου που χρησιμοποιείται και στον αλγόριθμο TidalTrust.

4.6.4 Εξατομικευμένη εμπιστοσύνη

Ο TrustWebRank ο οποίος παρουσιάζεται στο [66] είναι μια εξατομικευμένη μετρική εμπιστοσύνης η οποία χρησιμοποιεί την αρχή της τοπικότητας για να κατασκευάσει την (ενδεχομένως έμμεση) εμπιστοσύνη μεταξύ των κόμβων i και j από την υπάρχουσα άμεση εμπιστοσύνη μεταξύ όλων των κόμβων. Η μετρική καθορίζεται λαμβάνοντας υπόψη ένα σενάριο όπου διάφοροι κόμβοι έχουν σχέση εμπιστοσύνης ο ένας με τον άλλο και ότι επίσης οι κόμβοι αυτοί είναι σε θέση να αξιολογήσουν

ορισμένα αντικείμενα. Συνήθως κάθε κόμβος στοχεύει στη γνώση της άποψης άλλων σχετικά με άγνωστα αντικείμενα και τον καθορισμό αυτών που είναι αξιόπιστα. Φυσικά, αυτό υπονοεί ότι κάθε κόμβος θα πρέπει επίσης να αξιολογήσει την εμπιστοσύνη των άλλων. Δεδομένου ότι σε ένα κοινωνικό δίκτυο ένας κόμβος γνωρίζει άμεσα μόνο λίγους κόμβους (γείτονες), πρέπει να αξιολογήσει τους άλλους κόμβους με έναν έμμεσο τρόπο.

Στη συνέχεια, παρουσιάζουμε εν συντομία τους βασικούς ορισμούς που απαιτούνται για τον υπολογισμό της έμμεσης εμπιστοσύνης μεταξύ κάθε ζεύγους κόμβων. Συνήθως το σενάριο αυτό αντιπροσωπεύεται από ένα γράφημα όπου η απευθείας εμπιστοσύνη μεταξύ δύο κόμβων i και j αναπαρίσταται με μια ακμή η οποία τους συνδέει. Ορίζουμε το συγκεκριμένο δίκτυο ως ένα γράφο $G(V, L, T)$, όπου V είναι το σύνολο των κόμβων, L είναι το σύνολο των ακμών και T είναι το σύνολο των ετικετών (τιμές εμπιστοσύνης). Συνεπώς, έχοντας $i, j \in V$, η ακμή (i, j) σημαίνει ότι ο i εμπιστεύεται τον j με το επίπεδο $T_{ij} \in T$. Επίσης, ορίζεται N_iτο σύνολο των γειτόνων του i, ως $\forall j \in V: \exists (i, j) \in L$.

Θεωρούμε $T_{ij} \in T$ (ορίζεται [0,1]) τον πίνακα εμπιστοσύνης ο οποίος περιγράφει το παραπάνω δίκτυο όπου $T_{ij} = 0$ εάν $(i, j) \notin L$. Η εξίσωση (1) αναπαριστά έναν στοχαστικό πίνακα S ο οποίος προκύπτει από την κανονικοποίηση του T [66].

$$S_{ij} = \frac{T_{ij}}{\sum_{k \in N_i} T_{ik}} \text{ (1)}$$

θεωρώντας ότι η κανονικοποίηση κρύβει το επίπεδο εμπιστοσύνης του κόμβου i στους γείτονες του στην πραγματικότητα είτε ισχύει $T_{ij} = 0.01, \forall\, j \in N_i$ είτε $T_{ij} = 1.0, \forall\, j \in N_i$ τα αποτελέσματα για το S_{ij} είναι τα ίδια. Η εμπιστοσύνη του κόμβου i στον κόμβο j ορίζεται ως εξής [66]:

$$\check{T}_{ij} = \sum_{k \in N_i} S_{ik} \cdot \check{T}_{kj} \text{ (2)}$$

όπου κάθε στοιχείο αναπαριστά είτε την απευθείας εμπιστοσύνη, αν υπάρχει, είτε την έμμεση εμπιστοσύνη. Η εξίσωση (2) λύνει το πρόβλημα της κεντρικότητας αλλά δεν λαμβάνει υπόψη με τον σωστό τρόπο της απευθείας εμπειρίας. Το επόμενο βήμα είναι η εισαγωγή της απευθείας εμπειρίας η οποία αναπαρίσταται μέσω του S_{ij} [66].

$$\check{T}_{ij} = S_{ij} + \beta \sum_{k \in N_i} S_{ik} \cdot \check{T}_{kj}\ \forall i, j\ where\ \beta \in [0,1] \text{ (3)}$$

$$\check{T} = S + \beta S \cdot \check{T} \text{ (4)}$$

Η εξίσωση (4) για να επιλυθεί απαιτεί την αντιστροφή του πίνακα το οποίο δημιουργεί υπολογιστικά προβλήματα ειδικά σε μεγάλα γραφήματα. Εντούτοις, είναι εφικτή η επίλυση των παραπάνω εξισώσεων χάρη στην επαναληπτική μέθοδο της εξίσωσης (5) [66].

$$\check{T}_{ij}^{(k+1)} = S_{ij} + \beta \sum_{l \in N_i} S_{il} \cdot \check{T}_{ij}^{(k)}\ where\ \check{T}_{ij}^{(k)}\ is\ computed\ at\ step\ k \text{ (5)}$$

Είναι δυνατό να συμπεράνουμε ότι υπό ορισμένες συνθήκες υπάρχει μια μοναδική μη τετριμμένη λύση στην εξίσωση (4). Ο πρώτος όρος (S_{ij}) λαμβάνει υπόψη την άμεση εμπιστοσύνη και αφού $\beta < 1$, υπερισχύει στην έμμεση εμπιστοσύνη.

Προτείνεται επίσης μια κατανεμημένη υλοποίηση του TrustWebRank η οποία προέρχεται από ερμηνεία της εξίσωσης 5 σε ένα κατανεμημένο πλαίσιο. Στον προτεινόμενο αλγόριθμο τόσο η αξιολόγηση όσο και η αποθήκευση κατανέμονται με αποτέλεσμα κάθε κόμβος i να είναι υπεύθυνος για τον υπολογισμό και την αποθήκευση της μη κατευθυνόμενης εμπιστοσύνης $\check{T}_{ij}$ προς τον κόμβο j. Ένας τέτοιος υπολογισμός πραγματοποιείται από τον κόμβο i ανταλλάσσοντας μηνύματα με τη γειτονιά του. Ο κατανεμημένος αλγόριθμος TrustWebRank παρουσιάζεται στον Αλγόριθμο 1 που ακολουθεί [66]:

Algorithm 1 Distributed TrustWebRank Algorithm

each node $i \in V$ **do:**
  $k = 0$
  repeat
    Query all nodes $l \in N_i$ for $T_{lj}^{(k)}$
    Compute $\tilde{N}_i^{(k+1)}$ as $\{j \in V : T_{lj}^{(k)} \neq 0\} \bigcup N_i$
    Compute $\tilde{T}_{ij}^{(k+1)} = S_{ij} + \beta \sum_{l \in N_i} S_{il} \cdot \tilde{T}_{lj}^{(k)}$ where $j \in \tilde{N}_i^{(k+1)}$
    Compute $\delta = ||\tilde{T}_{ij}^{(k+1)} - \tilde{T}_{ij}^{(k)}||$
    $k = k + 1$
  until $\delta < \varepsilon$

Εικόνα 4-17 Αλγόριθμος (1) Trust WebRank

Για χάρη της σαφήνειας, καθορίζουμε την έμμεση γειτονιά $\check{N}_i^{(k)}$ ενός κόμβου i στο βήμα k ως την ένωση των έμμεσων γειτόνων i και των έμμεσων γειτόνων $\check{N}_i^{(k-1)}$ στον βήμα $k-1$ για όλους τους κόμβους $l \in N_i$. Να σημειώσουμε ότι λόγω της κατανεμημένης φύσης του αλγορίθμου, κάθε κόμβος αρχικά γνωρίζει μόνο τους απευθείας γείτονες του με αποτέλεσμα να ισχύει $\check{N}_i^{(0)} \equiv N_i$.

Συνεπώς, στο πρώτο βήμα ο κόμβος είναι σε θέση να υπολογίσει μόνο τις αξίες εμπιστοσύνης των παρακείμενων κόμβων του (i.e. $\check{T}_{ij}^{(1)} = S_{ij}$). Με δεδομένο όμως ότι οι επαναλήψεις συνεχίζουν να εκτελούνται, ένας κόμβος λαμβάνει πληροφορίες από κόμβους οι οποίοι είναι πολύ μακριά από αυτόν. Επομένως, γενικά το $\check{N}_i^{(k)}$ενός κόμβου i μεγαλώνει με το πέρασμα του χρόνου και κάποια στιγμή περιλαμβάνει όλους τους κόμβους στο V, i.e $|\check{N}_i^{(k)}| = |V|$. Όπως φαίνεται και στον Αλγόριθμο 1 μετά τον υπολογισμό της μη κατευθυνόμενης γειτονιάς κάθε κόμβος συνεχίζει τον υπολογισμό της έμμεσης εμπιστοσύνης προς εκείνους τους κόμβους που περιλαμβάνονται στο $\check{N}_i^{(k+1)}$.

Το τελευταίο βήμα του αλγορίθμου είναι ο υπολογισμός του τοπικού υπολοίπου δ, το οποίο ορίζεται ως $|\check{T}_{ij}^{(k+1)} - \check{T}_{ij}^{(k)}|$. Ένας κόμβος θεωρείται ότι συγκλίνει όταν το τοπικό υπόλοιπο δ γίνει μικρότερο από ένα δεδομένο κατώφλι ε. Είναι αξιοσημείωτο ότι δεν συγκλίνουν ταυτόχρονα όλοι οι κόμβοι λόγω των τοπικών κριτηρίων σύγκλισης.

Συνοψίζοντας, η εργασία [66] προτείνει μια κατανεμημένη εφαρμογή της μετρικής TrustWebRank. Ο αλγόριθμος στοχεύει στη μετακίνηση από την κεντρικότητα προς την τοπικότητα δικαιολογώντας την κοινωνική συμπεριφορά όπου κάθε χρήστης παρέχει εμπιστοσύνη για άλλους χρήστες. Εντούτοις, η αρχική διατύπωση της μετρικής TrustWebRank δεν ισχύει άμεσα σε ένα κατανεμημένο πλαίσιο. Ο αλγόριθμος που προτείνεται υπολογίζει αυτή τη μετρική χρησιμοποιώντας μόνο τις τοπικές πληροφορίες κόμβων, επιτρέποντας κατά συνέπεια την εφαρμογή του TrustWebRank. Η απόδοση του κατανεμημένου TrustWebRank πραγματοποιήθηκε χρησιμοποιώντας ένα σύνολο δεδομένων που εξήχθη από το Epinions.com, το οποίο είναι ένα γνωστό δίκτυο συστάσεων αρκετά μεγάλο για να επιτρέπει μη τετριμμένα πειράματα αντικατοπτρίζοντας την πραγματική εμπιστοσύνη των χρηστών.

4.6.5 Σχετικές εργασίες

Στην εργασία [67], αρχικά εισάγεται μια πολύπλοκη κοινωνική δομή δικτύου η οποία λαμβάνει υπόψη πληροφορίες εμπιστοσύνης και κοινωνικές σχέσεις και μια νέα έννοια την ποιότητα εμπιστοσύνης (QoT – Quality of Trust). Για την επιλογή της βέλτιστης κοινωνικής διαδρομής εμπιστοσύνης με περιορισμούς στην ποιότητα εμπιστοσύνης (QoT) σε σύνθετα κοινωνικά δίκτυα προτείνεται ο H_OSTP ένας αποτελεσματικός ευριστικός αλγόριθμος. Τα αποτελέσματα των πειραμάτων που πραγματοποιούνται σε ένα πραγματικό σύνολο δεδομένων κοινωνικού δικτύου αποδεικνύουν ότι ο H_OSTP ξεπερνά σημαντικά

τις υπάρχουσες μεθόδους τόσο σε χρόνο όσο και σε εκτέλεση για την επιλογή της βέλτιστης διαδρομής εμπιστοσύνης.

Τα κοινωνικά δίκτυα έχουν σημαντικό αντίκτυπο στον τρόπο με τον οποίο οι χρήστες του Διαδικτύου επικοινωνούν, αναζητούν και μοιράζονται δεδομένα. Πολλοί πιστεύουν ότι αυξάνοντας τις online αγορές μέσω των κοινωνικών δικτύων μπορούμε να βελτιώσουμε την εμπιστοσύνη και να αυξήσουμε την ικανοποίηση των χρηστών. Στην εργασία [68], γίνεται μια λεπτομερής μελέτη των Overstock δημοπρασιών μέσω της αντίστοιχης ιστοσελίδας η οποία έχει ενσωματώσει πρόσφατα κοινωνικές συνδέσεις σε προφίλ χρηστών. Χρησιμοποιώντας τα στοιχεία σχετικά με τις συνδέσεις μεταξύ των περίπου 400.000 χρηστών του Overstock, αξιολογήθηκαν οι επιπτώσεις των κοινωνικών συνδέσμων στις επιχειρηματικές συναλλαγές. Τα αποτελέσματα δείχνουν ότι ενώ η πλειοψηφία των χρηστών δεν ασχολείται με την κοινωνική δικτύωση, όσοι συναλλάσσονται με φίλους από το κοινωνικό τους δίκτυο αποκομίζουν εν γένει σημαντικά οφέλη με τη μορφή της μεγαλύτερης ικανοποίησης.

Οι M. Lesani και S. Bagheri στο [69] θεωρούν το πρόβλημα της εμπιστοσύνης σε ένα κοινωνικό δίκτυο ότι μπορεί να συναντήσει αντιφατικές πληροφορίες. Προτείνουν, ένα ασαφές γλωσσικό επίπεδο για να καθορίσουν την εμπιστοσύνη σε άλλους χρήστες και να αναπτύξουν έναν αλγόριθμο ο οποίος βασίζεται σε αυτά. Αυτός ο αλγόριθμος υπολογίζει την εμπιστοσύνη από την ισχυρότερη και πιο σύντομη διαδρομή καθώς επιτελεί μια αναζήτηση πρώτα σε πλάτος (breadth-first like search) μέσα από τους κόμβους για τον εντοπισμό των συντομότερων διαδρομών. Ο αλγόριθμος αυτός είναι σε θέση να χειριστεί συγκρουόμενες αξίες εμπιστοσύνης με χρήση γλωσσικής έκφρασης (π.χ. χαμηλή, μεσαία, υψηλή) η οποία αποτελεί πιο εύκολο τρόπο για τους χρήστες στην αντιστοίχιση της εμπιστοσύνης. Προσομοιώνεται και συγκρίνεται με τον αλγόριθμο TidalTrust με τα αποτελέσματα να δείχνουν ότι ο ασαφής αλγόριθμος παρέχει πιο ακριβείς πληροφορίες. Ωστόσο, λόγω του ότι χρησιμοποιεί τον αλγόριθμο TidalTrust ως βάση έχει τα ίδια μειονεκτήματα με αυτόν.

Στην εργασία [70], έχει αναπτυχθεί ένα επίσημο πλαίσιο των συστημάτων διάδοσης εμπιστοσύνης, εισάγοντας την τυπική και υπολογιστική αντιμετώπιση της διάδοσης της δυσπιστίας. Αποδεικνύεται επίσης ότι μέσω ενός μικρού αριθμού απόψεων εμπιστοσύνης/δυσπιστίας ανά άτομο, δίνεται η δυνατότητα της πρόβλεψης της εμπιστοσύνης μεταξύ δύο ατόμων στο σύστημα με μεγάλη ακρίβεια. Η εργασία αυτή αποτελεί την πρώτη προσέγγιση η οποία ενσωματώνει τη δυσπιστία σε έναν υπολογισμό διάδοσης εμπιστοσύνης.

Το SocialTrust, το οποίο προτάθηκε από τους J. Caverlee et al. στο [71] είναι ένα πλαίσιο συγκέντρωσης εμπιστοσύνης με σκοπό την καθιέρωση της εμπιστοσύνης στα κοινωνικά δίκτυα. Ένα βασικό χαρακτηριστικό αυτού του μηχανισμού είναι η δυναμική αναθεώρηση της εμπιστοσύνης η οποία διαφοροποιεί την ποιότητα σχέσης από την εμπιστοσύνη και περιλαμβάνει έναν εξατομικευμένο μηχανισμό ανατροφοδότησης στο χρήστη με στόχο την προσαρμογή στις εξελίξεις του κοινωνικού δικτύου. Το SocialTrust ενημερώνει τις αξίες εμπιστοσύνης μέσω της δυναμικής αναθεώρησης των αξιολογήσεων εμπιστοσύνης, σύμφωνα με τις εξής τρεις συνιστώσες: την τρέχουσα βαθμολογία, την ιστορικότητα και την προσαρμογή στην αλλαγή.

Στον αλγόριθμο RN-Trust οι M. Taherian et al. στο [72] προτείνουν την έννοια της «αντίστασης» σε ένα δίκτυο, η οποία είναι παρόμοια με την ιδέα της ηλεκτρικής αντίστασης, για την προσομοίωση των δικτύων εμπιστοσύνης. Παρόμοια με τον τρόπο που ο αλγόριθμος TidalTrust αποδίδει τιμές εμπιστοσύνης, οι τιμές εμπιστοσύνης σε αυτόν τον αλγόριθμο είναι συνεχείς τιμές του εύρους [0,1]. Αυτός ο αλγόριθμος εμπιστοσύνης επιλύει πολλά προβλήματα. Για παράδειγμα, δεν χρειάζεται να αγνοήσει την εμπιστοσύνη η οποία προέρχεται από τη μεγαλύτερη διαδρομή παρά μόνο αυτή της συντομότερης διαδρομής δίνοντας έτσι μια πληρέστερη επισκόπηση στον καθορισμό της αξίας εμπιστοσύνης. Αυτό ανταποκρίνεται στη μεγάλη αδυναμία του αλγόριθμο TidalTrust όπου ορισμένα στοιχεία κατά μήκος της διαδρομής τα οποία επηρεάζουν την ακρίβεια του υπολογισμού είναι πιθανό να αγνοούνται. Επιπλέον, ο αλγόριθμος είναι πολύ απλός και η πολυπλοκότητα στο χρόνο του είναι πολυωνυμική, επομένως εξαιρετικά επεκτάσιμη.

4.7 Εύρεση ειδικών

Η *εύρεση ειδικών* (finding experts) σύμφωνα με το [73] είναι ένα από τα πιο σημαντικά θέματα της εξόρυξης γνώσης σε κοινωνικά δίκτυα. Διεκπεραιώνει την εργασία εντοπισμού των σωστών ατόμων με τις κατάλληλες δεξιότητες και γνώσεις για ένα συγκεκριμένο θέμα. Απαντά στο ερώτημα: *«Ποιοι είναι οι ειδικοί στο θέμα X;»* Τυπικά, ένα κοινωνικό δίκτυο μπορεί να οριστεί ως ένα γράφημα $G = (V, E)$, όπου $v \in V$ αντιπροσωπεύει ένα άτομο του κοινωνικού δικτύου και $e_{ij}^{t} \in E$ αντιπροσωπεύει μια σχέση τύπου t μεταξύ των ατόμων v_i και v_j. Το έργο του εντοπισμού ειδικών ορίζεται ως εξής: δεδομένου ενός ερωτήματος q, να βρεθεί ένα υποσύνολο ατόμων του κοινωνικού δικτύου με απεικόνιση αυτών σε μία λίστα κατάταξης.

Αλγόριθμοι που βασίζονται στη δομή των συνδέσμων όπως για παράδειγμα οι PageRank και HITS, μπορούν να χρησιμοποιηθούν για την ανάλυση των σχέσεων σε ένα κοινωνικό δίκτυο, οι οποίοι θα μπορούσαν να βελτιώσουν την απόδοση του εντοπισμού ειδικών. Ωστόσο, τόσο ο PageRank όσο και ο HITS παρουσιάζουν ένα κοινό πρόβλημα το οποίο ονομάζεται *topic drift* και καθιστά τους πιο εσωτερικούς συνδέσμους στο δίκτυο να τείνουν να κυριαρχήσουν [73].

Εικόνα 4-18 Πρόβλημα εύρεσης ειδικών

Ο όρος *ειδικός* ή *εμπειρογνώμονας* χρησιμοποιείται για να χαρακτηρίσει ένα άτομο/φορέα με υψηλό βαθμό δεξιοτήτων ή γνώσεων σε ένα συγκεκριμένο θέμα. Στην ιδανική περίπτωση, δεδομένου μιας εργασίας και ενός συνόλου υποψηφίων, κάποιος επιθυμεί να εντοπίσει αποτελεσματικά το σωστό σύνολο ειδικών οι οποίοι θα εκτελέσουν αποτελεσματικά τη συγκεκριμένη εργασία. Η πληθώρα στοιχείων τα οποία παρακολουθούν την παρουσία ατόμων και την τεχνογνωσία επιβάλλει αρκετές προκλήσεις στο πρόβλημα εντοπισμού ειδικών.

Ας αναλογιστούμε για παράδειγμα τα δεδομένα από μια μεγάλη εταιρεία η οποία καταγράφει την εσωτερική δραστηριότητα των εργαζομένων (πχ. έγγραφα, σεμινάρια, ξένες γλώσσες, διπλώματα ευρεσιτεχνίας, ηλεκτρονικά μηνύματα κτλ.). Λαμβάνοντας υπόψη ένα τέτοιο σύνολο δεδομένων, οι διαχειριστές χρειάζεται να μετρούν το βαθμός της εμπειρίας των εργαζομένων σε σχέση με διαφορετικά θέματα. Αυτή η γνώση επιτρέπει στους διαχειριστές να προσδιορίσουν τους κατάλληλους εργαζόμενους οι οποίοι θα μπορούσαν να αποδειχθούν πολύτιμοι για ένα συγκεκριμένο έργο. Ως εκ τούτου, μια σημαντική πτυχή του εντοπισμού των ειδικών είναι η εύρεση διαφορετικών ατόμων από διαθέσιμα δεδομένα όπως παρουσιάζεται στην εργασία [74].

4.7.1 Διάδοση εύρεσης ειδικών

Το πρόβλημα εντοπισμού ειδικών μπορεί να οριστεί ως η πιθανότητα ένας υποψήφιος c να είναι κατάλληλος για μια δεδομένη εργασία q. Εντούτοις, ο απευθείας υπολογισμός της πιθανότητας $P(c \mid q)$

δεν είναι εφικτός. Στην εργασία [75] παρουσιάζεται μια ειδική διαδικασία διάδοσης (propagation process) για την εύρεση ειδικών. Χρησιμοποιούνται οι συσχετίσεις μεταξύ των υποψηφίων με σκοπό τη διάδοση της πιθανότητας από τους υποψήφιους οι οποίοι είναι πιθανώς ειδικοί σε άλλους υποψηφίους.

Με τη διαδικασία αυτή υπολογίζεται το $P(c_y|c_x)$ το οποία εκφράζει την πιθανότητα ένας υποψήφιος c_y να είναι ειδικός με βάση τον ειδικό c_x . Έτσι, προκύπτει ένα σύνολο S από τους N υποψήφιους οι οποίοι έχουν τις περισσότερες πιθανότητες να είναι ειδικοί.

Στη συνέχεια, αναμένουμε ότι οι υποψήφιοι οι οποίοι έχουν δυνατές συσχετίσεις με ειδικούς, θα είναι και οι ίδιοι ειδικοί για αυτό αναθέτουμε ένα σκορ εξειδίκευσης (expertise score) από τους ειδικούς στο S σε αυτούς τους υποψήφιους. Για έναν υποψήφιο c_x στο σύνολο S και για έναν υποψήφιο c_y ο οποίος πρόκειται να εξεταστεί ισχύει ότι όσο λιγότερο πιθανό είναι ο c_x να είναι ειδικός και όσο η συσχέτισή τους $a(c_x, c_y)$, είναι αβέβαιη τόσο λιγότερο πιθανό είναι και ο c_y να είναι ειδικός. Το γεγονός αυτό δείχνει ότι όσο ο c_x διαδίδει σε άλλους υποψήφιους σύμφωνα με τις συσχετίσεις του ισχύει ότι: εάν ένας ειδικός c_x έχει μια πιθανότητα εξειδίκευσης $P(c_x)$ και έχει επίσης ω συσχετισμένους υποψήφιους, κάθε ένας από τους ω υποψήφιους c_y έχει μια συσχέτιση $a(c_x, c_y)$ η οποία θα λάβει ένα σκορ κλάσματος από τον c_x. Ο υπολογισμός του $P(c_y|c_x)$ δίνεται από την παρακάτω εξίσωση [75]:

$$P(c_y|c_x) = \frac{a(c_y,c_x)}{\sum_{c_i \epsilon \omega} a(c_i,c_x)} \beta P(c_x)$$

όπου β ελέγχει πόσο αντιδρά η επίδραση. Σε αυτή την περίπτωση, το πραγματικό σκορ διάδοσης του c_y θα είναι το άθροισμα των κλασματικών σκορ που έχει λάβει μέσω των συσχετισμένων υποψηφίων του στο S. Διαισθητικά, όσο αυξάνεται το σκορ εξειδίκευσης ενός υποψηφίου τόσο πιο πιθανό είναι να είναι ο υποψήφιος αυτός ειδικός. Η παραπάνω προσέγγιση είναι αρκετά όμοια με τον αλγόριθμο PageRank [73]. Εντούτοις, ο PageRank βασίζεται σε μια αμοιβαία ενίσχυση μεταξύ των σελίδων και του σκορ το οποίο υπολογίζεται επαναληπτικά μέχρι τη σύγκλιση.

4.7.2 Δημιουργία σχέσεων ανάμεσα σε ειδικούς

Στην εργασία [75] γίνεται μια προσπάθεια εύρεσης σχέσεων μεταξύ των υποψηφίων τόσο από τις ιστοσελίδες όσο και από τα μηνύματα ηλεκτρονικού ταχυδρομείου. Στις ιστοσελίδες τα πρόσωπα συνυπάρχουν σε κάποιο τοπικό περιεχόμενο το οποίο ενδέχεται να υποδηλώνει ότι συνδέονται κάτω από ένα θέμα ενώ η ύπαρξη της ηλεκτρονικής αλληλογραφίας υποδηλώνει επίσης την πιθανή συσχέτιση μεταξύ των προσώπων.

Μέσω της ανάλυσης του γραφήματος είμαστε σε θέση να προσδιορίσουμε την ισχύ της συσχέτισης μεταξύ δύο προσώπων. Η μεγάλη ισχύς υποδηλώνει ότι δύο πρόσωπα έχουν κοινά ενδιαφέροντα και συχνή επικοινωνία. Ένας διαισθητικός τρόπος προσδιορισμού της ισχύς της συσχέτισης μεταξύ δύο υποψηφίων είναι η μέτρηση της συνύπαρξής τους σε ένα έγγραφο. Τυποποιούμε την έννοια της συνύπαρξης δύο υποψηφίων c_x και c_y σε ένα έγγραφο d μέσω μιας δυαδικής συνάρτησης CO για όλα τα έγγραφα στην συλλογή των ιστοσελίδων [75]:

$$CO(c_y,c_x,d) = \begin{cases} 0 \text{ εάν } c_y, c_x \text{ δεν συνυπάρχουν στο } d \\ 1 \text{ εάν } c_y, c_x \text{ συνυπάρχουν στο } d \end{cases} \quad (3)$$

Η συσχέτιση μεταξύ των υποψηφίων c_x και c_y προκύπτει [75]:

$$a(c_y, c_x) = \sum_d CO(c_y, c_x, d) \quad (4)$$

Ωστόσο, η συνάρτηση της συνύπαρξης που ορίσαμε παραπάνω δεν είναι αρκετά ευαίσθητη για να εκπροσωπήσει τη συσχέτιση με ακρίβεια. Με δεδομένο ότι μπορεί να υπάρχουν πολλά θέματα μέσα σε ένα έγγραφο, η περαιτέρω απόσταση μεταξύ δύο υποψηφίων καθιστά λιγότερο πιθανό να είναι υπό το ίδιο θέμα. Αυτή η παρατήρηση σημαίνει ότι αποδυναμώνεται η ισχύς της ένωσης των υποψηφίων. Για να επιτευχθεί αυτό περιγράφεται το πιθανό σχήμα το οποίο εμπλέκει τον αντίστροφο του αριθμού των λέξεων μεταξύ των συνυπάρξεων ως η ισχύ της συνύπαρξης.

Στην εργασία [75] προτείνονται επίσης αρκετές στρατηγικές για τη δημιουργία συσχέτισης μεταξύ υποψηφίων μέσω της ανάλυσης των προτύπων ηλεκτρονικής επικοινωνίας. Ο πιο απλός τρόπος της εκτίμησης της ισχύς της σύνδεσης μεταξύ δύο υποψηφίων είναι η μέτρηση του ποσού της ηλεκτρονικής τους αλληλογραφίας. Οι υποψήφιοι c_x και c_y σχετίζονται εάν εμφανίζονται μαζί στο *from, to* ή *cc* πεδίου ενός μηνύματος ηλεκτρονικού ταχυδρομείου *e*. Συνεπώς, προκύπτει μια δυαδική συνάρτηση *EC(c_x , c_y ,e)* για την αναπαράσταση αυτής της σύνδεσης [75]:

$$EC(c_y, c_x, e) = \begin{cases} 0 \text{ εάν } c_y, c_x \text{ δεν εμφανίζονται στα πεδία } from, to, cc \text{ ενός μηνύματος } e \\ 1 \text{ εάν } c_y, c_x \text{ εμφανίζονται στα πεδία } from, to, cc \text{ fields ενός μηνύματος } e \end{cases}$$

Τα μηνύματα ηλεκτρονικού ταχυδρομείου αποτελούν μια δενδρική δομή. Η ρίζα του δέντρου είναι το πρώτο μήνυμα το οποίο έστειλε κάποιος υποψήφιος και στη συνέχεια το δέντρο επεκτείνεται στις απαντήσεις άλλων ατόμων σε αυτό το μήνυμα ή στην προώθηση του μηνύματος σε άλλους με σκοπό τη συνέχιση της επικοινωνίας. Με δεδομένο ότι οι περισσότεροι υποψήφιοι συμμετέχουν σε μηνύματα ηλεκτρονικού ταχυδρομείου, θεωρείται ότι υποψήφιοι που σχετίζονται, εμφανίζονται στο ίδιο μήνυμα και είναι σε θέση να αντιμετωπίσουν το πρόβλημα ότι πολλές φορές ο πίνακας απλών μηνυμάτων είναι διάσπαρτος. Χρησιμοποιείται η δυαδική συνάρτηση $TC(c_y, c_x, t)$ για να εκφράσει τη σύνδεσή τους:

$$TC(c_y, c_x, t) = \begin{cases} 0 \text{ εάν } c_y, c_x \text{ δεν εμφανίζονται στα πεδία } from, to, cc \text{ ενός "δένδρου" } t \\ 1 \text{ εάν } c_y, c_x \text{ εμφανίζονται στα πεδία } from, to, cc \text{ ενός "δένδρου" } t \end{cases} \quad (6)$$

Ανακεφαλαιώνοντας, στην εργασία [75], αναλύεται ένας αλγόριθμος διάδοσης της τεχνογνωσίας με στόχο την προσαρμογή της σχέσης τεχνογνωσίας μεταξύ των υποψηφίων. Τα κοινωνικά δίκτυα μπορούν να κατασκευαστούν από δύο πηγές: από τη μία πλευρά, οι άνθρωποι αποστέλλουν μηνύματα ηλεκτρονικού ταχυδρομείου ο ένας στον άλλο με αποτέλεσμα η σχέση τεχνογνωσίας να περιέχεται στα πρότυπα της επικοινωνίας και από την άλλη πλευρά, λαμβάνοντας υπόψη στατιστικές σχέσεις που προέρχονται από συνυπάρξεις ατόμων στις ιστοσελίδες.

Ενώ από τη μια όταν κάποιος απαντά σε πολλές ερωτήσεις σημαίνει ότι έχει υψηλή τεχνογνωσία από την άλλη κάποιος που κάνει πολλές ερωτήσεις δείχνει ότι έχει έλλειψη γνώσεων σε πολλά θέματα. Για το λόγο αυτό οι Zhang *et al.* στο [76] προτείνουν το *z-score* ως ένα μέτρο το οποίο συνδυάζει την ερώτηση και τα πρότυπα απάντησης ως εξής: εάν ένας χρήστης θέτει $n = q + a$ όπου q αποτελούν ερωτήσεις και a αποτελούν απαντήσεις, θα θέλαμε να μετρήσουμε πόσο διαφορετική είναι αυτή η συμπεριφορά από έναν τυχαίο χρήστη ο οποίος απαντά με πιθανότητα $p = 0.5$ και θέτει νέες ερωτήσεις με πιθανότητα $1 - p = 0.5$. Θα περιμέναμε από έναν τυχαίο χρήστη να θέσει $n * p = {}^{n}/_{2}$ απαντήσεις με μια σταθερή απόκλιση $\sqrt{n * p * (1-p)} = {}^{\sqrt{n}}/_{2}$. Το z-score μετρά πόσες τυπικές αποκλίσεις πάνω ή κάτω από την αναμενόμενη τυχαία τιμή βρίσκεται ένας χρήστης: $z = \frac{a - {}^{n}/_{2}}{{}^{\sqrt{n}}/_{2}} = \frac{a-q}{\sqrt{a+q}}$ [76]

Εάν ένας χρήστης θέτει ερωτήσεις και δίνει απαντήσεις σχετικά με την ίδια συχνότητα, το z-score του θα είναι κοντά στο 0. Εάν απαντά σε ερωτήσεις περισσότερο από το να θέτει ερωτήσεις το z-score θα είναι θετικό, αλλιώς αρνητικό.

Υπάρχει ένα πιθανό πρόβλημα στην καταμέτρηση του αριθμού των απαντήσεων που δίνει κάποιος ή του αριθμού οι οποίοι βοήθησαν. Ένας χρήστης ο οποίος απαντά στις 100 ερωτήσεις αρχάριων θα μπορούσε να πάρει την ίδια κατάταξη ως ειδικός με κάποιον που απαντά στις 100 ερωτήσεις έμπειρων χρηστών. Οι Zhang *et al.* στο [76] χρησιμοποίησαν τους αλγορίθμους PageRank και HITS στα δίκτυα ερωτήσεων-απαντήσεων τα οποία βασίζονται σε κοινότητες όπως τα *Yahoo!Answers* και *GoogleGroups*. Σε αυτά τα δίκτυα ένας κόμβος αναπαριστά έναν χρήστη και μια ακμή ξεκινά από έναν χρήστη ο οποίος έθεσε την αρχική ερώτηση και καταλήγει σε όποιον έχει απαντήσει σε αυτήν. Όταν ένας χρήστης A απαντά σε ενός άλλου χρήστη την ερώτηση B αυτό συχνά υποδεικνύει ότι ο A έχει περισσότερες γνώσεις στο συγκεκριμένο θέμα από ότι ο B. Επιπρόσθετα, εάν ο χρήστης B απαντήσει ερωτήσεις από τον C, τότε το σκορ εξειδίκευσης του θα έπρεπε να ενισχυθεί επειδή είναι σε θέση να απαντήσει μια ερώτηση από κάποιον ο οποίος έχει ένα συγκεκριμένο επίπεδο εξειδίκευσης. Η απλή αυτή ιδέα οδηγεί σε έναν αλγόριθμο όμοιου στυλ με τον PageRank ο οποίος ονομάζεται ExpertiseRank [76]:

Θεωρούμε ότι ο χρήστης A έχει απαντήσει ερωτήσεις για τους χρήστες $U_1, \ldots, U_n$, τότε το σκορ κατάταξης εξειδίκευσης του υποδηλώνεται από ER(A), το οποίο υπολογίζεται ως εξής [76]:

$$ER(A)=(1-a) + a\left(\sum_{i=1}^{n} \frac{ER(U_i)}{C(U_i)}\right)$$

Στην παραπάνω εξίσωση, το $C(U_i)$ είναι ο συνολικός αριθμός των χρηστών οι οποίοι έχουν βοηθήσει τον U_i (υπενθυμίζουμε ότι υπάρχει μια ακμή από τον U_i σε κάθε έναν από τους βοηθούς του) και α είναι ο συντελεστής απόσβεσης (damping factor) ο οποίος ορίζεται ανάμεσα στο 0 και 1.

Οι Zhang *et al.* στο [73] ασχολούνται με την εύρεση ειδικών σε ακαδημαϊκά δίκτυα συνεργασίας. Αρχικά, υπολογίζεται το σκορ εξειδίκευσης από τις προσωπικές πληροφορίες των υποψηφίων χρησιμοποιώντας πιθανολογικά μοντέλα ανάκτησης πληροφοριών. Στη συνέχεια, χρησιμοποιείται ένα μοντέλο διάδοσης για την ενημέρωση των σκορ. Η διαίσθηση πίσω από το μοντέλο διάδοσης είναι παρόμοια με αυτή που συναντάμε στους αλγορίθμους PageRank και HITS - εάν ένα άτομο έχει στενές συνδέσεις με πολλούς άλλους ειδικούς σε ένα θέμα τότε είναι πιθανό ότι το πρόσωπο αυτό είναι επίσης ένας ειδικός σε αυτό το θέμα. Το σκορ εξειδίκευσης ενός χρήστη v_i υποδηλώνεται ως $s(v_i)$ και ενημερώνεται χρησιμοποιώντας τον παρακάτω κανόνα [73]:

$$s(v_i)^{t+1} = s(v_i)^t + \sum_{v_j \in U} \sum_{e \in R_{ji}} w((v_j, v_i, e) s(v_j)^t$$

Στην παραπάνω εξίσωση το $w((v_j, v_i, e)$ είναι ο συντελεστής διάδοσης και $e \in R_{ji}$ είναι ένα είδος σχέσεων από το πρόσωπο v_j στο v_i. Το U υποδηλώνει το σύνολο των γειτόνων του v_i και το R_{ji} είναι το σύνολο όλων των σχέσεων μεταξύ του v_j και v_i .

4.7.3 Σχηματισμός ομάδας ειδικών

Στην ενότητα αυτή αναλύουμε το πρόβλημα *σχηματισμού ομάδας ειδικών* (expert-team formation) σε κοινωνικά δίκτυα. Ο στόχος είναι η εύρεση μιας ομάδας ειδικών στο δίκτυο οι οποίοι θα μπορούν να εκτελέσουν από κοινού μια εργασία με αποτελεσματικό τρόπο.

Οι Lappas et al. στο [74] αναλύουν τις υπολογιστικές πτυχές της διαδικασίας σχηματισμού ομάδας. Στην εργασία τους, θεωρούν ότι υπάρχει μια ομάδα από n ειδικούς $X = \{1, \ldots, n\}$, όπου κάθε ειδικός i έχει ένα σύνολο δεξιοτήτων X_i. Επίσης, γίνεται η υπόθεση ότι οι ειδικοί οργανώνονται σε ένα σταθμισμένο και μη κατευθυνόμενο κοινωνικό γράφημα $G = (X, E)$. Τα βάρη των ακμών του G θα πρέπει να ερμηνευτούν ως εξής: μια χαμηλού βάρους ακμή μεταξύ των κόμβων i και j εκφράζει ότι ο ειδικός i και ο ειδικός j μπορούν να συνεργαστούν ή να επικοινωνήσουν πολύ πιο εύκολα από ότι δύο ειδικοί οι οποίοι συνδέονται μέσω μιας ακμής με υψηλό βάρος. Τα βάρη αυτά μπορούν να ερμηνευτούν με

διαφορετικούς τρόπους σε διάφορες περιοχές εφαρμογών. Λαμβάνοντας υπόψη μια εργασία T η οποία απαιτεί μια σειρά δεξιοτήτων, ο στόχος είναι να βρεθεί ένα σύνολο ατόμων $X' \subseteq X$ το οποίο θα είναι σε θέση να ολοκληρώσει με επιτυχία την εργασία αυτή. Στη διατύπωση του προβλήματος, τα άτομα του X' είναι απαραίτητο να έχουν συλλογικά όλες τις δεξιότητες για την εκτέλεση της T καθώς και να είναι σε θέση να εργαστούν αποτελεσματικά ως ομάδα.

Οι Lappas *et al.* μετρούν την αποτελεσματικότητα της συνεργασίας με την έννοια του *κόστους επικοινωνίας* (communication cost) του υπογράφου στο G το οποίο περιλαμβάνει μόνο τα μέλη της ομάδας X. Ορίζουμε αυτόν τον υπογράφο ως $G[X']$. Οι επίσημοι ορισμοί του προβλήματος είναι οι ακόλουθοι [74]:

Ορισμός Σχηματισμού Ομάδας (Team Formation): Με δεδομένο ένα σύνολο n ατόμων στο σύνολο $X = \{1, \ldots, n\}$, έναν γράφο $G(X, E)$ και μια εργασία T, να βρεθεί το υποσύνολο $X' \subseteq X$, έτσι ώστε $(\cup_{i \in X'} X_i) \cap T = T$ και να ελαχιστοποιηθεί το επικοινωνιακό κόστος $Cc(X')$.

Στον παραπάνω ορισμό το $(\cup_{i \in X'} X_i) \cap T = T$ σημαίνει ότι οι δεξιότητες που τίθενται από τα μέλη της ομάδας X' ικανοποιούν τις απαιτήσεις για την εργασία T. Το επικοινωνιακό κόστος είναι σκόπιμα απροσδιόριστο. Οι Lappas *et al.* εστίασαν στο επικοινωνιακό κόστος μέσω της μοντελοποίησης του, ως διάμετρο του υπογράφου G ο οποίος ορίζεται από τα μέλη του X'. Μια δεύτερη εκδοχή του επικοινωνιακού κόστους μετράται σαν το κόστος του ελάχιστα επικαλυπτόμενου δέντρου (*minimum spanning tree)* το οποίο καλύπτει όλους τους κόμβους στο X'.

Διάμετρος R: δεδομένου ενός γράφου $G(X, E)$ και ενός συνόλου ατόμων $X' \subseteq X$, ορίζουμε τη διάμετρο επικοινωνιακού κόστους του X', ως $Cc - R(X')$, να είναι η διάμετρος του υπογράφου $G[X']$. Υπενθυμίζουμε ότι η διάμετρος ενός γράφου είναι η μεγαλύτερη διαδρομή ανάμεσα σε δύο κόμβους.

Ελάχιστα επικαλυπτόμενο δέντρο Mst: δεδομένου ενός γράφου $G(X, E)$ και ενός συνόλου ατόμων $X' \subseteq X$, ορίζουμε το ελάχιστα επικαλυπτόμενο δέντρο επικοινωνιακού κόστους του X', ως $Cc - Mst\,(X')$, να είναι το κόστος του ελάχιστα επικαλυπτόμενου δέντρου του υπογράφου $G[X']$. Υπενθυμίζουμε ότι το κόστος ενός επικαλυπτόμενου δέντρου είναι το άθροισμα των βαρών των ακμών του.

Το πρόβλημα σχηματισμού ομάδας με συνάρτηση επικοινωνιακού κόστους $Cc - R$ ονομάζεται *Diameter-Tf πρόβλημα*. Ομοίως, το πρόβλημα σχηματισμού ομάδας με συνάρτηση επικοινωνιακού κόστους $Cc - Mst$ ονομάζεται *Mst-Tf πρόβλημα*.

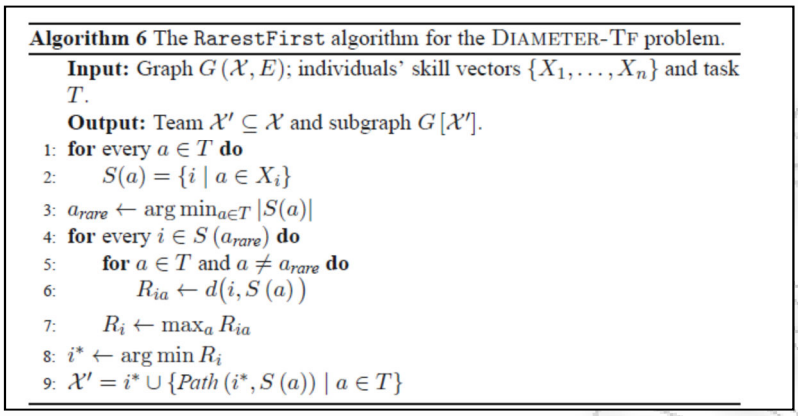
Για το πρόβλημα Diameter-T_F οι Lappas *et al.* προτείνουν τον Αλγόριθμο RarestFirst στο [74] του οποίου ο ψευδοκώδικας περιγράφεται στην εικόνα 4.19. Αρχικά, για κάθε δεξιότητα a η οποία απαιτείται για μια εργασία T, ο αλγόριθμος υπολογίζει το $S(a)$, την υποστήριξη του a, η οποία είναι το σύνολο των ατόμων στο X οι οποίοι έχουν αυτή τη δεξιότητα. Στη συνέχεια ο αλγόριθμος διαλέγει τη δεξιότητα $a_{rare} \in T$ με τη χαμηλότερη υποστήριξη $S(a_{rare})$ σημειώνοντας ότι τουλάχιστον ένα άτομο από το σύνολο $S(a_{rare})$ χρειάζεται να περιλαμβάνεται στη λύση.

Ανάμεσα σε όλους τους υποψηφίους του συνόλου $S(a_{rare})$, ο αλγόριθμος διαλέγει εκείνον ο οποίος οδηγεί στη μικρότερη διάμετρο του υπογράφου όταν συνδέεται με το κοντινότερο άτομο όλων των ομάδων υποστήριξης $S(a)$ όπου $a \in T$ και $a \neq a_{rare}$.

Για την κατανόηση του Αλγορίθμου 6 στο [74] ακολουθούν ορισμένες διευκρινήσεις. Για κάθε δύο κόμβους $i, i' \in X$, ορίζουμε τη συνάρτηση της απόστασης $d(i, i')$ ως το βάρος της πιο σύντομης διαδρομής ανάμεσα στους κόμβους i και i' του G. Αυτή η συνάρτηση απόστασης ανάμεσα στους κόμβους είναι μια μετρική και με αυτό τον τρόπο ικανοποιεί την τριγωνική ανισότητα.

Για κάθε ζευγάρι κόμβων, χρησιμοποιούμε το $Path(i, i')$ το οποίο αναπαριστά το σύνολο των κόμβων οι οποίοι είναι κατά μήκος της συντομότερης διαδρομής από το i στο i'. Ορίζουμε επίσης την απόσταση μεταξύ ενός κόμβου $i \in X$ και ενός συνόλου κόμβων $X' \subseteq X$ να ισούται με $d(i, X') = min_{i' \in X'} d(i, i')$.

Σε αυτή την περίπτωση, χρησιμοποιούμε το $Path(i, X')$ για την αναπαράσταση του συνόλου των κόμβων οι οποίοι εμφανίζονται κατά μήκος της πιο σύντομης διαδρομής από τον κόμβο i στον κόμβο j να ισούται με $argmin_{i' \in X'}\ d(i, i')$.



Algorithm 6 The RarestFirst algorithm for the DIAMETER-TF problem.

Input: Graph $G(\mathcal{X}, E)$; individuals' skill vectors $\{X_1, \ldots, X_n\}$ and task T.

Output: Team $\mathcal{X}' \subseteq \mathcal{X}$ and subgraph $G[\mathcal{X}']$.

1: **for** every $a \in T$ **do**
2: $S(a) = \{i \mid a \in X_i\}$
3: $a_{rare} \leftarrow \arg\min_{a \in T} |S(a)|$
4: **for** every $i \in S(a_{rare})$ **do**
5: **for** $a \in T$ and $a \neq a_{rare}$ **do**
6: $R_{ia} \leftarrow d(i, S(a))$
7: $R_i \leftarrow \max_a R_{ia}$
8: $i^* \leftarrow \arg\min R_i$
9: $\mathcal{X}' = i^* \cup \{Path(i^*, S(a)) \mid a \in T\}$

Εικόνα 4-19 Αλγόριθμος (6) Rarest First για το Diameter-TF πρόβλημα

Στη γραμμή 6 του αλγορίθμου το $d(i, S(a))$ εκφράζει απλά την ελάχιστη απόσταση των δύο κόμβων ως $min_{i' \in S(a)}\ d(i, i')$.

Επίσης, το $Path(i^*, S(a))$ στη γραμμή 9 αναφέρεται στο σύνολο των κόμβων του γραφήματος οι οποίοι βρίσκονται κατά μήκος της συντομότερης διαδρομής από τον κόμβο i^* στον κόμβο i', όπου i' έτσι ώστε $i' \in S(a)$ και $d(i^*, S(a)) = d(i^*, i')$. Ενώ όλα τα ζεύγη της συντομότερης διαδρομής έχουν υπολογιστεί εκ των προτέρων και οι κατακερματισμένοι πίνακες (hash tables) χρησιμοποιούνται για την αποθήκευση των δεξιοτήτων κάθε ατόμου και ένα άλλο σύνολο κατακερματισμένων πινάκων χρησιμοποιείται για την αποθήκευση των ατόμων οι οποίοι κατέχουν ένα συγκεκριμένο χαρακτηριστικό ο χρόνος εκτέλεσης του αλγορίθμου RarestFirst είναι $O(|S(a_{rare}) \mid \times n)$. Ο RarestFirst είναι ένας αλγόριθμος προσέγγισης για το Diameter-T_F πρόβλημα και αυτό οφείλεται κυρίως στο γεγονός ότι η απόσταση d είναι μια μετρική.

Για κάθε συνάρτηση απόστασης d ενός γραφήματος η οποία ικανοποιεί την τριγωνική ανισότητα, το κόστος $Cc - R$ της λύσης X', δίνεται από τον αλγόριθμο RarestFirst για μια συγκεκριμένη εργασία και είναι τουλάχιστον διπλάσιο από το κόστος $Cc - R$ της βέλτιστης λύσης X^*. Συνεπώς ισχύει $Cc - R(X') \leq 2 \cdot Cc - R(X*)$.

4.7.4 Σχετικές εργασίες

Τα έγγραφα μηνυμάτων ηλεκτρονικού ταχυδρομείου φαίνεται ότι προσφέρονται ιδιαίτερα για την εργασία της εύρεσης ειδικών, καθώς οι άνθρωποι συνήθως επικοινωνούν αυτά για τα οποία γνωρίζουν. Επιπλέον, επειδή οι άνθρωποι απευθύνουν ρητά ηλεκτρονικά μηνύματα ο ένας στον άλλο, τα κοινωνικά δίκτυα είναι πιθανό να περιέχονται στα πρότυπα αυτής της επικοινωνίας. Για να βρεθούν απαντήσεις σε αυτά τα ερωτήματα στην εργασία [77], συγκρίνονται δύο αλγόριθμοι για τον προσδιορισμό της εμπειρίας από ένα ηλεκτρονικό μήνυμα: μια προσέγγιση βασισμένη στο περιεχόμενο η οποία λαμβάνει υπόψη μόνο το κείμενο του μηνύματος και ένας αλγόριθμος κατάταξης γράφου (HITS) ο οποίος λαμβάνει υπόψη τόσο το κείμενο του μηνύματος όσο και τα πρότυπα επικοινωνίας. Σε πειράματα που πραγματοποιήθηκαν φάνηκε ότι για ένα σχετικά μικρό δείγμα μηνυμάτων ηλεκτρονικού ταχυδρομείου τα οποία συλλέχτηκαν από δύο διαφορετικούς οργανισμούς, ο αλγόριθμος κατάταξης γράφου έχει καλύτερα αποτελέσματα στον εντοπισμό ειδικών σε σχέση με το αλγόριθμο που βασίζεται στο περιεχόμενο.

Οι Campbell et al. στο [78] χρησιμοποιούν τον σύνδεσμο μεταξύ των δημιουργών και αποδεκτών μηνυμάτων για τη βελτίωση του αποτελέσματος εύρεσης ειδικών. Παρουσιάζουν ένα end-to-end σύστημα το οποίο δέχεται ως δεδομένα εισόδου τα εισερχόμενα μηνύματα του χρήστη και εξάγει το κοινωνικό δίκτυο του χρήστη και τις πληροφορίες των επαφών του. Το σύστημα εντοπίζει τα άτομα του ηλεκτρονικού μηνύματος μοναδικά, διαπιστώνει την παρουσία τους στο διαδίκτυο και συμπληρώνει αυτόματα τα πεδία ενός καταλόγου διευθύνσεων επικοινωνίας χρησιμοποιώντας τυχαία πεδία σε ένα είδος πιθανολογικού μοντέλου κατάλληλο για την εξαγωγή τέτοιου είδους πληροφοριών. Επιπλέον, μια σειρά από λέξεις κλειδιά τεχνογνωσίας εξάγονται και συσχετίζονται με κάθε άτομο.

Στην εργασία [79] παρουσιάζεται το Author-Recipient-Topic μοντέλο, ένα μπεϋζιανό δίκτυο για ανάλυση κοινωνικού δικτύου το οποίο ανακαλύπτει θέματα συζήτησης στις σχέσεις αποστολέα/αποδέκτη στο περιεχόμενο ενός μηνύματος ηλεκτρονικού ταχυδρομείου. Το μοντέλο αυτό συνδυάζει για πρώτη φορά την απευθείας συνδεσιμότητα γραφήματος της ανάλυσης κοινωνικών δικτύων με τη συσταδοποίηση των λέξεων to, from από πιθανολογικά γλωσσικά μοντέλα. Το μοντέλο θα μπορούσε να αποτελέσει ένα χρήσιμο συστατικό σε συστήματα αιτήσεων δρομολόγησης, εύρεσης ειδικών, σύστασης μηνυμάτων, καθορισμό προτεραιοτήτων και κατανόησης αλληλεπιδράσεων σε έναν οργανισμό.

ΠΡΑΚΤΙΚΟ ΜΕΡΟΣ

Στο Θεωρητικό Μέρος περιγράψαμε τα βασικά χαρακτηριστικά των κοινωνικών δικτύων και αναλύσαμε αλγορίθμους σε ότι αφορά θέματα ανίχνευσης κοινοτήτων, πρόβλεψης συνδέσμου, κατηγοριοποίησης κόμβων, εμπιστοσύνης, ανάλυση κοινωνικής επιρροής και εύρεσης ειδικών. Ένα συμπέρασμα που προκύπτει από την παραπάνω ανάλυση είναι ότι στο μεγαλύτερο ποσοστό των περιπτώσεων τόσο στις μεθόδους προσέγγισης των προβλημάτων (μαθηματικά μοντέλα) όσο και στους ίδιους τους αλγόριθμους το κοινωνικό δίκτυο απεικονίζεται ως ένα γράφημα και οι προτεινόμενες μέθοδοι/αλγόριθμοι ανάλυσης/εξόρυξης γνώσης από το κοινωνικό δίκτυο στηρίζονται στη θεωρία των γραφημάτων.

Στο Πρακτικό Μέρος αυτής της εργασίας θα μελετήσουμε ένα πραγματικό κοινωνικό δίκτυο απεικονίζοντάς το ως μια βάση δεδομένων και θα προσπαθήσουμε να εξάγουμε συμπεράσματα εξόρυξης γνώσης μέσω του SQL Server 2008 και συγκεκριμένα της υπηρεσίας Analysis Services εφαρμόζοντας αλγορίθμους συσταδοποίησης, κανόνων εξόρυξης γνώσης και κατηγοριοποίησης.

5. Βασικές έννοιες του Analysis Services

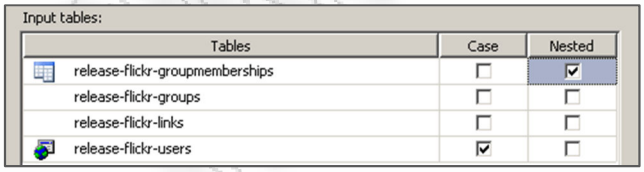
Οι υπηρεσίες ανάλυσης του Microsoft SQL Server περιέχουν αρκετούς τύπους αλγορίθμων ορισμένοι από τους οποίους είναι:

- Αλγόριθμοι κατηγοριοποίησης οι οποίοι προβλέπουν μια ή περισσότερες μεταβλητές βάσει άλλων χαρακτηριστικών του συνόλου δεδομένων. Ένα παράδειγμα αλγορίθμου κατηγοριοποίησης είναι ο Microsoft Decision Trees Algorithm.
- Αλγόριθμοι συσταδοποίησης οι οποίοι διαιρούν τα δεδομένα σε ομάδες ή συστάδες αντικειμένων τα οποία παρουσιάζουν όμοιες ιδιότητες. Ένα παράδειγμα είναι ο Microsoft Clustering Algorithm.
- Αλγόριθμοι συσχέτισης οι οποίοι ανακαλύπτουν συσχετίσεις μεταξύ διαφορετικών χαρακτηριστικών του συνόλου δεδομένων. Ένα παράδειγμα είναι ο Microsoft Association Algorithm.

Για τη δημιουργία μοντέλων εξόρυξης γνώσης μέσω των Analysis Services είναι απαραίτητη η κατανόηση δύο βασικών εννοιών:

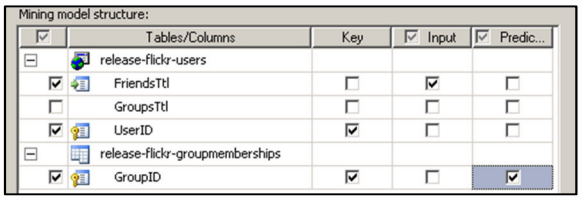
i. Case Table: απεικονίζει τις περιπτώσεις που πρόκειται να αναλυθούν σε ένα μοντέλο. Κάθε περίπτωση εκφράζει μια εγγραφή (πχ. ο πίνακας Users) και είναι πιθανό να αποτελείται από επιπλέον χαρακτηριστικά (πχ. το φύλο του χρήστη, η ηλικία του κτλ) . Αφού επιλεχθεί ο Case πίνακας πάντα θα πρέπει να ορίζεται ένα κλειδί (case key) το οποίο σε πολλές περιπτώσεις μπορεί να είναι το primary key του πίνακα.

ii. Nested Table: σε μια σχεσιακή βάση επιπλέον (αναλυτικές) πληροφορίες για εγγραφή απεικονίζονται σε έναν άλλο πίνακα τον οποίο εδώ αποκαλούμε Nested (πχ. τα groups που είναι εγγεγραμμένος ένας χρήστης).



Εικόνα 5-1 Επιλογή Case και Nested Table

Δημιουργώντας ένα νέο μοντέλο κάθε φορά ορίζουμε τους πίνακες Case και Nested όπως φαίνεται στην εικόνα 5.1. Στη συνέχεια έχουμε τη δυνατότητα να ορίσουμε στήλες κλειδιά (key), στήλες εισόδου (input) και στήλες πρόβλεψης (predict) όπως απεικονίζεται στην εικόνα 5.2. Οι στήλες κλειδιά είναι απαραίτητο να οριστούν σε κάθε μοντέλο κάτι που δεν ισχύει για τις στήλες εισόδου και πρόβλεψης.



Εικόνα 5-2 Ορισμός στηλών Key, Input και Predict

6. Μελέτη περίπτωσης FlickR

6.1 Σύνολο δεδομένων κοινωνικού δικτύου

Το **FlickR** (www.flickr.com) αρχικά ήταν μια ιστοσελίδα, η οποία δημιουργήθηκε για να φιλοξενεί φωτογραφίες και βίντεο. Στη συνέχεια όμως αναπτύσσοντας χαρακτηριστικά όπως το *"Find your Friends in FlickR"* έγινε μια προσπάθεια ενδυνάμωσης του κοινωνικού περιεχομένου του ιστότοπου με αποτέλεσμα πλέον να θεωρείται ένα κοινωνικό δίκτυο.

Το σύνολο δεδομένων που θα χρησιμοποιηθεί σε αυτή την εργασία για να δημιουργηθούν τα μοντέλα εξόρυξης γνώσης προέρχεται από ένα crawl που πραγματοποιήθηκε στις 9 Ιανουαρίου 2007 από τους A. Mislove et. al. [6] στο κοινωνικό δίκτυο FlickR και περιέχει πάνω από 1.8 εκατομμύρια χρήστες και 22 εκατομμύρια συνδέσμους. Επίσης, συγκεντρώθηκαν δεδομένα για τα Groups στα οποία είναι εγγεγραμμένοι οι χρήστες (group membership). Το αποτέλεσμα είναι τέσσερα διακριτά αρχεία τα οποία έχουμε διαθέσιμα προς επεξεργασία και τα βασικά τους χαρακτηριστικά είναι:

i. **Users:** περιέχει μια λίστα όλων των UserID των χρηστών [1.846.198 χρήστες]

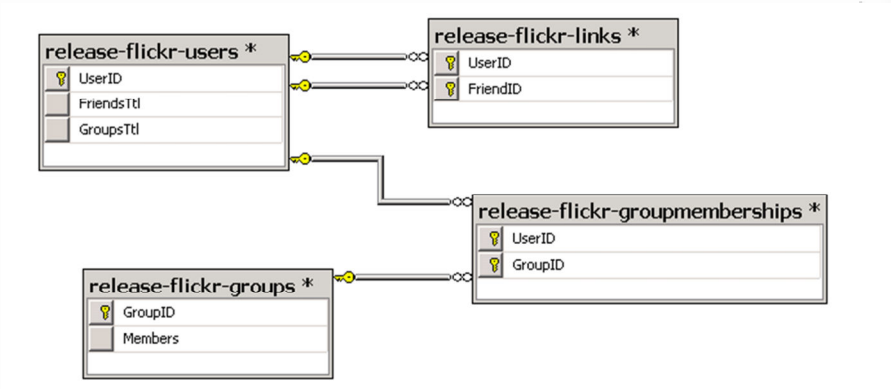
ii. **Links:** περιέχει μια λίστα όλων των συνδέσμων μεταξύ των χρηστών. Απεικονίζει τη φιλία μεταξύ των χρηστών. [22.613.981 σύνδεσμοι]

iii. **Groups:** περιέχει μια λίστα όλων των GroupID των ομάδων [103.648 groups].

iv. **Groupmemberships:** περιέχει μια λίστα όλων των συνδέσμων μεταξύ χρηστών και ομάδων. Εκφράζει ποιος χρήστης ανήκει σε ποια groups. [8.545.307]

6.2 Σχεδίαση βάσης δεδομένων

Στην παρακάτω εικόνα παρουσιάζεται το σχεδιάγραμμα της βάσης δεδομένων FLICKR.



Εικόνα 6-1 Σχεδιάγραμμα Βάσης Δεδομένων FlickR

Όπως παρατηρούμε η ΒΔ αποτελείται από 4 πίνακες:

- **USERS** (UserID, FriendsTtl, GroupsTtl)

  (Τα πεδία FriendTtl και GroupsTtl δεν υπήρχαν στο αρχικό σύνολο δεδομένων για τον πίνακα USERS. Υπολογίστηκαν στη συνέχεια με σκοπό να χρησιμοποιηθούν τόσο στην κατανόηση των δεδομένων όσο και στην κατασκευή μοντέλων εξόρυξης γνώσης.)

- **LINKS** (UserID, FriendID)

- **GROUPS** (GroupID, Members)

  (Το πεδίο Members δεν υπήρχε στο αρχικό σύνολο δεδομένων, υπολογίστηκε στη συνέχεια με σκοπό να χρησιμοποιηθεί στην κατανόηση των δεδομένων)

- **GROUPMEMBERSHIPS** (UserID, GroupID)

Δυστυχώς, λόγω πολιτικής απορρήτου που ισχύει για τα κοινωνικά δίκτυα τα δεδομένα που έχουμε διαθέσιμα και παρουσιάστηκαν παραπάνω είναι περιορισμένα (πχ. στον πίνακα Users δόθηκε μόνο το UserID). Ιδανικά θα θέλαμε να είχαμε διαθέσιμες πληροφορίες όπως το φύλο του χρήστη, την ηλικία του, το επίπεδο μόρφωσης του, την περιγραφή του group κτλ. Η έλλειψη αυτών των πληροφοριών καθιστά την προσπάθεια εξόρυξης γνώσης μέσω των μοντέλων που αναλύονται παρακάτω αρκετά δύσκολη.

6.3 Εξερευνώντας τα δεδομένα

Με την ολοκλήρωση της εισαγωγής των δεδομένων βάσει των τεσσάρων αρχείων που έχουμε διαθέσιμα θέτουμε ορισμένα ερωτήματα με στόχο την απεικόνιση δείγματος για κάθε έναν από τους πίνακες.

Για την καλύτερη κατανόηση των δεδομένων μας κρίνουμε απαραίτητο τη δημιουργία δύο νέων στηλών στον πίνακα των `Users`. Η πρώτη στήλη ονομάζεται `FriendsTtl` και εκφράζει για κάθε χρήστη το σύνολο των φίλων του και η δεύτερη στήλη ονομάζεται `GroupsTtl` και εκφράζει το σύνολο των `Groups` στα οποία αυτός ο χρήστης είναι εγγεγραμμένος. Συνεπώς, με ένα απλό ερώτημα προς τη βάση μπορούμε εύκολα να εντοπίσουμε ποιοι είναι οι 10 πιο δημοφιλής χρήστες του κοινωνικού δικτύου ως προς το σύνολο των φίλων τους.

```
SELECT TOP 10 [UserID]
      ,[FriendsTtl]
      ,[GroupsTtl]
  FROM [FLICKR].[dbo].[release-flickr-users]
  ORDER BY FriendsTtl DESC
```

	UserID	FriendsTtl	GroupsTtl
1	3574	26185	490
2	526	19257	789
3	928	18877	1380
4	1432461	14398	354
5	1918	12919	656
6	4531	12212	700
7	7241	12051	104
8	259	11102	134
9	3041	11060	1237
10	50540	11046	114

Εικόνα 6-2 Δείγμα δημιουργημένου πίνακα Users (Ταξινόμηση βάσει FriendsTtl)

Επίσης, με το παρακάτω ερώτημα μπορούμε να εντοπίσουμε ποιοι είναι οι 10 χρήστες του κοινωνικού δικτύου οι οποίοι έχουν εγγραφεί στα περισσότερα `Groups`.

```
SELECT TOP 10 [UserID]
      ,[FriendsTtl]
      ,[GroupsTtl]
  FROM [FLICKR].[dbo].[release-flickr-users]
  ORDER BY GroupsTtl DESC
```

	UserID	FriendsTtl	GroupsTtl
1	19528	2240	2186
2	3425	475	2084
3	982662	9827	2058
4	29072	1130	1728
5	49	6475	1633
6	5226	972	1606
7	928	18877	1380
8	42396	134	1359
9	37301	2264	1346
10	195540	3	1322

Εικόνα 6-3 Δείγμα δημιουργημένου πίνακα Users (Ταξινόμηση βάσει GroupsTtl)

```
SELECT TOP 10 [UserID]
      ,[FriendID]
  FROM [FLICKR].[dbo].[release-flickr-links]
```

	UserID	FriendID
1	1	2
2	1	3
3	1	4
4	1	5
5	1	6
6	1	7
7	1	8
8	1	9
9	1	794106
10	1	908381

Εικόνα 6-4 Δείγμα δημιουργημένου πίνακα Links

Για τον ίδιο λόγο που δημιουργήσαμε τις δύο νέες στήλες στον πίνακα `Users` δημιουργούμε μια νέα στήλη `Members` στον πίνακα `Groups`. Η στήλη αυτή μας δίνει την πληροφορία πόσα μέλη αριθμεί το κάθε Group. Οπότε, με το παρακάτω ερώτημα εντοπίζουμε ποια είναι τα 10 πιο δημοφιλή Groups του δικτύου από τη στιγμή που περιέχουν τα περισσότερα μέλη.

```
SELECT TOP 10 [GroupID]
      ,[Members]
  FROM [FLICKR].[dbo].[release-flickr-groups]
  ORDER BY Members DESC
```

	GroupID	Members
1	471	34989
2	172	30310
3	508	27543
4	228	22097
5	135	21567
6	156	21427
7	295	17473
8	227	16146
9	904	15412
10	3	14820

Εικόνα 6-5 Δείγμα δημιουργημένου πίνακα Groups (Ταξινόμηση βάσει Members)

```
SELECT TOP 10 [UserID]
      ,[GroupID]
  FROM [FLICKR].[dbo].[release-flickr-groupmemberships]
```

	UserID	GroupID
1	1	1
2	1	2
3	2	3
4	2	19
5	2	54
6	2	57
7	2	109
8	2	135
9	2	139
10	2	156

Εικόνα 6-6 Δείγμα δημιουργημένου πίνακα Groupmemberships

Ένα συνηθισμένο φαίνομενο στα κοινωνικά δίκτυα είναι η πρόταση πιθανών φίλων (*People you may know*). Για παράδειγμα θεωρούμε ότι δυο χρήστες οι οποίοι δεν είναι μεταξύ τους φίλοι είναι πιθανό να σχετίζονται με έναν τρίτο χρήστη ο οποίος είναι φίλος και με τους δύο. Αναφερόμαστε σε αυτό ως τη συντομότερη διαδρομή (shortest path) αλλά πιο συχνά χρησιμοποιούμε τον όρο «degrees of separation».

Μέσω του παρακάνω ερωτήματος είμαστε σε θέση να εντοπίσουμε τη δυντομότερη διαδρομή σύνδεσης μεταξύ δύο οποιονδήποτε χρηστών.

```
WITH TC (UserID, FriendID, Distance, path) AS
(SELECT UserID, FriendID, 1,
CAST('.' + CAST(UserID AS varchar (10)) + '.' +
CAST(FriendID AS varchar (10)) + '.' AS varchar (900))
FROM dbo.SampleLinks
UNION ALL
SELECT P.UserID, C.FriendID, P.Distance + 1,
CAST(P.path + CAST(C.FriendID AS varchar (10)) + '.' AS varchar (900))
FROM dbo.SampleLinks AS C
JOIN TC AS P ON C.UserID = P.FriendID
WHERE P.path NOT LIKE '%.' + CAST(C.FriendID AS varchar (10)) + '.%')
SELECT AP.*
FROM TC AS AP -- All Paths
JOIN (SELECT UserID, FriendID, MIN(Distance) AS Min_Distance
FROM TC
GROUP BY UserID, FriendID) AS MD
ON AP.UserID = MD.UserID
AND AP.FriendID = MD.FriendID
AND AP.Distance = MD.Min_Distance
ORDER BY UserID, FriendID;
```

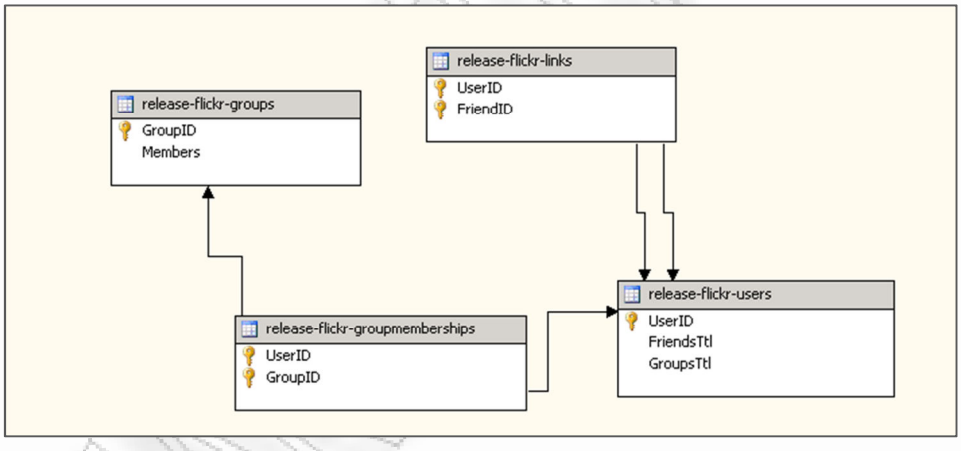
Για παράδειγμα, δύο χρήστες που είναι φίλοι έχουν 1 βαθμό διαχωρισμού (degree of separation), ενώ δύο χρήστες που έχουν έναν κοινό φίλο έχουν 2 βαθμούς διαχωρισμού. Όπως φαίνεται στην παρακάτω εικόνα οι UserID=1 και UserID=10 μοιράζονται έναν κοινό φίλο UserID=3 οπότε απαιτούνται 2 «βήματα» για να φτάσουμε από τον έναν στον άλλο.

	UserID	FriendID	Distance	path
1	1	2	1	.1.2.
2	1	3	1	.1.3.
3	1	4	1	.1.4.
4	1	5	1	.1.5.
5	1	6	1	.1.6.
6	1	7	1	.1.7.
7	1	8	1	.1.8.
8	1	9	1	.1.9.
9	1	10	2	.1.3.10.
10	1	11	2	.1.6.11.
11	1	12	2	.1.6.12.
12	1	13	2	.1.6.13.
13	1	14	2	.1.6.14.
14	1	15	2	.1.6.15.

Εικόνα 6-7 Ερώτημα για την εύρεση απόστασης μεταξύ δύο χρηστών

7. Αποτελέσματα μοντέλων εξόρυξης γνώσης

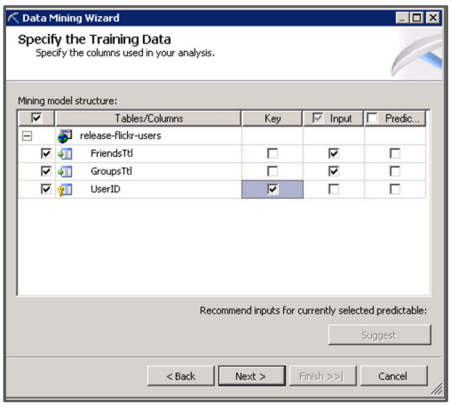
Στην εικόνα 7.1 απεικονίζεται το διαθέσιμο DataSource βάσει του οποίου θα δημιουργηθούν τα μοντέλα εξόρυξης γνώσης χρησιμοποιώντας το εργαλείο Microsoft Visual Studio.



Εικόνα 7-1 Διαθέσιμο DataSource για την κατασκευή των μοντέλων εξόρυξης γνώσης

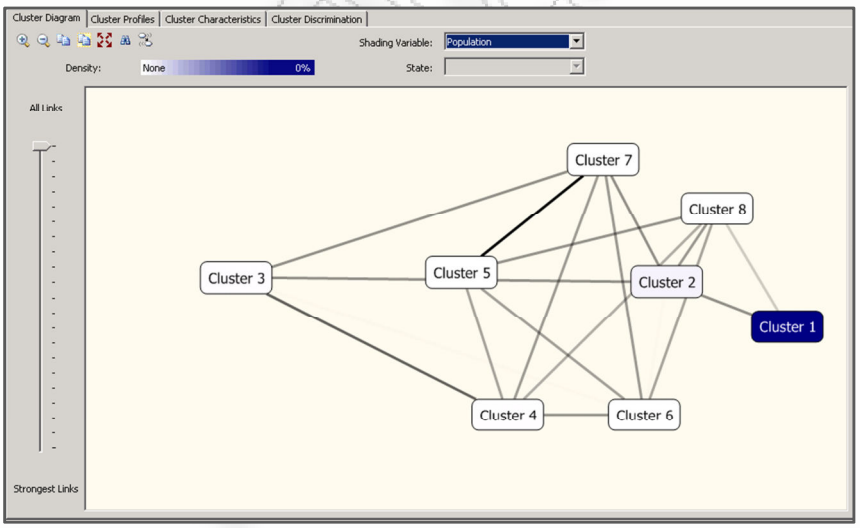
7.1 Συσταδοποίηση

Στόχος του μοντέλου που ακολουθεί είναι να εξετάσουμε τον τρόπο κατανομής των χρηστών σε συστάδες βάσει του αριθμού των φίλων τους και του αριθμού των groups στα οποία έχουν εγγραφεί. Για το σκοπό αυτό θα χρησιμοποιήσουμε τον αλγόριθμο *Microsoft Clustering*. Μια υπόθεση που θα περιμέναμε να επιβεβαιωθεί με αυτό το μοντέλο είναι ότι χρήστες που έχουν λίγους φίλους ανήκουν σε μικρό αριθμό Groups και αντίστοιχα χρήστες με πολλούς φίλους και μεγάλη δραστηριότητα στο δίκτυο θα είναι εγγεγραμμένοι σε μεγάλο αριθμό groups. Στην εικόνα 7.2 απεικονίζεται ο ορισμός του μοντέλου συσταδοποίησης.



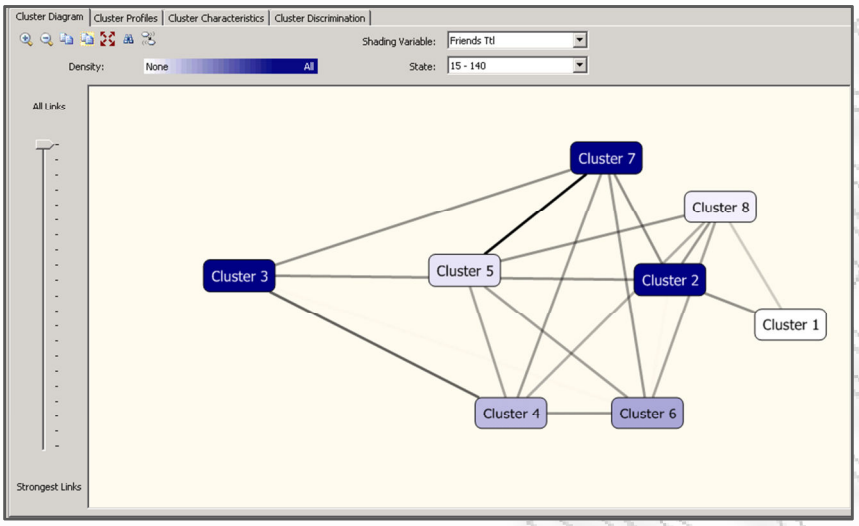
Εικόνα 7-2 Ορισμός μοντέλου συσταδοποίησης

Με την ολοκλήρωση της διαδικασίας Process του μοντέλου μέσω του Cluster Diagram κάθε συστάδα απεικονίζεται ως ένας κόμβος. Αυτοί οι κόμβοι εμφανίζονται διάσπαρτοι και ομαδοποιούνται αυτόματα με βάση τις ομοιότητες τους. Το αποτέλεσμα είναι ένα διάγραμμα το οποίο υποδεικνύει ποιοι κόμβοι είναι όμοιοι ή ανόμοιοι και τη σχετική ισχύ αυτών των ομοιοτήτων. Παρατηρούμε στο δικό μας παράδειγμα ότι το μεγαλύτερο μέρος του ποσοστού των χρηστών ανήκει στην συστάδα 1 και ότι οι ομοιότητες μεταξύ των συστάδων δεν είναι τόσο ισχυρές και για το λόγο αυτό οι ακμές που συνδέουν τις συστάδες εμφανίζονται με αχνό χρώμα.



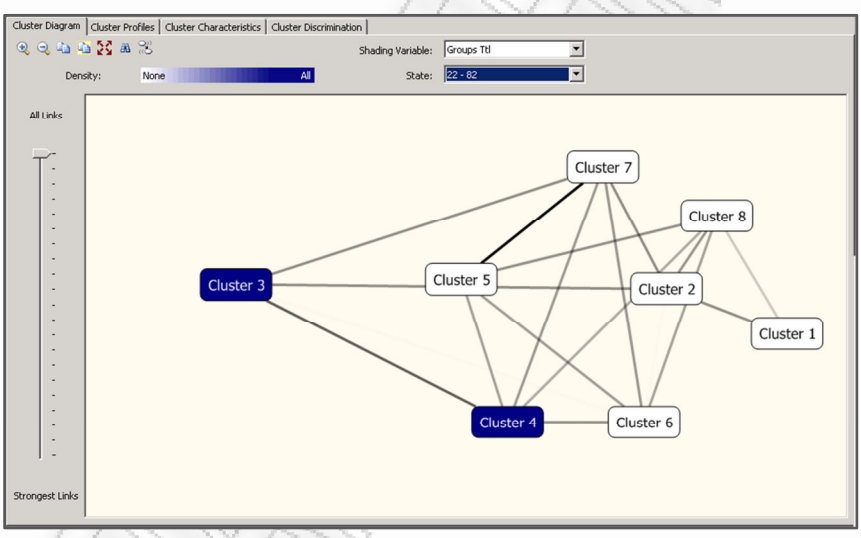
Εικόνα 7-3 Cluster Diagram All

Χρησιμοποιώντας τη λειτουργία της προβολής στο Cluster Diagram, είναι εύκολο να ζητήσουμε πιο στοχευμένες ερωτήσεις σχετικά με το μοντέλο μας. Για παράδειγμα "σε ποιες ομάδες περιέχονται χρήστες που έχουν συνολικό αριθμό φίλων μεταξύ 15 και 140;" όπως φαίνεται στην εικόνα 7.4.



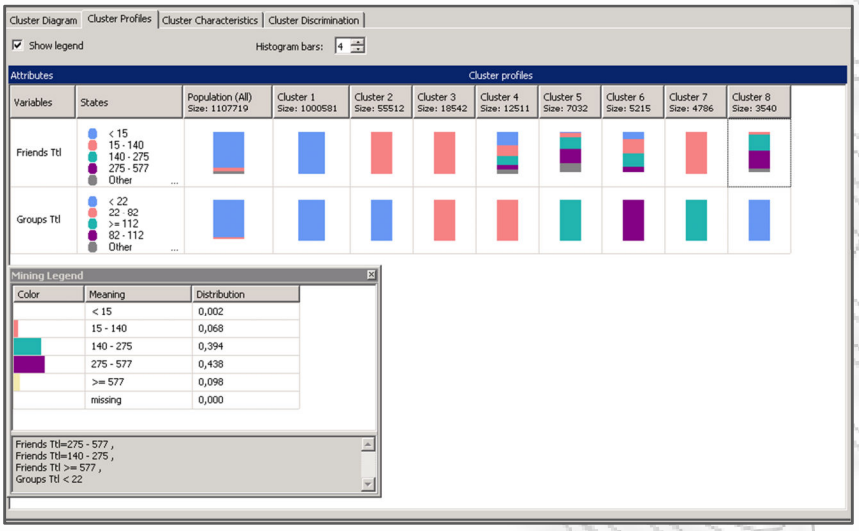
Εικόνα 7-4 Cluster Diagram (FriendsTtl 15-140)

Ένα άλλο ερώτημα θα μπορούσε να είναι "σε ποιες ομάδες περιέχονται χρήστες που είναι εγγεγραμμένοι σε συνολικό αριθμό groups μεταξύ 22 και 82;"



Εικόνα 7-5 Cluster Diagram (GroupsTtl 22-82)

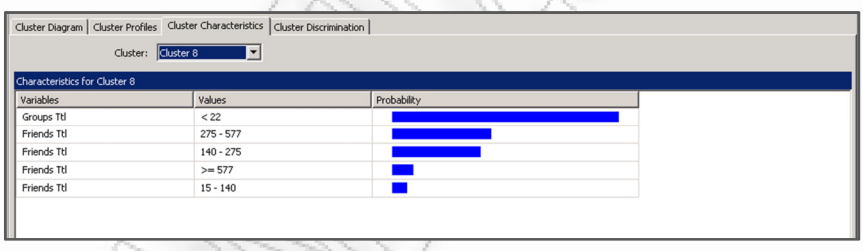
Η επιλογή Cluster Profiles απεικονίζει όλες τις συστάδες οι οποίες έχουν ανιχνευθεί ως στήλες και κάθε χαρακτηριστικό το οποίο επιλέχθηκε στο μοντέλο ως γραμμές. Στη συγκεκριμένη περίπτωση τα δύο χαρακτηριστικά (Friends Total, Groups Total) που χρησιμοποιούμε παίρνουν διακριτές τιμές και για το λόγο αυτό το αποτέλεσμα είναι οι τιμές τους να εμφανίζονται ομαδοποιημένες.



Εικόνα 7-6 Cluster Profiles

Παρατηρώντας την εικόνα 7.6 συμπεραίνουμε ότι το μεγαλύτερο μέρος των χρηστών που ανήκουν στη συστάδα 1 έχει μικρό αριθμό φίλων <15 και ανήκει σε λίγα Groups <22 γεγονός που επιβεβαιώνει την αρχική υπόθεση που κάναμε. Ενδιαφέρον όμως παρουσιάζει η συστάδα 8 στην οποία ανήκουν χρήστες με πολλούς φίλους οι οποίοι όμως είναι εγγεγραμμένοι σε μικρό αριθμό Groups <22. Επιλέγοντας οποιοδήποτε κελί στην περιοχή εμφανίζονται πληροφορίες μέσω του mining legend για τα περιεχόμενα της συστάδας και για το ποσοστό κατανομής των χρηστών βάσει του επιλεγμένου χαρακτηριστικού.

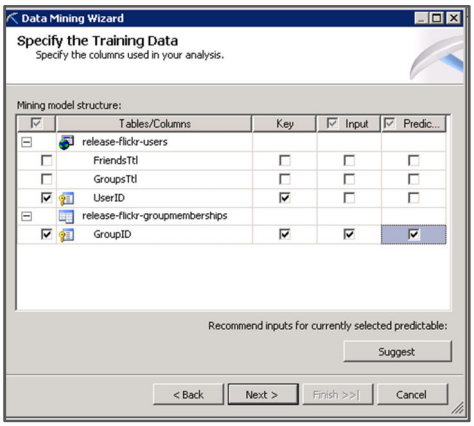
Μέσω της επιλογής Cluster Characteristics για την επιλεγμένη συστάδα εμφανίζεται η πιθανότητα να ανήκει σε αυτή κάποιος χρήστης βάσει των τιμών στις επιλεγμένες μεταβλητές.



Εικόνα 7-7 Cluster Characteristics

7.2 Κανόνες συσχέτισης

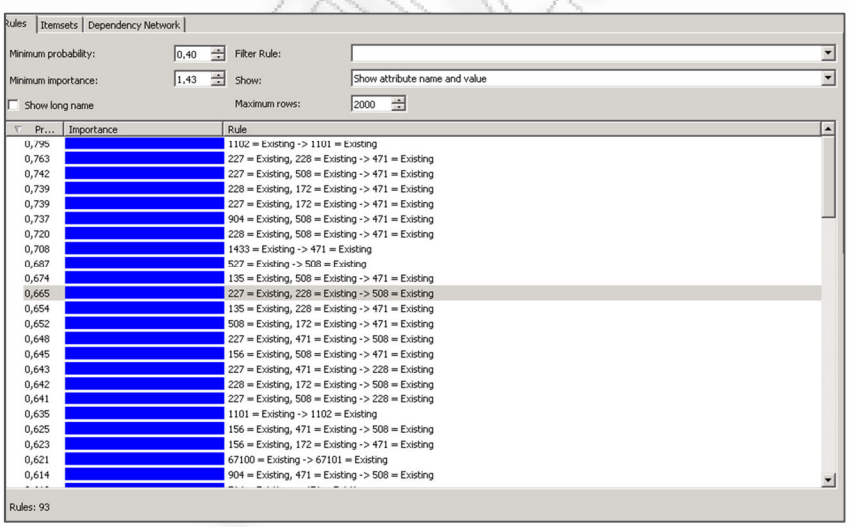
Στόχος του μοντέλου που ακολουθεί είναι να εξετάσουμε αν χρησιμοποιώντας τον αλγόριθμο *Microsoft Association Rules* είμαστε σε θέση να συμπεράνουμε αν κάποιος χρήστης που ανήκει στο *Group A* υπάρχει πιθανότητα να ανήκει και στο *Group B*. Στην εικόνα 7.8 απεικονίζεται ο ορισμός του συγκεκριμένου μοντέλου.



Εικόνα 7-8 Ορισμός μοντέλου κανόνων συσχέτισης (1)

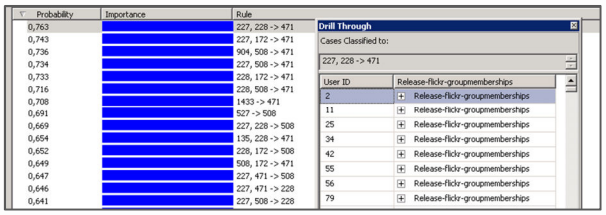
Τα αποτελέσματα της εκτέλεσης του αλγορίθμου σχολιάζονται παρακάτω. Συγκεκριμένα, στην περιοχή Rules απεικονίζονται οι κανόνες συσχέτισης που προέκυψαν αξιολογώντας κάθε κανόνα με βάση την πιθανότητα και τη σπουδαιότητά του. Το μέτρο της σπουδαιότητας εκφράζει το πόσο χρήσιμος είναι ένας κανόνας. Συνεπώς, όσο υψηλότερος ο βαθμός σπουδαιότητας τόσο καλύτερη είναι η ποιότητα του κανόνα.

Παρατηρώντας την εικόνα 7.9 διαπιστώνουμε ότι με τα αποτελέσματα του αλγορίθμου είμαστε σε θέση να απαντήσουμε στο ερώτημα που θέσαμε. Δηλαδή, εύκολα διαπιστώνουμε για παράδειγμα ότι αν ένας χρήστης ανήκει στα Groups {227,228} τότε με πιθανότητα 76% θα ανήκει και στο Group {471} ή όταν ανήκει στο Group {1102} με πιθανότητα 80% θα ανήκει και στο Group {1101}.



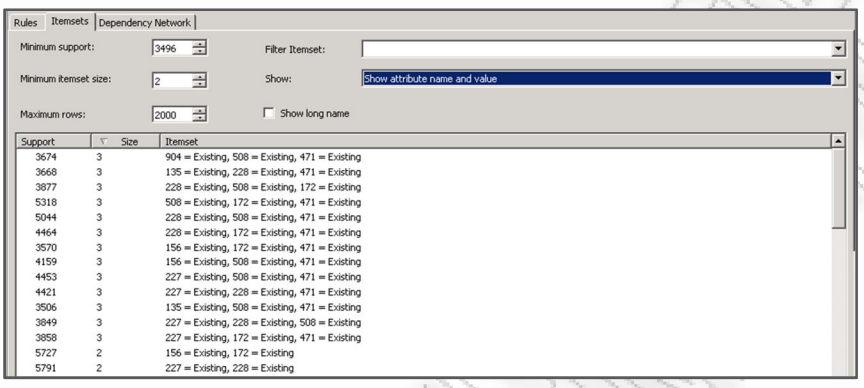
Εικόνα 7-9 Κανόνες συσχέτισης μεταξύ Groups

Μέσω της επιλογής Drill Through σε έναν επιλεγμένο κανόνα έχουμε τη δυνατότητα να εντοπίσουμε τους σχετικούς χρήστες όπως φαίνεται στην παρακάτω εικόνα:



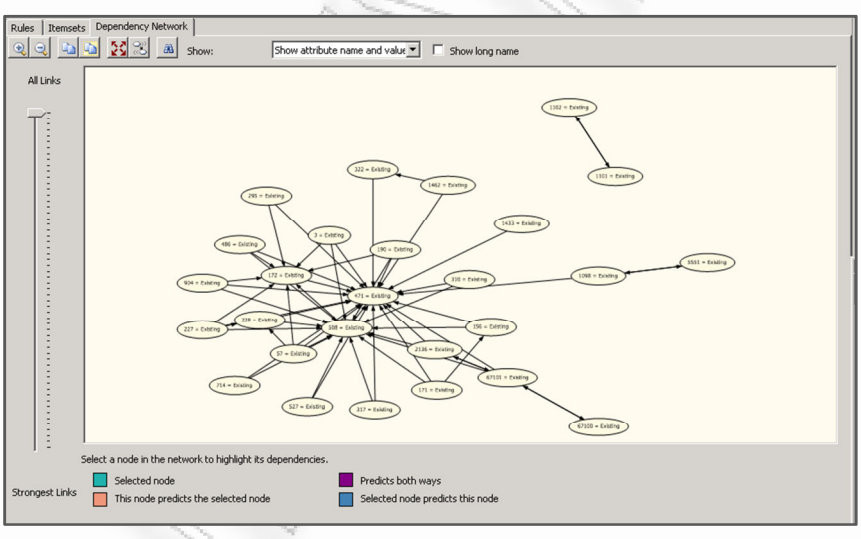
Εικόνα 7-10 Drill Through για εύρεση χρηστών

Στην περιοχή Itemsets απεικονίζονται τα στοιχεία (GroupsID) που εμφανίζονται συχνά μαζί.



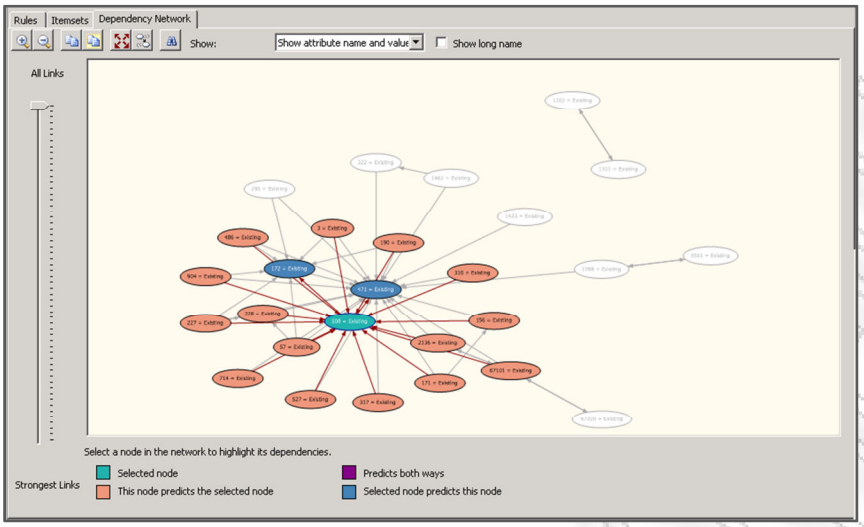
Εικόνα 7-11 Itemsets

Η τρίτη περιοχή είναι το Dependency Network όπου κάθε κόμβος αποτελεί μια περίπτωση και κάθε ακμή εκφράζει έναν κανόνα συσχέτισης.

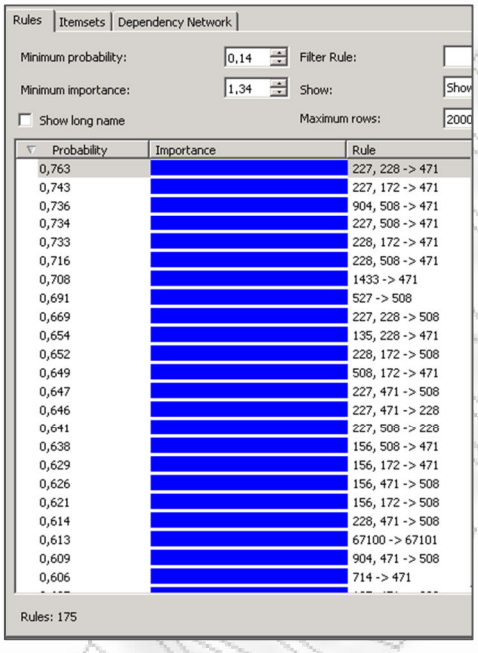


Εικόνα 7-12 Dependency Network All

Επιλέγοντας έναν κόμβο μπορούμε να δούμε με γραφικό τρόπο τις συσχετίσεις με άλλους κόμβους



Εικόνα 7-13 Dependency Network – Εμφάνιση συσχετίσεων κόμβων



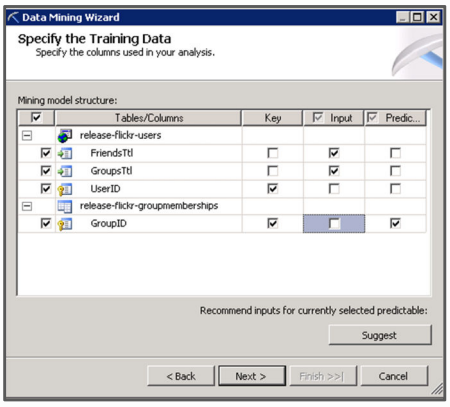
Τα δέκα πιο δημοφιλή group του δικτύου παρουσιάζονται στην παρακάτω εικόνα:

	GroupID	Members
1	471	34989
2	172	30310
3	508	27543
4	228	22097
5	135	21567
6	156	21427
7	295	17473
8	227	16146
9	904	15412
10	3	14820

Με μια γρήγορη ματιά από τα αποτελέσματα του αλγορίθμου συμπεραίνουμε ότι οι κανόνες που προέκυψαν περιέχουν κατά κύριο λόγο τα παραπάνω Groups.

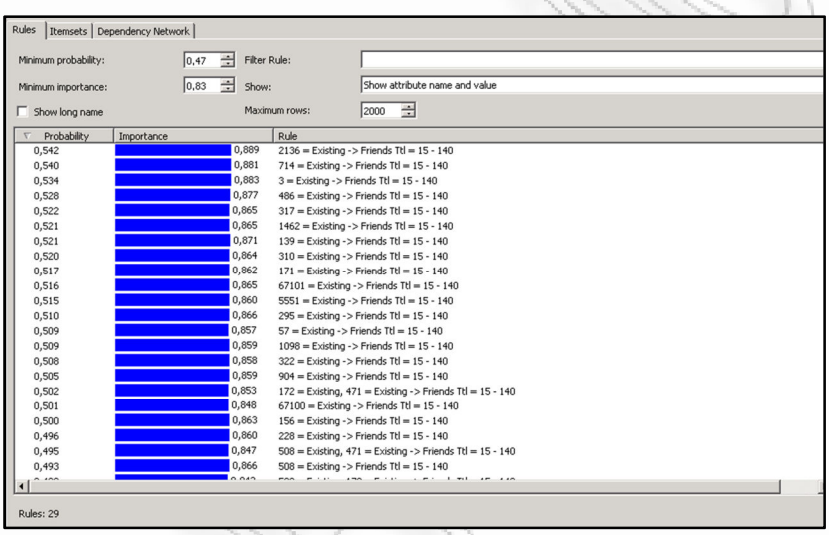
Εικόνα 7-14 Σύγκριση αποτελεσμάτων

Μέσω ενός άλλου μοντέλου κανόνων συσχέτισης ο ορισμός του οποίου απεικονίζεται στην εικόνα 7.15 θα εξάγουμε συμπεράσματα σχετικά με τη συσχέτιση των groups στα οποία ένας χρήστης είναι εγγεγραμμένος και στο συνολικό αριθμό φίλων του.



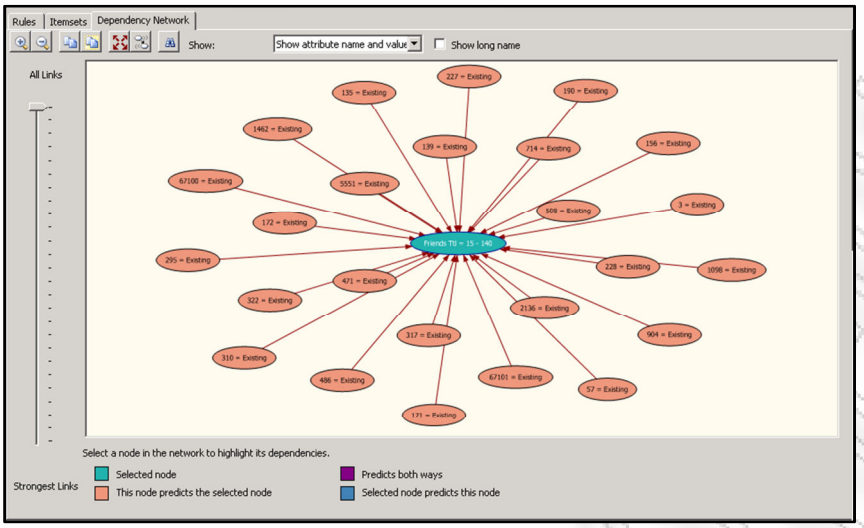
Εικόνα 7-15 Ορισμός μοντέλου κανόνων συσχέτισης (2)

Συμπεραίνουμε όπως φαίνεται παρακάτω ότι με πιθανότητα 50% όταν κάποιος χρήστης ανήκει στα Groups {172, 471} τότε ο αριθμός φίλων του θα κυμαίνεται από 15 έως 140.



Εικόνα 7-16 Κανόνες συσχέτισης μεταξύ Groups και FriendsTtl

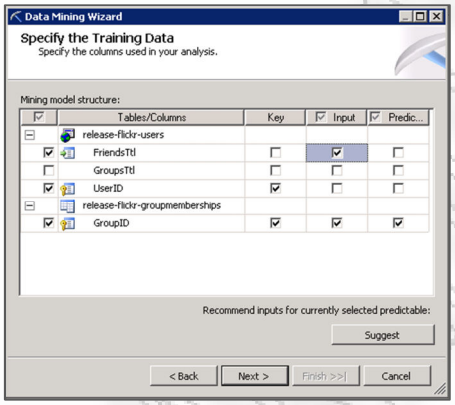
Η παραπάνω συσχέτιση απεικονίζεται γραφικά και στο Dependency Network.



Εικόνα 7-17 Dependency Network Groups και FriendsTtl

7.3 Κατηγοριοποίηση

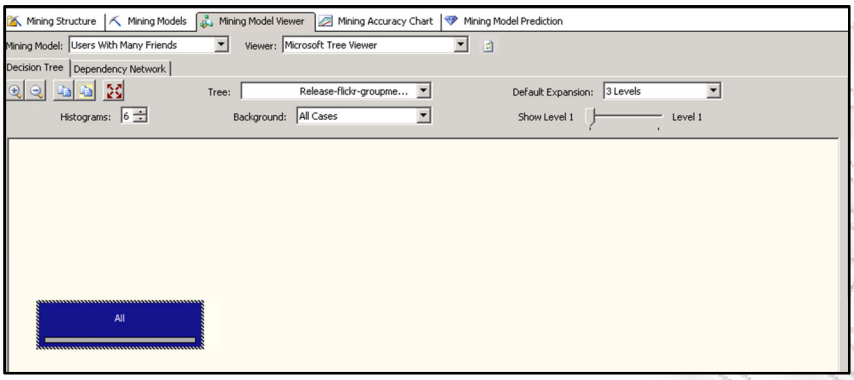
Στόχος του μοντέλου που ακολουθεί είναι να εξετάσουμε αν χρησιμοποιώντας τον αλγόριθμο *Microsoft Decision Tree* είμαστε σε θέση να απαντήσουμε στο ερώτημα: «χρήστες που έχουν πολλούς φίλους τείνουν να ανήκουν σε ένα συγκεκριμένο group;». Στην εικόνα 7.18 απεικονίζεται ο ορισμός του συγκεκριμένου μοντέλου.



Εικόνα 7-18 Ορισμός μοντέλου κατηγοριοποίησης

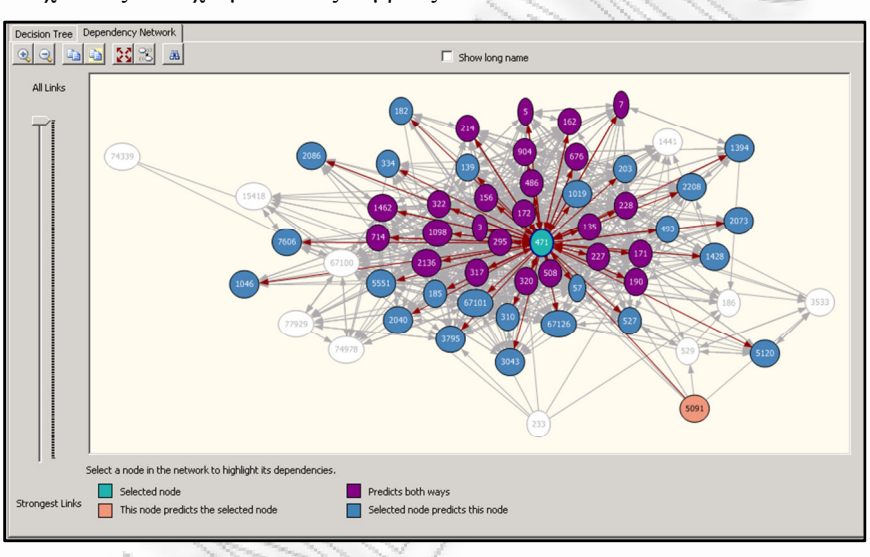
Τα αποτελέσματα σε ένα δέντρο απόφασης είναι αρκετά εύκολο να ερμηνευτούν. Κάθε μονοπάτι από τη ρίζα ενός επιλεγμένου κόμβου αποτελεί έναν κανόνα. Έχουμε τη δυνατότητα να χρησιμοποιήσουμε την αναπτυσσόμενη λίστα δέντρων απόφασης διότι ένα μοντέλο μπορεί να περιέχει ένα σύνολο από δέντρα.

Όμως όπως φαίνεται και στην παρακάτω εικόνα δεν ήταν εφικτή η δημιουργία ενός δέντρου απόφασης για το συγκεκριμένο μοντέλο με αποτέλεσμα να μην είμαστε σε θέση να απαντήσουμε στο ερώτημα που θέσαμε παραπάνω.



Εικόνα 7-19 Αδυναμία δημιουργίας Decision Tree

Με βάση το Dependency Network του ίδιου μοντέλου παρατηρούμε ότι υπάρχει ένας μεγάλος αριθμός συσχετίσεων μεταξύ διαφόρων Groups. Η συγκεκριμένη εικόνα ταυτίζεται αρκετά με αυτή του αλγορίθμου Κανόνων Συσχέτισης. Με τον ίδιο τρόπο επιλέγοντας έναν κόμβο παρατηρούμε τις συσχετίσεις που έχει με άλλους κόμβους.



Εικόνα 7-20 Dependency Network κατηγοριοποίησης

8. ΣΥΜΠΕΡΑΣΜΑΤΑ

Συμπερασματικά κύριος στόχος αυτής της εργασίας ήταν η αναλυτική παρουσίαση της θεωρίας ανάλυσης των κοινωνικών δικτύων σε ότι αφορά μεθόδους εξόρυξης γνώσης. Στο πλαίσιο αυτού του στόχου παρουσιάστηκαν όλα τα απαραίτητα σημεία για την κατανόηση βασικών αλγορίθμων ανίχνευσης κοινοτήτων, κατηγοριοποίησης κόμβων, εύρεσης ειδικών, κοινωνικής επιρροής, εμπιστοσύνης και πρόβλεψης συνδέσμου. Με την ανάλυση αυτή διαπιστώσαμε ότι προκύπτουν χρήσιμα συμπεράσματα τόσο για τη δομή αυτών των δικτύων όσο και για τη συμπεριφορά τους τα οποία χρησιμοποιούνται από πολλές επιστήμες όπως η κοινωνιολογία, η εγκληματολογία, το μάρκετινγκ κτλ.

Παράλληλα, η εργασία αυτή περιέχει και μια μελέτη περίπτωσης πραγματικού κοινωνικού δικτύου (FlickR) με στόχο τη δημιουργία μοντέλων εξόρυξης γνώσης μέσω των SQL Server Analysis Services. Στην προσπάθεια αυτή τα αποτελέσματα είναι αρκετά περιορισμένα και ο βασικός λόγος είναι ότι λόγω πολιτικής απορρήτου που ισχύει για τα κοινωνικά δίκτυα τα δεδομένα που έχουμε διαθέσιμα περιορίζονταν σε ένα μόνο ID. Ουσιαστικά, μέσω των μοντέλων που δημιουργήσαμε μπορέσαμε να απαντήσουμε σε ερωτήματα όπως ποια είναι η κατανομή των χρηστών σε συστάδες βάσει του αριθμού των φίλων τους και το σύνολο των groups που οι χρήστες είναι εγγεγραμμένοι, εντοπίσαμε συσχετίσεις μεταξύ των groups με αποτέλεσμα να γνωρίζουμε με συγκεκριμένο βαθμό πιθανότητας όταν ο χρήστης ανήκει σε ένα group αν θα ανήκει και σε κάποιο άλλο. Ιδανικά, αν είχαμε διαθέσιμες πληροφορίες όπως το φύλο του χρήστη, την ηλικία του, το επίπεδο μόρφωσης του, την περιγραφή του group που ανήκει κτλ. θα είχαμε τη δυνατότητα να δημιουργήσουμε μοντέλα εξόρυξης γνώσης από τα οποία θα προέκυπταν περισσότερα και πιο χρήσιμα συμπεράσματα.

Ολοκληρώνοντας θεωρούμε ότι πρέπει να αναφέρουμε τόσο τα θετικά όσο και τα αρνητικά σημεία της προσπάθειας μας. Το βασικό πλεονέκτημα της μελέτης θεωρούμε ότι είναι η εκτενής και αναλυτική παρουσίαση των αλγορίθμων εξόρυξης γνώσης σε κοινωνικά δίκτυα μέσα από τους οποίους δίνεται η δυνατότητα στον αναγνώστη να κατανοήσει τις ιδιότητες και τον τρόπο μελέτης αυτών των δικτύων που εξελίσσονται με τόσο έντονους ρυθμούς. Από την άλλη θεωρούμε ότι το βασικό μειονέκτημα της μελέτης είναι ότι λόγω της έλλειψης δεδομένων τα συμπεράσματα που προέκυψαν στο πρακτικό μέρος είναι αρκετά περιορισμένα.

9. ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Ι. Παπαρίζος, "Αλγόριθμος Εξόρυξης Γνώσης από Δεδομένα Δομής, Περιεχομένου και Χρήσης του Παγκόσμιου Ιστού", Διπλωματική εργασία, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2009.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", in American Association for Artificial Intelligence, 1996.

[3] Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας και Η. Σακελλαρίου, Τεχνητή Νοημοσύνη, 3η έκδοση, Β. Γκιούρδας Εκδοτική, 2006.

[4] Μ. Ρήγκου, "Αποδοτικοί Αλγόριθμοι Εξατομίκευσης Βασισμένοι σε Εξόρυξη γνώσης από Δεδομένα χρήσης web", Διδακτορική διατριβή, Πανεπιστήμιο Πατρών, 2005.

[5] A. Mislove, "Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems", PhD Thesis submitted to Rice Univercity, Houston, 2009.

[6] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks", in Proceeding of the 5th ACM/USENIX Internet Measurement Conference (IMC'07), 2007.

[7] N. Papailiou, D. Apostolou, and G. Mentzas, "Social Networks for Knowledge Management in Management Consulting Firms", in Proceeding of the 4th Conference Professional Knowledge Management Experiences and Visions, GITO-Verlag, 2007.

[8] Y. Huang, "Supporting Meaningful Social Networks", PhD Thesis submitted to Univercity of Southampton, 2009.

[9] C. Aggarwal, "An Introduction to Social Network Data Analytics", in Social Network Data Analytics, Edited by C. Aggarwal, Springer, 2011.

[10] O. Serrat, "Social Network Analysis", in Knowledge Solutions, 2009.

[11] J. Shi and J. Malik, "Normalized cuts and image segmentation", in Proceeding of the Conference on Computer Vision and Pattern Recognition (CVPR '97), 1997.

[12] J. Newman and M. Girvan, "Finding and evaluating community structure in networks", in Physical Review E, 69:026113, 2004.

[13] U. Luxburg, "A tutorial on spectral clustering", inStatistics and Computing, 2007.

[14] J. Ruan and W. Zhang, "An efficient spectral algorithm for network community discovery and its applications to biological and social networks", in Proceeding of the 2007 Seventh IEEE International Conference on Data Mining, October 2007.

[15] S. White and P. Smyth, "A Spectral Clustering Approach to Finding Communities in Graphs" in SIAM International Conference on Data Mining, 2005.

[16] J. Newman, A. Clauset and C. Moore, "Finding community structure in very large networks", in Physics Review E70, 066111, 2004.

[17] D. Spielman and N. Srivastava, "Graph sparsification by effective resistances" in Proceedings of the 40th annual ACM symposium on Theory of computing, 2008.

[18] J. Bagrow and E. Bollt, "Local Method for detecting communities", in Physics Review E72, 046108, 2005.

[19] S. Parthasarathy, Y. Ruan, and V. Satuluri, "Community Discovery in Social Networks: Applications, Methods and Emerging Trends", in Social Network Data Analytics, Edited by C. Aggarwal, Springer, 2011.

[20] G. Yu, "Social network analysis based on BSP clustering algorithm", in Communication of IIMA, December 2007.

[21] J. Pujol, J. Béjar, and J. Delgado, "Clustering algorithm for determining community structure in large networks", in Physical Review, 2006.

[22] M. Lipczak and E. Milios, "Agglomerative genetic algorithm for clustering in social networks", in Proceeding of the 11th Annual conference on Genetic and evolutionary computation, Montreal, Québec, Canada, July 2009.

[23] M. Handcock, A.Raftery, and J.Tantrum, "Model-based clustering for social networks", in Journal of the Royal Statistical Society, series A (Statistics in Society), vol. 170, 2007.

[24] C. Pizzuti, "Community detection in social networks with genetic algorithms", in Proceeding of the 10th Annual Conference on Genetic and Evolutionary Computation, Atlanta, Georgia, USA, July 2008.

[25] B. Thompson and D. Yao, "The union-split algorithm and cluster-based anonymization of social networks", in Proceeding of the 4th ACM Symposium on Information, Computer and Communications Security (ASIACCS), Sydney, Australia. March 2009.

[26] J. Bagrow, "Evaluating Local Community Methods in Networks", Department of Physics, Clarkson University, Potsdam, 2008.

[27] S. Bhagat, G. Cormode, and S. Muthukrishnan, "Node Classification in Social Networks", in Social Network Data Analytics, Edited by C. Aggarwal, Springer, 2011.

[28] S. Macskassy and F. Provost, “Classification in Networked Data: A toolkit and a univariate case study”, Technical Report CeDER Working Paper 04-08, Stern School of Business, New York University, 2006.

[29] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher and T. Eliassi-Rad, “Collective classification in network data”, AI Magazine, (2008).

[30] J. Neville, and D. Jensen, “Iterative classification in relational data”, in Workshop on Statistical Relational Learning, AAAI, 2000.

[31] R. Heatherly, M. Kantarcioglu and, B. Thuraisingham, “Social network classification incorporating link type values”, in Intelligence and Security Informatics Conference, 2009.

[32] S. Tang, J. Yuan, X. Mao, X. Li, W. Chen, and G. Dai, “Relationship Classification in Large Scale Online Social Networks and Its Impact on Information Propagation”, in IEEE INFOCOM 2011.

[33] Q. Lu and L. Getoor, “Link-based classification using labeled and unlabeled data”, in ICML Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, 2003.

[34] J. Surma and A. Furmanek, “Improving marketing response by data mining in social network”, in Advances in Social Networks Analysis and Mining (ASONAM), International Conference, 2010.

[35] V. Tuulos and H. Tirri, “Combining Topic Models and Social Networks for Chat Data Mining”, in Proceedings of IEEE/WIC/ACM International Conference, 2004.

[36] M. Thelwall, D. Wilkinson and S. Uppal, “Data Mining Emotion in Social Network Communication: Gender differences in MySpace”, in Journal of the American Society for Information Science and Technology, vol. 61, issue 1, January 2010.

[37] X. Zhu, Z. Ghahramani, and J.D. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions", in Proceeding of ICML, 2003.

[38] S. Yoo, Y. Yang, F. Lin, and I.C. Moon, “Mining social networks for personalized email prioritization”, in ACM SIGKDD Conference, Paris, France, June 2009.

[39] S. Volkova, “Link Prediction in Social Networks”, Independent Project Final Report.

[40] L. Nowell and J. Kleinberg, “The link prediction problem for social networks”, in Proceeding of the 12th international conference on Information and knowledge management, 2003.

[41] M. Hasan, V. Chaoji, S. Salem and M. Zaki, “Link Prediction Using Supervised Learning”, in Proceeding of SDM 06 workshop on Link Analysis, Counterterrorism and Security, 2006

[42] H. Kashima, T. Kato, Y. Yamanishi, M. Sugiyama and K. Tsuda, "Link Propagation: a fast semi-supervised learning algorithm for link prediction" in Proceeding of the 2009 SIAM International Conference on Data Mining, 2009

[43] T. Tylenda, A. Ralitsa, and S. Bahadur, "Towards time-aware link prediction in evolving social network", in Proceeding of SNA-KDD, 2009.

[44] H. Song, T. Cho, V. Dave, Y. Zhang, and L. Qiu, "Scalable proximity Estimation and Link Prediction in Online Social Networks", in Proc. of the Internet Measurement Conference (IMC '09), 2009

[45] B. Tasker, M. Wong, P. Abbeel, and D. Koller, "Link Prediction in Relational Data" in Proceeding of Neural Information Processing Systems (NIPS '03), 2003.

[46] J. Kunegis, A. Lommatzsch and C. Bauckhage, "The slashdot zoo: mining a social network with negative edges", in Proceeding of WWW 2009 MADRID, Madrid, Spain, April 2009.

[47] Bringmann, M. Berlingerio, F. Bonchi, and A. Gionis, "Learning and Predicting the Evolution of Social Networks" in IEEE Intelligent Systems, vol. 25, no. 4, 2010.

[48] P. Domingos, "Mining social networks for viral marketing", in IEEE Intelligent Systems 20, 2005.

[49] A. Goyal, F. Bonchi, and L.V.S. Lakshmanan, "Learning infuence probabilities in social networks", in Proceeding of 3rd ACM International Conference on Web Search and Data Mining, New York, USA, February 2010.

[50] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks", in Proceeding of the 19th international conference on World Wide Web (WWW '10), 2010.

[51] J. Scripps, P.N. Tan, and A.-H. Esfahanian, "Measuring the effects of preprocessing decisions and network forces in dynamic network analysis", in Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09), 2009.

[52] L. Tang and H. Liu, "Relational learning via latent social dimensions", in Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09), 2009.

[53] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network", in Proceeding of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'03), 2003.

[54] S. Tang, J. Yuan, X. Mao, X. Li, W. Chen, and G. Dai, "Relationship Classification in Large Scale Online Social Networks and Its Impact on Information Propagation", in IEEE INFOCOM, 2011.

[55] P. Domingos and M. Richardson, "Mining the network value of customers", in Proceeding of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining, 2001.

[56] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing", in Proceeding of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'02), 2002.

[57] Y. Kim and J. Srivastava, "Impact of social influence in e-commerce decision making", in Proceeding of the 9th international conference on Electronic commerce, 2007.

[58] J. Hartline, V. Mirrokni, and M. Sundararajan, "Optimal marketing strategies over social networks", in Proceeding of the 17th international conference on World Wide Web, 2008.

[59] A.M. Rashid, G. Karypis, and J. Riedl, "Influence in Ratings-Based Recommender Systems: An Algorithm-Independent Approach", in Proceeding of SDM, 2005.

[60] S. Adali, R. Escriva, M. Goldberg, M. Hayvanovych, M. Malik, B. Szymanski, W. William, G. Williams, "Measuring behavioral trust in social networks", in Intelligence and Security Informatics (ISI), IEEE International Conference, 2010.

[61] E. Wang, "A Survey of Web-based Social Network Trust", in ITEC 810 final report.

[62] J. Golbeck and J. Hendler, "Inferring binary trust relationships in Web-based social networks", in ACM Transactions on Internet Technology, vol. 6, Issue 4, 2006.

[63] J. Golbeck, "Computing and Applying Trust in Web-based Social Networks" PhD Thesis submitted to University of Maryland, 2005.

[64] C. Bizer, "Web Information Quality Assessment Policy Language Specification", 2006.

[65] U. Kuter and J. Golbeck, "SUNNY: A New Algorithm for Trust Inference in Social Networks Using Probabilistic Confidence Models", in Proceeding of Artificial intelligence (AAAI '07), 2007.

[66] V. Carchiolo, A. Longheu, M. Malgeri, G. Mangioni and G. Torrisi, "A Distributed Algorithm for Personalized Trust Evaluation in Social Networks", in Proceeding of the 4th International Symposium on Intelligent Distributed Computing (IDC '10) 2010.

[67] G. Liu, Y. Wang, M. Orgun, and E. Lim, "A Heuristic Algorithm for Trust-Oriented Service Provider Selection in Complex Social Networks", in Proceeding of IEEE SCC, 2010.

[68] G. Swamynathan, C. Wilson, B. Boe, K. Almeroth, and B. Zhao, "Do social networks improve e-commerce? A study on social marketplaces", in Proceeding of the first workshop on Online social networks (WOSP '08) ACM, 2008.

[69] M. Lesani and N. Montazeri, "Fuzzy trust aggregation and personilized trust inference in virtual social networks", in Computational Intelligence, vol. 25, issue 2, 2009.

[70] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of Trust and Distrust", in Proceedings of the 13th international conference on World Wide Web, 2004.

[71] J. Caverlee, L. Liu, and S. Webb, "Towards Robust Trust Establishment in Web-Based Social Networks with SocialTrust", in Proceeding of the 17th international conference on World Wide Web, 2008.

[72] M. Taherian, M. Amini, and R. Jalili, "Trust Inference in Web-Based Social Networks Using Resistive Networks. Internet and Web Applications and Services", in ICIW '08 Third International Conference, 2008.

[73] J. Zhang, J. Tang, and J. Li, "Expert finding in a social network", in Advances in Databases: Concepts, Systems and Applications, vol. 4443, 2010.

[74] T. Lappas, K. Liu, and E. Terzi, "A survey of algorithms and systems for expert location in social networks", in Social Network Data Analytics, Edited by C. Aggarwal, Springer, 2011.

[75] Y. Fu, X. Rongjing, L. Yiqun, Z. Min, and M. Shaoping, "Finding experts using social network analysis", in Proceeding of the IEEE/WIC/ACM International Conference on Web Intelligence, 2007.

[76] J. Zhang, M. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms", in Proceeding of the 16th International Conference on World Wide Web (WWW'07), 2007.

[77] C. Campbell, P. Maglio, A. Cozzi, and B. Dom, "Expertise identification using email communications", in Proceeding of the twelfth international conference on Information and knowledge management CIKM '03, (2003).

[78] A. Culotta, R. Bekkerman, and A. McCallum, "Extracting social networks and contact information from email and the web", in Proceeding of Conference on Email and Spam (CEAS'04), 2004.

[79] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks", in Proceeding of the 19th international joint conference on Artificial intelligence, 2005.