Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Πρόγνωση Δεσμών σε Κοινωνικά Δίκτυα με Χρήση Τεχνικών Μηχανικής Μάθησης** **Link Prediction in Social Networks through the utilization of Machine Learning Techniques**
Ονοματεπώνυμο Φοιτητή	**Σπυρίδων Παπαδόπουλος**
Πατρώνυμο	**Δημήτριος**
Αριθμός Μητρώου	**ΜΠΣΠ/15063**
Επιβλέπων Καθηγητής	**Γεώργιος Τσιχριντζής, Καθηγητής**

Ημερομηνία Παράδοσης **Δεκέμβριος 2019**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Γεώργιος Τσιχριντζής	Διονύσιος Σωτηρόπουλος	Σακκόπουλος Ευάγγελος
Καθηγητής	Επίκουρος Καθηγητής	Επίκουρος Καθηγητής

Περιεχόμενα

1. Abstract

Social network is a social structure which is consisting of social groups and links between these groups. A social network can be depicted as a graph where the nodes represent the groups or the participants, such as people or organizations. Due to rapid development of Internet, communication and collaboration between people has become easier. In recent years, digital social networks such as Facebook, Twitter and Weibo have become an important part of our daily lives. These platforms allow us to exchange large amount of information. Given the enormous amounts of data, social networks have direct reflection to human society. However, mining and analyzing social network data is a non-trivial task, which faces two challenges: imperfection and dynamism. First, all the received data are incomplete and secondly all social networks are particularly dynamic. Forecasting, for lost or unconnected links to current social networks that have recently been added or deleted, is very important for future social networks. This problem is commonly known as link prediction, and over the past decade, several efforts have been made by psychologists, computer scientists, physicists and economists to resolve this problem.

2. Εισαγωγή

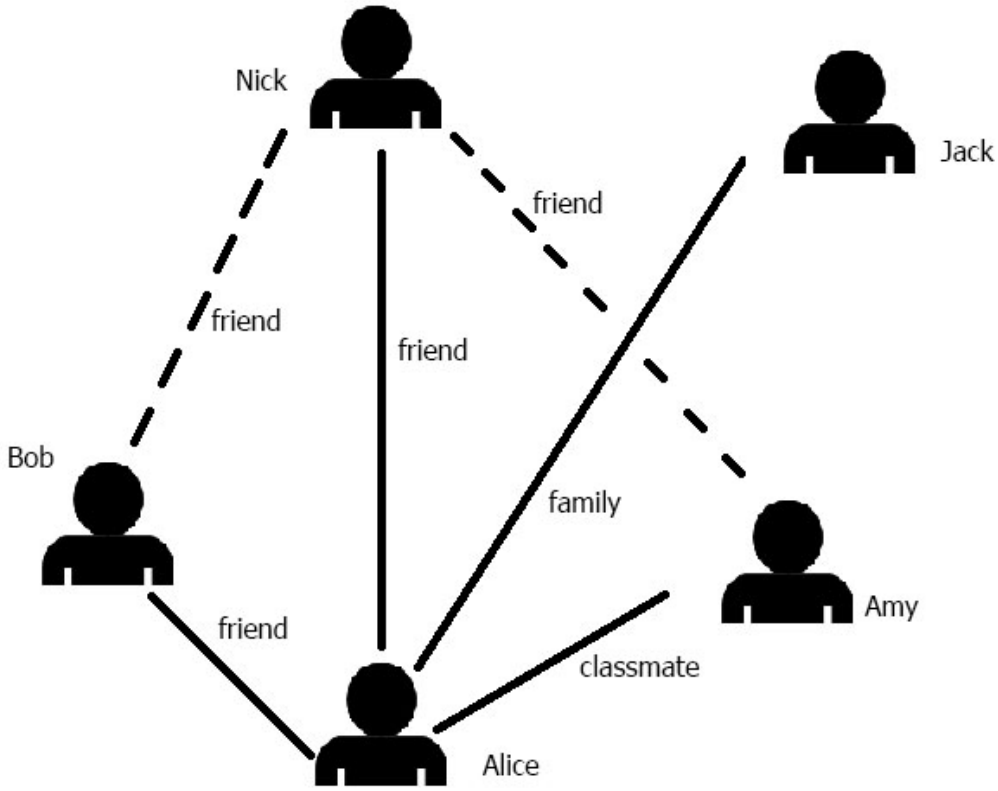
Ένα κοινωνικό δίκτυο είναι μια κοινωνική δομή που αποτελείται από ένα σύνολο κοινωνικών φορέων και ένα σύνολο δεσμών μεταξύ των φορέων αυτών. Ένα κοινωνικό δίκτυο μπορεί να απεικονιστεί ως ένα γράφημα, όπου οι κόμβοι αντιπροσωπεύουν τους φορείς / συμμετέχοντες (όπως άτομα, οργανώσεις, κ.α.), ενώ τα άκρα (δηλαδή οι συνδέσεις) αντιστοιχούν στις σχέσεις / αλληλεπιδράσεις μεταξύ των φορέων. Με τη ραγδαία ανάπτυξη του διαδικτύου, η επικοινωνία και η συνεργασία μεταξύ των ανθρώπων έχουν γίνει πιο εύκολες. Τα τελευταία χρόνια, τα ψηφιακά κοινωνικά δίκτυα όπως το Facebook, το Twitter και το Weibo, έχουν γίνει ένα σημαντικό μέρος της καθημερινής μας ζωής και αποτελούν πλατφόρμες που επιτρέπουν την ανταλλαγή πληροφοριών μεταξύ τους. Δεδομένου των τεράστιων ποσοτήτων δεδομένων τα κοινωνικά δίκτυα έχουν κάποια εμφανή χαρακτηριστικά όπως η υψηλή ποιότητα και η άμεση αντανάκλαση της πραγματικής ανθρώπινης κοινωνίας, με αποτέλεσμα πολλοί ερευνητές από διαφορετικές περιοχές ή κλάδους να δίνουν όλο και μεγαλύτερη προσοχή στα κοινωνικά δίκτυα.

Ωστόσο, η εξόρυξη και ανάλυση δεδομένων των κοινωνικών δικτύων είναι μια μη τετριμμένη εργασία, η οποία αντιμετωπίζει δύο προκλήσεις: την ατέλεια και τη δυναμική. Πρώτον, σχεδόν όλα τα ληφθέντα δεδομένα των κοινωνικών δικτύων δεν είναι πλήρη, δεδομένου ότι μόνο ένα μέρος των πληροφοριών μπορούν να συλλεχθούν από τις πλατφόρμες κοινωνικής δικτύωσης. Δεύτερον, τα κοινωνικά δίκτυα είναι ιδιαίτερα δυναμικά, το οποίο μπορεί να οδηγήσει τους κόμβους και τις ακμές να εμφανιστούν ή να χαθούν στο μέλλον. Ως εκ τούτου, η πρόβλεψη που γίνεται, για χαμένες ή για απαρατήρητες συνδέσεις στα τρέχουσα κοινωνικά δίκτυα που πρόσφατα έχουν προστεθεί ή διαγραφεί, είναι πολύ σημαντική για τα μελλοντικά κοινωνικά δίκτυα, όχι μόνο για την κατανόηση της εξέλιξης των κοινωνικών δικτύων, αλλά και για την ολοκλήρωση της τρέχουσας κατάστασης των κοινωνικών δικτύων. Αυτό το πρόβλημα είναι κοινώς γνωστό ως πρόβλεψη δεσμών, ενώ την τελευταία δεκαετία, έχουν καταβληθεί αρκετές προσπάθειες από ψυχολόγους, επιστήμονες πληροφορικής, φυσικούς και οικονομολόγους για την επίλυση αυτού του προβλήματος.

3. Πρόβλημα

Θεωρήστε ένα κοινωνικό δίκτυο G(V, E) τη χρονική στιγμή t, όπου V και E είναι σύνολα κόμβων και δεσμών, αντίστοιχα. Η πρόβλεψη δεσμών στοχεύει στην πρόβλεψη νέων δεσμών ή διαγραμμένων δεσμών μεταξύ των κόμβων για τη χρονική στιγμή t'(t'>t), ή χαμένων δεσμών ή απαρατήρητους δεσμούς, στο υπάρχον δίκτυο.

Αυτό το πρόβλημα μπορεί να εξηγηθεί μέσω ενός απλού κοινωνικού δικτύου, το οποίο αποτελείται από 5 ανθρώπους, όπως φαίνεται στο Σχήμα 1. Στο παράδειγμα αυτό, παρατηρούμε ότι τη χρονική στιγμή t υπάρχουν σταθεροί δεσμοί και οι διακεκομμένοι δεσμοί υποδεικνύουν ότι δημιουργήθηκαν νέοι δεσμοί κατά τη χρονική διάρκεια [t, t'].

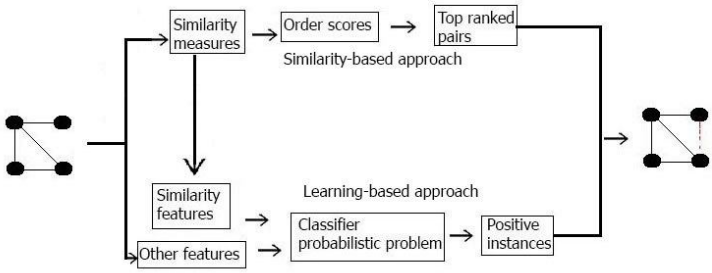


Σχήμα 1. Ένα παράδειγμα για την κατανόηση του προβλήματος σύνδεσης.

Τη χρονική στιγμή t, η Alice και ο Bob είναι φίλοι, όπως και η Alice με τον Nick είναι επίσης φίλοι. Κατά τη χρονική στιγμή t', ίσως η Alice σύστησε τον Bob στον Nick, έγιναν και αυτοί φίλοι. Κατ 'ανάλογο τρόπο, ο Nick και η Amy θα γίνουν φίλοι τη χρονική στιγμή t'. Ο στόχος του προβλήματος της πρόβλεψης δεσμών εδώ είναι να προβλέψει την εμφάνιση της φιλίας που προστέθηκε πρόσφατα μεταξύ των ατόμων.

Για να λυθεί το πρόβλημα της πρόβλεψης χρειάζεται να καθορίσουμε τις πιθανότητες της δομής ή της διάσπασης των δεσμών μεταξύ όλων των ζευγαριών κόμβων. Συνήθως, αυτές οι πιθανότητες μπορούν να υπολογιστούν από παραπλήσιες ή από σχετικές κατατάξεις μεταξύ ζευγαριών κόμβων.

Χρησιμοποιούμε ένα γενικό πλαίσιο για την απεικόνιση της λύσης της πρόβλεψης δεσμών, όπως φαίνεται στο Σχήμα 2. Για ένα αρχικό κοινωνικό δίκτυο, υπάρχουν δύο τρόποι για να προβλέψουμε την εξέλιξη των δεσμών: 1^{ov} με την προσέγγιση της ομοιότητας και 2^{ov} με την προσέγγιση της μάθησης. Εδώ προβλέπουμε τους νέους / χαμένους / απαρατήρητους δεσμούς ως παραδείγματα.



Σχήμα 2. Πλαίσιο γενική πρόβλεψης δεσμών

Μια προσέγγιση ομοιότητας είναι να υπολογίσουμε τις ομοιότητες σε μη συνδεδεμένα ζεύγη κόμβων σε ένα κοινωνικό δίκτυο. Κάθε δυναμικό ζεύγος κόμβων (x, y) θα εκχωρηθεί ως αποτέλεσμα. όπου το υψηλότερο αποτέλεσμα σημαίνει και μεγαλύτερη πιθανότητα ότι τα X και Y θα συνδέονται στο μέλλον, και αντιστρόφως. Στη συνέχεια, βάση του πίνακα κατάταξης των αποτελεσμάτων κατά φθίνουσα σειρά μπορούμε να πούμε ότι οι δεσμοί που βρίσκονται στη κορυφή της λίστας είναι πολύ πιθανό να εμφανιστούν.

Η προσέγγιση της μάθησης αντιμετωπίζει το πρόβλημα της πρόβλεψης ως μια δυαδική ταξινόμηση διεργασιών. Όπως αναφέρουν οι Hasan, Chaoji, Salem, Zaki, "η πρόβλεψη δεσμών σε ένα κοινωνικό δίκτυο είναι ένα σημαντικό πρόβλημα, ενώ παράλληλα μας βοηθάει στο να κατανοήσουμε και να αναλύσουμε τα social groups, τονίζοντας ότι η λύση στο πρόβλημα μπορεί να δοθεί από το μοντέλο ταξινόμησης" [1]. Ως εκ τούτου, ορισμένα τυπικά μοντέλα μηχανικής

μάθησης, όπως το μοντέλο ταξινόμησης και πιθανοτήτων, μπορεί να χρησιμοποιηθεί για την επίλυση αυτού του προβλήματος.

Κάθε μη-συνδεδεμένο ζεύγος κόμβων αντιστοιχεί σε μια εμφάνιση με τα χαρακτηριστικά γνωρίσματα που περιγράφουν τους κόμβους και την κλάση. Αν υπάρχει ένας πιθανός δεσμός που συνδέει ένα ζεύγος κόμβων, τότε αυτό το ζευγάρι χαρακτηρίζεται ως θετικό, αλλιώς είναι αρνητικό. Για τις προσεγγίσεις της μάθησης, τα χαρακτηριστικά αποτελούνται από δύο μέρη: το ένα είναι το χαρακτηριστικό ομοιότητας από τις προσεγγίσεις ομοιότητας, και το άλλο μέρος είναι τα χαρακτηριστικά που προέρχονται από το κοινωνικό δίκτυο, όπως η γραπτή ενημέρωση των χαρακτηριστικών και η γνώση του τομέα. Η πρόβλεψη δεσμών για τη διαγραφή / εξαφάνιση δεσμών μπορεί να επιλυθεί με ανάλογο τρόπο.

Υπάρχουν πολλές αναφορές για την πρόβλεψη δεσμών, οι οποίες εστιάζουν στις τεχνικές γενικής πρόβλεψης, σε ειδικά προβλήματα πρόβλεψης δεσμών, και χρησιμοποιούν τις υπάρχουσες τεχνικές πρόβλεψης για να αντιμετωπίσουν διάφορες εφαρμογές. Η παρούσα εργασία προτείνει μια νέα κατηγορία πρόβλεψης με δύο προοπτικές: Η προοπτική της τεχνική και η προοπτική του προβλήματος.

Η κατηγορία δεν περιέχει εφαρμογές πρόβλεψης, δεδομένου ότι βασίζεται σε τεχνικές πρόβλεψης και προβλήματα, και θα πρέπει να αντιμετωπιστούν ξεχωριστά.

Οι τεχνικές πρόβλεψης, μπορούν να χωριστούν σε τέσσερα επίπεδα:

(1) Σύμφωνα με τις βασικές πληροφορίες δικτύου που χρησιμοποιούνται στην πρόβλεψη, το πρώτο και υψηλότερο επίπεδο που αποτελείται από κόμβο, τοπολογία και κοινωνική θεωρία.

(2) Στο δεύτερο επίπεδο, η τοπολογία χωρίζεται σε γειτονιά, τη διαδρομή, ενώ η κοινωνική θεωρία είναι επίσης χωρισμένη σε γειτονιά, τριάδα και δομημένη τρύπα.

(3) Το τρίτο επίπεδο περιλαμβάνει δημοφιλής βασικές τεχνικές πρόβλεψης δεσμών βασισμένες σε κόμβους, γειτονιές, μονοπάτια και κοινωνική θεωρία.

(4) Στο τέταρτο επίπεδο, η βασική τεχνική πρόβλεψης και εξωτερική πληροφορία, η οποία περιλαμβάνει βάρη, χαρακτηριστικά και αρχείο γνώσης, παρέχει χαρακτηριστικά για περίπλοκες τεχνικές μάθησης.

4. Βιβλιογραφική επισκόπηση

4.1. Τεχνικές πρόβλεψης δεσμών

Υπάρχουν αρκετές μετρήσεις πρόβλεψης δεσμών, οι οποίες χρησιμοποιούν πληροφορίες κόμβων, τοπολογίας και κοινωνικής θεωρίας για να υπολογίσουν τις ομοιότητες των ζευγαριών κόμβων. Επιπλέον, οι μέθοδοι της μάθησης είναι πιο πολύπλοκοι, αλλά στηρίζονται στις βασικές μετρήσεις και την εξωτερική πληροφορία. Παρακάτω παρουσιάζονται μερικές μετρήσεις και μέθοδοι πρόβλεψης.

4.1.1. Μετρήσεις βάση κόμβων

Ο υπολογισμός της ομοιότητας μεταξύ ενός ζευγαριού κόμβων αποτελεί μία διαισθητική λύση για τη πρόβλεψη δεσμών. Βασίζεται στην απλή ιδέα: όσο πιο απλό είναι το ζευγάρι, τόσο πιο πιθανή είναι η σύνδεση μεταξύ τους, και αντίστροφα. Αυτό συνάδει με το γεγονός ότι οι χρήστες έχουν την τάση να δημιουργούν σχέσεις με τους ανθρώπους που έχουν ίδια εκπαίδευση, θρησκεία, συμφέροντα και τοποθεσία. Μπορεί να μετρηθεί η βάση της ομοιότητας, στην οποία κάθε μη συνδεδεμένο ζεύγος κόμβων (X,Y) έχει μια βαθμολογία που σηματοδοτεί την ομοιότητα μεταξύ χ και y. Μια υψηλή βαθμολογία υποδεικνύει υψηλή πιθανότητα ότι τα X και Y θα συνδέονται στο μέλλον, ενώ μια χαμηλή βαθμολογία υποδεικνύει επίσης υψηλή πιθανότητα ότι τα X και Y δεν θα συνδέονται. Ως εκ τούτου, χρησιμοποιώντας την κατάταξη των βαθμών ομοιότητας, μπορούμε να προβλέψουμε τις εμφανίσεις ή χαμένες συνδέσεις στο μέλλον ή τις μη ορατές συνδέσεις στα τρέχουσα δίκτυα.

Σε ένα πρακτικό κοινωνικό δίκτυο, ένας κόμβος έχει συνήθως κάποια χαρακτηριστικά, όπως το προφίλ σε ένα ψηφιακό κοινωνικό δίκτυο, το όνομα του mail στα δίκτυα ηλεκτρονικού ταχυδρομείου, και δημοσιευμένες ακαδημαϊκές εργασίες σε κοινωνικά δίκτυα. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν άμεσα για τον υπολογισμό της ομοιότητας μεταξύ δύο κόμβων. Δεδομένου ότι στις περισσότερες περιπτώσεις οι τιμές των κόμβων είναι γραπτής μορφής, οι μετρήσεις της ομοιότητας βάση κειμένου και string συνήθως χρησιμοποιούνται εδώ.

Αυτό έχει ως συμπέρασμα, οι μετρήσεις που βασίζονται σε κόμβους να χρησιμοποιούν χαρακτηριστικά και ενέργειες, οι οποίες αντανακλούν στα προσωπικά ενδιαφέροντα και τη κοινωνική συμπεριφορά, ώστε να υπολογιστούν οι ομοιότητες μεταξύ των ζευγαριών των κόμβων. Ως εκ τούτου, οι μετρήσεις αυτές είναι χρήσιμες στην πρόβλεψη δεσμών εάν μπορούμε να ανακτήσουμε τα χαρακτηριστικά των χρηστών και τις ενέργειές τους στα κοινωνικά δίκτυα.

4.1.2. Μετρήσεις βάση τοπολογίας

Ακόμα και σε ένα απλό δίκτυο χωρίς κόμβους ή χαρακτηριστικά, υπάρχουν αρκετές μετρήσεις διαθέσιμες για να υπολογίσουμε τις ομοιότητες δύο κόμβων. Οι περισσότερες μετρήσεις βασίζονται στις τοπολογικές πληροφορίες και ονομάζονται μετρήσεις τοπολογίας. Εδώ θα παρουσιάσουμε μία συστηματική εξήγηση διάσημων μετρήσεων τοπολογίας στη πρόβλεψη δεσμών. Σύμφωνα με τα χαρακτηριστικά των μετρήσεων, μπορούν να χωριστούν σε μετρήσεις γειτόνων, μονοπατιών και τυχαίων ακολουθιών.

4.1.2.1. Μετρήσεις γειτόνων

Σε ένα κοινωνικό δίκτυο, οι άνθρωποι τείνουν να δημιουργήσουν νέες σχέσεις με ανθρώπους, με τους οποίους βρίσκονται κοντά. Οι γείτονες αφορούν τους κοντινούς ανθρώπους του χρήστη.

Common Neighbors (CN): Η CN μέτρηση είναι η πιο διαδεδομένη που χρησιμοποιείται για το πρόβλημα της πρόβλεψης δεσμών κυρίως λόγω της απλότητάς της. Για δύο κόμβους x, y η CN καθορίζεται από τους αριθμούς των κόμβων όπου και οι x, y θα δημιουργήσουν. Ο τύπος του CN είναι ο παρακάτω:

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

Jaccard Coe_cient (JC): Θεωρεί τις υψηλότερες τιμές των ζευγών κόμβων σε ένα σύνολο αριθμών γειτόνων.

$$JC(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Sorensen Index (SI): Η μέτρηση αυτή μας δείχνει ότι οι χαμηλές ενδείξεις των κόμβων θα μπορούσαν να έχουν υψηλές πιθανότητες δεσμών

$$SI(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| + |\Gamma(y)|}$$

Salton Cosine Similarity (SC): SC είναι ένα κοινό συνημίτονο για τη μέτρηση ομοιοτήτων μεταξύ δύο κόμβων x, y.

$$SC(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x) \times \Gamma(y)|}}$$

Hub Promoted (HP): Η HP αφορά τη τοπολογία των κόμβων x,y και η τιμή καθορίζεται από το χαμηλό δείκτη των κόμβων.

$$HP(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min(|\Gamma(x)|,|\Gamma(y)|)}$$

Hub Depressed (HD): Είναι ίδια μέτρηση με την HP αλλά η τιμή καθορίζεται από τους υψηλούς δείκτες των κόμβων.

$$HD(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max(|\Gamma(x)|,|\Gamma(y)|)}$$

Leicht-Holme-Nerman (LHN): Αυτή η μέτρηση μας δίνει υψηλή ομοιότητα των ζεύγων κόμβων οι οποίοι έχουν κοινούς γείτονες σε σύγκριση με τον αναμενόμενο αριθμό γειτόνων.

$$LHN(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| \, x \, |\Gamma(y)|}$$

Parameter-Dependent (PD): Για να βελτιώσουμε την ακρίβεια της πρόβλεψης προτείνεται η μέτρηση PD. λ είναι μία ελεύθερη μεταβλητή, όπου λ=0, PD μετατρέπεται σε CN. Εάν λ=0,5 και λ=1, τότε μετατρέπεται σε Salton και LHN, αντίστοιχα.

$$PD(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{(|\Gamma(x)| \, x \, |\Gamma(y)|)^{\lambda}}$$

Adamic-Adar Coefficient (AA): Χρησιμοποιείται για τη μέτρηση μεταξύ δύο ιστοσελίδων, που χρησιμοποιούνται στα κοινωνικά δίκτυα.

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

Preferential Attachment (PA): Υποδεικνύει ότι οι νέοι δεσμοί είναι πιο πιθανό να συνδέουν υψηλό αριθμό κόμβων, παρά χαμηλό.

$$PA(x, y) = |\Gamma(x)| \times |\Gamma(y)|$$

Resource Allocation (RA) : Η RA μέτρηση μοιάζει με την AA. Όχι μόνο χρησιμοποιούν γείτονες, αλλά επίσης θεωρούν γείτονες των γειτόνων.

$$RA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|}$$

4.1.2.2. Μετρήσεις μονοπατιού

Εκτός από τις πληροφορίες των κόμβων και των γειτόνων, τα μονοπάτια μεταξύ δύο κόμβων μπορούν να χρησιμοποιηθούν για τον υπολογισμό ομοιοτήτων των ζευγών κόμβων.

Local Path (LP): Η LP μέτρηση χρησιμοποιεί τις πληροφορίες των κοντινών μονοπατιών σε βάθος 2 και 3. Τα μονοπάτια σε βάθος 2 είναι πιο σχετικά από αυτά σε βάθος 3.

$$LP = A^2 + aA^3$$

Katz: Η μέτρηση αυτή βασίζεται στο σύνολο των κόμβων και μετράει όλα τα μονοπάτια μεταξύ των κόμβων.

$$Katz(x, y) = \sum_{l=1}^{\infty} \beta^l \times |path_{x,y}^l| = \beta A + \beta^2 A^2 + \beta^3 A^3 + \dots$$

Relation Strength Similarity (RSS): Πρόκειται για μία ασύμμετρη μέτρηση που μπορεί να χρησιμοποιηθεί για τα βάρη των κοινωνικών δικτύων.

$$RSS(x, y) = \sum_{l=1}^{l} R_{pl}^*(x, y)$$

$$R_{pl}^*(x, y) = \begin{cases} \Pi_{k=1}^{K} R(zk, zk+1) & K \leq r \\ 0 & otherwise \end{cases}$$

FriendLink (FL): Η μέτρηση αυτή είναι μία ομοιότητα μεταξύ των κόμβων x, y. Μας παρέχει μεγαλύτερη ακρίβεια και γρηγορότερη πρόβλεψη.

$$FL(x, y) = \sum_{l=1}^{l} \frac{1}{i-1} \; x \; \frac{|path_{x,y}^{i}|}{\Pi_{j=2}^{i}(n-j)}$$

Vertex Collocation Pro_le (VCP): Προτείνεται για ανάλυση δεσμών και πρόβλεψης.

4.1.2.3. Μέτρηση τυχαίων ακολουθιών

Οι κοινωνικές αλληλεπιδράσεις μεταξύ των κόμβων στα κοινωνικά δίκτυα μπορούν επίσης να αποτελέσουν πρότυπο από τυχαία ακολουθία.

Hitting Time (HT): Είναι ο αναμενόμενος αριθμός βημάτων που απαιτείται για μία τυχαία ακολουθία από τον κόμβο x στον y.

$$HT(x, y) = 1 + \sum_{\omega \in \Gamma(x)} P_{x,\omega} \; HT(\omega, y)$$

Commute Time (CT): Από τη στιγμή που η HT μέτρηση δεν είναι συμμετρική η CT χρησιμοποιείται για να μετρήσει τα απαιτούμενα βήματα τόσο από το x στ y, όσο και από το y στο x.

$$CT(x, y) = HT(x, y) + HT(y, x) = m(L_{x,x}^{+} + L_{y,y}^{+} - 2L_{x,y}^{+})$$

Cosine Similarity Time (CST): Βασίζεται στο L+ υπολογίζοντας την ομοιότητα δύο διανυσμάτων.

$$CST(x, y) = \frac{L^{+}_{x,y}}{\sqrt{L^{+}_{x,x}L^{+}_{y,y}}}$$

SimRank: Καθορίζεται βάση της υπόθεσης ότι δύο κόμβοι είναι όμοιοι εάν συνδέονται σε ίδιους κόμβους.

$$simRank(x, y) = \begin{cases} 1 & x = y \\ \gamma \times \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} simRank(a,b)}{|\Gamma(x)|x\,|\Gamma(y)|} & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

Rooted PageRank (RPR): Είναι μία παραλλαγή του PageRank, που αποτελεί κορυφαίο αλγόριθμο και χρησιμοποιείται στη μηχανή αναζήτησης.

$$RPR = (1 - \epsilon)\,(I - \epsilon D^{-1}A)^{-1}$$

PropFlow: Είναι ίδιο με το Rooted PageRank αλλά πιο τοπικό.

$$PF(x, y) = PF(\alpha, x)\,\frac{w_{xy}}{\sum_{k \in \Gamma(x)} w_{xk}}$$

4.1.3. Μετρήσεις βάση κοινωνικής θεωρίας

Τα τελευταία χρόνια όλο και περισσότερες εργασίες έχουν πραγματοποιηθεί πάνω στις κλασικές θεωρίες των social. Σε σχέση με τις προηγούμενες μετρήσεις, οι μετρήσεις πρόβλεψης δεσμών, που βασίζονται στη θεωρία των social, μπορούν να βελτιώσουν την απόδοσή τους με τη βοήθεια επιπρόσθετων πληροφοριών κοινωνικής αλληλεπίδρασης, ειδικά όταν αναφερόμαστε για κοινωνικά δίκτυα μεγάλης κλίμακας.

Σε ένα ενδιαφέρον δίκτυο, η ομοφυλία μπορεί να βοηθήσει όχι μόνο στην πρόβλεψη δεσμών μεταξύ ενός χρήστη και των υπηρεσιών του, αλλά επίσης και στην πρόβλεψη δεσμών μεταξύ δύο χρηστών που έχουν κοινά ενδιαφέροντα.

4.1.4. Μετρήσεις βάση μάθησης

Feature-based Classification: Στη πρόβλεψη δεσμών, κάθε μη συνδεδεμένο ζεύγος κόμβων που αντιστοιχεί σε μία περίπτωση περιλαμβάνει μία κλάση και χαρακτηριστικά που περιγράφουν ένα ζεύγος κόμβων. Επομένως, το ζεύγος κόμβων μπορεί να χαρακτηριστεί ως θετικό εάν υπάρχει σύνδεση μεταξύ των κόμβων, αλλιώς το ζεύγος χαρακτηρίζεται ως αρνητικό.

Για να δημιουργήσουμε μία αποτελεσματική κατάταξη για την πρόβλεψη σύνδεσης, είναι κρίσιμο να ορίσουμε και να εξάγουμε μια σειρά κατάλληλων χαρακτηριστικών από τα κοινωνικά δίκτυα. Τα χαρακτηριστικά, που παρέχονται από τη βάση κόμβων, τοπολογίας και της κοινωνικής θεωρίας που βασίζεται, είναι δημοφιλή και σημαντικά για την ταξινόμηση των μοντέλων μάθησης. Για παράδειγμα, η μετρική VCP μπορεί να θεωρηθεί ως ένα είδος ειδικού χαρακτηριστικού το οποίο περιγράφει τοπικές πληροφορίες τοπολογίας. Σύμφωνα με τους Lichtenwalter και Chawla “η VCP αποτελεί μία νέα μέθοδο για την ανάλυση των δεσμών και έχει την ιδιότητα να χειρίζεται πολλαπλές σχέσεις” [2]. Επιπλέον, πολλές μελέτες δείχνουν ότι η χρήση χαρακτηριστικών των κόμβων και των συνδέσεων μπορούν να βελτιώσουν σημαντικά την απόδοση της πρόβλεψης συνδέσμου.

Probabilistic Graph Model: Σε ένα κοινωνικό δίκτυο, σε μια σύνδεση μεταξύ κάθε ζεύγους κόμβων μπορεί να ανατεθεί μια τιμή πιθανότητας όπως μια τοπολογική ομοιότητα ή πιθανότητα μετάβασης σε τυχαία ακολουθία. Αυτό είναι το Probabilistic Graph. Υπάρχουν πολλές μέθοδοι μάθησης της πρόβλεψης συνδέσμου που έχουν προταθεί από την αξιοποίηση του μοντέλου Probabilistic Graph. Μελέτες δείχνουν ότι πολλά δίκτυα παρουσιάζουν ιεραρχική δομή, όπου οι κόμβοι χωρίζονται σε ομάδες που μπορούν να υποδιαιρούνται σε περαιτέρω ομάδες των ομάδων, και ούτω καθεξής σε πολλαπλές κλίμακα.

Matrix Factorization: Το μοντέλο αυτό συνδυάζει τα λανθάνοντα γνωρίσματα με σαφή χαρακτηριστικά για τους κόμβους και τις συνδέσεις στο γράφημα μέσω ενός διγραμμικού μοντέλου παλινδρόμησης. Τα λανθάνοντα γνωρίσματα μπορούν επίσης να συνδυαστούν με τα αποτελέσματα των άλλων μοντέλων πρόβλεψης συνδέσμου. Το μοντέλο βελτιστοποιεί για την AUC άμεσα, προκειμένου να ξεπεραστεί το πρόβλημα ανισορροπίας, το οποίο αναφέρεται στο φαινόμενο της θετικής σύνδεσης για ένα πολύ μικρό ποσοστό του συνόλου των συνδέσεων, και στο φαινόμενο της αρνητικής σύνδεσης για το μεγαλύτερο μέρος των περιπτώσεων.

4.1.5. Σύνολο δεδομένων και εργαλεία

Σχεδόν όλες οι ενέργειες πρόβλεψη δεσμού χρειάζονται να επιβεβαιώσουν τις μεθόδους τους σε μία βάση δεδομένων. Το σύνολο δεδομένων είναι σημαντικό για την αναπαραγωγή και σύγκριση διαφορετικών μεθόδων πρόβλεψης δεσμού. Η δημιουργία και η συλλογή δεδομένων είναι χρονοβόρα και απαιτεί έντονη εργασία. Ωστόσο, δεν είναι όλα τα σύνολα δεδομένων δημόσια και διαθέσιμα. Επίσης, οφείλουμε να επισημάνουμε και μερικά από τα μειονεκτήματα των datasets. Καταρχάς μερικά από αυτά έχουν θόρυβο, που πρέπει να καθαριστεί πριν χρησιμοποιηθούν. Επίσης, αυτά τα datasets δεν είναι πλούσια και αρκετά διαφορετικά ως προς το μέγεθος και τον τύπο δικτύου.

Ωστόσο, υπάρχουν αρκετές μετρήσεις και μέθοδοι πρόβλεψης που προτείνονται και ελάχιστοι είναι αυτοί που λειτουργούν. Είναι πολύ σημαντικό να επιλέξουμε τις κατάλληλες μετρήσεις ή μεθόδους για μία πρόβλεψη.

5. Προβλήματα πρόβλεψης σύνδεσης

Υπάρχουν πολλές εργασίες που επικεντρώνονται στην επίλυση ειδικών προβλέψεων δεσμών, οι οποίες μπορούν να χωριστούν σε τέσσερις κατηγορίες: πρόβλεψη χρονικής σύνδεσης, πρόβλεψη σε ετερογενή δίκτυα, πρόβλεψη σε χαμένη σύνδεση και κλιμάκωση της πρόβλεψης.

5.1. Πρόβλεψη χρονικής σύνδεσης

Τα τελευταία χρόνια, η έρευνα για την πρόβλεψη σύνδεσης έχει εξελιχθεί με διάφορες πτυχές. Η μία είναι να θεωρήσουμε το χρόνο στο μοντέλο, το οποίο μπορεί να ονομαστεί ως πρόβλεψη χρονικής σύνδεσης [3, 4]. Ένα κοινωνικό δίκτυο, με το χρόνο, μπορεί να οργανωθεί ως τρίτης τάξης tensor, ή πολυδιάστατο array. Ένα tensor Z του μεγέθους MxNxT μπορεί να οριστεί ως

$$Z(i, j, t) = \begin{cases} 1 & \text{εάν η κορυφή } i \text{ ενώνεται με την κορυφή } j \text{ τη χρονική στιγμή } t \\ 0 & \text{αλλιώς} \end{cases}$$

Μπορεί να απαντήσει συγκεκριμένες ερωτήσεις όπως «Ποιος είναι πιο πιθανό να δημοσιεύσει στο ICDM το επόμενο έτος». Με δεδομένο κοινωνικό δίκτυο για τους χρόνους 1 έως T, θα πρέπει να προβλέψει τις συνδέσεις στο χρόνο T + 1.

5.2. Πρόβλεψη σύνδεσης σε Ετερογενή Δίκτυα

Τα περισσότερα έργα πρόβλεψης δεσμών εστιάζουν στα ομοιογενή δίκτυα, στα οποία μόνο ένα τύπος κόμβων ή δεσμών υπάρχει. Ωστόσο, πολλά κοινωνικά δίκτυα περιέχουν διαφορετικούς τύπους συνδέσεων και διαφορετικά είδη κόμβων, οι οποίοι μπορεί να έχουν διαφορετικές τυπολογίες ή μορφή σύνδεσης και επιρροή.

Ετερογενή σχέσεις, όπως η φιλία, η οικογένεια και η συναδελφικότητα, έχουν χαρακτηριστεί ως ακαθόριστες στα κοινωνικά δίκτυα. Μερικές νέες multi-relational link prediction (MRLP) μέθοδοι για ετερογενή πληροφοριακά δίκτυα προτείνονται ώστε να ξεπεραστούν κοντινές παραδοσιακές μέθοδοι πρόβλεψης. Το κλειδί του MRLP είναι ένα κατάλληλο σχέδιο βάρους για διαφορετικό τύπο συνδυασμών. Τα βάρη καθορίζονται από την μέτρηση της εμφάνισης κάθε μοναδικής triad census με τρεις κόμβους. Η triad census παρέχει τη πιθανότητα κάθε δομής, η οποία κατ' επέκταση μεταφράζεται σε πιθανότητα, όπου μερική τριάδα είναι κλειστή.

5.3. Πρόβλεψη σύνδεσης σε ανακόλουθες ή χαμένες συνδέσεις

Ο σχηματισμός και η διάλυση της σύνδεσης είναι δύο θεμελιώδεις διαδικασίες της αλλαγής σύνδεσης και της εξέλιξης σε δυναμικά δίκτυα. Σύνδεσμοι σε κοινωνικά δίκτυα μπορεί να εμφανιστούν ή να εξαφανιστούν. Για παράδειγμα, ο χρήστης Α στο Twitter διακόπτει τη σχέση με έναν άλλο χρήστη Β. Στα κοινωνικά δίκτυα, καλούμε αυτήν τη συμπεριφορά “unfollow”. Από όσο γνωρίζουμε, έχουν γίνει πολλές προσπάθειες στη μελέτη του σχηματισμού της σύνδεσης για την πρόβλεψη νέων μελλοντικών συνδέσεων, αλλά μόνο λίγοι δίνουν σημασία στη διάλυση της σύνδεσης, δηλαδή, προβλέποντας την εξαφάνιση της σύνδεσης στο μέλλον.

Έρευνες έχουν δείξει ότι οι χρήστες του Twitter συχνά κάνουν “unfollow”, και μερικοί παράγοντες που οδηγούν σε αυτό το αποτέλεσμα είναι η αβεβαιότητα των σχέσεων, η διάρκεια μίας σχέσης και η επικάλυψη της σχέσης.

5.4. Κλιμάκωση της πρόβλεψης δεσμών

Η επεκτασιμότητα και η αποτελεσματικότητα είναι και οι δύο σημαντικές για τα τεράστια κοινωνικά δίκτυα του πραγματικού κόσμου. Ο Sarkar πρότεινε μια μη παραμετρική πρόβλεψη συνδέσμων για δυναμικά δίκτυα [5] στα οποία το μοντέλο τους μπορεί να φιλοξενήσει γειτονιές με πολύ διαφορετικές προοπτικές εξέλιξης. Παρέχει επίσης τη δυνατότητα εκμάθησης βασισμένη τόσο σε τοπολογικές όσο και σε άλλα διαθέσιμα εξωτερικά χαρακτηριστικά. Επίσης, προσάρμοσαν τον ευαίσθητο αλγόριθμο κατακερματισμού της περιοχής για να επιλύσουν την επεκτασιμότητα της πρόβλεψης σύνδεσης σε μεγάλα δίκτυα και μακρές ακολουθίες χρόνου.

Ο Song ανέπτυξε δύο καινοτόμες μεθόδους που προσεγγίζουν με αποτελεσματικότητα και ακρίβεια μια μεγάλη οικογένεια μέτρων εγγύτητας, κάτι το οποίο αποτελεί πρόκληση για μία μαζική κλίμακα και δυναμικά online κοινωνικά δίκτυα. Στη συνέχεια, η προτεινόμενη εκτίμηση εγγύτητας χρησιμοποιείται για την πρόβλεψη της σύνδεσης και επιτυγχάνει υψηλή ακρίβεια στη πρόβλεψη συνδέσεων συνδυάζοντας πολλαπλά μέτρα εγγύτητας. Το σημαντικό στα νέα ευρήματα του Song είναι ότι “αφορούν 1) την αποτελεσματικότητα διαφορετικών μέτρων προσέγγισης που ποικίλλουν σε διαφορετικά δίκτυα και 2) τον συνδυασμό πολλαπλών μέτρων προσέγγισης που αποδίδουν με ακρίβεια” [6]. Ως εκ τούτου ένας νέος αλγόριθμος προτείνεται για να επιτρέπει την πρόβλεψη συνδέσεων σχεδόν σε πραγματικό χρόνο σε εξαιρετικά δυναμικά κοινωνικά δίκτυα.

Για να χειριστούμε την πρόβλεψη δεσμών σε μαζικά εξελισσόμενα δίκτυα με αραιές συνδετικότητες και μη γραμμικά μεταβατικά μοτίβα, ο Li πρότεινε ένα πλαίσιο «βαθιάς» μάθησης που ονομάζεται υποθετική χρονική περιορισμένη μηχανή Boltzmann [7], η οποία προβλέπει συνδέσεις που βασίζονται σε μεμονωμένη μεταβατική μεταβολή και επιρροή που εισήχθησαν από τους τοπικούς γείτονες. Είναι ανθεκτικό στο θόρυβο και έχει την εκθετική ικανότητα να καταγράφει μη γραμμικές διακυμάνσεις. Εκτός από τα υπολογιστικά πλεονεκτήματα, αυτή η μέθοδος επινοεί δύο τύπους κατευθυνόμενων συνδέσεων σε κρυφές μεταβλητές: χρονικές συνδέσεις και γειτονικές συνδέσεις, που δύο γνωστές παραδοχές. Κάθε κόμβος έχει ένα μοναδικό μεταβατικό μοτίβο και η συμπεριφορά ενός κόμβου επηρεάζεται από τους τοπικούς γείτονές του.

6. Μεθοδολογία Dataset

Τα dataset παίζουν σημαντικό ρόλο για την αναπαραγωγή και τη σύγκριση διαφορετικών μεθόδων πρόβλεψης δεσμών.

Έχει παρατηρηθεί μία αλγοριθμική εξίσωση μέσω της οποίας περιγράφεται η σχέση του dataset, σύμφωνα με την οποία, μεγάλη σημασία έχει ότι κάθε άκρη του dataset σχετίζεται με το χρόνο. Αυτό έχει ως αποτέλεσμα ένα dataset να αποτελεί ουσιαστικά ένα σύνολο ακρών, βασισμένων στο χρόνο, και αυτό μπορεί να αποτυπωθεί μέσω του παρακάτω τύπου:

$$E = \{e_{ij}(t) = (v_i, v_j,\ t) :\ t_{min} \leq t \leq t_{max}\} \qquad (1)$$

Όπου t_{min} και t_{max} η παλιότερη και πιο πρόσφατη χρονική ένδειξη αντίστοιχα.

Η ανάλυσή μας θα πρέπει να συνταχθεί μέσω μία ακολουθίας N μη γραμμικές χρονικές περιόδους $\{T_1, \dots\dots, T_N\}$, της ίδιας χρονικής διάρκειας δt, που καλύπτουν όλο το χρονικό φάσμα

$$T = [t_{min}, t_{max}] \quad (2)$$

Το σύνολο των N χρονικών περιόδων μπορεί να καθοριστεί από την ακολουθία N + 1 χρονικών στιγμών $\{t_0, \dots\dots, t_N\}$ όπου,

$$t_j = t_{min} + j * \delta t,\ 0 \leq j \leq N \qquad (3)$$

Όπου $\delta t = \frac{\Delta T}{N}$ και $\Delta T = t_{max} - t_{min}$.

Είναι εύκολο να συμπεράνουμε ότι η χρονική περίοδος j-th μπορεί να καθοριστεί ως:

$$T_j = \begin{cases} [\,t_{j-1}, t_j), & 1 \leq j \leq N-1 \\ [t_{j-1}, t_j], & j = N \end{cases} \qquad (4)$$

Για κάθε χρονική περίοδο T_j, με $1 \leq j \leq N$, μπορούμε να θεωρήσουμε τα έμμεσα υπογραφήματα του δικτύου, που συμβολίζονται ως:

$$G[t_{j-1}, t_j] = (V[t_{j-1}, t_j], E[t_{j-1}, t_j]) \qquad (5)$$

όπου

$$E[t_{j-1}, t_j] = \{e_{ij}(t) : t \in T_j\} \quad (6)$$

Το σύνολο των κορυφών V $[t_{j-1}, t_j]$ για κάθε χρονική περίοδο μπορεί να οριστεί ως ένα σύνολο κόμβων που εμφανίζονται στις άκρες των ακμών που αφορούν το σύνολο E$[t_{j-1}, t_j]$.

Μία πιο ακριβή περιγραφή για την εξέλιξη του δικτύου μεταξύ διαδοχικών χρονικών περιόδων μπορεί να θεωρηθεί ένα σύνολο κόμβων που εξακολουθούν να υπάρχουν κατά τη χρονική μετάβαση από T_j σε T_{j+1}, και μπορεί να χαρακτηριστεί από τον τύπο:

$$V^*[t_{j-1}, t_{j+1}] = V[t_{j-1}, t_j] \cap V[t_j, t_{j+1}], 1 \leq j \leq N \quad (7)$$

Σε αυτή τη σχέση μας ενδιαφέρει να περιορίσουμε τα σύνολα E$[t_{j-1}, t_j]$ και E$[t_j, t_{j+1}]$ μεταξύ των κοινών κόμβων V*$[t_{j-1}, t_{j+1}]$ όπως:

$$E^*[t_{j-1}, t_j] = \{(u, v) \in E[t_{j-1}, t_j] : u \in V^*[t_{j-1}, t_{j+1}] \wedge u \in V^*[t_{j-1}, t_{j+1}]\} \quad (8)$$

$$E^*[t_j, t_{j+1}] = \{(u, v) \in E[t_j, t_{j+1}] : u \in V^*[t_{j-1}, t_{j+1}] \wedge u \in V^*[t_{j-1}, t_{j+1}]\} \quad (9)$$

Σύμφωνα με τα παραπάνω, μπορούμε εύκολα να συμπεράνουμε ότι κατά τη διάρκεια των διαδοχικών χρονικών περιόδων T_j και T_{j+1}, το σύνολο όλων των πιθανών ακμών μεταξύ των κορυφών στους κοινούς κόμβους V*$[t_{j-1}, t_{j+1}]$, δίνεται από τον τύπο:

$$E^0[t_{j-1}, t_{j+1}] = V^*[t_{j-1}, t_{j+1}] \times V^*[t_{j-1}, t_{j+1}], 1 \leq j \leq N - 1 \quad (10)$$

Ωστόσο το υποσύνολο των ακμών που ανταποκρίνονται στο σύνολο είναι:

$$E^*[t_{j-1}, t_{j+1}] = E^*[t_{j-1}, t_j] \cup E^*[t_j, t_{j+1}] \quad (11)$$

Στην περίπτωση της πρόβλεψης δεσμών, η σχέση $E^*[t_{j-1}, t_j]$ θα μας χρησιμέψει ως ένα δοκιμαστικό σύνολο, ενώ η $E^*[t_j, t_{j+1}]$ θα χρησιμοποιηθεί για τεστ. Έτσι κάθε μία από τις προηγούμενες μετρήσεις

Sx : X ∈ {GD, CN, JC, A, PA} (12)

μπορούν να χρησιμοποιηθούν ώστε να υλοποιηθεί ένας απλός μηχανισμός κατηγοριοποίησης που μας δίνει μία εκτίμηση για το πραγματικό σύνολο ακμών $E^*[t_{j-1}, t_{j+1}]$, σύμφωνα με την παρακάτω εξίσωση:

$$\hat{E_X} * [t_{j-1}, t_{j+1}] = \{(u, v) \in E^0[t_{j-1}, t_{j+1}] : S_X(u,v) \in R_X\} \quad (13)$$

Όπου R_Xτο εύρος τιμών για τη βαθμολόγηση S_X.

Μέτρα Ταξινομητικής Ακρίβειας

Η πρόγνωση ακρίβειας για κάθε κατηγοριοποίηση μπορεί να οριστεί μέσα από το σύνολο των ακμών E και του τύπου:

$$ACC(R_X, E) = \lambda * TPR(R_X, E) + (1-\lambda) * TNR(R_X, E) \quad (14)$$

Όπου

$$TPR(R_X, E) = \frac{|\hat{E_X} * [t_{j-1}, t_{j+1}] \cap E|}{|E|} \quad (15)$$

$$TNR(R_X, E) = 1 - \frac{|\hat{E_X} * [t_{j-1}, t_{j+1}]| - |\hat{E_X} * [t_{j-1}, t_{j+1}] \cap E|}{|E^0[t_{j-1}, t_{j+1}]| - |E|} \quad (16)$$

$$\lambda = \frac{|E|}{|E^0[t_{j-1}, t_{j+1}]|} \quad (17)$$

Μέσα από τα αποτελέσματα βλέπουμε ότι η πρόγνωση δεσμών είναι ένα σύνθετο task το οποίο απαιτεί συνεχόμενα πειράματα ώστε να προσδιορίσουμε την καλύτερη προσέγγιση, ενώ η απόδοση εξαρτάται από διάφορα πράγματα, όπως το μέγεθος των γραφημάτων και από το πώς κατανέμεται.

7. Αποτελέσματα

Έπειτα από αρκετές δοκιμές, οι οποίες πραγματοποιήθηκαν με το εργαλείο weka, παρατηρήσαμε ότι σε όλες τις περιπτώσεις το ποσοστό των σωστά ταξινομημένων περιπτώσεων είναι αρκετά υψηλό, ενώ το απόλυτο σφάλμα αγγίζει το 100%.

Παρακάτω αναλύουμε τα αποτελέσματα βάση του percentage split και κλάσεων.

Όταν αυτό είναι στο 50% ο χρόνος απόκρισης είναι 10.83 sec και τα αποτελέσματα είναι τα εξής:

Correctly Classified Instances	1139394	99.8454 %
Incorrectly Classified Instances	1764	0.1546 %
Total Number of Instances	1141158	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1,000	1,000	0,998	1,000	0,999	?	0,500	0,998	-1
	0,000	0,000	?	0,000	?	?	0,500	0,002	1
Weighted Avg.	0,998	0,998	?	0,998	?	?	0,500	0,997	

=== Confusion Matrix ===

A	b	<-- classified as
1139394	0 |	a = -1
1764	0 |	b = 1

Όταν το percentage split είναι στο 60% ο χρόνος απόκρισης είναι 10.83 sec και τα αποτελέσματα είναι τα εξής:

Correctly Classified Instances	911505	99.8443 %
Incorrectly Classified Instances	1421	0.1557 %
Total Number of Instances	912926	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1,000	1,000	0,998	1,000	0,999	?	0,500	0,998	-1
	0,000	0,000	?	0,000	?	?	0,500	0,002	1
Weighted Avg.	0,998	0,998	?	0,998	?	?	0,500	0,997	

=== Confusion Matrix ===

a	b	<-- classified as
911505	0 |	a = -1
1421	0 |	b = 1

Όταν το percentage split είναι στο 65% ο χρόνος απόκρισης είναι 10.83 sec και τα αποτελέσματα είναι τα εξής:

Correctly Classified Instances	797556	99.8429 %
Incorrectly Classified Instances	1255	0.1571 %
Total Number of Instances	798811	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1,000	1,000	0,998	1,000	0,999	?	0,500	0,998	-1
	0,000	0,000	?	0,000	?	?	0,500	0,002	1
Weighted Avg.	0,998	0,998	?	0,998	?	?	0,500	0,997	

=== Confusion Matrix ===

a	b	<-- classified as
797556	0 |	a = -1
1255	0 |	b = 1

8. Σύνοψη

Στα κοινωνικά δίκτυα η πρόβλεψη των δεσμών είναι σημαντική για την εξόρυξη και την ανάλυση της εξέλιξης των κοινωνικών δικτύων γιατί προβλέπει χαμένες συνδέσεις στα υπάρχοντα δίκτυα και νέες ή διασπώμενες συνδέσεις σε μελλοντικά δίκτυα. Τα τελευταία χρόνια, έχουν πραγματοποιηθεί πολλά projects σχετικά με την πρόβλεψη δεσμών σε κοινωνικά δίκτυα. Η παρούσα εργασία αντιμετωπίζει το πρόβλημα της πρόγνωσης δεσμών με τη χρήση τεχνικών μηχανικής μάθησης και παρουσιάζει μια συστηματική κατηγορία για τις τεχνικές πρόβλεψης δεσμών, όπως και τα προβλήματα της πρόβλεψης. Επίσης, αναφέρονται διάφορες εξισώσεις σχετικές με την πρόγνωση δεσμών και η χρήση των νευρονικών δικτύων πάνω στο πρόβλημά μας. Τέλος, παρουσιάζουμε διάφορα αποτελέσματα και συμπεράσματα που προκύπτουν μέσα από την έρευνά μας.

9. Βιβλιογραφία

[1] Hasan M A, Chaoji V, Salem S, et al. Link prediction using supervised learning. In: Proceedings of SDM'06 Workshop
on Link Analysis, Counter terrorism and Security, Bethesda, USA, 2006.

[2] Lichtenwalter R N, Chawla N V. Vertex collocation profiles: subgraph counting for link analysis and prediction. In:
Proceedings of the 21st World Wide Web Conference (WWW'12), Lyon, France, 2012. 1019-1028

[3] Dunlavy D M, Kolda T G, Acar E. Temporal link prediction using matrix and tensor factorizations. ACM Transactions on Knowledge Discovery from Data, 2011, 5: 1-27

[4] O'Madadhain J, Hutchins J, Smyth P. Prediction and ranking algorithms for event-based network data. ACM SIGKDD Explorations, 2005, 7: 23-30

[5] Sarkar P, Chakrabarti D, Jordan M. Nonparametric link prediction in dynamic networks. In: Proceedings of the 29th
International Conference on Machine Learning (ICML'12), Edinburgh, Scotland, 2012

[6] Song H H, Cho T W, Dave V, et al. Scalable proximity estimation and link prediction in online social networks. In:
Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, Chicago, USA, 2009. 322-335

[7] Li X, Du N, Li H, et al. A deep learning approach to link prediction in dynamic networks. In: Proceedings of the
2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, 2014. 289-297