

Κοινωνικά Δίκτυα και Ανάκτηση Πληροφορίας



Άγκο Μπεσιάννα 1059662
Ζεκυριά Αθανασία 1059660

Διδάσκοντας: Μακρής Χρήστος

Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής
Πανεπιστήμιο Πατρών
Ελλάδα
2021

Περιεχόμενα

Εισαγωγή	3
1) Πρόβλεψη Συνδέσμων σε Ανομοιογενή Δίκτυα	5
1.1 Πρόβλεψη Συνδέσμων σε Ετερογενή και Ομογενή Δίκτυα.....	5
1.2 Metric Learning (Μετρική Μάθηση)	5
1.3 SPML	6
1.4 Περιορισμοί που συναντώνται στην πρόβλεψη συνδέσμων με SPML	7
1.5 Σύγκριση της προτεινόμενης μεθόδου.....	9
1.6 Πειράματα	10
2) Πρόβλεψη Συνδέσμων σε Ανομοιογενή Δίκτυα	11
2.1 Πρόβλεψη Μελλοντικών Σχέσεων Εμπιστοσύνης και Περιορισμοί.....	11
2.2 PIN-TRUST (προτεινόμενη μέθοδος).....	11
2.3 Διαδικασία Πειράματος	16
3) Ανάλυση Δυναμικών Δικτύων	18
3.1 Ανάλυση της δυναμικής εξέλιξης των κοινοτικών διακυμάνσεων.....	18
3.2 Τα στάδια	19
3.3 Πείραμα.....	23
4) Πρόβλεψη κοινωνικών δράσεων στα μεγάλης κλίμακας δίκτυα.....	22
4.1 Μαθησιακές αλληλεπιδράσεις για την κοινωνική πρόβλεψη σε δίκτυα μεγάλης κλίμακας	23
4.2 Τρόποι αντιμετώπισης προβλημάτων μεγάλου μέγεθους δεδομένων.....	24
4.3 Πείραμα.....	26
5) Κοινωνική Διάσταση του Διαδικτύου	29
Βιβλιογραφία:	31

Εισαγωγή:

Στην σύγχρονη εποχή του διαδικτύου, τα μέσα κοινωνικής δικτύωσης φέρουν κεντρικό ρόλο στην επικοινωνία των ανθρώπων. Ο τεράστιος όγκος πληροφοριών που μοιράζεται καθημερινά καθιστά, τη χρήση μηχανισμών ανάκτησης πληροφορίας, απαραίτητη. Η δυνατότητα αυτοματοποίησης της εξαγωγής χρήσιμων πληροφοριών σε εύλογο χρονικό διάστημα, με τη χρήση αλγορίθμων τεχνικής μάθησης, απλοποιεί την λήψη εκάστοτε αποφάσεων.

Η ανάλυση πληροφορίας στα κοινωνικά δίκτυα μπορεί να πραγματοποιηθεί είτε με απλούς τρόπους εξαγωγής πληροφορίας, είτε με αναπτυσσόμενα νευρωνικά δίκτυα [1], ικανά να ανταποκριθούν σε πολύπλοκα προβλήματα, όπως την εξόρυξη συγκεκριμένης πληροφορίας και όχι απλώς του ύφους του μηνύματος. Η ικανότητά τους να εξάγουν μεγάλο αριθμό χαρακτηριστικών και να βρίσκουν συσχετίσεις ανάμεσα στις μεταβλητές αυξάνει αισθητά τη χρήση τους στα κοινωνικά δίκτυα. Στην παρούσα εργασία, εμβαθύνουμε σε τέσσερις σειρές πειραμάτων που αποσκοπούν:

1. Στην πρόβλεψη συνδέσμων σε ετερογενή δίκτυα [2]:

Η πρόβλεψη των συνδέσμων σε ετερογενή δίκτυα αποτελεί ένα multi-tasking, metric learning (MTML) πρόβλημα. Ένα ετερογενές κοινωνικό δίκτυο χαρακτηρίζεται από πολλαπλούς τύπους συνδέσεων που καθιστούν πιο ενεργό το έργο της πρόβλεψης συνδέσεων σε τέτοια δίκτυα. Τα τελευταία χρόνια, έχουν προταθεί μέθοδοι συλλογικής πρόβλεψης των δεσμών για τη λύση του προβλήματος στα ετερογενή δίκτυα. Αυτές οι μέθοδοι αποτυπώνουν τη συσχέτιση μεταξύ διαφορετικών τύπων συνδέσεων και χρησιμοποιούν αυτές τις πληροφορίες στην πρόβλεψη συνδέσεων. Για κάθε τύπου σύνδεσης, μαθαίνουμε ένα αντίστοιχο μέτρο απόστασης, το οποίο χρησιμοποιεί χαρακτηριστικά δικτύου και κόμβων. Αυτά τα ειδικά μέτρα απόστασης τύπου μαθαίνονται με «ζευγαρωτό» τρόπο με τη χρήση της ρύθμισης Multi-Task Structure Preserving Metric Learning (MT-SPML). Η παραπάνω μέθοδος επεκτείνεται ώστε να λαμβάνονται υπόψη οι συσχετίσεις των εργασιών, η ανθεκτικότητα σε μη ενημερωμένα στοιχεία και η διανομή μη στατικών πτυχίων μεταξύ δικτύων. Τα πειράματα στο δίκτυο Flickr και DBLP καταδεικνύουν την αποτελεσματικότητα της προτεινόμενης προσέγγισής έναντι των ανταγωνιστικών βασικών γραμμών.

2. Στην πρόβλεψη εμπιστοσύνης μεταξύ των χρηστών [3]:

Με την πλέον συχνή εμφάνιση απατεώνων σε ιστοσελίδες, η σημασία της πρόβλεψης εμπιστοσύνης έχει αυξηθεί. Οι περισσότερες μέθοδοι που χρησιμοποιούνται, αξιοποιούν μόνο ρητές και έμμεσες πληροφορίες εμπιστοσύνης, αλλά δεν λαμβάνουν υπόψη τη δυσπιστία. Για τον λόγο αυτό, προτείνεται το PIN-TRUST, μια ρηξικέλευθη μέθοδος για τη διαχείριση και των τριών τύπων πληροφοριών αλληλεπίδρασης: την ρητή εμπιστοσύνη, την

έμμεση εμπιστοσύνη και τη ρητή δυσπιστία. Έχει σχεδιαστεί προσεκτικά ώστε να λαμβάνει υπόψη θετικές, έμμεσες και αρνητικές πληροφορίες. Είναι γραμμική ως προς το μέγεθος των εισροών και κυρίως, αποτελεσματική και ακριβής. Τα εκτενή πειράματα με πραγματικό σύνολο δεδομένων (Epinions.com) από κόμβους 100 χιλιάδων και άκρα 1 εκατομμυρίου, επιβεβαιώνουν ότι το PIN-TRUST είναι κλιμακούμενο και ξεπερνά τις υπάρχουσες μεθόδους όσον αφορά την ακρίβεια πρόβλεψης.

3. Την **προσωρινή ανάλυση των δυναμικών δικτύων** [4]:

Η ανάλυση των δυναμικών δικτύων είναι ένα δημοφιλές ζήτημα σήμερα λόγω της μεγάλης διαθεσιμότητας δεδομένων, που αναδύονται συνεχώς, και σχετίζεται με την ανακάλυψη γνώσεων μέσω της μελέτης της συνεχούς εξέλιξης των δεδομένων. Το πρόβλημα της μελέτης του ποσοστού ποικιλίας και της εξέλιξης των κοινοτήτων με την πάροδο του χρόνου οδήγησε στη δημιουργία του αλγορίθμου VACS. Η δομή του οποίου χωρίζεται σε δύο φάσεις. Η πρώτη περιλαμβάνει τον εντοπισμό της εξελικτικής ομαδοποίησης των κοινοτήτων και της αντιστοίχισης μεταξύ στιγμιότυπων και η δεύτερη το προτεινόμενο μοντέλο βελτιστοποίησης και το ποσοστό διακύμανσης για κάθε κοινότητα. Τέλος, μελέτες συνόλων δεδομένων όπως DBLP και IMDB έδειξαν ότι η μέθοδος αυτή έχει μεγάλη αποτελεσματικότητα στην ανίχνευση της ποικιλίας των κοινοτήτων.

4. Στην **πρόβλεψη κοινωνικών δράσεων στα μεγάλης κλίμακας δίκτυα** [5]:

Τα κοινωνικά δίκτυα παρέχουν σε όλους τους χρήστες τη δυνατότητα κοινωνικής δικτύωσης και διάδοσης πληροφοριών. Για την εκμετάλλευση των κοινωνικών ενεργειών και την αξιοποίηση των αλληλεπιδράσεων ανάμεσα στους χρήστες, καθώς και την ανακάλυψη των κοινωνικών τους δεσμών, δημιουργήθηκε μια δομή γραφημάτων (MLRGs) που παρουσιάζουν κρυμμένους παράγοντες. Επίσης, προτάθηκε ένας, κατά προσέγγιση, αλγόριθμος ο οποίος διερευνά τους συντελεστές των γραφημάτων για την βελτιστοποίηση των παραμέτρων του μοντέλου. Ο αλγόριθμος αυτός αποδείχθηκε ότι ήταν γρήγορος μετά από τη χρήση της αρχιτεκτονικής Hadoop MapShine αλλά και αρκετά έγκυρος έπειτα από πειράματα με σύνολα δεδομένων κοινωνικής δικτύωσης όπως Epinion+ και Mobile.

5. Ταυτόχρονα, ασχοληθήκαμε με μία πέμπτη μελέτη που εμβαθύνει στην **κοινωνική διάσταση του διαδικτύου** [6]. Αυτή η κοινωνική διάσταση είναι κάθε κοινωνική πληροφορία που περιβάλλει τα έγγραφα, αλλά και το κοινωνικό πλαίσιο κάθε χρήστη. Η μέθοδος που προτείνεται, επεκτείνει την δημιουργία μοντέλων IR και αποτελεί μια νέα λειτουργία κατάταξης που ονομάζεται SoPRa. Βασίζεται σε ταυτότητες για την εξαγωγή των κοινωνικών πλαισίων, όμως μπορεί να χρησιμοποιηθεί σε οποιαδήποτε κοινωνικά στοιχεία, όπως σχόλια, αξιολογήσεις, αναρτήσεις. Τα αποτελέσματα της

πειραματικής αξιολόγησης της τεχνικής αυτής παρουσιάζει τα οφέλη της εξατομικευμένης αναζήτησης.

1)Πρόβλεψη Συνδέσμων σε Ανομοιογενή Δίκτυα

1.1 Πρόβλεψη Σύνδεσης σε Ετερογενή και Ομογενή Δίκτυα

Η πρόβλεψη συνδέσεων στα ομογενή δίκτυα περιλαμβάνει:

1)μη εποπτευόμενες προσεγγίσεις:

- α)τοπικοί δείκτες που βασίζονται σε γείτονες και
- β)ο παγκόσμιος προγνωστικός που βασίζεται σε διαδρομές δικτύου

2)εποπτευόμενες προσεγγίσεις:

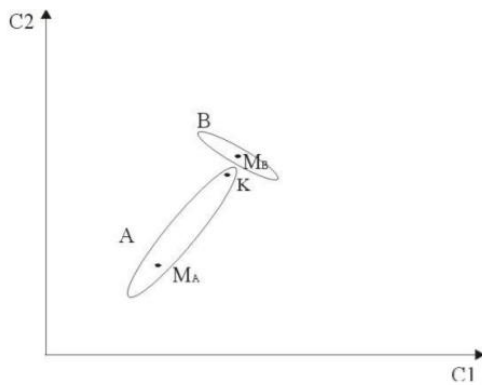
- α)μέθοδοι που βασίζονται στις δυνατότητες ταξινόμησης,
- β)το πιθανοτικό μοντέλο γραφήματος και
- γ)μέθοδοι που βασίζονται σε παραγοντική μήτρα.

Η πρόβλεψη συνδέσεων στα ετερογενή δίκτυα περιλαμβάνει:

- α)μια νέα επέκταση του μέτρου Adamic/Adar [7],
- β)ένα εποπτευόμενο πλαίσιο μάθησης που μπορεί αποτελεσματικά να μάθει τη δυναμική των κοινωνικών δικτύων παρουσία βοηθητικών δικτύων [8],
- γ)ένα σχέδιο χαρακτηριστικών για την κατασκευή μια πλούσιας ποικιλίας λειτουργιών που βασίζονται σε διαδρομές χρησιμοποιώντας πολλαπλές πηγές και μια αποτελεσματική στρατηγική επιλογής χαρακτηριστικών βασισμένη στην δομημένη αραιότητα (structured sparsity) [8],
- δ)χαρακτηριστικά μέτρα απόστασης, τα οποία χρησιμοποιούν δυνατότητες δικτύου και κόμβου [9]και
- γ)ο Multi-Relational Influence Propagation(MRIP).

1.2 Metric Learning (Μετρική Μάθηση)

Η μετρική μάθηση έχει αποδειχθεί αποτελεσματική για την ταξινόμηση, την δημιουργία συμπλέγματος και την κατάταξη. Οι αλγόριθμοι εκμάθησης μετρικών συνήθως λειτουργούν βελτιστοποιώντας την απόσταση από τον στόχο υπό διάφορους περιορισμούς. Μια κλάση συναρτήσεων απόστασης με αποτελεσματικές ιδιότητες γενίκευσης είναι η απόσταση Mahalanobis.



Σχήμα 1.2.1 Δύο κλάσεις *pixels* A και B με μέσους ραδιομετρικούς όρους M_A και M_B . Ενώ το *pixel* K λόγω της μικρότερης απόστασής του από το κέντρο M_B θα ταξινομείτο στην κλάση B, με την χρήση της απόστασης Mahalanobis ταξινομείται (σωστά) στην κλάση A. [10]

Δυστυχώς, οι περισσότεροι αλγόριθμοι μετρικής μάθησης δεν εφαρμόζονται με ευκολία όταν η είσοδος είναι δίκτυο αντί για σημεία με ετικέτες κλάσεων. Ο περιορισμός αυτός προσπερνάται με τη Διατήρηση της Μετρικής Εκμάθησης με δομή SPML – αλγόριθμο για την εκμάθηση μετρικής απόστασης Mahalanobis από ένα δίκτυο, ώστε οι γνωστές αποστάσεις να συνδέονται με την εγγενή δομή συνδεσιμότητας του δικτύου. Η SPML μαθαίνει ένα μετρικό που είναι η διατήρηση της δομής: ένας αλγόριθμος συνδεσιμότητας όπως οι πλησιέστεροι γείτονες που αποδίδει τη σωστή συνδεσιμότητα όταν εφαρμόζεται, χρησιμοποιώντας τις αποστάσεις από το διδαγμένο μετρικό. Με αποτέλεσμα να προβλέπει πρότυπα συνδέσμων από τα χαρακτηριστικά των κόμβων με μεγαλύτερη ακρίβεια από τις τυπικές τεχνικές. [11]

1.3 SPML

Όταν ετερογενές δίκτυο αποσυντίθεται σε μεμονωμένα δίκτυα αλληλεπίδρασης, καθένα από τα επιμέρους δίκτυα λαμβάνει ένα αντίστοιχο μέτρο απόστασης τύπου συνδέσμου. Τα δεδομένα δικτύου περιέχουν δομικές πληροφορίες που κωδικοποιούνται κυρίως σε συνδέσμους. Για να προσδιοριστεί ένα κατάλληλο μέτρο απόστασης για αυτά τα δίκτυα, αξιοποιούνται κατά το στάδιο εκμάθησης μετρικών τα χαρακτηριστικά αλλά και η δομή δικτύου με την βοήθεια του προτεινόμενου αλγόριθμου SPML. Όπως προαναφέρθηκε, η SPML μαθαίνει ένα μετρικό απόστασης Mahalanobis για χαρακτηριστικά κόμβων, χρησιμοποιώντας τη δομή δικτύου ως επίβλεψη. Ο στόχος της SPML είναι να μάθει το μετρικό M της Mahalanobis από ένα δίκτυο $G = (V, E)$, ώστε το M να διατηρήσει δομή γειτονίας.

Ουσιαστικά, η μέθοδος SPML, δεδομένου ενός αλγορίθμου συνδεσιμότητας, μαθαίνει ένα μετρικό, το εφαρμόζει στα δεδομένα εισόδου, χρησιμοποιώντας το μαθημένο μετρικό M και παράγει τον πίνακα γειτονικής εισόδου.

Ο αλγόριθμος SPML επεκτείνεται στις πολλαπλές εργασίες (Multi-Tasking SPML). Η επέκταση αυτή μαθαίνει από πολλαπλές σχετικές εργασίες σε πολλαπλά δίκτυα μέσω μιας κοινόχρηστης ενδιάμεσης παραμέτρου. Η είσοδος στο MT-SPML είναι ένα σύνολο δικτύων $G = \{G^1, \dots, G^Q\}$. Κάθε δίκτυο απεικονίζεται ως $G^q = \{V^q, E^q\}$ όπου οι κόμβοι όλων των δικτύων μοιράζονται τον

ίδιο χώρο δυνατοτήτων. Η MT-SPML αντιμετωπίζει κάθε δίκτυο ως εργασία και μαθαίνει ένα συγκεκριμένο αντίστοιχο μετρικό. Η συσχέτιση εργασιών πραγματοποιείται μέσω του κοινού μετρικού και τα επιμέρους μετρικά κωδικοποιούν ορισμένες πληροφορίες εργασιών.

Ο στόχος της MT-SPML μπορεί να συμπυκνωθεί στην εκμάθηση ενός μετρικού για κάθε εργασία και για κάθε δίκτυο.

1.4 Περιορισμοί που συναντώνται στην πρόβλεψη συνδέσμων με SPML

1. Όταν τα δεδομένα εισόδου περιέχουν μεγάλο μέρος μη ενημερωτικών χαρακτηριστικών, οι γνωστές μέθοδοι metric learning, συμπεριλαμβανομένων και των MT-SPML και SPML, αποτυγχάνουν να προσδιορίσουν τα σχετικά χαρακτηριστικά. Έχει προταθεί η ιδέα του **Robust Metric Learning** [12] που επιβάλλει την αδυναμία της ομάδας στον μαθημένο μετασχηματισμό, βελτιστοποιώντας παράλληλα τη δομημένη πρόβλεψη εξόδου κατάταξης. Αυτό επεκτείνεται στο MT-SPML, εφαρμόζοντας μια μικτή κανονικοποίηση πάνω στο μαθημένο μετρικό σύστημα.

Πιο συγκεκριμένα, εάν κάποια διάσταση εισόδου j δεν είναι ενημερωτική, η αντίστοιχη σειρά και στήλη M πρέπει να καταστείλει αυτό το χαρακτηριστικό ($M_j = M_{\cdot j} = 0$). Αντιθέτως, αυτό δεν πρέπει να εφαρμόζεται σε πληροφοριακά χαρακτηριστικά. Έτσι, γίνεται χρήση μιας σειράς (ή στήλης) καταχωρήσεων του M , ώστε οι γραμμές που αντιστοιχούν σε πληροφοριακά χαρακτηριστικά να είναι γεμάτες και να μην επιδέχονται καταστολή. Στο πλαίσιο αυτής της μικτής κανονιστικής ρύθμισης $L_{2,1}$, το πρόβλημα της SPML μπορεί να ξαναγραφτεί ως

$$\min_M \frac{\lambda}{2} \|M\|_F^2 + \beta \|M\|_{2,1} + \frac{1}{|S|} \sum_{i,j,k \in S} \max(D_M(x_i, x_j) - D_M(x_i, x_k) + 1, 0) \quad (1)$$

Οι περιορισμοί αναγράφονται με βάση τις απώλειες έναντι των (i, j, k) , καθένα εκ των οποίων αποτελείται από έναν κόμβο, τον γείτονα και τον μη γείτονα του. Το σύνολο όλων αυτών υποδηλώνεται με $S = \{(i, j, k) \mid A_{ij} = 0\}$.

2. Το γεγονός ότι τα μεμονωμένα μέτρα απόστασης τύπου συνδέσμου μαθαίνονται με ζευγαρωμένο τρόπο καθιστά το πρόβλημα ως πρόβλημα πολλαπλής μετρικής μάθησης. Συγκεκριμένα, ορισμένες από τις εργασίες ενδέχεται να μην είναι σχετικές, με αποτέλεσμα, η παραδοχή ότι όλες οι εργασίες σχετίζονται να επηρεάσει την εκτέλεση άλλων εργασιών.

Προτείνεται αλγόριθμος που μαθαίνει ένα μετρικό ομοιότητας και ένα σύνολο λειτουργιών βαθμολογίας που παρέχουν από κοινού μια μέθοδο βάσει δομής, βάσει απόστασης με σκοπό την πρόβλεψη συνδέσεων.

Ο αλγόριθμος συνδεσιμότητας (**Degree Distributional Metric Learning DDML**) χρησιμοποιεί μια συνάρτηση προτιμήσεων g που λαμβάνει ένα διανυσματικό

χαρακτηριστικό ενός κόμβου x και έναν βαθμό στόχου k και παρατηρείται από τον πίνακα $B \in \mathbb{R}^{d \times n}$ (b_k είναι η k' σειρά του B). Η βαθμολογία υπολογίζεται ως

$$g(k | x; B) = \sum_{k'=1}^k x^T b_{k'}$$

Έπειτα, η βαθμολογία ενός γραφήματος G ορίζεται ως άθροισμα όλων των αποστάσεων άκρων και των συναρτήσεων προτιμήσεων κάθε κόμβου. Αυτό φαίνεται ακολούθως (A είναι ο πίνακας γειτονίας για το γράφημα G):

$$F(A | X; M, B) = \sum_{A_{ij}} D_M(x_i, x_j) - \sum_i g\left(\sum_j A_{ij} | x_i; B\right)$$

$$\begin{aligned} & \min_{M, B} \frac{\lambda}{2} \|M\|_F^2 + \beta \|M\|_{2,1} \\ & + \frac{1}{|S|} \sum_{i,j,k} \max(F(A | X; M, B) - F(A^{(i,j,k)} | X; M, B) + 1, 0) \end{aligned} \quad (2)$$

Το $A^{(i,j,k)}$ είναι το ψευδές γράφημα που παράγεται με την εναλλαγή των κόμβων i και j και του άκρου μεταξύ των κόμβων i, k . Παραπάνω είναι η στοχαστική έκδοση του DDML όπου η συνάρτηση προτιμήσεων γίνεται με παραμέτρους μόνο μέχρι τον καθορισμένο μέγιστο βαθμό, εξαλείφοντας με αυτόν τον τρόπο, την εξάρτηση του χρόνου εκτέλεσης από το μέγεθος του γραφήματος.

3. Η κατανομή των βαθμών σε ένα μεμονωμένο δίκτυο αλληλεπίδρασης ενδέχεται να διαφέρει από το ένα δίκτυο στο άλλο. Προτείνεται μία μέθοδος που ενσωματώνει την ιδέα της **Κατανομής Βαθμών Μετρικής Μάθησης** στη συνολική ρύθμιση της πολλαπλής μάθησης, κατανέμοντας τους βαθμούς σύμφωνα με τα χαρακτηριστικά των κόμβων.

Για τη ρύθμιση αυτή, χρησιμοποιείται διαφορετική παράμετρος από αυτή που χρησιμοποιείται στην MT-SPML. Χρησιμοποιείται ένας πίνακας διακύμανσης εργασιών Σ για να γίνουν γνωστές οι σχέσεις μεταξύ εργασιών. Παρατίθεται η εξίσωση 2 βάσει της φόρμας πολλαπλών εργασιών:

$$\begin{aligned}
& \min_{M_1, M_2, \dots, M_Q, B_1, B_2, \dots, B_Q, \Sigma} \sum_{q=1}^Q \frac{\lambda_q}{2} \|M_q\|_F^2 + \sum_{q=1}^Q \beta_q \|M_q\|_{2,1} \\
& \quad + \frac{\lambda}{2} \text{tr}(\mathbf{T}\Sigma^{-1}\mathbf{T}^T) + \sum_{q=1}^Q \\
& \frac{1}{|S_q|} \sum_{i,j,k \in S_q} \max(F(A_q | X_q; M_q, B_q) - F(A_q^{(i,j,k)} | X_q; M_q, B_q) + 1, 0) \\
& \quad s.t. M_i \succeq 0 \forall i = 1 \dots Q, \Sigma \succeq 0, \text{tr}(\Sigma) = 1 \\
& \quad T = (\text{vec}(M_1), \dots, \text{vec}(M_Q))
\end{aligned} \tag{3}$$

Το $\text{vec}()$ δηλώνει τον τελεστή που μετατρέπει έναν πίνακα σε διάνυσμα. Η χρήση αυτής της διατύπωσης στην εγκατάστασή μας δίνει ένα βασικό πλεονέκτημα: η εξίσωση 3 αποσυντίθεται σε προβλήματα μεταφοράς Q τα οποία έχουν την ακόλουθη δομή:

$$\begin{aligned}
& \min_{M_q, B_q, \Sigma} \frac{\lambda_q}{2} \|M_q\|_F^2 + \beta_q \|M_q\|_{2,1} + \frac{\lambda}{2} \text{tr}(\mathbf{T}\Sigma^{-1}\mathbf{T}^T) \\
& \quad + \frac{1}{|S_q|} \sum_{i,j,k \in S_q} \max(F(A_q | X_q; M_q, B_q) - F(A_q^{(i,j,k)} | X_q; M_q, B_q) + 1, 0) \\
& \quad s.t. M_q \succeq 0, \Sigma \succeq 0, \text{tr}(\Sigma) = 1 \\
& \quad T = (\text{vec}(M_1), \dots, \text{vec}(M_Q))
\end{aligned} \tag{4}$$

1.5 Σύγκριση της προτεινόμενης μεθόδου με τις ήδη υπάρχουσες

SPML:

- ST-SPML**: ανταποκρίνεται σε μεμονωμένα δίκτυα αλλά όχι συσχετίσεις τύπου συνδέσμου.
- Pooled-ST-SPML**: αποτελεί τρόπο ανταλλαγής γνώσεων μεταξύ εργασιών αλλά δεν ασχολείται με τις διαφορές μεταξύ εργασιών.
- MT-SPML**: πρόκειται για την παραλλαγή της ST-SPML για πολλαπλές εργασίες και δίκτυα.

SVM προσεγγίσεις: υιοθετώντας την προσέγγιση που προτείνεται στο [13]

κατασκευάζονται παραδείγματα κατάρτισης λαμβάνοντας υπόψη την κατά ζεύγη διαφορά των χαρακτηριστικών μεταξύ δύο κόμβων. Οι ετικέτες είναι δυαδικές: το 1 αντιπροσωπεύει την ύπαρξη ενός συνδέσμου μεταξύ ενός ζεύγους κόμβων και το 0 την απουσία. Η βαθμολογία ταξινόμησης χρησιμοποιείται για την μέτρηση της απόστασης.

- ST-SVM**: ένα LIBLINEAR SVM για κάθε δίκτυο ανεξάρτητα.
- Pooled-SVM**: ένα SVM για όλα τα δίκτυα ομαδοποιώντας όλα τα δεδομένα μαζί.
- MT-SVM**: μαθαίνει από κοινού ένα κοινό όριο απόφασης για όλες τις εργασίες και ένα συγκεκριμένο όριο για κάθε εργασία ξεχωριστά, τα οποία, κατά τη διάρκεια της δοκιμής, αποτελούν το τελικό μοντέλο ταξινόμησης για κάθε εργασία.

Μέθοδοι πρόβλεψης ετερογενών συνδέσμων: το σύστημα συγκρίνεται με τον αλγόριθμο Πολλαπλασιασμού Πολλαπλών Σχισιακών Επιρροών (MRIP), ο οποίος παρουσιάζει την επίδραση που αναπαράγεται μεταξύ των ετερογενών σχέσεων.

Προτεινόμενη προσέγγιση: συγκρίνονται οι επιδόσεις της προτεινόμενης μεθόδου με τις προαναφερθείσες μεθόδους:

Link-Type	ST-SVM	Pooled-SVM	MT-SVM	SPML	Pooled-SPML	MT-SPML	MRIP	Combined Approach
user-user	0.36	0.31	0.39	0.43	0.44	0.61	0.54	0.77
user-flickr_group	0.29	0.31	0.32	0.48	0.55	0.59	0.61	0.68
flickr_group-image	0.33	0.36	0.38	0.52	0.59	0.61	0.64	0.72

Link-Type	ST-SVM	Pooled-SVM	MT-SVM	SPML	Pooled-SPML	MT-SPML	MRIP	Combined Approach
author-author	0.47	0.40	0.49	0.59	0.61	0.63	0.71	0.79
document-conference	0.49	0.51	0.52	0.59	0.57	0.60	0.64	0.76
document-document	0.50	0.54	0.58	0.62	0.64	0.65	0.71	0.78

Πίνακας 1.5.1 AUC για διάφορους τύπους συνδέσμων - Flickr (πάνω πίνακας), DBLP (κάτω πίνακας)

Παρέχονται επίσης αποτελέσματα σχετικά με τις επιδόσεις διάφορων παραλλαγών, όπως προαναφέρθηκαν, της προτεινόμενης μεθόδου στα σύνολα δεδομένων δοκιμών:

-**Robust MT-SPML:** σύνθεση MT-SPML με μικτή κανονικοποίηση

-**Robust-Degree_{perf}-MT-SPML:** λειτουργίες προτίμησης

-**Robust-Covariance-MT-SPML:** Robust MT-SPML με μετρική εκμάθηση πολλαπλών εργασιών με καθορισμό παραμέτρων συνδιασμού

-**Συνδυασμένη Προσέγγιση:** συνδυασμός των παραπάνω προσεγγίσεων

Link-Type	Robust-MT-SPML	Robust-Degree _{perf} -MT-SPML	Robust-Covariance-MT-SPML	Combined Approach
user-user	0.67	0.73	0.69	0.77
user-flickr_group	0.61	0.64	0.62	0.68
flickr_group-image	0.67	0.69	0.66	0.72

Link-Type	Robust-MT-SPML	Robust-Degree _{perf} -MT-SPML	Robust-Covariance-MT-SPML	Combined Approach
author-author	0.64	0.72	0.70	0.79
document-conference	0.62	0.73	0.68	0.76
document-document	0.66	0.74	0.69	0.78

Πίνακας 1.5.2 AUC για διάφορους τύπους συνδέσμων - Flickr (πάνω πίνακας), DBLP (κάτω πίνακας)

Όπως φαίνεται από τους πίνακες, η προτεινόμενη μέθοδος υπερβαίνει κάθε βάση. Κατά μέσο όρο η μέθοδος δίνει βελτίωση AUC κατά 9,6% σε σχέση με τη μέθοδο αναφοράς με τις καλύτερες επιδόσεις.

1.6 Πειράματα

Χρησιμοποιήθηκαν δύο σύνολα δεδομένων (Flickr, DBLP) για την συγκριτική αξιολόγηση της απόδοσης του συστήματος.

FLICKR

Το Flickr αποτελεί κοινωνικό δίκτυο και περιέχει τις ακόλουθες 3 οντότητες: χρήστης, ομάδα, εικόνα (user, flickr-group, image). Το πείραμα αποσκοπεί στη πρόβλεψη των σχέσεων: χρήστης-χρήστης, χρήστης-ομάδα και ομάδα-εικόνα. Τα δεδομένα ελήφθησαν με την υποβολή ερωτήματος στο δημόσιο API του Flickr στο χρονικό διάστημα 2013-2014 και στη συνέχεια, ανιχνεύτηκαν 100 ομάδες με λιγότερες από 20 εικόνες η κάθε μία, επιλέγοντας 10 δημοφιλείς ετικέτες και επιλέγοντας τις 10 πρώτες ομάδες για κάθε ετικέτα. Έπειτα, συλλέχθηκαν πληροφορίες 10.058 χρηστών αυτών των ομάδων.

Για τα πειράματα d=6843: το Flickr περιέχει πληροφορίες χρονικής σήμανσης. Η

ακρίβεια της πρόβλεψης συνδέσμων υπολογίζεται βάσει της περιοχής υπό της καμπύλης ROC (AUC) σε ένα σύνολο προσημασμένων παραδειγμάτων που προκύπτουν από το σύνολο δοκιμών. Όλοι οι σύνδεσμοι αντιμετωπίζονται ως θετικά παραδείγματα, ενώ οι μη υφιστάμενοι σύνδεσμοι ως αρνητικά. Τίθενται 2 θέματα: ο αριθμός των μη υφιστάμενων συνδέσμων ενδέχεται να είναι τεράστιος και οι ελλείπουσες συνδέσεις δεν αντιπροσωπεύουν απαραίτητα αρνητικές πληροφορίες. Έτσι, η απόδοση πρόβλεψης συνδέσμου αξιολογείται ξεχωριστά για κάθε τύπο συνδέσμου με χρήση AUROC.

Η προτεινόμενη μέθοδος δίνει μια μέση βελτίωση της AUC κατά 10,3% πάνω στην πιο ανταγωνιστική βάση.

DBLP

Το DBLP [8] αποτελεί ετερογενές δίκτυο και περιέχει τις 3 οντότητες: έγγραφα, συγγραφείς, συνέδρια (documents, authors, conferences). Το σύνολο των δεδομένων περιέχει 28,569 έγγραφα, 702 συγγραφείς και 20 συνέδρια. Το πείραμα αποσκοπεί στην πρόβλεψη των σχέσεων: συντάκτης-συντάκτης, έγγραφο-έγγραφο και έγγραφο-συνέδριο.

Τα δεδομένα επιλέχθηκαν μεταξύ 1990-2000 ως σύνολο κατάρτισης και τα δεδομένα μεταξύ 2001-2005 ως σύνολο δοκιμής. Ακόμα και αν το δίκτυο περιέχει και άλλους τύπου συνδέσεων, τους αγνοούμε. Η βασική μετεπεξεργασία, όπως η κατάργηση λέξεων διακοπής, το φιλτράρισμα tf-idf, εκτελείται με όρους.

Για τα πειράματα $d=7389$: ο πίνακας 2 παρέχει τους αριθμούς AUC για τις διάφορες μεθόδους και τύπους συνδέσεων.

Η προτεινόμενη μέθοδος δίνει μια μέση βελτίωση της AUC κατά 9% πάνω από την πιο ανταγωνιστική βάση.

2) Πρόβλεψη Συνδέσμων σε Ανομοιογενή Δίκτυα

2.1 Πρόβλεψη μελλοντικών σχέσεων εμπιστοσύνης και περιορισμοί

Γίνεται πλέον αντιληπτή και η ανάγκη εύρεσης αξιόπιστων χρηστών στα μεγάλα κοινωνικά δίκτυα, για τον λόγο αυτό, έχουν προταθεί μέθοδοι πρόβλεψης εμπιστοσύνης που προβλέπουν μελλοντικές σχέσεις εμπιστοσύνης χρηστών. Οι μέθοδοι αυτοί χρησιμοποιούν πληροφορίες αλληλεπίδρασης χρηστών: ρητή σχέση εμπιστοσύνης, ρητή σχέση δυσπιστίας και αξιολόγηση. Ωστόσο, οι υπάρχουσες μέθοδοι αντιμετωπίζουν τα εξής προβλήματα:

1. Μη καθολική εξέταση πληροφοριών αλληλεπίδρασης.
2. Σε περίπτωση που ο χρήστης δεν έχει επισυνάψει ήδη σχέσεις, υπάρχει χαμηλή ακρίβεια πρόβλεψης για τον χρήστη αυτό.
3. Δεν λαμβάνεται υπόψη η αμοιβαιότητα εμπιστοσύνης.

2.2 PIN-TRUST (προτεινόμενη μέθοδος)

Η PIN-TRUST αποτελεί μια νέα μέθοδο εμπιστοσύνης που δημιουργεί πρώτα ένα δίκτυο εμπιστοσύνης που έχει δύο τύπους κατευθυνόμενων άκρων για να εξετάσει

κάθε είδους πληροφορίες αλληλεπίδρασης. Το θετικό άκρο δημιουργείται όταν ο χρήστης δημιουργεί θετική σχέση με άλλο χρήστη (αξιοπιστία, θετική αξιολόγηση, ενώ το αρνητικό όταν ο χρήστης δημιουργεί μια σχέση δυσπιστίας.

Η μέθοδος PIN-TRUST χρησιμοποιεί την έννοια της διάδοσης πεποιθήσεων (BP) για την εκμετάλλευση των δικτυακών αποτελεσμάτων. Για τη χρήση της BP απαιτείται η **μοντελοποίηση ενός δικτύου εμπιστοσύνης(1)**, ο **καθορισμός πινάκων διάδοσης(2)** και η τελική **απόφαση προηγούμενων πεποιθήσεων(3)**.

Μοντέλο PIN: παρουσιάζει τις συμπεριφορές των χρηστών σε ένα κοινωνικό δίκτυο για να εξετάσει όλους τους τύπους πληροφοριών αλληλεπίδρασης.

Δυνατότητα Κλιμάκωσης: η χρονική πολυπλοκότητα PIN-TRUST είναι γραμμική.

Αποτελεσματικότητα: η PIN-TRUST υπερβαίνει τις υπάρχουσες μεθόδους με σχετική βελτίωση έως 19,7 και 50,4% στην ακρίβεια πρόβλεψης σε σύγκριση με την ITD και την ABIT L.

	PIN-TRUST	ITD	ABIT_L
Trust relationship and rating	✓	✓	✓
Distrust relationship	✓	-	-
Trust reciprocation	✓	-	-
Scalability	✓	✓	✓
Accuracy (in worst case)	61.4%	51.3%	40.8%

Πίνακας 2.2.1 συγκρίσεις ήδη υπάρχουσών μεθόδων και της αξιοπιστίας PIN

Πιο συγκεκριμένα:

>>Κατασκευάζεται ένα δίκτυο εμπιστοσύνης με κόμβους τους χρήστες και κατευθυνόμενα άκρα τις αλληλεπιδράσεις. Όπως προαναφέρθηκε, οι κατευθύνσεις των άκρων αντιπροσωπεύουν την φύση της αλληλεπίδρασης. Οι σχέσεις εμπιστοσύνης-δυσπιστίας ταξινομούνται ως ρητές πληροφορίες, δεδομένου ότι δηλώνονται ρητά από το χρήστη, ενώ η αξιολόγηση αποτελεί έκφραση σιωπηρής εμπιστοσύνης. Αν και οι αξιολογήσεις υπάγονται στα θετικά άκρα, ο βαθμός εμπιστοσύνης εξαρτάται από την ποιότητα αλλά και την ποσότητα αυτών.

>>Κατασκευάζεται ένας πίνακας πολλαπλασιασμού για τα θετικά και ένας για τα αρνητικά άκρα. Οι πίνακες αυτοί βασίζονται στην έννοια της “homophily influence”(ανθρωποφιλικής ροής) [14]που αναφέρεται στην τάση ενός κόμβου να συνδέεται με παρόμοιους κόμβους στο δίκτυο του.

(a) Positive edge			
		Destination	
		Trust.	Untrust.
Source	Trust.	$0.5 + \varepsilon \times (t + f(r))$	$0.5 - \varepsilon \times (t + f(r))$
	Untrust.	$0.5 - \varepsilon \times (t + f(r))$	$0.5 + \varepsilon \times (t + f(r))$

(b) Negative edge			
		Destination	
		Trust.	Untrust.
Source	Trust.	$0.5 - \varepsilon$	$0.5 + \varepsilon$
	Untrust.	$0.5 + \varepsilon$	$0.5 - \varepsilon$

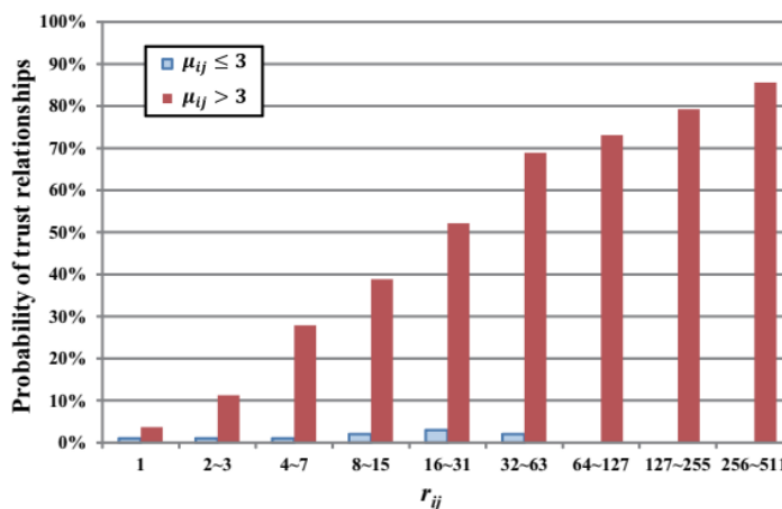
Πίνακας 2.2.2 ενδείξεις των πινάκων πολλαπλασιασμού στα θετικά και αρνητικά άκρα

Το ε ελέγχει την επίδραση της πηγής στον προορισμό.

Το t είναι μια δυαδική παράμετρος που αναπαριστά την ύπαρξη εμπιστοσύνης: $t=1$ για σχέση αξιοπιστίας, $t=0$ για σχέση δυσπιστίας.

Το $f(r)$ είναι ο βαθμός εμπιστοσύνης βάσει του αριθμού βαθμολογιών και τη μέση βαθμολογία.

Στο σύνολο δεδομένων, οι αξιολογήσεις παρέχονται ως ακέραιος αριθμός 1 έως 5.



Σχήμα 2.2.1 ο άξονας x είναι ο αριθμός διαβαθμίσεων ενός ζεύγους χρηστών, ενώ ο άξονας y η πιθανότητα το ζεύγος να έχει σχέση αξιοπιστίας

Παρατηρείται ότι οι σχέσεις εμπιστοσύνης σπάνια εμφανίζονται σε όλα τα διαστήματα των κλειδιών, βάσει αυτής της παρατήρησης υπολογίζεται το $f(r)$:

r_{ij}	$f(r)$	
	$\mu_{ij} \leq 3$	$\mu_{ij} > 3$
1	0.01	0.05
2-3	0.01	0.10
4-7	0.01	0.30
8-15	0.02	0.40
16-31	0.03	0.55
32-63	0.02	0.70
64-127	0	0.75
127-255	0	0.80
256-511	0	0.85

Πίνακας 2.2.3 εκτίμηση του βαθμού αξιοπιστίας με τον αριθμό αξιολογήσεων και τη μέση βαθμολογία αξιολόγησης μεταξύ δύο χρηστών.

>>Εξετάζονται δύο πτυχές για τον καθορισμό των περασμένων πεποιθήσεων για τους κόμβους: (1)η άποψη του διαχειριστή και (2)η επιλογή της πλειοψηφίας. Οι προηγούμενες πεποιθήσεις στον κόμβο-στόχο ορίζονται ως $\langle 0,5+\alpha, 0,5-\alpha \rangle$ για αξιόπιστες και αναξιόπιστες καταστάσεις, αντιστοίχως. Προκειμένου να δοθεί υψηλή αξιοπιστία σε έναν κόμβο που πολλοί κόμβοι εμπιστεύονται ευρέως, ορίζουμε την προηγούμενη εμπιστοσύνη για κάθε κόμβο, εκτός του κόμβου-στόχου, σε αξιόπιστη κατάσταση υψηλότερη από εκείνη σε μη αξιόπιστη κατάσταση. Συγκεκριμένα, οι προηγούμενες καταστάσεις του κόμβου ορίζονται ως $\langle 0,5+\beta, 0,5-\beta \rangle$. $\alpha > \beta$ ώστε η γνώμη του στόχου να θεωρηθεί μεγαλύτερη.

Parameter	Definition
$0.5 + \alpha$	Prior belief of a target node
$0.5 + \beta$	Prior beliefs of the other nodes
ϵ	Influence in propagation matrices
$f(r)$	Trust degree based on rating information
RF	Factor to adjust influence of trust reciprocation

Πίνακας 2.2.4 PIN-TRUST παράμετροι

>>Για να ληφθεί υπόψη η αμοιβαιότητα εμπιστοσύνης ορίζεται μια έννοια αντίστροφων άκρων που έχουν τις αντίθετες κατευθύνσεις των αντίστοιχων κανονικών άκρων τους. Ως αμοιβαιότητα εμπιστοσύνης [15]ορίζεται η αμφίδρομη εμπιστοσύνη που αντιπροσωπεύει συνήθως μία ισχυρή σχέση μεταξύ των συνδεδεμένων χρηστών σε σύγκριση με τη μονόπλευρη εμπιστοσύνη. Οι ισχυρές και σταθερές τέτοιες σχέσεις είναι πολύ χρήσιμες για την ανάπτυξη εφαρμογών έρευνας και προτάσεων επόμενης γενιάς, με σκοπό την παροχή πιο εξατομικευμένων και ακριβών αποτελεσμάτων.

		(a) Reverse positive edge	
		Source	
Positive		Trust.	Untrust.
Destn.	Trust.	$0.5 + \varepsilon \times (t + f(r)) \times RF$	$0.5 - \varepsilon \times (t + f(r)) \times RF$
	Untrust.	$0.5 - \varepsilon \times (t + f(r)) \times RF$	$0.5 + \varepsilon \times (t + f(r)) \times RF$

		(b) Reverse negative edge	
		Source	
Negative		Trust.	Untrust.
Destn.	Trust.	$0.5 - \varepsilon \times RF$	$0.5 + \varepsilon \times RF$
	Untrust.	$0.5 + \varepsilon \times RF$	$0.5 - \varepsilon \times RF$

Πίνακας 2.2.5 Πολλαπλασιασμός πινάκων ανάποδων άκρων

Το RF είναι μια παράμετρος που ελέγχει την αμοιβαιότητα εμπιστοσύνης, καθώς δεν επιστρέφουν όλοι οι χρήστες καλή αξιολόγηση.

Πείραμα:

Algorithm 1: PIN-TRUST	Algorithm 2: Functions
Input: $U = \{u_1, u_2, \dots, u_n\}$: Set of users (nodes) $S = \{Trustworthy, Untrustworthy\}$: Set of possible states of nodes P : Adjacency matrix with positive edges N : Adjacency matrix with negative edges PR : Adjacency matrix with reverse positive edges NR : Adjacency matrix with reverse negative edges $A = \{P, N, PR, NR\}$: Set of adjacency matrices ϕ : Prior beliefs of each node $\psi = \{\psi_P, \psi_N, \psi_{PR}, \psi_{NR}\}$: Set of propagation matrices for each adjacency matrix Output: Belief score matrix B begin $B = \text{NewMatrix}(U, S)$ foreach $x \in A$ do $MSG = \text{NewMatrix}(U, S)$ while not converged do foreach $i \in U$ do $msg = \text{NewMessage}(S)$ foreach $k \in \text{Neighbors}(x, i)$ do $msg = \text{MultiplyMessage}(msg, MSG[k][i], S)$ foreach $j \in \text{Neighbors}(x, i)$ do $msg_j = \text{DivideMessage}(msg, MSG[j][i], S)$ $MSG[j][i] = \text{PropagateMessage}(i, msg_j, \phi, x, S)$ foreach $i \in U$ do $msg = \text{NewMessage}(S)$ foreach $j \in \text{Neighbors}(x, i)$ do $msg = \text{MultiplyMessage}(msg, MSG[j][i], S)$ foreach $p \in S$ do $msg[p] = msg[p] \times \phi_i[p]$ $msg = \text{NormalizeMessage}(msg, S)$ foreach $p \in S$ do $B[i][p] = msg[p]$ end while end foreach end	function $\text{NewMatrix}(U, S)$: begin return a matrix with $ U $ rows and $ S $ columns, having each element filled with 1 end function $\text{NewMessage}(S)$: begin return a vector, having $ S $ elements filled with 1 end function $\text{Neighbors}(x, i)$: begin return a set of neighbors of node i in the adjacency matrix x end function $\text{MultiplyMessage}(msg, MSG[k][i], S)$: begin foreach $p \in S$ do $msg[p] = msg[p] \times MSG[k][i][p]$ return msg end function $\text{DivideMessage}(msg, MSG[j][i], S)$: begin foreach $p \in S$ do $msg[p] = msg[p] \div MSG[j][i][p]$ return msg end function $\text{PropagateMessage}(i, msg_j, \phi, x, S)$: begin $msg = \text{NewMessage}(S)$ $sum = 0$ foreach $p \in S$ do foreach $q \in S$ do $sum = sum + (\phi_i[q] \times \psi_x[q][p] \times msg_j[q])$ $msg[p] = sum$ return msg end function $\text{NormalizeMessage}(msg, S)$: begin $sum = 0$ foreach $p \in S$ do $sum = sum + msg[p]$ foreach $p \in S$ do $msg[p] = msg[p] \div sum$ return msg end

2.3 Διαδικασία Πειράματος

Στο πείραμα συγκρίνεται η μέθοδος PIN-TRUST με τις μεθόδους ITD και ABIT_L ενώ γίνεται χρήση, του συνόλου δεδομένων Epinions.com: 131,828 χρήστες, 717,667 σχέσεις εμπιστοσύνης, 123,705 σχέσεις δυσπιστίας και 13,668,319 αξιολογήσεις (σε κλίμακα 1-5).

Οι χρήστες-στόχοι επιλέχθηκαν από τις επόμενες ομάδες: χρήστες που έχουν αξιολογήσει πάνω από 100 άλλους χρήστες, ή/και έχουν πάνω από 50 σχέσεις εμπιστοσύνης, χρήστες που έχουν αξιολογήσει από 50 μέχρι 100 άλλους χρήστες, ή/και έχουν από 20 μέχρι 50 σχέσεις εμπιστοσύνης και χρήστες που έχουν αξιολογήσει από 20 μέχρι 50 άλλους χρήστες, ή/και έχουν από 10 μέχρι 20 σχέσεις εμπιστοσύνης.

Ένας τυχαίος χρήστης επιλέγεται και διαγράφονται οι ήδη υπάρχουσες σχέσεις εμπιστοσύνης του. Η PIN-TRUST υπολογίζει τη βαθμολογία πεποίθησης κάθε κόμβου και οι άλλες μέθοδοι την πιθανότητα δημιουργίας σχέσεων εμπιστοσύνης μεταξύ κόμβου και κάθε άλλου κόμβου στο δίκτυο. Οι βαθμολογίες ταξινομούνται με φθίνουσα σειρά και οι υψηλότεροι κόμβοι από την ταξινομημένη λίστα επιλέγονται ως αποτελέσματα πρόβλεψης.

2.4 Αποτελέσματα

Η συγκεκριμένη μελέτη εστιάζει σε 4 βασικά σημεία: **1) την ισχύ των τιμών των παραμέτρων, 2) την ακρίβεια εμπιστοσύνης με την μοντελοποίηση των σχέσεων δυσπιστίας, 3) την ακρίβεια της μεθόδου PIN-TRUST σε σχέση με τις υπάρχουσες μεθόδους και 4) την κλιμάκωση (scaling) της μεθόδου.**

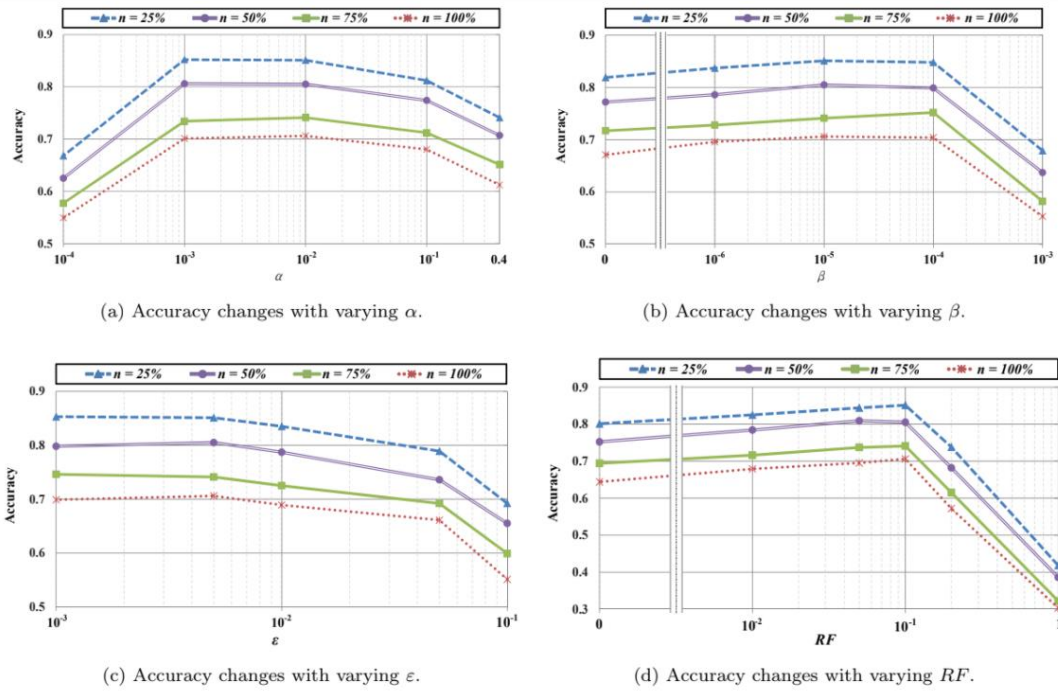
1) Τα αποτελέσματα αποδεικνύουν ότι η μέθοδος είναι πιο ακριβής όταν:

-> $\alpha = 10^{-2}$ αλλά υπάρχει ακόμα ακρίβεια όταν $10^{-1} \geq \alpha \geq 10^{-3}$,

-> $10^{-4} \geq \beta \geq 10^{-6}$ με μια μικρή πτώση στην ακρίβεια όταν $\beta=0$,

-> $10^{-2} \geq \epsilon \geq 10^{-3}$

-> $10^{-1} \geq RF \geq 10^{-2}$

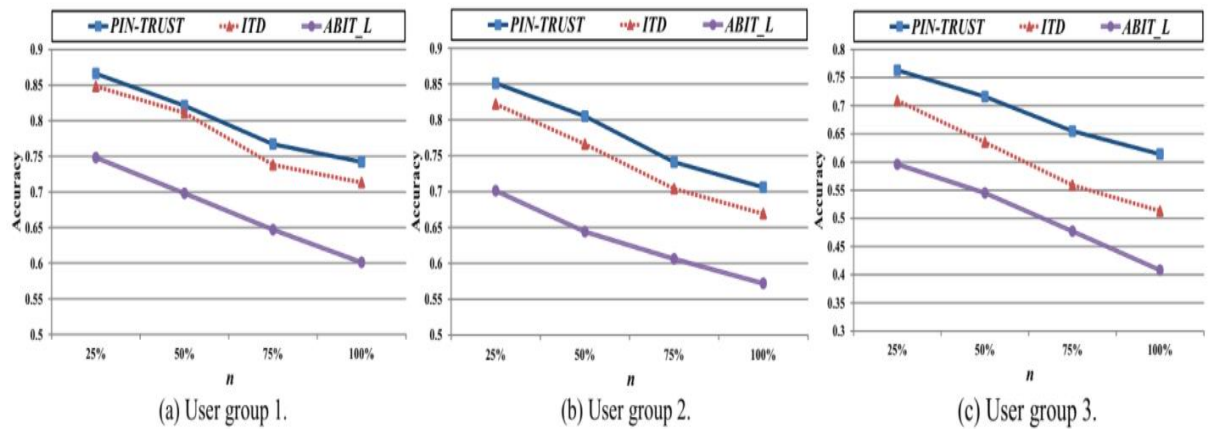


2) Παρατηρείται ότι σε σχέση με τις παραπάνω παραμέτρους, η σχέση δυσπιστίας δεν παίζει τόσο σημαντικό ρόλο στη πρόβλεψη εμπιστοσύνης.

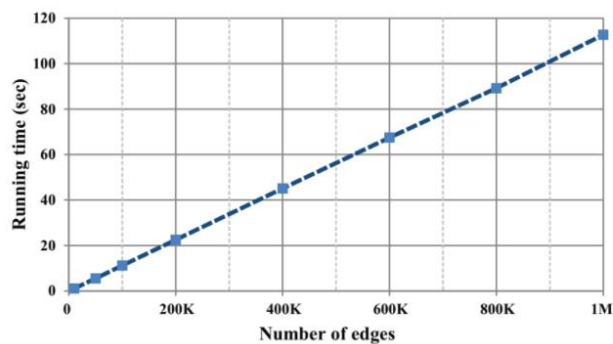
	$n=25\%$	$n=50\%$	$n=75\%$	$n=100\%$
PIN-TRUST with distrust relationships	0.851	0.805	0.741	0.706
PIN-TRUST without distrust relationships	0.851	0.787	0.720	0.673

Πίνακας 2.4.1 Αποτελέσματα ακρίβειας με και χωρίς την σχέση δυσπιστίας.

3) Στην πρώτη ομάδα, η ITD είναι συγκρίσιμη με την PIN-TRUST, πράγμα λογικό, καθώς ελέγχει και την σχέση αξιοπιστίας αλλά και τις αξιολογήσεις, ενώ η ABIT_L δείχνει χαμηλότερη ακρίβεια. Στην δεύτερη ομάδα, η διαφορά ακρίβειας μεταξύ της προτεινόμενης μεθόδου αμβλύνεται καθώς μειώνονται οι αξιολογήσεις. Στην τρίτη ομάδα, με την μείωση των διαδραστικών πληροφοριών είναι εμφανές ότι η PIN-TRUST αποτελεί την ακριβέστερη τεχνική.



4) Γίνεται αντιληπτό ότι με τη πάροδο του χρόνου και την επέκταση των πληροφοριών η μέθοδος γίνεται πιο ακριβής, ικανοποιώντας έτσι τις προσδοκίες κλιμάκωσής της.



3) Ανάλυση Δυναμικών Δικτύων

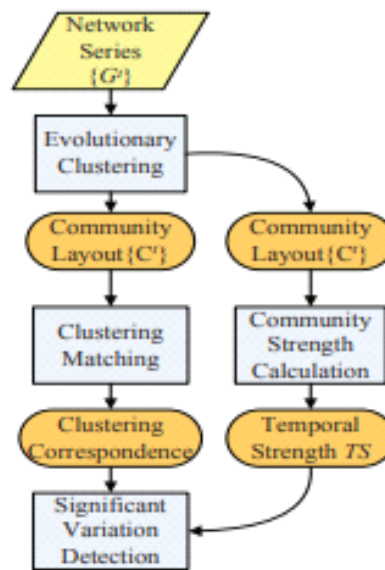
3.1 Ανάλυση της δυναμικής εξέλιξης των κοινωνικών διακυμάνσεων

Πρόσφατα, έχουν σημειωθεί σημαντικές πρόοδοι στον εντοπισμό των ιδιοτήτων των δυναμικών δικτύων, οι οποίες εκμεταλλεύονται τις αλλαγές οντοτήτων και συνδέσμων στο δίκτυο για να σχεδιάσουν τεχνικές ενσωμάτωσης δικτύου. Σε σύγκριση με ευρέως προτεινόμενες μεθόδους ενσωμάτωσης στατικού δικτύου, η ενσωμάτωση δυναμικού δικτύου επιχειρεί να κωδικοποιήσει κόμβους ως χαμηλής διάστασης απεικονίσεις πυκνότητας που διατηρούν αποτελεσματικά τις δομές του δικτύου και την προσωρινή δυναμική, η οποία είναι επωφελής για πολυάριθμες μεταγενέστερες εργασίες μάθησης μηχανής [16].

Ο εντοπισμός κοινοτήτων μέσα σε μεγάλο μέγεθος δεδομένων θεωρείται ένα δημοφιλές θέμα σήμερα, με πολλούς ερευνητές να ασχολούνται με αυτό, καθώς όλο και περισσότερα δεδομένα αναδύονται με την πάροδο του χρόνου. Κατά τη δημιουργία συμπλέγματος τα δεδομένα εξελίσσονται και ξεπερνούν τις παραδοσιακές στατικές δημιουργίες καθώς είναι ισχυρά σε μακροπρόθεσμες διακυμάνσεις[17].

Το πρόβλημα της μελέτης της δημιουργίας συμπλεγμάτων αυτών των δεδομένων , για τα δυναμικά δίκτυα, μπορεί να αντιμετωπιστεί σε δύο στάδια:

- Στο πρώτο στάδιο, με εξελικτική ομαδοποίηση και υπολογισμό της κοινοτικής ισχύος (Temporal Community Strength).
- Στο δεύτερο στάδιο χρησιμοποιώντας ένα μοντέλο βελτιστοποίησης, τον αλγόριθμο VACS ο οποίος περιλαμβάνει πληροφορίες από την εξέλιξη της δημιουργίας συμπλεγμάτων , τη χρονική ομαλότητα και τη μείωση του θορύβου.



Διάγραμμα 3.1.1 The Flow of Whole Framework

3.2 Τα στάδια

Πρώτο στάδιο :

Notation	Meaning
$W_{N \times N}^t$	weight matrix at time t
$C_{N \times K_t}^t$	membership matrix at time t
$M_{K_{t_1} \times K_{t_2}}$	clustering matching matrix from t_1 to t_2
$TS_{K \times T}$	$TS_{z,t}$: temporal strength of community z at t
$R_{K^t \times 1}^t$	variational rate of communities at time t
$t = 1, \dots, T$	index for snapshot
$k = 1, \dots, K_t$	index for community at time t

Πίνακας 3.2.1 Πίνακας συμβολισμού

Καθορίζεται πρώτα η χρονική κοινοτική ισχύ z σε G_t που μπορεί να οριστεί ως:

$$Strength(z) = \sum_{i \in V^t} \sum_{j \in V^t} w_{ij} * \sum_{k \in z} \sum_{l \in z} w_{kl} - \left(\sum_{k \in z} \sum_{m \in V^t} w_{km} \right)^2, \quad (1)$$

Συμπερασματικά, οι κοινότητες με μεγάλη ισχύ έχουν μεγάλη εσωτερική ένταση και λίγες εξωτερικές άκρες.

Έπειτα, γίνεται λόγος για την ομαλοποιημένη κοινοτική ισχύ σε σχέση με το μέγεθος της κοινότητας.

$$NorStrength(z) = \frac{Strength(z)}{|z|}, \quad (2)$$

Μπορεί να αναδιατυπωθεί και με άλλους τύπους όπως για παράδειγμα:

$$NorStrength(z) = \frac{C_{.z}^{t^T} (\sum (W^t) * W^t - D) C_{.z}^t}{|z|} \quad (3)$$

where $D = W^t * \mathbf{1}_{|V^t|, |V^t|} * W^{t^T}$,

Είτε:

$$\begin{aligned} \min_{TS,t} J(TS,t) &= \phi \sum_z \log\left(\frac{1}{TS_{z,t}}\right) NorStrength(z) \\ &+ (1-\phi) \sum_z \log\left(\frac{1}{TS_{z,t}}\right) NorHStrength(z) \quad (4) \\ s.t. \quad &\sum TS_{z,t} \leq \mu_t, TS_{z,t} \geq 0, \forall z = 1, \dots, K_t, \end{aligned}$$

Όσον αφορά την εξελικτική ομαδοποίηση (δημιουργία συμπλεγμάτων), ολόκληρη η διαδικασία παρουσιάζεται στον παρακάτω αλγόριθμο:

Input: A series of network $\{G^t\}_{t=1,\dots,T}$
Output: community membership $\{C^t\}_{t=1,\dots,T}$

```

1 Initialize  $C^0, S^0$  and  $S^{0'}$ ;
2 for  $t \leftarrow 0$  to  $T - 1$  do
3   while not convergence do
4     Update  $S^t$  by Eq. 12;
5     Update  $S^{t'}$  by Eq. 13;
6     Update  $C^t$  by Eq. 11;
7    $\{C^{t+1}, S^{t+1}, S^{t+1'}\} = \{C^t, S^t, S^{t'}\}$ ;
8 return  $\{C^t\}_{t=1,\dots,T}$ ;

```

Αλγόριθμος 3.2.1 Evolutional Clustering (Εξελικτική Δημιουργία Συμπλέγματος)

Η αντιστοίχιση συμπλεγμάτων είναι απαραίτητη για τη διατήρηση συνέπειας. Έχει παρατηρηθεί ότι μια κοινότητα μπορεί να χωριστεί μεμονωμένα ενώ πολλές άλλες να συγχωνευτούν, κατά τη διάρκεια της εξέλιξης. Επίσης, ακόμη και εάν έχει προετοιμαστεί η διαδικασία με προηγούμενη δημιουργία συμπλέγματος, τα συμπλέγματα σε διάφορες χρονικές σημάνσεις μπορεί να μην συμφωνούν με τις ετικέτες τους (π.χ. το πρώτο να μην είναι το C_t αλλά το C_{t+1} και το αντίστροφο).

Δεύτερο στάδιο :

Για την ανάλυση του ποσοστού εξέλιξης των κοινοτήτων δημιουργήθηκε ο αλγόριθμος VACS.

Input: Community Strength TS , Clustering Matching Matrices
Output: a, R

```

1 Initialize  $R$ ;
2 while not converge do
3   Calculate  $TC$  and  $TP$  by Eq.24:
    $TC = TS_{t-1} - M_1 * (TS_{t-1} \circ R)$ ;
    $TP = M_1 * TS_{t-1} - M_1 * ((M_2 * TS_{t-2}) \circ R)$ ;
4   Calculate  $a$  by Eq.21:
    $S = \theta * TC \circ TC + (1 - \theta) * TP \circ TP$ ;
    $a = \mu_1 S \oslash \text{sum}(S)$ ;
5   Calculate  $DN$  by Eq.23:
    $LGA = \log(1 \oslash a)$ ;  $DC = M_1 * \text{diag}(TS_{t-1})$ ;
    $DP = M_1 * \text{diag}(M_2 * TS_{t-2})$ ;
    $DN = LGA^T * (2\theta * DC_{:,i} \circ DC_{:,i} + 2(1 - \theta)DP_{:,i} \circ DP_{:,i})$ ;
6   for each community  $i$  do
7     Calculate  $NN$  by Eq.25 and Eq.26:
      $TCC = \text{diag}(TC) * DC + DC * \text{diag}(R) \circ DC$ ;
      $TPP = \text{diag}(TP) * DP + DP * \text{diag}(R) \circ DP$ ;
     Calculate  $\lambda$  using Eq.22 and constraints in Eq.17;
      $NN = LGA^T * (2\theta TCC_{:,i} + 2(1 - \theta)TPP_{:,i}) - \lambda$ ;
8   Calculate  $R_i = NN/DN$ ;
9 return  $a, R$ ;

```

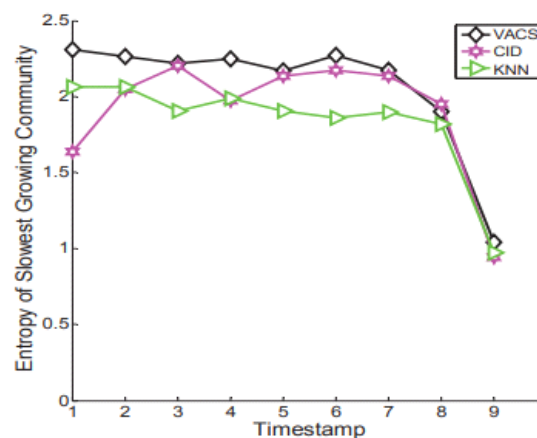
Αλγόριθμος 3.2.2 Ανάλυση Ανάπτυξης Κοινότητας

Method	Average Jaccard Coefficient (Fastest Growing)	Average Entropy (Slowest Growing)
VACS	0.6006	2.0658
KNN	0.5468	1.8307
CID	0.4429	1.9114

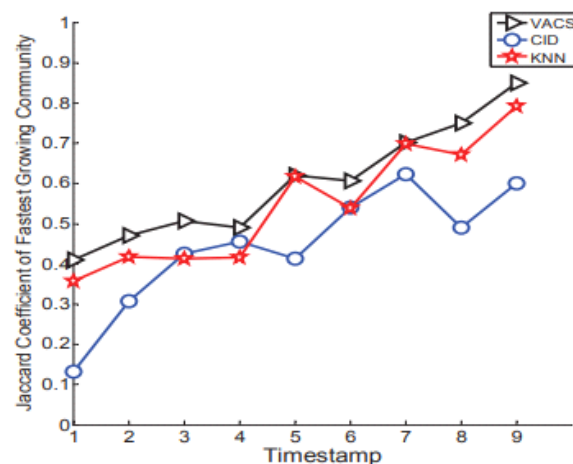
Πίνακας 3.2.2 Απόδοση Συνθετικού Συνόλου Δεδομένων

Friendship Level	Redefined	Weight
Blank	Do Not Know	0
Facebook All Tagged Photos	Not Familiar	1
Political Discuss Socialize Twice Per Week	Acquaintance	2
Close Friend	Friend	3

Πίνακας 3.2.3 Βάρη για Διάφορες Κατηγορίες Φιλίας



(b) Entropy on Slowest Growing Community



(a) Jaccard Coefficient on Fastest Growing Community

Ο Πίνακας 3.2.2 παρουσιάζει τη μέση τιμή σε δέκα στιγμιότυπα. Παρατηρείται ότι η μέθοδος ξεπερνά τις γραμμές βάσης τόσο στην ταχύτερη όσο και στην πιο αργή αναπτυσσόμενη ανίχνευση της κοινότητας.

Η συνολική αύξηση του σχήματος 3, στοιχείο (α), και η συνολική μείωση του σχήματος 3, στοιχείο (β), απεικονίζουν το καλύτερο κοινοτικό αποτέλεσμα ανίχνευσης, το οποίο καταδεικνύει σαφώς την αποτελεσματικότητά του.

3.3 Πείραμα

Οι περιπτωσιολογικές μελέτες για τα σύνολα δεδομένων **DBLP** και **IMDB** δείχνουν επίσης πολύ εντυπωσιακά αποτελέσματα με ενδιαφέρουσες πληροφορίες.

DBLP:

Εφαρμόζεται η προτεινόμενη μέθοδος σε ένα υποσύνολο του συνόλου δεδομένων DBLP, εμπλέκονται 144.179 έγγραφα κατά την περίοδο 1991-2000. Αρχικά ανιχνεύεται η πιο σταθερά αναπτυσσόμενη κοινότητα από το 1991-2000. Σε αυτό το σύνολο δεδομένων, υποτίθεται ότι η κοινότητα εξελίσσεται με σχετική ισχυρή εσωτερική ένταση, επομένως, είναι λιγότερο πιθανό να βρει δραματική αλλαγή μέλους, κατά την εξέλιξη.

IMDB:

Σε αυτό το πείραμα εντοπίζονται οι πιο γρήγορα αναπτυσσόμενες movie star κοινότητες, εφαρμόζοντας την προτεινόμενη μέθοδο σε ένα υποσύνολο του συνόλου δεδομένων IMDB1. Αυτό το σύνολο δεδομένων αφορά κωμικές ταινίες και τηλεοπτικά έργα από το 1991 έως το 2002, μαζί με τις πληροφορίες των αντίστοιχων φορέων που συμμετέχουν. Κατασκευάζεται ένα δίκτυο όπου το άκρο (v_i , v_j) θα οριστεί ως συχνότητα συνεργασίας μεταξύ του παράγοντα i και του παράγοντα j , κατά τη διάρκεια αυτής της περιόδου. Συνολικά 700 αστέρες ταινιών (movie stars) συμμετέχουν σε αυτό το σύνολο δεδομένων.

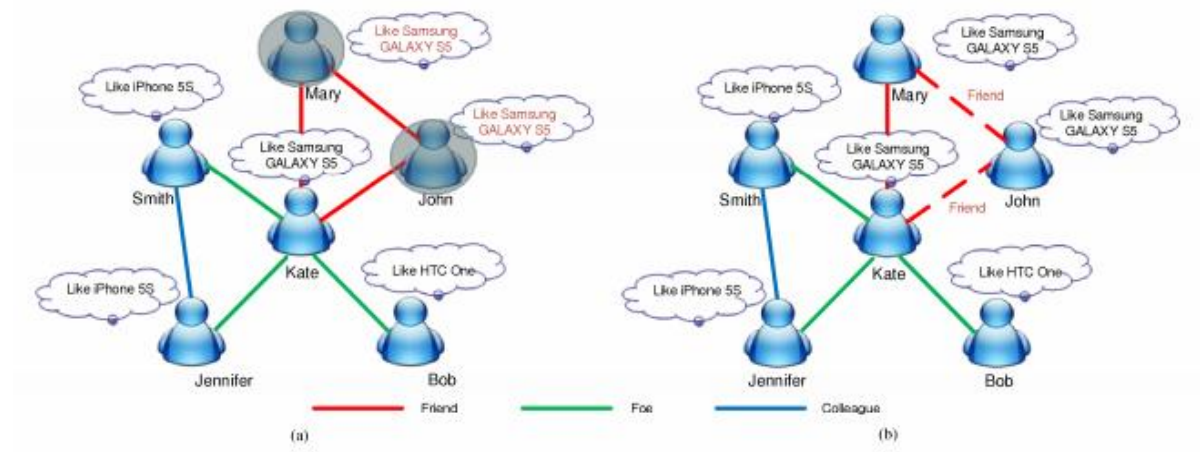
4) Πρόβλεψη κοινωνικών δράσεων στα μεγάλης κλίμακας δίκτυα

4.1 Μαθησιακές αλληλεπιδράσεις για την κοινωνική πρόβλεψη σε δίκτυα μεγάλης κλίμακας

Η τεράστια αύξηση των χρηστών παγκοσμίως στις υπηρεσίες κοινωνικής δικτύωσης, όπως (Facebook, Twitter, κλπ.), έχει αλλάξει δραματικά την σημερινή καθημερινή ζωή.

Η κοινωνική πρόβλεψη αποτελεί ένα ερευνητικό κατόρθωμα διότι στοχεύει στην ερμηνεία των κοινωνικών ενεργειών και δεσμών ανάμεσα στους χρήστες.

- Οι κοινωνικές ενέργειες ορίζονται ως δραστηριότητες, γνώμες, προτιμήσεις ή συμφέροντα που σχετίζονται με χρήστες σε κοινωνικά δίκτυα.
- Κοινωνικός δεσμός ορίζεται ως οποιαδήποτε λεπτή σχέση μεταξύ ζευγών ατόμων σε ένα κοινωνικό δίκτυο.



(a) Παράδειγμα κοινωνικών ενεργειών ενώ (b) παράδειγμα κοινωνικών δεσμών

Οι κοινωνικές ενέργειες και οι κοινωνικοί δεσμοί επηρεάζονται από πολλούς παράγοντες όπως:

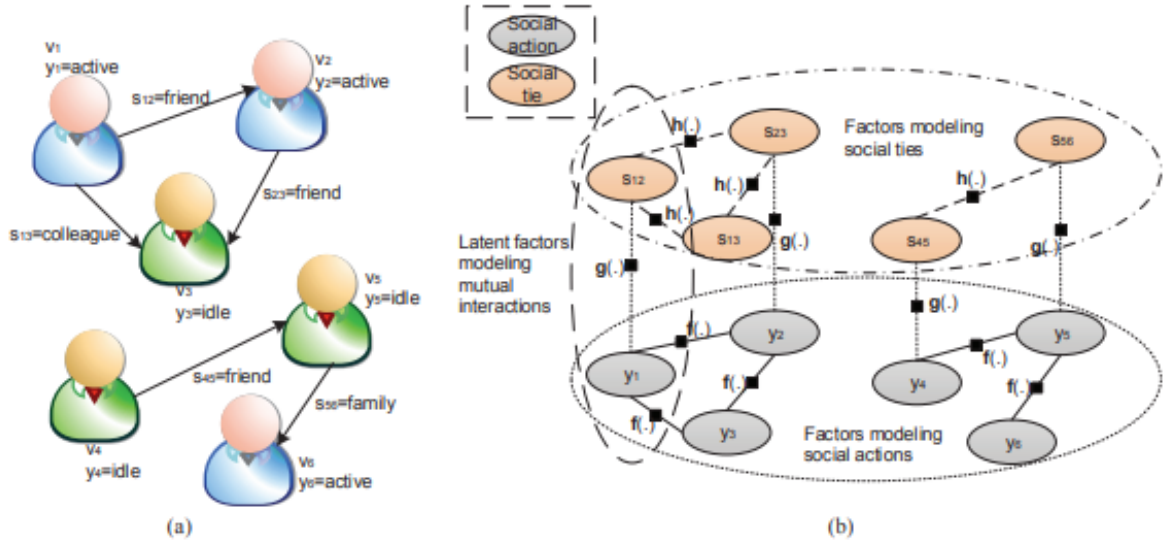
- Η θεωρία της ομοφιλίας (Homophily theory) όπου σύμφωνα με παρατηρήσεις κοινωνικών επιστημόνων παρόμοιοι άνθρωποι τείνουν να συνδέονται μεταξύ τους και να ακολουθούν συμπεριφορές φίλων τους (after all 'birds of a feather flock together').
- Η θεωρία της ετεροφιλίας (Heterophily theory) που σχετίζεται με τη 'structural hole theory' όπου αναφέρεται ότι τα άτομα που ανήκουν σε μία ομάδα τείνουν να έχουν σχέσεις και με μέλη άλλων ομάδων [18].

4.2 Τρόποι αντιμετώπισης προβλημάτων μεγάλου μέγεθους δεδομένων

Έχουν δημιουργηθεί πολλά προβλήματα που σχετίζονται με το τεράστιο μέγεθος των δεδομένων στα κοινωνικά δίκτυα λόγω της μεγάλης καθημερινής συγκέντρωσης χρηστών καθώς των αλληλεπιδράσεων μεταξύ κοινωνικών δεσμών και ενεργειών, τα οποία αντιμετωπίζονται από ερευνητές με τους εξής τρόπους:

A)MLRG graphs

Η δομή ενός γραφήματος μπορεί να οπτικοποιήσει αποτελεσματικότερα τα κοινωνικά δίκτυα. Τα MLRG είναι αμοιβαία λανθάνοντα τυχαία γραφήματα για ταυτόχρονη πρόβλεψη κοινωνικής δεσμών και κοινωνικών ενεργειών.



Είναι αξιοσημείωτο ότι το προτεινόμενο μοντέλο MLRG είναι ουσιαστικά διαφορετικό από τα τυπικά μοντέλα τυχαίων γραφημάτων (ERGM) και από τα Linking user-centered social networks graphs κυρίως σε δύο πτυχές. Πρώτον, το μοντέλο MLRG ορίζει λανθάνοντες παράγοντες για την ανάληψη αμοιβαίων και δυναμικών αλληλεπιδράσεων μεταξύ κοινωνικών δεσμών και κοινωνικών ενεργειών. Δεύτερον, σε σύγκριση με τα προηγούμενα μοντέλα, η MLRG παρέχει ένα ενιαίο πλαίσιο για την ταυτόχρονη αντιμετώπιση τους [19].

Β)Χρήση ενός κατά προσέγγιση αλγόριθμου που εξερευνά τους συντελεστές του γραφήματος για την αποτελεσματική βελτιστοποίηση των παραμέτρων.

Input: The social graph G , number of iterations T .
Output: Optimized parameters $\theta' = \{\alpha', \beta', \gamma'\}$.
Initialize pairwise potentials $\psi(y_e, s_e, G_e)$;
Initialize factorized distribution for $Q_{\theta'}(y, s)$;
repeat
 Map process: Calculate marginals on all slave nodes in parallel;
 Each slave node calculates the marginal expectation $E_{Q_{\theta'}}\{\log P_{\theta'}(y, s, G)|r_m\}$ according to Eq. (8);
 Each slave node calculates the marginal probability $Q_{\theta'_m}(r_m)$ according to Eq. (9) and Eq. (10);
 Reduce process: Calculate objective function and update all parameters on the master node;
 Master node calculates and maximizes objective function $\mathcal{L}(\theta')$ according to Eq. (6);
 Master node updates and optimizes all parameters $\theta' = \{\alpha', \beta', \gamma'\}$ such that $\theta' = \arg \max \mathcal{L}(\theta')$;
until converge;

Αλγόριθμος 4.2.1 Κατανεμημένος αλγόριθμος μάθησης για MLRG

4.3 Πείραμα

Χρησιμοποιήσαμε τα ακόλουθα δύο σύνολα δεδομένων για την πειραματική μας μελέτη:

Epinion+ :

Οι κοινωνικές ενέργειες εδώ είναι οι αξιολογήσεις των χρηστών και χρησιμοποιούμε very dislike, dislike, neutral, like, very like για να αξιολογήσουμε από το 1 έως το 5 στα προϊόντα, αντίστοιχα. Οι κοινωνικοί δεσμοί είναι οι σχέσεις εμπιστοσύνης και δυσπιστίας σε αυτό το δίκτυο.

Dataset statistics	
Nodes	75879
Edges	508837
Nodes in largest WCC	75877 (1.000)
Edges in largest WCC	508836 (1.000)
Nodes in largest SCC	32223 (0.425)
Edges in largest SCC	443506 (0.872)
Average clustering coefficient	0.1378
Number of triangles	1624481
Fraction of closed triangles	0.0229
Diameter (longest shortest path)	14
90-percentile effective diameter	5

Πίνακας 4.3.1 [20]

Mobile:

Τα κινητά θεωρούνται σήμερα τα σημαντικότερα εργαλεία επικοινωνίας . Οι κοινωνικές ενέργειες διαμορφώνονται με την κλήση ή την αποστολή σύντομων μηνυμάτων μεταξύ τους κατά τη διάρκεια μερικών μηνών. Οι κοινωνικοί δεσμοί που ερευνήσαμε είναι σχέσεις φίλων, οικογενειών και συναδέλφων.

Statistics	Epinion+	Mobile
Num. of Nodes	126,785	3,268
Num. of Actions	816,946	30,776
Num. of Ties	620,159	18,489
Avg. Num. of Actions	6.44	9.42
Max. Num. of Actions	932	1278
Avg. Num. of Ties	4.89	5.66
Max. Num. of Ties	956	247
ACC	0.14	0.45
Diameter	15	11

Πίνακας 4.3.2 Στατιστικά του *Epinion+* και του *Mobile*.

Για την αξιολόγηση των επιδόσεων στα δύο παραπάνω dataset χρησιμοποιούμε:

- **Ακρίβεια (P)**: ο αριθμός των ορθώς προβλεπόμενων ενεργειών (ή δεσμών) διαιρούμενος με τον συνολικό αριθμό προβλεπόμενων ενεργειών (ή δεσμών).
- **Ανάκληση (R)**: αριθμός των σωστά προβλεπόμενων ενεργειών (ή δεσμών) διά του συνολικού αριθμού δράσεων (ή δεσμών) στο σύνολο δεδομένων.
- **Μέτρηση (F)**: ισοροπημένος αρμονικός μέσος ακρίβειας και ανάκλησης και χρησιμοποιήσαμε
$$F_{\beta=1} = \frac{2PR}{P+R},$$
 η ισοροπημένη βαθμολογία που σταθμίζει την ακρίβεια και την ανακαλεί ομοιόμορφα.
- **Αποδοτικότητα**: τον χρόνο λειτουργίας των συγκριθέντων μοντέλων, συμπεριλαμβανομένου του χρόνου εκμάθησης και του χρόνου παρεμβολής.

Συγκρίναμε την προσέγγισή μας με τις ακόλουθες εναλλακτικές μεθόδους πρόβλεψης κοινωνικών ενεργειών και δημιουργίας κοινωνικών δεσμών:

- **SVM**: Αυτό το μοντέλο εξετάζει την πρόβλεψη κοινωνικής ενέργειας και της αναφοράς κοινωνικού δεσμού ως δύο ξεχωριστά προβλήματα ταξινόμησης και τα επιλύει ανεξάρτητα χωρίς να λαμβάνονται υπόψη οι αμφίδρομες αλληλεπιδράσεις τους.
- **ERGM**: Αυτό είναι το παραδοσιακό μοντέλο τυχαίου γραφήματος εκθετικής οικογένειας για την πρόβλεψη κοινωνικής ενέργειας και την κοινωνική παρεμβολή σύνδεσης ανεξάρτητα. Σημειώστε ότι αυτό το μοντέλο έχει ενσωματωμένο συντελεστή λανθάνοντος $\phi \cdot h(\cdot)$.
- **DCRF**: Το μοντέλο αυτό είναι ένα δυναμικό και παραγοντικό CRF χρησιμοποιείται για την από κοινού επίλυση των δύο εργασιών. Αυτό το μοντέλο προτάθηκε αρχικά για επισήμανση και διαχωρισμό δεδομένων ακολουθίας.

Method	Precision	Recall	F1-score
SVM	71.48	69.05	70.24
ERGM	72.08	69.76	70.90
DCRF	73.78	71.22	72.48
MLRG	74.96	73.89	74.42

Πίνακας 4.3.3 Συγκριτική απόδοση της πρόβλεψης κοινωνικής δράσης στο *Epinion+*.

Actions	SVM			ERGM			DCRF			MLRG		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
Calling	81.07	76.13	78.52	81.99	80.57	81.27	90.12	82.73	86.27	90.66	89.24	89.94
Sending msg	82.76	78.34	80.49	79.32	78.64	78.98	88.78	80.44	84.40	88.75	87.91	88.33
Overall	82.45	76.56	79.40	80.69	79.70	80.19	89.67	81.32	85.29	89.55	88.70	89.12

Πίνακας 4.3.4 Συγκριτική απόδοση πρόβλεψης κοινωνικής δράσης στο σύνολο πληροφοριών *Mobile*.

Ties	SVM			ERGM			DCRF			MLRG		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
Trust	83.77	76.68	80.07	75.93	83.62	79.59	83.45	76.99	80.09	84.26	82.91	83.58
Distrust	83.01	75.49	79.07	75.03	83.05	78.84	82.78	76.16	79.33	82.95	80.44	81.68
Overall	83.19	76.02	79.44	75.33	83.22	79.08	83.25	76.58	79.78	83.45	82.69	83.07

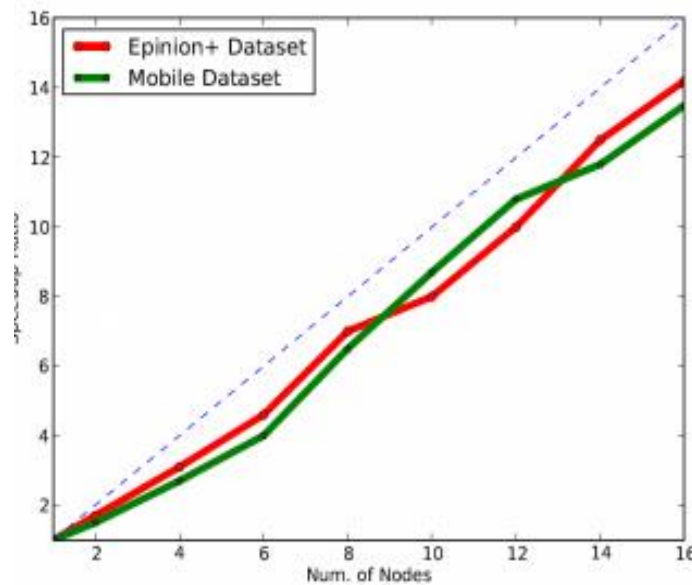
Πίνακας 4.3.5 Συγκριτική απόδοση της ανακάλυψης κοινωνικής σύνδεσης στο *Epinion+*.

Ties	SVM			ERGM			DCRF			MLRG		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
Friend	81.92	76.73	79.24	80.93	79.25	80.08	89.46	79.75	84.33	89.73	88.78	89.25
Family	81.04	75.72	78.29	79.81	78.03	78.91	88.82	78.42	83.30	88.71	87.52	88.11
Colleague	81.15	75.59	78.27	79.87	78.14	79.00	89.01	78.67	83.52	88.64	87.39	88.01
Overall	81.37	76.06	78.63	80.19	78.44	79.31	89.19	78.96	83.76	89.05	87.91	88.48

Πίνακας 4.3.6 Συγκριτική απόδοση της ανακάλυψης κοινωνικής σύνδεσης στο *Mobile*.

Method	Running Time	Epinion+ Dataset	Mobile Dataset
SVM		3.86hrs	29.25mins
ERGM		2.40hrs	18.20mins
DCRF		15.48hrs	1.69hrs
MLRG		9.25hrs	65.27mins
MLRG with Speedup		39.10mins	4.83mins

Πίνακας 4.3.7 Σύγκριση απόδοσης των διαφορετικών μοντέλων στα σύνολα δεδομένων *Epinion+* και *Mobile*. Ο χρόνος λειτουργίας περιλαμβάνει τόσο χρόνο εκμάθησης όσο και χρόνο παρεμβολής.



Διάγραμμα 4.3.1 απόδοση επιτάχυνσης για τους καταναμημένους αλγόριθμους μάθησης των MLRG.

5) Κοινωνική Διάσταση του Διαδικτύου

Το διαδίκτυο γίνεται όλο και πιο περίπλοκο με την κοινωνικοποίηση και την αλληλεπίδραση ατόμων και αντικειμένων, το κοινωνικό δίκτυο. Όλα αυτά τα στοιχεία εμπλουτίζουν το κοινωνικό πλαίσιο των ιστοσελίδων και μπορούν να χρησιμοποιηθούν για την πρόβλεψη επιλογής από τον εκάστοτε χρήστη. Τα κλασσικά μοντέλα της Ανάκτησης Πληροφοριών (Information Retrieval-IR) προσαρμόζονται βάσει: του κοινωνικού πλαισίου που περιβάλλει ιστοσελίδες και το κοινωνικό πλαίσιο των χρηστών.

Η εξέλιξη του διαδικτύου είναι γνωστή ως κοινωνικό διαδίκτυο όπου οι χρήστες μέσω της σύνδεσης τους σε πλατφόρμες του Παγκόσμιου Ιστού μοιράζονται ή σχολιάζουν πληροφορίες καθώς και προσθέτουν tags σε αυτά. Έρευνες αναφέρουν ότι τα tags είναι πολύ χρήσιμα καθώς βελτιώνουν την ποιότητα της αναζήτησης. Η εκμετάλλευση των πληροφοριών αυτών δεν θα πρέπει να παραβιάζει την ιδιωτική ζωή των χρηστών, οι οποίες δεν περιέχουν ευαίσθητες πληροφορίες σε σχέση με τους χρήστες. Τέλος, οι κοινωνικοί πόροι είναι συχνά προσιτοί στο κοινό, καθώς τα κοινωνικά δίκτυα παρέχουν στους API πρόσβαση στα δεδομένα του.

Λαμβάνοντας υπόψη το κοινωνικό πλαίσιο του Διαδικτύου, προτείνονται μέθοδοι βελτίωσης του μοντέλου IR που περιλαμβάνουν:

- Μια λειτουργία κοινωνικής εξατομικευμένης κατάταξης που ονομάζεται SoPRa

Η SoPRa κατατάσσει τα έγγραφα σύμφωνα με:

- 1) Μια βαθμολογία αντιστοίχισης κειμένου των εγγράφων του ερωτήματος,
- 2) Μια βαθμολογία κοινωνικής αντιστοίχισης των εγγράφων του ερωτήματος,
- 3) Μια βαθμολογία κοινωνικού ενδιαφέροντος

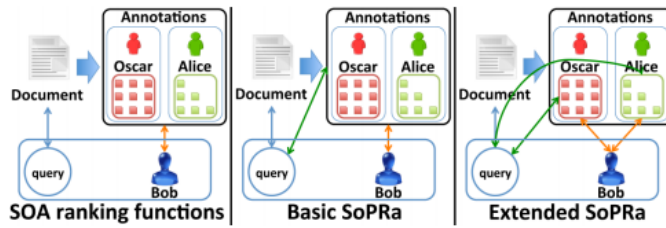


Figure 1: Illustration of the basic differences between the different approaches.

- Μια μέθοδο για τη ζύγιση των προφίλ χρηστών και των κοινωνικών εγγράφων

Το VSM έχει αποδείξει την αποτελεσματικότητά του και ένα από τα βασικά σημεία του είναι η στάθμιση των όρων, επομένως εάν θεωρήσουμε τις ιστοσελίδες ή τους χρήστες ως έγγραφα και τις σημειώσεις ως όρους έχουμε τον εξής τύπο:

$$w_t^d = tf_t \times \log\left(\frac{|R|}{|R_t|}\right), w_t^u = uf_t \times \log\left(\frac{|U|}{|U_t|}\right)$$

- Επέκταση της λειτουργίας SoPRa με την προσθήκη ετικετών

Οι λειτουργίες κατάταξης IR ενδέχεται να βελτιωθούν λαμβάνοντας υπόψη κάθε χρήστη που σχολιάζει, βελτιώνοντας έτσι και τη βαθμολογία ενός εγγράφου.

Επέκταση SoPRa:

$$Rank(d, q, u) = \gamma \times \sum_{u_k \in U_d} Cos(\vec{p_{u_k}}, \vec{p_u}) \times Cos(\vec{p_u}, \vec{T_{u_k, d}}) + (1 - \gamma) \times \left[\beta \times \sum_{u_k \in U_d} Cos(\vec{p_{u_k}}, \vec{p_u}) \times Cos(\vec{q}, \vec{T_{u_k, d}}) + (1 - \beta) \times Sim(\vec{q}, \vec{d}) \right]$$

- Εντατική αξιολόγηση SoPRa

Για να επιτευχθεί αποτελεσματική αξιολόγηση :

- 1) Αφαιρούμε τις πολύ προσωπικές παρατηρήσεις ή ανούσιες
- 2) Χρησιμοποιούμε τον αλγόριθμο του Porter για να εξαλειφθούν οι διαφορές μεταξύ ορών που έχουν ίδια ρίζα
- 3) Εκτελούμε το εργαλείο γραμμής εντολών cURL για να κατεβάσουμε όλες τις διαθέσιμες σελίδες καταργώντας εκείνες που δεν είναι διαθέσιμες πλέον
- 4) Καταργούμε όλες τις μη αγγλικές ιστοσελίδες

Βιβλιογραφία:

- [1] Π. Αθανάσιος, *Ανάκτηση πληροφορίας από μέσα κοινωνικής δικτύωσης σε καταστάσεις έκτακτης ανάγκης*, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2019.
- [2] S. N. a. S. Chaudhury, *Link Prediction in Heterogeneous Social Networks*, Indian Institute of Technology, Delhi.
- [3] C. F. S.-W. K. U. K. J. H. Min-Hee Jang, *PIN-TRUST: Fast Trust Propagation Exploiting Positive,,* Hanyang University, Carnegie Mellon University, Seoul National University.
- [4] N. D. J. G. A. Z. Xiaowei Jia, *Analysis on Community Variational Trend in Dynamic*, State University of New York at Buffalo, Buffalo, NY, USA.
- [5] J. X. Xiaofeng Yu, *Learning Interactions for Social Prediction in Large-scale*, Universal Business Park, 10 Jiu XianQiao Road.
- [6] H. H. M. B. Mohamed Reda, *SoPRa: A New Social Personalized RankingFunction for Improving Web Search*.
- [7] Z. G. T. S. J. a. L. K. S. A. M. I. Jordan, *introduction to variational methods for graphical methods*, 1999.
- [8] X. Y. a. W. Lam, «Bidirectional integration of pipeline models,» σε *Proceedings of AAAI-10*, Atlanta, 2010, p. 1045–1050.
- [9] A. T. L. M. L. H. E. C. R. S. N. Z. Gong, «Joint link prediction and attribute inference using a social-attribute network,» 2014.
- [10] Κ. Γ. Περάκης, «Ταξινομήσεις».
- [11] B. H. T. J. Blake Shaw, "Learning a Distance Metric from a Network," Columbia University.
- [12] G. R. G. L. a. B. M. Daryl Lim, «Robust structural metric learning,» σε *volume 28 of JMLR*, 2013, p. 615–623.
- [13] A. P. J.-M. R. M. S. François Fouss, «Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation,» 2007, p. 355–369.
- [14] H. L. Reza Zafarani, « Influence and Homophily,» σε *Social Media Mining*, Cambridge University Press.
- [15] E.-P. L. H.-H. T. J. J. Viet-An Nguyen, «DoYou Trust to Get Trust? A Study of Trust Reciprocity Behaviors andReciprocal Trust Prediction,» Singapore Management University †Nanyang Technological University.
- [16] C. L. B. Y. C. Z. Z. T. Yu Xie, «A Survey on Dynamic Network Embedding,» 2020.

- [17] J. M. K. A. O. H. I. Kevin S. Xu, «Adaptive Evolutionary Clustering,» Israel, 2018.
- [18] C. L. D. M.-S. Oriol Barranco, *Heterophily in social groups formation: a social network analysis*, Springer Science+Business Media B.V., part of Springer Nature 2018, 2018.
- [19] R. B. R. A. H. T. B. A. Amina Amara, *Simulating the merge between user-centered graphs of social networks*, Tunisia, 2016.
- [20] R. A. P. D. Matthew Richardson, *Trust Management for the Semantic Web*, University of Washington.