ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών
«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Μεθοδολογίες Πρόγνωσης Ακμών σε Κοινωνικά Δίκτυα με Χρήση Νευρωνικών Δικτύων Γράφων** **Link Prediction in Social Networks with Graph Neural Networks**
Ονοματεπώνυμο Φοιτητή	**Μάρκος-Δημήτριος Καράτζιας**
Πατρώνυμο	**Αστέριος**
Αριθμός Μητρώου	**ΜΠΣΠ/16012**
Επιβλέπων	**Γεώργιος Τσιχριντζής, Καθηγητής**

Ημερομηνία Παράδοσης **Σεπτέμβριος 2022**

Τριμελής Εξεταστική Επιτροπή

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

ΠΕΡΙΛΗΨΗ

Ο σκοπός αυτής της διπλωματικής εργασίας είναι η μελέτη του προβλήματος της πρόγνωσης ακμών στα μέσα κοινωνικής δικτύωσης με χρήση νευρωνικών δικτύων γράφων.

Τα Νευρωνικά Δίκτυα Γράφων είναι μοντέλα σύνδεσης που χρησιμοποιούν τη μετάδοση μηνυμάτων μεταξύ κόμβων γραφημάτων για να αντικατοπτρίζουν την εξάρτηση γράφου. Τα Νευρωνικά δίκτυα γράφων, σε αντίθεση με τα κανονικά νευρωνικά δίκτυα, διατηρούν μια κατάσταση που τους επιτρέπει να αντιπροσωπεύουν πληροφορίες από το άμεσο περιβάλλον τους με αυθαίρετο βάθος.

Η πρόβλεψη σύνδεσης είναι ένα βασικό πρόβλημα για δεδομένα δομημένα σε δίκτυο. Τα ευρετικά πρόβλεψης συνδέσμων χρησιμοποιούν ορισμένες συναρτήσεις βαθμολογίας, όπως κοινούς γείτονες και δείκτη Katz, για τη μέτρηση της πιθανότητας συνδέσεων. Έχουν λάβει ευρείες πρακτικές χρήσεις λόγω της απλότητα, της ερμηνευτικότητας, και για ορισμένα από αυτά, της επεκτασιμότητάς τους.

Αρχικά γίνεται μια αναφορά στα κοινωνικά δίκτυα, τη δομή τους και την ανάλυση αυτών καθώς επίσης αναφέρουμε και άλλα ζητήματα που αφορούν στην έρευνα της ανάλυσης κοινωνικών δικτύων. Έπειτα περιγράφουμε το πρόβλημα της πρόγνωσης ακμών, αλλά πριν περάσουμε στις τεχνικές πρόβλεψης γίνεται μια γενικότερη αλλά και εκτενής ταυτόχρονα αναφορά στα νευρωνικά δίκτυα σε γράφους, τον σχεδιασμό αλλά και το που αυτά βρίσκουν εφαρμογές στο πραγματικό κόσμο.

ABSTRACT

The purpose of this dissertation is to study the problem of link prediction on social media using graph neural networks.

Graph Neural Networks are connection models that use the transmission of messages between graph nodes to reflect graph dependency. Graphic neural networks, unlike normal neural networks, maintain a state that allows them to represent information from their immediate environment with arbitrary depth.

Link prediction is a major problem for network-structured data. Link prediction heuristics use certain scoring functions, such as common neighbors and the Katz index, to measure link probability. They have received wide practical uses due to their simplicity, interpretability, and for some of them, their scalability.

Initially, a reference is made to social networks, their structure, and their analysis as well as other issues related to the research of social network analysis. Then we describe the problem of link prediction, but before moving on to the link prediction techniques, a more general but at the same time extensive reference is made to neural networks in graphs, their general design and how these are applied in the real world.

Περιεχόμενα

Εισαγωγή

Ο όρος **μέσα κοινωνικής δικτύωσης** [1] αναφέρεται στα μέσα αλληλεπίδρασης και επικοινωνίας ομάδων ανθρώπων μέσω διαδικτυακών κοινοτήτων. Κάποια από τα πιο διαδεδομένα μέσα κοινωνικής δικτύωσης είναι το Facebook, το Instagram, το Twitter, το TikTok, κ.α. Επιπλέον ενισχύουν την δια δραστικότητα μεταξύ ανθρώπων οι οποίοι συνδέονται μεταξύ τους, δημιουργούν, μοιράζονται ή ανταλλάσσουν πληροφορίες συμμετέχοντας σε αυτά.

Κοινωνική δικτύωση είναι η συγκέντρωση ή συμμετοχή των ατόμων σε ομάδες διαδικτυακές ή μη. Ένα κοινωνικό δίκτυο απαρτίζεται από ένα σύνολο ανθρώπων, οργανισμών ή άλλων κοινωνικών ομάδων καθώς και από τις σχέσεις που προκύπτουν μεταξύ αυτών. Τα Κοινωνικά δίκτυα συναντώνται αρχής γενομένης, από τους πρώτους ανθρώπους έως στο παγκόσμιο ιστό και έχουν μελετηθεί ενδελεχώς από τους κοινωνιολόγους.

Τα διαδικτυακά κοινωνικά δίκτυα ορίζονται ως υπηρεσίες βασισμένες στο διαδίκτυο που επιτρέπουν στα άτομα να δημιουργήσουν ένα προφίλ ιδιωτικό ή δημόσιο μέσα σε ένα οριοθετημένο σύστημα και να επικοινωνήσουν με άλλους χρήστες με τους οποίους συνδέονται άμεσα ή έμμεσα. Οι όροι *μέσα κοινωνικής δικτύωσης* και *κοινωνικό δίκτυο* συχνά ταυτίζονται κάτω από τον όρο «κοινωνική δικτύωση». Υπάρχει όμως μια σημαντική διαφοροποίηση, ο όρος “μέσο κοινωνικής δικτύωσης” αναφέρεται στα μέσα (εργαλεία) διαμοιρασμού της πληροφορίας, των δεδομένων και της επικοινωνίας στο κοινό, ενώ ο όρος “κοινωνική δικτύωση” αναφέρεται στη δημιουργία και την αξιοποίηση των κοινοτήτων για τη διασύνδεση ανθρώπων με κοινά ενδιαφέροντα. Θα μπορούσε να ειπωθεί δηλαδή ότι ο όρος “μέσα κοινωνικής δικτύωσης” αναφέρεται στα εργαλεία - μέσα ενημέρωσης κοινωνικής δικτύωσης, ενώ ο όρος “κοινωνική δικτύωση” στη διαδικασία της κοινωνικής δικτύωσης.

Τα κοινωνικά δίκτυα διαθέτουν τα παρακάτω χαρακτηριστικά:

- Πολλαπλές μορφές περιεχομένου, όπως κείμενο, βίντεο, φωτογραφία, ήχο, κ.τ.λ.
- Υποστηρίζουν διαφορετικά επίπεδα αλληλεπίδρασης των χρηστών οι οποίοι μπορούν να δημιουργήσουν, να σχολιάσουν ή να παρακολουθήσουν
- Απλοποιούν καθώς και βελτιώνουν την ταχύτητα και το εύρος της διάδοσης των πληροφοριών
- Επιτρέπουν μονόδρομη και αμφίδρομη επικοινωνία καθώς αυτή μπορεί να πραγματοποιείται είτε σε πραγματικό χρόνο ή ασύγχρονη με την πάροδο του χρόνου
- Είναι ανεξάρτητα της συσκευής που μπορεί να χρησιμοποιήσει ο χρήστης. Η διείσδυση σε ένα μέσο κοινωνικής δικτύωσης μπορεί να γίνει με έναν υπολογιστή, ή μια κινητή συσκευή
- Επεκτείνει την εμπλοκή του χρήστη με τρεις τρόπους: με τη δημιουργία διαδικτυακών εκδηλώσεων σε πραγματικό χρόνο, με την απευθείας σύνδεση και αλληλεπίδραση σε δια ζώσης εκδηλώσεις, και τελευταία με την υποστήριξη ζωντανών εκδηλώσεων

Επιπροσθέτως τα μέσα κοινωνικής δικτύωσης, είναι ικανά να συντελέσουν στην προβολή αφενός της κοινής γνώμης, καθώς καθιστούν τους χρήστες δέκτες αλλά και εκδότες περιεχομένου μέσω της ανατροφοδότησης, ενώ παρέχουν κοινωνική και συναισθηματική υποστήριξη. Αφετέρου, συμβάλουν στην προώθηση συγκεκριμένων ιδεών, αξιών, ακόμη και προϊόντων, ώστε επιχειρήσεις, πολιτικοί, ομάδες συμφερόντων κ.α. να μάχονται για την αλλοίωσή της, στο σκληρό πλαίσιο του ανταγωνισμού.

Η μεγάλη ανάπτυξη και ταχεία εξάπλωση των μέσων κοινωνικής δικτύωσης δεν έχει περάσει απαρατήρητη και από την επιστημονική κοινότητα καθώς ολοένα και περισσότεροι ασχολούνται με τον κλάδο των κοινωνικών δικτύων και την ανάλυση αυτών με διαφορετικές μεθόδους και τεχνικές με σκοπό την εξόρυξη πληροφοριών.

Σχήμα 1: Εξέλιξη των μέσων κοινωνικής δικτύωσης

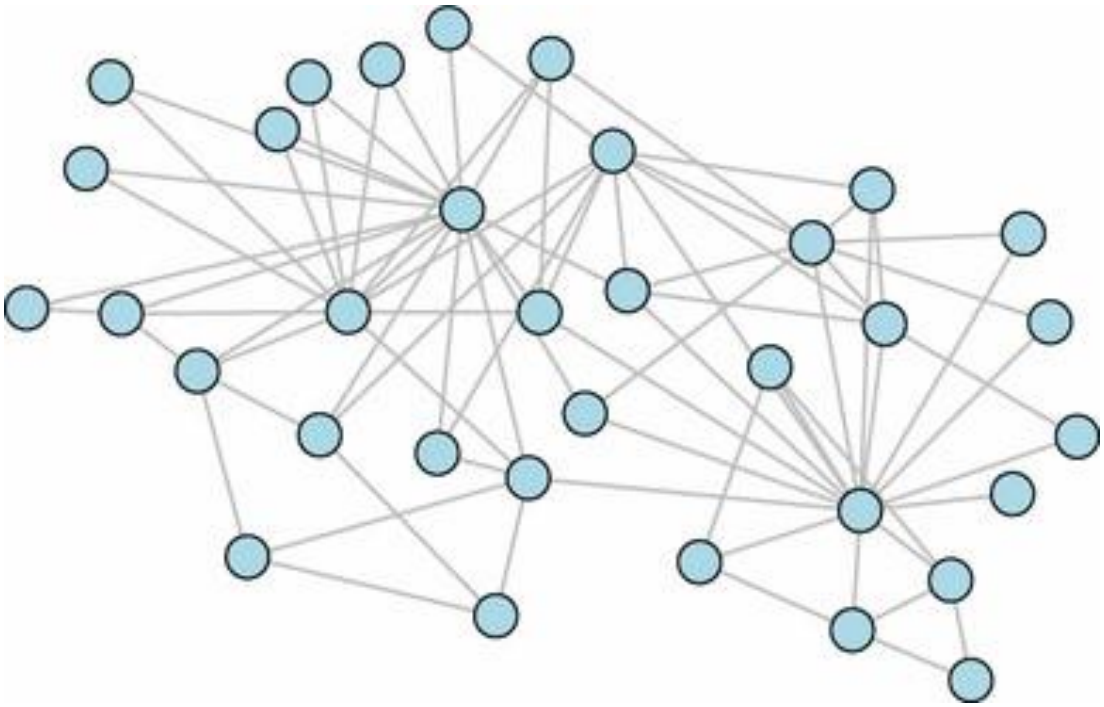
Η έννοια του κοινωνικού δικτύου και οι μέθοδοι ανάλυσης κοινωνικών δικτύων έχουν προσελκύσει σημαντικό ενδιαφέρον και περιέργεια από την κοινότητα της κοινωνικής και συμπεριφοριστικής επιστήμης τις τελευταίες δεκαετίες. Μεγάλο μέρος αυτού του ενδιαφέροντος μπορεί να αποδοθεί στην ελκυστική εστίαση της ανάλυσης των κοινωνικών δικτύων στις σχέσεις μεταξύ κοινωνικών οντοτήτων και στα πρότυπα και τις επιπτώσεις αυτών των σχέσεων. Πολλοί ερευνητές έχουν συνειδητοποιήσει ότι η προοπτική του δικτύου επιτρέπει νέο μοχλό για την απάντηση τυπικών ερωτημάτων έρευνας κοινωνικής και συμπεριφοριστικής επιστήμης, δίνοντας ακριβή επίσημο ορισμό σε πτυχές του πολιτικού, οικονομικού ή κοινωνικού δομικού περιβάλλοντος. Από την άποψη της ανάλυσης των κοινωνικών δικτύων, το κοινωνικό περιβάλλον μπορεί να εκφραστεί ως πρότυπα ή κανονικότητες στις σχέσεις μεταξύ αλληλοεπιδρώντων μονάδων.

Κύριο χαρακτηριστικό της ανάλυσης κοινωνικών δικτύων είναι η μελέτη των σχέσεων μεταξύ των οντοτήτων του δικτύου, αντί της επεξεργασίας των χαρακτηριστικών των ίδιων των ατόμων. Το πεδίο της ανάλυσης κοινωνικών δικτύων αποτελεί τη βάση πάνω στην οποία ορίζονται και θεμελιώνονται θεωρητικές έννοιες και αξιολογούνται μοντέλα και θεωρίες. Παρέχει επίσης τα εργαλεία για την αναπαράσταση των κοινωνικών δικτύων, τη μελέτη της δομή τους και την εύρεση των σημαντικότερων κοινωνικών ατόμων. Επίσης, δίνεται η δυνατότητα να αναγνωριστούν μοτίβα στις σχέσεις μεταξύ των ατόμων, χωρίς να απαιτείται η μελέτη των χαρακτηριστικών και της συμπεριφοράς μεμονωμένων μελών του δικτύου. [2]

1 Ανάλυση Κοινωνικών Δικτύων

Τι είναι ένα δίκτυο

- Με τον όρο δίκτυο αναφερόμαστε συνήθως σε οντότητες συνδεδεμένες μεταξύ τους. Είτε αυτές είναι κάποια ομάδα είτε κάποιο σύστημα.
- Διάφοροι τύποι δικτύων είναι οι εξής:
    - Δίκτυα επιχειρήσεων
    - Δίκτυα οικονομικά
    - Δίκτυα καταστημάτων
    - Δίκτυα κοινωνικά
    - Δίκτυα ηλεκτρονικών υπολογιστών



Σχήμα 2: Απεικόνιση Δικτύου

Κοινωνικό Δίκτυο

Ένα κοινωνικό δίκτυο είναι μία κοινωνική δομή που αποτελείται από ένα σύνολο παραγόντων, όπως άτομα ή οργανισμούς. Αναπαρίσταται συνήθως από κόμβους οι οποίοι είναι συνδεδεμένοι μεταξύ τους με έναν ή περισσότερους τύπους αλληλεξάρτησης, φιλία, συγγένεια, αντιπάθεια, συγκρούσεις ή διαδικτυακές επαφές.

Σχήμα 3: Απεικόνιση Κοινωνικού Δικτύου

Τι είναι η ανάλυση Κοινωνικών Δικτύων

Είναι μία τεχνική η οποία σχετίζεται με τη θεωρία δικτύων και αποσκοπεί στην παρατήρηση, μέτρηση και απεικόνιση των σχέσεων ανάμεσα στις οντότητες ενός δικτύου το οποίο μπορεί να απαρτίζεται από ανθρώπους, ομάδες, επιχειρήσεις, ηλεκτρονικούς υπολογιστές κα. Καθώς και στην εξόρυξη πληροφοριών μέσα από αυτό.

Το επιστημονικό πεδίο της Ανάλυσης Κοινωνικών Δικτύων

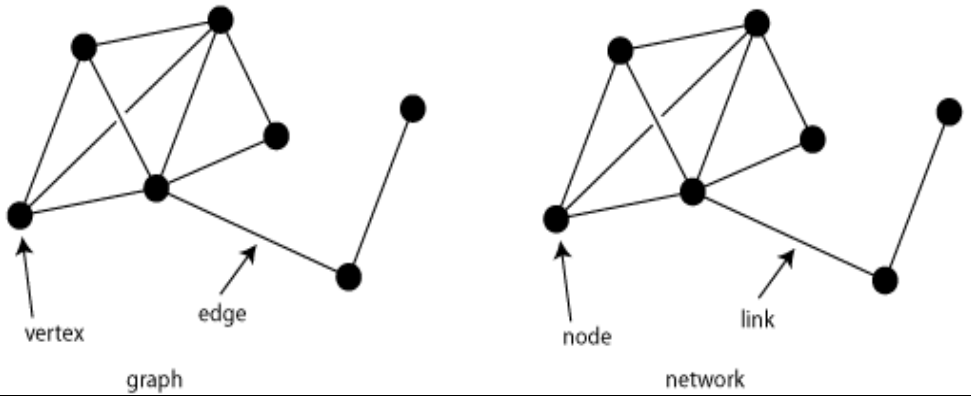
Η Ανάλυση Κοινωνικών Δικτύων είναι μία κοινωνική επιστήμη η οποία παρουσιάζει ιδιαίτερο ενδιαφέρον για τους κοινωνιολόγους. Η συμβολή των θετικών και κυρίως των φυσικών επιστημών, όπου τα δίκτυα πλέον περιγράφονται σαν Γράφοι έχει βοηθήσει σημαντικά στην κατανόησή τους.

1.1 Απεικόνιση Κοινωνικών Δικτύων

Η απεικόνιση κοινωνικών δικτύων γίνεται κυρίως με Γράφους και Πίνακες [3]

- **Γράφοι**

  Ένας γράφος ή γράφημα G=(V,E) αποτελείται από ένα σύνολο κόμβων (nodes), ή κορυφών V (vertices), και ένα σύνολο ακμών E (edges), που αποτελεί τις συνδέσεις των κόμβων. Συμβολίζουμε το πλήθος κόμβων και ακμών ως |V| και |E|, αντιστοίχως. Οι κόμβοι απεικονίζουν την οντότητα, τους ανθρώπους ή τις ομάδες, ενώ οι ακμές δείχνουν τις σχέσεις ή τις ροές μεταξύ τους.



Σχήμα 4: Γράφος & Δίκτυο

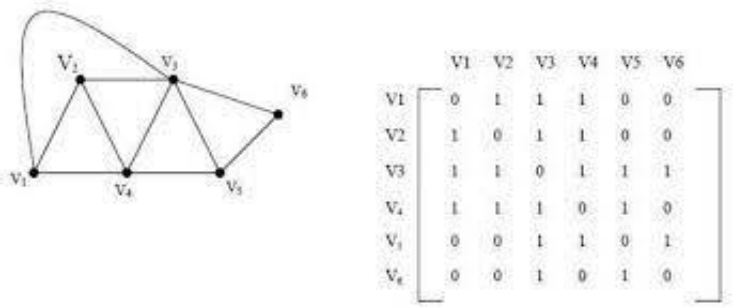
Αυτή η δομή είναι πολύ απλή, αποτελούμενη από έναν αριθμό κουκκίδων που ονομάζονται κορυφές στην ορολογία γραφημάτων και ονομάζονται κόμβοι στην ορολογία δικτύου. Μεταξύ ορισμένων από αυτές τις κουκκίδες υπάρχουν γραμμές που ονομάζονται ακμές στην ορολογία γραφημάτων ή σύνδεσμοι στην ορολογία δικτύου, όπως στο παραπάνω σχήμα. Έτσι, ένα γράφημα αποτελείται από κορυφές που συνδέονται με ακμές, ενώ ένα δίκτυο αποτελείται από κόμβους που συνδέονται με συνδέσμους. Η ορολογία δικτύου χρησιμοποιείται γενικά σε περιπτώσεις όπου θέλουμε να σκεφτούμε τη μεταφορά/αποστολή πραγμάτων κατά μήκος των συνδέσεων μεταξύ κόμβων, είτε αυτά τα πράγματα είναι φυσικά αντικείμενα (οδικά δίκτυα και σιδηροδρομικά δίκτυα) ή πληροφορίες (δίκτυα υπολογιστών και κοινωνικά δίκτυα).

Η ορολογία γραφήματος χρησιμοποιείται συχνότερα σε καταστάσεις όπου θέλουμε οι ακμές/οι σύνδεσμοι να αντιπροσωπεύουν άλλους τύπους σχέσεων μεταξύ των κορυφών/κόμβων. Ένα παράδειγμα που έχει τραβήξει την προσοχή πρόσφατα είναι το «γράφημα ενδιαφέροντος» στο οποίο οι κορυφές είναι άτομα και θέματα, και κάθε άκρη συνδέει ένα άτομο με ένα θέμα που τον ενδιαφέρει. Θα μπορούσαμε να πούμε ότι ένα κοινωνικό δίκτυο θα πρέπει πραγματικά να ονομάζεται γράφημα, καθώς συχνά το σκεφτόμαστε από την άποψη των σχέσεων μεταξύ των ανθρώπων και όχι των ενημερώσεων κατάστασης και των μηνυμάτων που αποστέλλονται μεταξύ τους. Στην πράξη, δεν υπάρχει ακριβής κανόνας για να αποφασιστεί ποιοι όροι θα χρησιμοποιηθούν, αλλά ευτυχώς δεν είναι πολύ δύσκολο να συμβαδίσουμε και με τους δύο τύπους ορολογίας.

- **Πίνακες**

  Χρησιμοποιούνται για να αναπαραστήσουν τους κόμβους ενός γράφου, οι οποίοι συνδέονται με άλλους κόμβους και αποτελούνται από τόσες γραμμές και στήλες όσοι είναι οι κόμβοι του δικτύου, με τα στοιχεία του να παριστάνουν τους δεσμούς ανάμεσα στους κόμβους.

  Η πιο απλή μορφή είναι οι δυαδικοί πίνακες (Σχήμα 5), όπου εάν υπάρχει σχέση ανάμεσα στους κόμβους V_i και V_j τότε το στοιχείο (V_i, V_j) είναι 1, αλλιώς είναι 0 (πίνακες γειτνίασης)



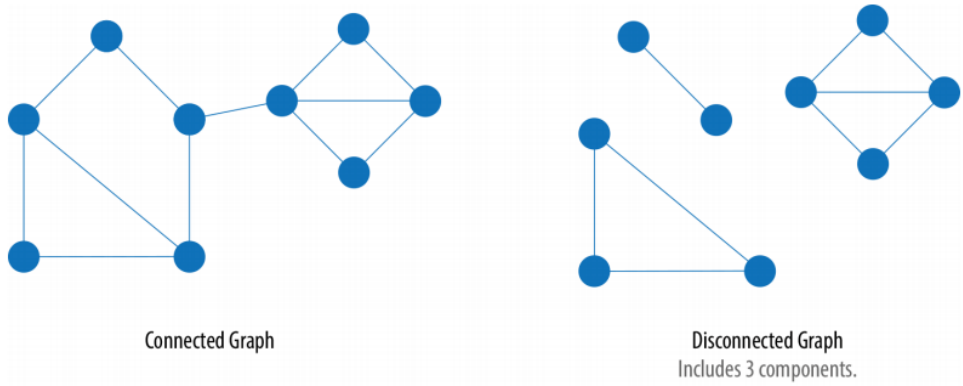
Σχήμα 5: Γράφος & Πίνακας Γειτνίασης

1.2 Δομή Δικτύων - Γράφων

Όπως προαναφέρθηκε ένα δίκτυο είναι ένα σύνολο κόμβων/nodes που είναι ή όχι συνδεδεμένοι μεταξύ τους με ακμές/links.

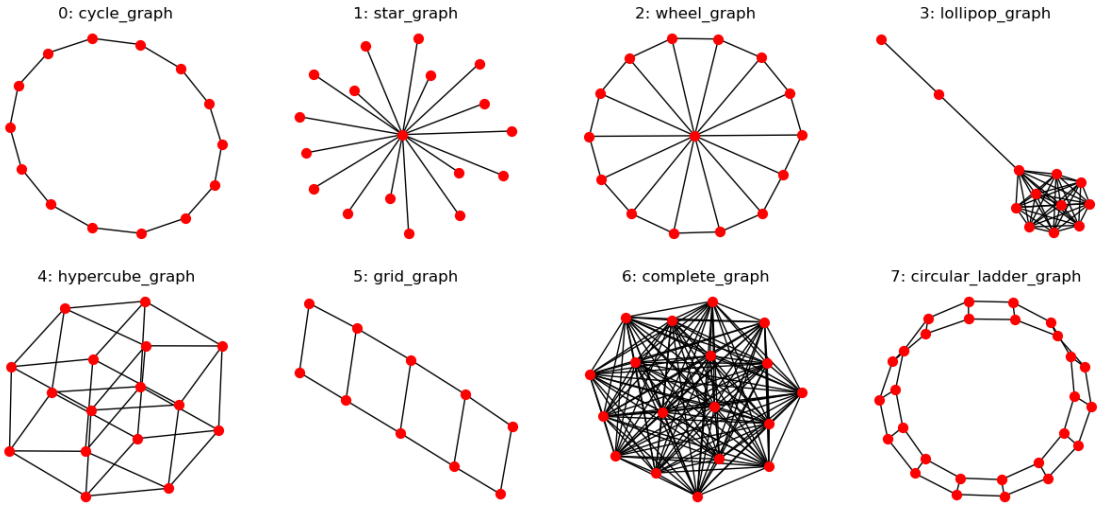
Παραδείγματα δομών δικτύων:

- Σύνδεση vs όχι – Σύνδεση



Σχήμα 6: Συνδεδεμένος vs Μη-Συνδεδεμένος Γράφος

- Διαφορετικές Δομές Δικτύων - Γράφων



Σχήμα 7: Σχήματα Γράφων

1.2.1 Μετρικές Δικτύων - Γράφων

➢ **Μέγεθος Δικτύου – Γράφου**

Ο αριθμός των ακμών του γράφου.

➢ **Διάμετρος**

Είναι το μήκος της μεγαλύτερης διαδρομής μεταξύ δυο οποιονδήποτε κόμβων. Η μακρύτερη δηλαδή απόσταση αυτών σε ένα Γράφο. Ονομάζεται και *εκκεντρικότητα*.

➢ **Γεωδαιτική απόσταση**

Το πλήθος των ακμών του συντομότερου μονοπατιού με το οποίο συνδέεται ένα ζεύγος κόμβων. Μπορεί να υπάρχουν πολλά μονοπάτια ελαχίστου μήκους που ενώνουν δύο κόμβους, αλλά μπορεί να μην υπάρχει και κανένα. Στην περίπτωση αυτή, το γράφημα δεν είναι συνεκτικό και η γεωδαιτική απόσταση των κόμβων που δεν συνδέονται είναι άπειρη. Σε κάθε άλλη περίπτωση, υπολογίζεται σε χρόνο O(n), όπου n ο αριθμός των ακμών στον Γράφο.

➢ **Συνοχή**

Ο βαθμός κατά τον οποίο τα μέλη του δικτύου συνδέονται μεταξύ τους με συνεκτικούς δεσμούς.

➢ **Δομική συνοχή**

Ο ελάχιστος αριθμός των μελών που εάν μετακινηθούν από το δίκτυο θα διακόψουν τη σύνδεση του δικτύου

➢ **Πυκνότητα**

Αντιπροσωπεύει την αναλογία μεταξύ των ακμών που υπάρχουν σε ένα γράφημα και του μέγιστου αριθμού ακμών που μπορεί να περιέχει το γράφημα. Συνήθως συμβολίζεται με D για Γράφο V κόμβων και E ακμών. Διακρίνουμε δυο περιπτώσεις. Την περίπτωση του κατευθυνόμενου γράφου και του μη - κατευθυνόμενου γράφου.

Μη-κατευθυνόμενος Γράφος V κόμβων:

$$D = \frac{2|E|}{|V|(|V|-1)}$$

Κατευθυνόμενος Γράφος V κόμβων:

$$D = \frac{|E|}{|V|(|V|-1)}$$

Κάποια τμήματα ενός δικτύου είναι πιο πυκνά στο εσωτερικό τους, δηλαδή υπάρχουν περισσότερες συνδέσεις σε σχέση με άλλα, αυτό συμβαίνει διότι οι κόμβοι τους είναι πιο δραστήριοι και επηρεάζουν τη ροή της πληροφορίας. Τα πυκνά τμήματα του Γράφου παρουσιάζουν ανθεκτικότητα και διατηρούν την συνοχή τους σε περίπτωση πρόσθεσης ή αφαίρεσης κόμβων ή ακμών.

- **Μήκος Μονοπατιού**
  Η απόσταση μεταξύ ενός ζεύγους κόμβων στο δίκτυο. *Μέσο μήκος μονοπατιού* είναι ο μέσος όρος των αποστάσεων μεταξύ όλων των ζευγαριών του δικτύου

1.2.2 Μέτρα Κεντρικότητας – Μετρικές Κόμβων

Στην ανάλυση γραφημάτων, η Κεντρικότητα [4] είναι μια πολύ σημαντική έννοια για τον εντοπισμό σημαντικών κόμβων σε ένα γράφημα. Χρησιμοποιείται για τη μέτρηση της σημασίας ή της «κεντρικότητας» όπως στο πόσο «κεντρικός» είναι ένας κόμβος στο γράφημα διαφόρων κόμβων. Το πόσο «κεντρικός» είναι ένας κόμβος εξαρτάται και από πια οπτική γωνία επιλέγουμε να ορίσουμε τη «σημασία» του. Έτσι έχουν προκύψει διαφορετικές κατηγορίες μετρικών κεντρικότητας. Παρακάτω αναφέρονται κάποιες από αυτές.

Κεντρικότητα Βαθμού (Degree Centrality)

Σε ένα Γράφο G(V, E), ορίζεται ως ο αριθμός των ακμών που προσπίπτουν σε έναν κόμβο (Freeman 1978). Σε ένα δίκτυο ο βαθμός ενός κόμβου Vi αντιπροσωπεύει τον αριθμό των ακμών που συνδέονται τον κόμβο που αντιστοιχεί στο Vi. Έστω N (Vi) το σύνολο των κόμβων που συνδέονται με το Vi, ο βαθμός κεντρικότητας [5] ενός κόμβου Vi δίνεται από:

$$C_D(V_i) = \frac{|N(V_i)|}{|V|-1}$$

Ιδιοδιανυσματική Κεντρικότητα (Eigenvector Centrality)

Σε ένα δίκτυο η σύνδεση με ένα κεντρικό - δημοφιλή κόμβο είναι σαφώς σημαντικότερη από την σύνδεση με ένα κόμβο χωρίς πολλές συνδέσεις. Σε ένα μέτρο κεντρικότητας λοιπόν είναι σημαντικό να υπολογίσουμε εκτός από το πλήθος των συνδέσεων και τη σημαντικότητα του κάθε κόμβου με τον οποίο υπάρχει σύνδεση. Η κεντρικότητα του ιδιοδιανύσματος [5] μετρά την κεντρική θέση ενός κόμβου ως συνάρτηση των κεντρικοτήτων των γειτόνων του. Εξηγεί την ιδέα ότι οι συνδέσεις με κόμβους υψηλής βαθμολογίας είναι πιο σημαντικές από αυτές με κόμβους με χαμηλή βαθμολογία. Έστω w_{ji} το βάρος της ακμής μεταξύ των κόμβων V_j και V_i και λ μια σταθερά, η κεντρική ιδιότητα του διανύσματος ενός κόμβου V_j δίνεται από:

$$C_E(V_i) = \frac{1}{\lambda} \sum_{V_j \in N(V_i)} w_{ji} \times C_E\,(V_j)$$

Κεντρικότητα Katz (Katz centrality)

Η κεντρικότητα Katz ενός κόμβου είναι ένα μέτρο της κεντρικότητας σε ένα δίκτυο. Παρουσιάστηκε για πρώτη φορά από τον Leo Katz το 1953 και χρησιμοποιείται για τη μέτρηση του σχετικού βαθμού επιρροής ενός κόμβου μέσα σε ένα κοινωνικό δίκτυο. Μετρά τη σχετική επίδραση κάθε κόμβου σε ένα δεδομένο δίκτυο λαμβάνοντας υπόψη τους άμεσους γειτονικούς κόμβους του καθώς και τους μη άμεσους γειτονικούς κόμβους που συνδέονται μέσω άμεσων γειτονικών κόμβων. Η κεντρικότητα Katz ενός κόμβου V_i υπολογίζεται ως:

$$C_{Katz}(V_i) = \alpha \sum_{j=1}^{n} A_{j,i} C_{Katz}\big(V_j\big) + \beta$$

όπου α είναι μια σταθερά που ονομάζεται συντελεστής απόσβεσης, που συνήθως θεωρείται μικρότερη από τη μεγαλύτερη ιδιοτιμή λ πχ $\alpha < 1/\lambda$ και το β είναι μια σταθερά πόλωσης, που ονομάζεται επίσης εξωγενές διάνυσμα, που χρησιμοποιείται για να αποφευχθούν οι τιμές μηδενικής κεντρικότητας. Με $\alpha \ \geq \lambda$, η κεντρική θέση τείνει να αποκλίνει. [6]

Βαθμική Κεντρικότητα Page (Pagerank centrality)

Η κεντρικότητα Katz έχει το εξής μειονέκτημα: αν ένα κόμβος έχει μεγάλη κεντρικότητα, τότε και οι υπόλοιποι κόμβοι συνδεδεμένοι με αυτόν έχουν και εκείνοι μεγάλη κεντρικότητα. Για παράδειγμα ο Παγκόσμιος Ιστός μπορεί να αναπαρασταθεί ως ένα κατευθυνόμενο γράφημα στο οποίο κάθε ιστοσελίδα αντιπροσωπεύεται από έναν κόμβο και οι ακμές που δείχνουν σε έναν κόμβο αντιπροσωπεύουν τους συνδέσμους που δείχνουν προς μια ιστοσελίδα και οι ακμές που δείχνουν μακριά από τον κόμβο αντιπροσωπεύουν τους συνδέσμους που δείχνουν προς άλλες ιστοσελίδες έτσι κάποιος κόμβος που συνδέεται με τη Google, ή την Yahoo που έχουν μεγάλη κεντρικότητα θα έχει και αυτός μεγάλη κεντρικότητα. Οι ιδρυτές της Google ανέπτυξαν έναν νέο αλγόριθμο για την αξιολόγηση της σημαντικότητας των σελίδων που συνδέονται με αυτή, που ήταν βελτίωση της κατά Katz κεντρικότητας, που τον ονόμασαν PageRank.

Η κεντρικότητα PageRank καθορίζει τη σημασία των κόμβων με βάση τον αριθμό και την ποιότητα των κόμβων που συνδέονται με αυτόν. Η πιο σχετική σελίδα θα αποφασιστεί όχι μόνο από τον βαθμό, αλλά και από τη σημασία του κόμβου που δείχνει προς τον εν λόγω κόμβο. Το PageRank βασίζεται βασικά στην πιθανότητα ενός ατόμου να σερφάρει στο διαδίκτυο κάνοντας τυχαία κλικ σε συνδέσμους, να σταματήσει. Η πιθανότητα το άτομο να συνεχίσει να κάνει κλικ στους συνδέσμους δίνεται από τον παράγοντα απόσβεσης και οι μελέτες δείχνουν ότι η ιδανική τιμή είναι 0,85. Ο τύπος δίνεται από:

$$PR_a = (1 - d) + d \sum_{b \in S_a} \frac{PR_b}{Ol_b}$$

Όπου S_a είναι ένα σύνολο από όλες τις ιστοσελίδες που δείχνουν σε μια σελίδα 'a' και Ol_b είναι οι έξω-σύνδεσμοι από τη σελίδα, d είναι ο συντελεστής απόσβεσης. [7]

Alpha κεντρικότητα (Alpha centrality)

Μια πιο γενικευμένη μορφή της κεντρικότητας Katz δημιούργησε ο P.Bonacich. Υπολογίζει την κεντρικότητα ενός κόμβου, λαμβάνοντας υπόψιν τη διαφορετική συνεισφορά κάθε άλλου με τον οποίο έρχεται σε επικοινωνία αλλά και τις επιρροές που δέχεται από το περιβάλλον εκτός δικτύου. Όρισε *α-centrality* $C_{i,j}(\alpha,\beta,n)$ ως ο συνολικός αριθμός εξασθενημένων διαδρομών μεταξύ των κόμβων i και j, με τα β και α να δίνουν τους συντελεστές εξασθένησης κατά μήκος των άμεσων ακμών (από i) και των έμμεσων ακμών (από τους ενδιάμεσους κόμβους) στη διαδρομή από i έως j, αντίστοιχα, και n είναι το μήκος της μεγαλύτερης διαδρομής.

Δεδομένου του πίνακα γειτνίασης του δικτύου A, ο πίνακας α-κεντρικότητας ορίζεται ως εξής:

$$C(\alpha,\beta,n) = \beta A + \beta\alpha_1 A^2 + \cdots + \beta\prod_{k=1}^{n}\alpha_k A^{n+1}$$

Ο πρώτος όρος δίνει τον αριθμό των μονοπατιών μήκους ένα (ακμές) από το i έως το j, ο δεύτερος δίνει τον αριθμό των μονοπατιών μήκους δύο κ.λπ. Παρόλο που το α_k κατά μήκος διαφορετικών ακμών σε ένα μονοπάτι θα μπορούσε κατ' αρχήν να είναι διαφορετικό, για λόγους απλότητας, τα θεωρούμε όλα ίσα:

$$\alpha_k = a, \qquad \forall k$$

Σε αυτή την περίπτωση, η σειρά συγκλίνει σε:

$$C(\alpha,\beta,n \to \infty) = \beta A(I-\alpha A)^{-1}$$

που κρατάει όσο $\alpha < 1/\lambda_1$ όπου λ_1 είναι η μεγαλύτερη χαρακτηριστική ρίζα του A. Ο υπολογισμός του λ_1 είναι δύσκολος, ειδικά για μεγάλα δίκτυα, τα οποία περιλαμβάνουν τα πιο πολύπλοκα δίκτυα του πραγματικού κόσμου. [8]

Κεντρικότητα Ενδιαμεσότητας (Betweenness Centrality)

Σε ένα δίκτυο η μετάδοση της πληροφορίας θέλουμε να γίνεται γρήγορα. Αυτό επιτυγχάνεται όταν η ροή της πληροφορίας ακολουθεί τη βέλτιστη - συντομότερη διαδρομή μεταξύ των συνδεδεμένων κόμβων. Πρέπει λοιπόν να ελέγξουμε τον κάθε κόμβο του δικτύου για τη συμμέτοχή του στις διαδρομές του δικτύου ώστε να ορίσουμε το πόσο ενεργός είναι στην μετάδοση της πληροφορίας. Η κεντρικότητα ενδιαμεσότητας ποσοτικοποιεί τον αριθμό των φορών που ένας κόμβος λειτουργεί ως γέφυρα κατά μήκος της συντομότερης διαδρομής μεταξύ δύο άλλων κόμβων. Έστω $\sigma(V_j,V_k)$, ο αριθμός των συντομότερων μονοπατιών από τον κόμβο V_j στον κόμβο V_k και $\sigma(V_j,V_k|V_i)$ ο αριθμός εκείνων των μονοπατιών που διέρχονται από τον κόμβο V_i. [5] Η ενδιάμεση κεντρικότητα ενός κόμβου V_i δίνεται από:

$$C_B(V_i) = \frac{\sum_{V_i \neq V_j \neq V_k \in V}\frac{\sigma(V_j,V_k|V_i)}{\sigma(V_j,V_k)}}{(|V|-1)(|V|-2)/2}$$

Κεντρικότητα Εγγύτητας (Closeness Centrality)

Η κεντρικότητα εγγύτητας σ_C βασίζεται στην ιδέα ότι οι κόμβοι με μικρή απόσταση από άλλους κόμβους μπορούν να διαδώσουν πληροφορίες πολύ παραγωγικά μέσω του δικτύου. Προκειμένου να υπολογιστεί η σ_C ενός κόμβου x συνοψίζονται οι αποστάσεις μεταξύ του κόμβου x και όλων των άλλων κόμβων του δικτύου. Χρησιμοποιώντας την αμοιβαία τιμή επιτυγχάνουμε ότι η τιμή της κεντρικότητας εγγύτητας αυξάνεται όταν μειώνεται η απόσταση σε έναν άλλο κόμβο, δηλαδή όταν βελτιώνεται η ενσωμάτωση στο δίκτυο. Αυτό μας δίνει το εξής:

$$\sigma_C(x) = \frac{1}{\sum_{i=1}^{n} d_G(x,i)}$$

όπου $d_G(x,i)$: Η απόσταση των κόμβων x και i. [4]

Συντελεστής Ομαδοποίησης (Clustering Coefficient)

Η μετρική αυτή υπολογίζει το κατά πόσο οι κόμβοι ενός Γράφου τείνουν να ομαδοποιούνται μεταξύ τους. Σε ένα πραγματικό δίκτυο πχ κοινωνικό δίκτυο έχει παρατηρηθεί ότι οι κόμβοι τείνουν να δημιουργούν στενά δεμένες ομάδες που χαρακτηρίζονται από μια σχετικά υψηλή πυκνότητα δεσμών. Παρακάτω αναλύονται οι προηγούμενοι προτεινόμενοι συντελεστές ομαδοποίησης για μη σταθμισμένα και σταθμισμένα δίκτυα με βάση τον συντελεστή συσχέτισης Pearson $\rho(i,j)$. Αυτά που βασίζονται στον συντελεστή μερικής συσχέτισης $\rho^{-partial}(i,j)$ υπολογίζονται αναλόγως. [9]

Συντελεστές ομαδοποίησης για μη σταθμισμένα δίκτυα

Για να κατασκευάσουμε ένα μη σταθμισμένο λειτουργικό δίκτυο, βάζουμε μια ακμή μεταξύ των κόμβων i και j $(1 \le i \ne j \le N)$ αν και μόνο αν $\rho(i,j) \ge \theta$, όπου θ είναι ένα προκαθορισμένο όριο. Το δίκτυο που δημιουργείται είναι μη κατευθυνόμενο. Συμβολίζουμε τον πίνακα γειτνίασης του δικτύου με $A = (\alpha_{ij})$, όπου $1 \le i,j \le N_{ROI}$. Με άλλα λόγια $\alpha_{ij} = 1$ αν (i,j) είναι ακμή, $\alpha_{ij} = 0$ αλλιώς.

Ο συντελεστής ομαδοποίησης αντιπροσωπεύει την αφθονία των συνδεδεμένων τριγώνων σε ένα δίκτυο (Watts και Strogatz, 1998). Ο συντελεστής τοπικής ομαδοποίησης του κόμβου i ορίζεται από:

$$C_i^{unw} = \frac{Αριθμός\ συνδεδεμένων\ τριγώνων\ συμπεριλαμβανομένου\ του\ κόμβου\ i}{\frac{k_i(k_i-1)}{2}}$$

$$= \frac{\sum_{1 \le j < \ell \le N_{ROI}} \alpha_{ij}\alpha_{i\ell}\alpha_{j\ell}}{\frac{k_i(k_i-1)}{2}} \qquad (1)$$

Όπου $k_i = \sum_{j=1}^{N_{ROI}} \alpha_{ij} = \sum_{j=1}^{N_{ROI}} \alpha_{ji}$ ο βαθμός του κόμβου i δηλαδή ο αριθμός των ακμών με τις οποίες ο κόμβος i είναι γειτονικός. Ο παρονομαστής της εξίσωσης (1) αντιπροσωπεύει τον μεγαλύτερο δυνατό αριθμό τριγώνων στα οποία ανήκει ο κόμβος i. Να σημειωθεί επίσης ότι $0 \le C_i^{unw} \le 1$ $(1 \le i \le N_{ROI})$ και ότι C_i^{unw} είναι απροσδιόριστο αν $k_i = 0\ or\ 1$. Ο συνολικός συντελεστής ομαδοποίησης για ολόκληρο το δίκτυο, που συμβολίζεται με C^{unw}, δίνεται από τον μέσο όρο των C_i^{unw} σε όλους τους κόμβους.

Εξαιρούμε τους κόμβους με $k_i \leq 1$ από τον υπολογισμό του C^{unw}, $0 \leq C^{unw} \leq 1$
Παρόμοια με άλλους τύπους δικτύων, τα περισσότερα δίκτυα εγκεφάλου, ανατομικά ή λειτουργικά, έχουν μεγάλες τιμές C^{unw} σε σύγκριση με τα τυχαιοποιημένα δίκτυα (Bullmore and Sporns, 2009, Bassett and Sporns, 2017).

Συντελεστές ομαδοποίησης για σταθμισμένα δίκτυα

Ένα σταθμισμένο λειτουργικό δίκτυο ορίζεται λαμβάνοντας υπόψη $\rho(i,j)$ ως το βάρος της ακμής (i,j) Επειδή δεν έχουμε καθιερωμένες μεθόδους για να αντιμετωπίσουμε αρνητικά σταθμισμένες ακμές (βλέπε Rubinov και Sporns, 2011) και είναι σύνηθες να απορρίπτουμε ακμές με αρνητική τιμή $\rho(i,j)$ (Rubinov and Sporns, 2010, Kaiser, 2011), ο σταθμισμένος πίνακας γειτνίασης δίνεται από $w_{ij} = \rho(i,j)$ αν $\rho(i,j) > 0$ και $w_{ij} = 0$ διαφορετικά. Ως σημεία αναφοράς, εξετάζουμε τρεις παραλλαγές του σταθμισμένου συντελεστή ομαδοποίησης που χρησιμοποιούνται συνήθως στη βιβλιογραφία (Saramäki et al., 2007, Rubinov and Sporns, 2010, 2011, Wang et al., 2017). Σημειώνουμε με (α_{ij}) τον πίνακα γειτνίασης του μη σταθμισμένου δικτύου που προκύπτει, αγνοώντας το βάρος των ακμών στο σταθμισμένο δίκτυο. Με άλλα λόγια, θέσαμε $\alpha_{ij} = 1$ αν $w_{ij} > 0$ (ισοδύναμα $\rho(i,j) > 0$) και $\alpha_{ij} = 0$ διαφορετικά.

Ο συντελεστής τοπικής ομαδοποίησης του κόμβου i που προτείνεται από τους Barrat et al., (2004) δίνεται από τον τύπο:

$$C_i^{wei,B} = \frac{1}{s_i(k_i-1)} \sum_{\substack{1\leq j,\ell\leq N_{\mathrm{ROI}} \\ j,\ell\neq i}} \frac{w_{ij}+w_{i\ell}}{2} a_{ij}a_{i\ell}a_{j\ell}, \qquad (2)$$

Όπου $s_i = \sum_{j=1}^{N_{\mathrm{ROI}}} w_{ij}$ είναι η ισχύς του κόμβου.

Πρέπει να σημειωθεί ότι $a_{ij}a_{i\ell}a_{j\ell} = 1$ αν και μόνο αν οι κόμβοι i, j και ℓ σχηματίζουν ένα τρίγωνο στο μη σταθμισμένο δίκτυο, $a_{ij}a_{i\ell}a_{j\ell} = 0$ διαφορετικά. Ο μέσος όρος του $C_i^{wei,B}$ σε όλους τους κόμβους ορίζει τον συνολικό σταθμισμένο συντελεστή ομαδοποίησης (global weighted clustering coefficient) που συμβολίζεται με $C^{wei,B}$.

Ο συντελεστής τοπικής ομαδοποίησης (local clustering coefficient) που προτείνεται από τους Onnela et al., (2005), δίνεται από (Saramäki et al., 2007):

$$C_i^{wei,O} = \frac{1}{k_i(k_i-1)} \sum_{\substack{1\leq j,\ell\leq N_{\mathrm{ROI}} \\ j,\ell\neq i}} \frac{\left(w_{ij}w_{i\ell}w_{j\ell}\right)^{1/3}}{max_{i'j'}w_{i'j'}}, \qquad (3)$$

Ο παράγοντας $max_{i'j'}w_{i'j'}$ ομαλοποιεί το $C_i^{wei,O}$ μεταξύ 0 και 1 και εμποδίζει την κλιμάκωση όταν η κλίμακα w_{ij} αλλάζει (δηλαδή, όταν w_{ij} για όλα τα $1 \leq i,j \leq N_{\mathrm{ROI}}$ πολλαπλασιάζεται με την ίδια σταθερά). Ο αντίστοιχος συνολικός συντελεστής ομαδοποίησης, που συμβολίζεται με $C^{wei,O}$, δίνεται από τον μέσο όρο των $C^{wei,O}$ σε όλους τους κόμβους.

Ο τοπικός συντελεστής ομαδοποίησης (local clustering coefficient) που προτείνεται από τους Zhang και Horvath (2005) γράφεται ως:

$$C_i^{wei,Z} = \frac{1}{max_{i'j'}w_{i'j'}} \frac{\sum_{\substack{1\leq j,\ell\leq N_{\mathrm{ROI}} \\ j,\ell\neq i}} w_{ij}w_{i\ell}w_{j\ell}}{\sum_{\substack{1\leq j,\ell\leq N_{\mathrm{ROI}} \\ j,\ell\neq i;j\neq\ell}} w_{ij}w_{i\ell}}, \qquad (4)$$

Ο αντίστοιχος συνολικός συντελεστής ομαδοποίησης, που συμβολίζεται με $C^{wei,Z}$ δίνεται από τον μέσο όρο των $C_i^{wei,Z}$ σε όλους τους κόμβους.

1.3 Υφιστάμενα Προβλήματα Κοινωνικών Δικτύων

1.3.1 Ανίχνευση Κοινοτήτων Δικτύων – Γράφων

Η εξέλιξη της επιστήμης των δικτύων έχει φέρει σημαντικές προόδους στην κατανόησή μας περί πολύπλοκων συστημάτων. Ένα από τα πιο σχετικά χαρακτηριστικά των Γράφων που αντιπροσωπεύουν πραγματικά συστήματα είναι η δομή της κοινότητας ή η ομαδοποίηση, δηλαδή η οργάνωση κορυφών σε συστάδες, με πολλές ακμές να ενώνουν κορυφές του ίδιου συμπλέγματος και συγκριτικά λίγες ακμές που ενώνουν κορυφές διαφορετικών συστάδων. Τέτοιες ομάδες, ή κοινότητες, μπορούν να θεωρηθούν ως αρκετά ανεξάρτητα διαμερίσματα ενός Γράφου, που παίζουν παρόμοιο ρόλο. Το Facebook, για παράδειγμα, είναι ένα μεγάλο κοινωνικό δίκτυο, όπου περισσότεροι από ένα δισεκατομμύριο άνθρωποι συνδέονται μέσω εικονικών γνωριμιών. Ένα άλλο παράδειγμα είναι το Διαδίκτυο, το φυσικό δίκτυο υπολογιστών, δρομολογητών και μόντεμ που συνδέονται μέσω καλωδίων ή ασύρματων σημάτων. Πολλά άλλα παραδείγματα προέρχονται από τη βιολογία, τη φυσική, τα οικονομικά, τη μηχανική, την επιστήμη των υπολογιστών, την οικολογία, το μάρκετινγκ, τις κοινωνικές και πολιτικές επιστήμες κ.λπ. Το πρόβλημα της ανίχνευσης κοινοτήτων είναι πολύ δύσκολο και δεν έχει λυθεί ακόμη ικανοποιητικά, παρά την τεράστια προσπάθεια των επιστημόνων που εργάζεται πάνω σε αυτό τα τελευταία χρόνια.

Ο εντοπισμός κοινοτήτων μπορεί να προσφέρει πληροφορίες σχετικά με τον τρόπο οργάνωσης του δικτύου. Μας επιτρέπει να εστιάσουμε σε περιοχές που έχουν κάποιο βαθμό αυτονομίας μέσα στο Γράφο. Βοηθά στην ταξινόμηση των κόμβων, με βάση τον ρόλο τους σε σχέση με τις κοινότητες στις οποίες ανήκουν. Για παράδειγμα, μπορούμε να διακρίνουμε τους κόμβους που είναι πλήρως ενσωματωμένοι στα συμπλέγματά τους από τους κόμβους στα όρια των συστάδων, οι οποίοι μπορεί να λειτουργήσουν ως διαμεσολαβητές μεταξύ των ενοτήτων και, στην περίπτωση αυτή, θα μπορούσαν να διαδραματίσουν σημαντικό ρόλο τόσο στη συγκράτηση των ενοτήτων μαζί, όσο και στη δυναμική διάδοση διαδικασιών σε όλο το δίκτυο.

Υπάρχουν πολλές μεθοδολογίες και αλγόριθμοι για την ανίχνευση κοινοτήτων σε Γράφους. Μπορούν να ομαδοποιηθούν σε κατηγορίες, με βάση διαφορετικά κριτήρια, όπως η πραγματική λειτουργική μέθοδος (Fortunato, 2010) ή η υποκείμενη έννοια της κοινότητας (Coscia et al., 2011). Στις περισσότερες εφαρμογές, ωστόσο, χρησιμοποιούνται κυρίως οι πιο δημοφιλείς μεθοδολογίες. Παρακάτω παρουσιάζεται μια ανάλυση ορισμένων εξ αυτών των μεθοδολογιών. [10][11]

Αριθμός Συστάδων

Η μόνη πληροφορία που είναι διαθέσιμη εκ των προτέρων σε οποιονδήποτε αλγόριθμο είναι η δομή του δικτύου, δηλαδή ποια ζεύγη κόμβων συνδέονται μεταξύ τους και ποια όχι. Οποιαδήποτε εικόνα σχετικά με τη δομή της κοινότητας υποτίθεται ότι παρέχεται ως αποτέλεσμα της διαδικασίας. Φυσικά, θα ήταν πολύτιμο να έχουμε εκ των προτέρων κάποιες πληροφορίες για την άγνωστη διαίρεση του δικτύου, καθώς θα μπορούσε κανείς να μειώσει σημαντικά τον τεράστιο όγκο των πιθανών λύσεων και να αυξήσει την πιθανότητα επιτυχούς αναγνώρισης των κοινοτήτων. Πως μπορούμε να γνωρίζουμε όμως ποιος είναι ο αριθμός των συστάδων που υπάρχουν;

Αναφέρουμε εν συντομία μερικές ευρετικές τεχνικές, για μεθόδους που βασίζονται σε στατιστικές αρχές. Ας ορίσουμε ως q των αριθμό των συστάδων. Έχει αποδειχθεί πρόσφατα ότι στο μοντέλο φυτεμένου διαμερίσματος (planted partition model) ο q μπορεί να συναχθεί σωστά μέχρι το όριο ανιχνευσιμότητας από τα φάσματα δύο πινάκων: του μη αναδρομικού πίνακα (non-backtracking matrix) **B** (Krzakala et al., 2013) και του πίνακα ροής (flow matrix) **F** (Newman, 2013). Είναι πίνακες $2m \times 2m$, όπου m είναι ο αριθμός των ακμών του γραφήματος. Κάθε άκρο θεωρείται και προς τις δύο κατευθύνσεις, αποδίδοντας κατευθυνόμενες άκρες $2m$ και υποδεικνύεται με τον συμβολισμό $i \to j$, που σημαίνει ότι η άκρη πηγαίνει από την κορυφή i στην κορυφή j. Τα στοιχεία τους:

$$B_{i\to j,r\to s} = \delta_{is}\left(1 - \delta_{jr}\right), \qquad (1)$$

και

$$F_{i\to j,r\to s} = \frac{\delta_{is}\left(1-\delta_{jr}\right)}{k_i - 1}, \qquad (2)$$

Στην εξίσωση (2) k_i είναι ο βαθμός της κορυφής i. Άρα τα στοιχεία του **F** είναι βασικά τα στοιχεία του **B**, κανονικοποιημένα ως προς τον βαθμό κόμβου. Αυτό γίνεται για να ληφθούν υπόψη οι ετερογενείς κατανομές βαθμών που παρατηρούνται στα περισσότερα πραγματικά δίκτυα. Και οι δύο πίνακες έχουν μη μηδενικά στοιχεία μόνο για κάθε ζεύγος ακμών που σχηματίζουν μια κατευθυνόμενη διαδρομή από την πρώτη κορυφή της μιας ακμής στη δεύτερη της άλλης ακμής. Για να γίνει αυτό, οι ακμές πρέπει να προσπίπτουν σε μία κορυφή. Στην πραγματικότητα, ο μη αναδρομικός πίνακας **B** είναι απλώς ο πίνακας γειτνίασης των (κατευθυνόμενων) άκρων του γραφήματος.

Ο πίνακας B οφείλεται σε μια σύνδεση με τις ιδιότητες των μη οπισθοδρομικών περιπάτων (non-backtracking walks). Ένας μη οπισθοδρομικός περίπατος (Angel et al., 2015) είναι μια διαδρομή κατά μήκος των άκρων ενός γραφήματος που επιτρέπεται να επιστρέψει σε μια κορυφή που είχε επισκεφτεί προηγουμένως μόνο αφού έχουν επισκεφθεί τουλάχιστον άλλες δύο κορυφές. Απαγορεύονται οι άμεσες επιστροφές όπως 1 -> 2 -> 1. Τα στοιχεία της k-ης δύναμης του B δίνουν τον αριθμό των μη οπισθοδρομικών βημάτων μήκους k από μια (κατευθυνόμενη) άκρη του γραφήματος σε μια άλλη και το ίχνος του πίνακα ισχύος τον αριθμό των κλειστών μη οπισθοδρομικών βημάτων μήκους k ξεκινώντας από οποιαδήποτε δεδομένη (κατευθυνόμενη) άκρη.

Μια αξιοσημείωτη ιδιότητα και των δύο πινάκων είναι ότι σε δίκτυα με ομοιογενείς ομάδες (δηλαδή παρόμοιου μεγέθους και εσωτερικής πυκνότητας ακμών) οι περισσότερες ιδιοτιμές, οι οποίες είναι γενικά σύνθετες, περικλείονται από έναν κύκλο με κέντρο στην αρχή και ότι ο αριθμός των ιδιοτιμών βρίσκεται εκτός του κύκλου είναι ένας καλός αντιπρόσωπος του αριθμού των κοινοτήτων του δικτύου (Krzakala et al., 2013, Newman, 2013). Για το B η ακτίνα του κύκλου δίνεται από την τετραγωνική ρίζα $\sqrt{c}$ της κύριας ιδιοτιμής c, η οποία μπορεί να αποκλίνει για δίκτυα με ετερογενείς κατανομές βαθμών. Για το F δίνεται από τετραγωνική ρίζα: $\sqrt{\langle k/(k-1)\rangle/\langle k\rangle}$, που δεν είναι ποτέ μεγαλύτερο από 1.

Δυστυχώς, ο υπολογισμός των ιδιοτιμών του μη οπισθοδρομικού πίνακα ή του πίνακα ροής είναι μακρύς. Και οι δύο είναι πίνακες $2m \times 2m$. Ο πίνακας γειτνίασης A έχει στοιχεία $n \times n$, επομένως τα B και F είναι μεγαλύτερα κατά συντελεστή $\langle k\rangle^2$, όπου $\langle k\rangle$ είναι ο μέσος βαθμός του δικτύου. Ένας κατά προσέγγιση αλλά αξιόπιστος υπολογισμός των φασμάτων απαιτεί χρόνο που κλιμακώνεται υπέρ-γραμμικά (περίπου τετραγωνικά) με το μέγεθος δικτύου n. Επομένως, το πρόβλημα είναι δυσεπίλυτο για γραφήματα με αριθμό ακμών της τάξης των εκατομμυρίων ή μεγαλύτερα. Επίσης, εάν οι κοινότητες έχουν διαφορετικά μεγέθη και πυκνότητες ακμών, όπως συμβαίνει στα περισσότερα δίκτυα που συναντώνται σε εφαρμογές, ο κύριος όγκος των ιδιοτιμών μπορεί να μην έχει κυκλικό σχήμα και μπορεί να γίνει προβληματικός ο εντοπισμός ιδιοτιμών που βρίσκονται εκτός του όγκου.

Εαν κάποιος μπορεί να προσδιορίσει ένα σύνολο υποσχόμενων τιμών q, από προκαταρκτικές πληροφορίες ή μέσω υπολογισμών όπως αυτοί που περιγράφονται παραπάνω, είναι προτιμότερο να εκτελούνται περιορισμένες εκδόσεις μεθόδων ομαδοποίησης, αναζητώντας λύσεις μόνο μεταξύ των κατατμήσεων με αυτούς τους αριθμούς των κοινοτήτων, παρά να αφήνουμε τις μεθόδους να ανακαλύψουν μόνες τους q, κάτι που μπορεί να οδηγήσει σε παραπλανητικά ή λάθος αποτελέσματα.

Συναινετική ομαδοποίηση

Αρκετές από τις τεχνικές ομαδοποίησης είναι στοχαστικές με αποτέλεσμα να μη δίνουν μια και μόνο απάντηση. Συνηθισμένες περιπτώσεις είναι όταν η επιθυμητή λύση αντιστοιχεί σε ακραία σημεία μιας συνάρτησης κόστους, και βρίσκεται μέσω τεχνικών προσέγγισης. Έτσι ανάλογα με τις τυχαίες τροφοδοτήσεις και την επιλογή των αρχικών συνθηκών παίρνουμε και τα ανάλογα

αποτελέσματα. Το ίδιο χαρακτηριστικό συναντάμε και σε άλλες τεχνικές που δεν βασίζονται στη βελτιστοποίηση. Διότι προκειμένου να γίνει τελικώς η επιλογή μεταξύ ισοδύναμων επιλογών που συναντώνται κατά τον υπολογισμό υιοθετούνται ορισμένοι κανόνες ισοπαλίας.

Για να ταξινομήσουμε λοιπόν ένα συγκεκριμένο διαμέρισμα και να απορρίψουμε όλα τα υπόλοιπα χρησιμοποιούμε κάποια αντικειμενικά κριτήρια. Για παράδειγμα, σε αλγόριθμους που βασίζονται στη βελτιστοποίηση, θα μπορούσε κανείς να επιλέξει τη λύση που δίνει τη μεγαλύτερη τιμή της συνάρτησης για βελτιστοποίηση. Για άλλες τεχνικές δεν υπάρχει ξεκάθαρο κριτήριο.

Μια αρκετά καλή προσέγγιση είναι ο συνδυασμός των πληροφοριών των διαφορετικών εξόδων σε ένα νέο διαμέρισμα (partition). Η ομαδοποίηση συναίνεσης (Goder και Filkov, 2008, Strehl και Ghosh, 2002, Topchy et al., 2005) βασίζεται σε αυτή την ιδέα. Ο στόχος είναι η αναζήτηση για ένα διαμέρισμα συναίνεσης (consensus partition), το οποίο είναι καλύτερο από τα διαμερίσματα εισόδου. Η συναινετική ομαδοποίηση είναι ένα δύσκολο πρόβλημα συνδυαστικής βελτιστοποίησης. Μια εναλλακτική άπληστη στρατηγική (Strehl και Ghosh, 2002) βασίζεται στον συναινετικό πίνακα, ο οποίος είναι ένας πίνακας που βασίζεται στη συνύπαρξη κόμβων σε κοινότητες των διαμερισμάτων εισόδου. Ο συναινετικός πίνακας χρησιμοποιείται ως είσοδος για την τεχνική ομαδοποίησης γραφημάτων που υιοθετήθηκε, οδηγώντας σε ένα νέο σύνολο διαμερισμάτων, τα οποία παράγουν έναν νέο συναινετικό πίνακα, έως ότου επιτευχθεί τελικά ένα μοναδικό διαμέρισμα, το οποίο δεν αλλάζει με περαιτέρω επαναλήψεις. Τα βήματα της διαδικασίας είναι τα εξής:

Το σημείο εκκίνησης είναι ένα δίκτυο-Γράφος G με n κόμβους και έναν αλγόριθμο ομαδοποίησης A.

1. Εφαρμόζουμε τον **A** στο G n_P φορές, δίνοντας διαμερίσματα n_P.
2. Υπολογίζουμε τον συναινετικό πίνακα **D**: όπου D_{ij} είναι ο αριθμός των διαμερισμάτων στα οποία οι κορυφές i και j του G έχουν εκχωρηθεί στην ίδια κοινότητα, διαιρούμενο με n_P.
3. Όλες οι καταχωρήσεις του **D** κάτω από ένα επιλεγμένο όριο τ μηδενίζονται.
4. Εφαρμόζουμε το **A** σε **D** n_P φορές, δίνοντας n_P διαμερίσματα.
5. Εάν τα διαμερίσματα είναι όλα ίσα, σταματάμε. Διαφορετικά, επιστρέφουμε στο 2.

Αλγόριθμος Συναινετικής Ομαδοποίησης

Δεδομένου ότι ο συναινετικός πίνακας είναι σταθμισμένος, ο αλγόριθμος **A** πρέπει να μπορεί να εφαρμόζεται σε σταθμισμένα δίκτυα, ακόμα κι αν το γράφημα είναι δυαδικό. Πολλοί από τους δημοφιλείς αλγορίθμους έχουν επεκτάσεις οι οποίες καλύπτουν την περίπτωση σταθμισμένων δικτύων-Γράφων.

Ως αποτέλεσμα της μεταβλητότητας των διαμερισμάτων είναι ότι τα ίδια τα διαμερίσματα γίνονται παράγοντας βελτίωσης της απόδοσης. Τα αποτελέσματα της παραπάνω διαδικασίας εξαρτώνται από την επιλογή του κατώτερου ορίου τ και τον αριθμό των διαμερισμάτων εισόδου n_P, τα οποία μπορούν να επιλεγούν δοκιμάζοντας την απόδοση σε δίκτυα αναφοράς (Lancichinetti και Fortunato, 2012). Τέλος η συναινετική ομαδοποίηση είναι μια τεχνική για την ανίχνευση κοινοτήτων με αξιόλογα αποτελέσματα σε εξελισσόμενα δίκτυα όπως για παράδειγμα τα κοινωνικά δίκτυα.

Φασματική ομαδοποίηση

Από τις ευρέως γνωστές μεθόδους ομαδοποίησης στην ανάλυση δεδομένων είναι η φασματική ομαδοποίηση. Είναι μια τεχνική απλή στην εφαρμογή της και μπορεί να επιλυθεί με απλές

μεθόδους γραμμικής άλγεβρας δίνοντας αποτελέσματα τα οποία σε πολλές περιπτώσεις ξεπερνούν σε ποιότητα άλλους παραδοσιακούς αλγορίθμους ομαδοποίησης όπως ο k-means. Έστω ότι έχουμε ένα σύνολο n οντοτήτων $x_1, x_2, \dots, x_n$ η φασματική ομαδοποίηση χωρίζει το σύνολο σε συστάδες χρησιμοποιώντας τα ιδιοδιανύσματα πινάκων. Αν υποθέσουμε ότι οι οντότητες είναι σημεία στο χώρο ή οι κόμβοι ενός Γράφου, η φασματική ομαδοποίηση μετασχηματίζει το αρχικό σύνολο οντοτήτων σε ένα σύνολο σημείων όπου οι συντεταγμένες τους είναι στοιχεία ιδιοδιανυσμάτων. Ο μετασχηματισμός του συνόλου μέσω αυτής της τεχνικής κάνει πιο εμφανείς τις ιδιότητές του. Έπειτα το σύνολο αυτό μπορεί να ομαδοποιηθεί μέσω συνηθισμένων τεχνικών όπως η ομαδοποίηση k-means.

Οι πίνακες Laplace αποτελούν τα βασικά εργαλεία που χρησιμοποιούνται στην φασματική ομαδοποίηση. Υπάρχουν διάφορες παραλλαγές των πινάκων αυτών με κάποιες βασικές ιδιότητες. Δεν υπάρχει μοναδικός ορισμός στο ποιος πίνακας είναι Laplace. Παρακάτω παρουσιάζεται η φασματική ομαδοποίηση με πίνακα Laplace σύμφωνα με τον Ulrike von Luxburg 2007.

Κάποιες συμβάσεις για τη συνέχεια: Υποθέτουμε πάντα ότι το G είναι ένα μη κατευθυνόμενο, σταθμισμένο γράφημα με πίνακα βάρους W, όπου $w_{ij} = w_{ji} \geq 0$. Τα ιδιοδιανύσματα ενός πίνακα, δεν υποθέτουμε απαραίτητα ότι είναι πάντα κανονικοποιημένα. Για παράδειγμα, το σταθερό διάνυσμα $\mathbb{1}$ και ένα πολλαπλάσιο του $\alpha\mathbb{1}$ για κάποιο $\alpha \neq 0$ θα θεωρηθούν ως τα ίδια ιδιοδιανύσματα. Οι ιδιοτιμές θα ταξινομούνται πάντα σε αύξουσα σειρά σύμφωνα με την πολλαπλότητά τους. Ο όρος «τα πρώτα k ιδιοδιανύσματα» αναφέρεται στα ιδιοδιανύσματα που αντιστοιχούν στις k μικρότερες ιδιοτιμές.

Το μη κανονικοποιημένο γράφημα Laplace

Ο μη κανονικοποιημένος γραφικός πίνακας Laplace ορίζεται ως:

$$L = D - W$$

Ο πίνακας L ικανοποιεί τις ακόλουθες ιδιότητες:

1. Για κάθε διάνυσμα $f \in \mathbb{R}^n$ έχουμε:

$$f'Lf = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left(f_i - f_j\right)^2$$

2. Το L είναι συμμετρικό και θετικά ημιορισμένο.
3. Η μικρότερη ιδιοτιμή του L είναι 0, και το αντίστοιχο ιδιοδιάνυσμα είναι το σταθερό διάνυσμα $\mathbb{1}$.
4. Το L έχει n μη αρνητικές, πραγματικές ιδιοτιμές $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

Το μη κανονικοποιημένο γράφημα Laplace οι ιδιοτιμές και τα ιδιοδιανύσματά του μπορούν να χρησιμοποιηθούν για να περιγράψουν πολλές ιδιότητες γραφημάτων. Ένα σημαντικό παράδειγμα για τη φασματική ομαδοποίηση είναι το ακόλουθο:

Έστω G ένα μη κατευθυνόμενο γράφημα με μη αρνητικά βάρη. Τότε η πολλαπλότητα k της ιδιοτιμής 0 του L ισούται με τον αριθμό των συνεκτικών συνιστωσών $A_1, \dots, A_k$ στο γράφημα. Ο ιδιοχώρος της ιδιοτιμής 0 εκτείνεται από τα διανύσματα δεικτών $\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_k}$ αυτών των συνιστωσών.

Για τις αποδείξεις των παραπάνω βλέπε [12]

Το κανονικοποιημένο γράφημα Laplace

Υπάρχουν δύο πίνακες οι οποίοι ονομάζονται κανονικοποιημένοι πίνακες Laplace. Και οι δύο πίνακες σχετίζονται στενά μεταξύ τους και ορίζονται ως:

$$L_{\mathrm{sym}} := D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$
$$L_{\mathrm{rw}} := D^{-1} L = I - D^{-1} W.$$

Ο πρώτος πίνακας συμβολίζεται με L_{sym} καθώς είναι συμμετρικός πίνακας και ο δεύτερος με L_{rw} καθώς σχετίζεται στενά με έναν τυχαίο περίπατο. Στη συνέχεια αναφέρουμε κάποιες ιδιότητες των L_{sym} και L_{rw}.

1. Για κάθε διάνυσμα $f \in \mathbb{R}^n$ έχουμε:

$$f' L_{sym} f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2$$

2. Το λ είναι ιδιοτιμή του L_{rw} με ιδιοδιάνυσμα u αν και μόνο αν το λ είναι ιδιοτιμή του L_{sym} με ιδιοδιάνυσμα $w = D^{1/2} u$.

3. Το λ είναι μια ιδιοτιμή του L_{rw} με ιδιοδιάνυσμα u αν και μόνο αν το λ και u λύνουν το γενικευμένο πρόβλημα ιδιοτιμών $Lu = \lambda Du$.

4. Το 0 είναι μια ιδιοτιμή του L_{rw} με το σταθερό διάνυσμα $\mathbb{1}$ να απότελεί ιδιοδιάνυσμα. Το 0 είναι μια ιδιοτιμή του L_{sym} με ιδιοδιάνυσμα $D^{1/2}\mathbb{1}$.

5. Τα L_{sym} και L_{rw} είναι θετικά ημιορισμένα και έχουν n μη αρνητικές πραγματικές ιδιοτιμές $0 = \lambda_1 \leq \cdots \leq \lambda_n$.

Όπως συμβαίνει με το μη κανονικοποιημένο γράφημα Laplace, η πολλαπλότητα της ιδιοτιμής 0 του κανονικοποιημένου γραφήματος Laplace σχετίζεται με τον αριθμό των συνεκτικών συνιστωσών του γράφου. Έτσι έχουμε:

Έστω G ένα μη κατευθυνόμενο γράφημα με μη αρνητικά βάρη. Τότε η πολλαπλότητα k της ιδιοτιμής 0 τόσο του L_{rw} όσο και του L_{sym} ισούται με τον αριθμό των συνεκτικών συνιστωσών $A_1, \ldots, A_k$ στο γράφημα. Για το L_{rw}, ο ιδιοχώρος του 0 εκτείνεται από τα διανύσματα δεικτών $\mathbb{1}_{A_i}$ αυτών των συνιστωσών. Για το L_{sym}, ο ιδιοχώρος του 0 εκτείνεται από τα διανύσματα $D^{1/2}\mathbb{1}_{A_i}$.

Για τις αποδείξεις των παραπάνω βλέπε [12]

Μέθοδοι που βασίζονται στη Στατιστική Συμπερασματολογία

Στην ανίχνευση κοινοτήτων μπορούν να συμβάλουν και τα εργαλεία που χρησιμοποιούνται στην Στατιστική Συμπερασματολογία (Statistical Inference). Το μοντέλο στοχαστικού μπλοκ (stochastic block model SBM) το οποίο είναι ένα παραγωγικό μοντέλο για τυχαία γραφήματα κοινοτήτων είναι αυτό που χρησιμοποιείτε κατά κόρον. Η μη κανονικοποιημένη μέγιστη πιθανότητα καταγραφής ότι ένα δεδομένο διαμέρισμα g ανήκει σε q ομάδες του δικτύου G δίνεται από το τυπικό μοντέλο στοχαστικού μπλοκ ως εξής:

$$\mathcal{L}_S(G \mid g) = \sum_{r,s=1}^{q} e_{rs} \log \left(\frac{e_{rs}}{n_r n_s} \right)$$

όπου e_{rs} είναι ο αριθμός των ακμών από την ομάδα r στην ομάδα s, n_r και n_s ο αριθμός των κόμβων στο $r(s)$ και το άθροισμα εφαρμόζεται σε όλα τα ζεύγη ομάδων και στην περίπτωση του $r = s$). Η συγκεκριμένη έκδοση του μοντέλου στοχαστικού μπλοκ δεν προσμετρά τον βαθμό ετερογένειας που υπάρχει στα αληθινά δίκτυα, έχοντας σαν αποτέλεσμα την όχι και τόσο καλή περιγραφή της δομής των ομάδων τους.

Εδώ θα έρθουν οι Karrer και Newman (2011) προτείνοντας ένα μοντέλο στοχαστικού μπλοκ διορθωμένου βαθμού (*degree-corrected stochastic block model* DCSBM). Στο μοντέλο αυτό οι βαθμοί των κόμβων παραμένουν σταθεροί κατά μέσο όρο, εισάγοντας κατάλληλες παραμέτρους. Η μη κανονικοποιημένη μέγιστη πιθανότητα καταγραφής:

$$\mathcal{L}_{DC}(G \mid g) = \sum_{r,s=1}^{q} e_{rs} \log\left(\frac{e_{rs}}{e_r e_s}\right)$$

όπου e_r (e_s) είναι το άθροισμα των βαθμών των κόμβων στο r (s).
Το μειονέκτημα του παραπάνω μοντέλου είναι ότι πρέπει να καθοριστεί εξ αρχής ο αριθμός q των ομάδων, που συνήθως στα αληθινά δίκτυα δεν τον γνωρίζουμε.

Άλλες μέθοδοι ανίχνευσης κοινοτήτων βασίζονται στη **βελτιστοποίηση**, στη **δυναμική**, στη **σημασία**, στη **δυναμική ομαδοποίηση** κ.α. Το πια μέθοδο θα ακολουθήσει κάποιος, μιας και οι περισσότερες τεχνικές υιοθετούν παρόμοιες ιδέες για τις κοινότητες, εξαρτάται από τον τρόπο που θέλουμε να αναζητηθούν οι ομάδες.

1.3.2 Διάχυση πληροφοριών στα κοινωνικά δίκτυα

Στη σημερινή εποχή με την συνεχόμενη εξέλιξη των κοινωνικών δικτύων, και τη δημιουργία ολοένα και περισσότερων μέσων, όπως π.χ. το Facebook το 2004, το YouTube το 2005, το Twitter το 2006 κ.α., οι τρόποι με τους οποίους οι άνθρωποι λαμβάνουν τις πληροφορίες έχουν αλλάξει. Παλαιότερα τα άτομα ήταν παθητικοί δέκτες της πληροφορίας, τώρα όμως ο καθένας έχει τη δύναμη να μεταδώσει και ο ίδιος τις πληροφορίες που επιθυμεί καθιστώντας τον εαυτό του ενεργό παράγοντα στη μετάδοση της πληροφορίας. Ο μεγάλος αυτός όγκος πληροφοριών καθώς και ο τρόπος που μεταδίδονται αποτελεί ένα από τα ερευνητικά πεδία της ανάλυσης κοινωνικών δικτύων, και αποκαλείται διάχυση πληροφορίας στα κοινωνικά δίκτυα.

Η ανάλυση κοινωνικών δικτύων παρέχει καινοτόμες τεχνικές για την ανάλυση των αλληλεπιδράσεων μεταξύ των οντοτήτων δίνοντας έμφαση στις κοινωνικές σχέσεις. Η διάχυση στο κοινωνικό δίκτυο μπορεί να αναφέρεται στη διάδοση πληροφοριών μεταξύ διασυνδεδεμένων κόμβων ή οντοτήτων σε ένα δίκτυο. Ο ρυθμός και η ένταση της διάχυσης εξαρτώνται από την τοπολογία του δικτύου και την προετοιμασία των παραμέτρων του δικτύου. Οι μεμονωμένοι κόμβοι λειτουργούν ως πηγή κινήτρων για άλλους στη διαδικασία διάχυσης. Ένα παράδειγμα είναι το μοντέλο επιδημίας, είναι ένα από τα βασικά μοντέλα διάχυσης που βοηθά στην ανάλυση της μετάδοσης μολυσματικών ασθενειών από το ένα άτομο στο άλλο μέσω κοινωνικών συνδέσεων.

Για την ανάλυση της δομής και των σχέσεων κοινωνικής αλληλεπίδρασης χρησιμοποιούνται έννοιες των γραφημάτων και των δικτύων. Το γράφημα περιέχει κόμβους, οι οποίοι αναπαρίστανται ως άτομα και οι ακμές υποδηλώνουν τις σχέσεις μεταξύ τους. Ένα δίκτυο που έχει περισσότερες συνδέσεις με άλλους δικτυωμένους κόμβους προωθεί την καλύτερη ανταλλαγή ιδεών μαζί με τη διάδοση πληροφοριών εντός του ίδιου δικτύου.

Ένα κοινωνικό δίκτυο είναι δυναμικό και έχει την τάση να εξελίσσεται συνεχώς. Για παράδειγμα τα μέσα κοινωνικής δικτύωσης που προαναφέραμε, ένα οδικό δίκτυο ο παγκόσμιος ιστός κ.α. είναι σύνθετα δίκτυα με εκατομμύρια κόμβους με ακανόνιστη δομή που αλλάζουν δυναμικά σε σχέση με το χρόνο. Η διάχυση πληροφορίας σε τέτοια δίκτυα επηρεάζεται από αρκετούς παράγοντες όπως η πυκνότητα των συνδέσεων δικτύου, η προσαρμοστικότητα ενός κόμβου σε μια νέα συμπεριφορά, η δομή του δικτύου, ακόμα και από την ελκυστικότητα της πληροφορίας και τις παραμέτρους του δικτύου. Πολλές είναι οι προσεγγίσεις για την ανάλυση της διάχυσης πληροφοριών σε κοινωνικές δομές. Οι περισσότερες μελέτες διερευνούν ποιοι παράγοντες επηρεάζουν τη διάχυση πληροφοριών, ποιες πληροφορίες διαχέονται πιο γρήγορα

και τον τρόπο με τον οποίο διαχέονται οι πληροφορίες. Σύμφωνα με την βιβλιογραφία που σχετίζεται με τα ζητήματα διάχυσης πληροφοριών, μπορούμε να τα κατηγοριοποιήσουμε σε επεξηγηματικά μοντέλα και σε μοντέλα πρόβλεψης. [13][14]

Επεξηγηματικά μοντέλα

Σε ένα κοινωνικό δίκτυο η πληροφορία μεταδίδεται μεταξύ αλληλεπιδράσεων των χρηστών ή κόμβων. Τα συνηθέστερα ερωτήματα που προκύπτουν είναι: ποιοι είναι οι κύριοι παράγοντες που επηρεάζουν τη διάχυση πληροφοριών; ποιος κόμβος έχει τη μεγαλύτερη επιρροή; γιατί οι πληροφορίες διαχέονται με τον τρόπο που διαχέονται; Για παράδειγμα, ορισμένοι κόμβοι αρνούνται να δεχτούν πληροφορίες, κάποιοι αρνούνται να διαδώσουν πληροφορίες και μερικοί αποδέχονται και διαδίδουν πληροφορίες. Τα μοντέλα αυτά σκοπεύουν να εξετάσουν τη διαδικασία διάχυσης πληροφοριών και να αποσαφηνίσουν τους παράγοντες που την επηρεάζουν σε μια προσπάθεια να εξηγήσουν αυτό το φαινόμενο. Το μοντέλο της επιδημίας που προαναφέραμε μας χρησιμεύει στην προκυμμένη περίπτωση διότι η διαδικασία διάχυσης πληροφοριών μπορεί να θεωρηθεί με τον ίδιο τρόπο όπως μια διαδικασία εξάπλωσης επιδημίας. Έτσι τα υπάρχοντα μοντέλα επιδημίας μπορούν να μας βοηθήσουν. Τα βασικά μοντέλα είναι το μοντέλο **SI** (Suceptible Infected), το μοντέλο **SIS** (Suceptible Infected Suceptible), το μοντέλο **SIR** (Suceptible Infected Removed) και το μοντέλο **SIRS** (Suceptible Infected Removed Suceptible).

SI model

Έστω N ο συνολικός αριθμός ατόμων, το N χωρίζεται σε δύο κατηγορίες σε ευαίσθητα S (susceptible) και σε μολυσμένα I (infected). Σε μια χρονική στιγμή t το $s(t)$ είναι το ποσοστό του συνολικού πληθυσμού που είναι ευαίσθητο, και το $i(t)$ το ποσοστό των μολυσμένων και το λ αντιπροσωπεύει το ημερήσιο ποσοστό επαφής δηλαδή οι ευαίσθητοι χρήστες που έχουν μολυνθεί από τους μολυσμένους χρήστες στο συνολικό πληθυσμό, όπου $s(t) + i(t) = 1$. Έτσι έχουμε $Ns(t) + Ni(t) = N$. Συμπεραίνουμε ότι θα μολυνθούν $\lambda s(t)$ ευαίσθητοι χρήστες. Εάν οι μολυσμένοι χρήστες είναι $Ni(t)$, τότε θα υπάρχουν $\lambda s(t)Ni(t)$, ευαίσθητοι χρήστες που θα μολυνθούν την ημέρα. Το λNsi αντιπροσωπεύει την αύξηση του αριθμού των ασθενών ανά ημέρα. Δηλαδή $N\frac{di}{dt} = \lambda Nsi$, $\frac{di}{dt} = \lambda si$ και $s + i = 1$. Τη χρονική στιγμή $t = 0$, η αναλογία των ασθενών είναι i_0 και το μοντέλο **SI** ορίζεται από τις παρακάτω εξισώσεις:

$$\frac{di}{dt} = \lambda i(1 - i), \qquad (1)$$

$$i(0) = i_0, \qquad (2)$$

SIS model

Το προηγούμενο μοντέλο SI δεν είναι ιδιαίτερα πρακτικό διότι δε λαμβάνει υπόψιν του τους ασθενείς που θεραπεύονται. Το μοντέλο SIS χρησιμοποιεί τις ίδιες μεταβλητές με το SI, επιπροσθέτως υποθέτει ότι m είναι το ποσοστό των ασθενών που θεραπεύονται. Η αύξηση της αλλαγής στον αριθμό των ασθενών μπορεί να εκφραστεί ως $N\frac{di}{dt} = \lambda Nsi - \mu Ni$, όπου λNsi είναι η αύξηση του αριθμού των ασθενών ανά ημέρα και μNi είναι η αύξηση του αριθμού των θεραπευμένων ασθενών ανά ημέρα. Το μοντέλο **SIS** ορίζεται από τις παρακάτω εξισώσεις:

$$\frac{di}{dt} = \lambda i(1 - i) - \mu i, \qquad (3)$$

$$i(0) = i_0, \qquad (4)$$

Για $\mu = 0$ το μοντέλο SIS είναι στην ουσία το μοντέλο SI.

SIR model

Το μοντέλο αυτό το οποίο καθιερώθηκε από τους Kermack και McKendrick λαμβάνει υπόψιν του ότι όταν ένας ασθενής θεραπεύεται μπορεί από εδώ και πέρα να μην νοσήσει ξανά. Έτσι διαιρεί τον συνολικό πληθυσμό σε τρεις κατηγορίες σε S και σε I όπως τα προηγούμενα μοντέλα SI και SIS, αλλά και σε R (Removed) δηλαδή όσους έχουν αποκτήσει ανοσία. Έτσι $s(t) + i(t) + r(t) = 1$. Επίσης υποθέτει ότι $s(0) = s_0$, $i(0) = i_0$, $r(0) = 0$ και $\frac{ds}{dt} + \frac{di}{dt} + \frac{dr}{dt} = 0$. Η ημερήσια αύξηση του αυξανόμενου αριθμού των άνοσων χρηστών εκφράζεται με $N\frac{dr}{dt} = \mu N i$. Το μοντέλο **SIR** ορίζεται από τις παρακάτω εξισώσεις:

$$\frac{ds}{dt} = -\lambda si, \qquad (5)$$

$$\frac{di}{dt} = \lambda i(1 - i) - \mu i, \qquad (6)$$

$$\frac{dr}{dt} = \mu i, \qquad (7)$$

SIRS model

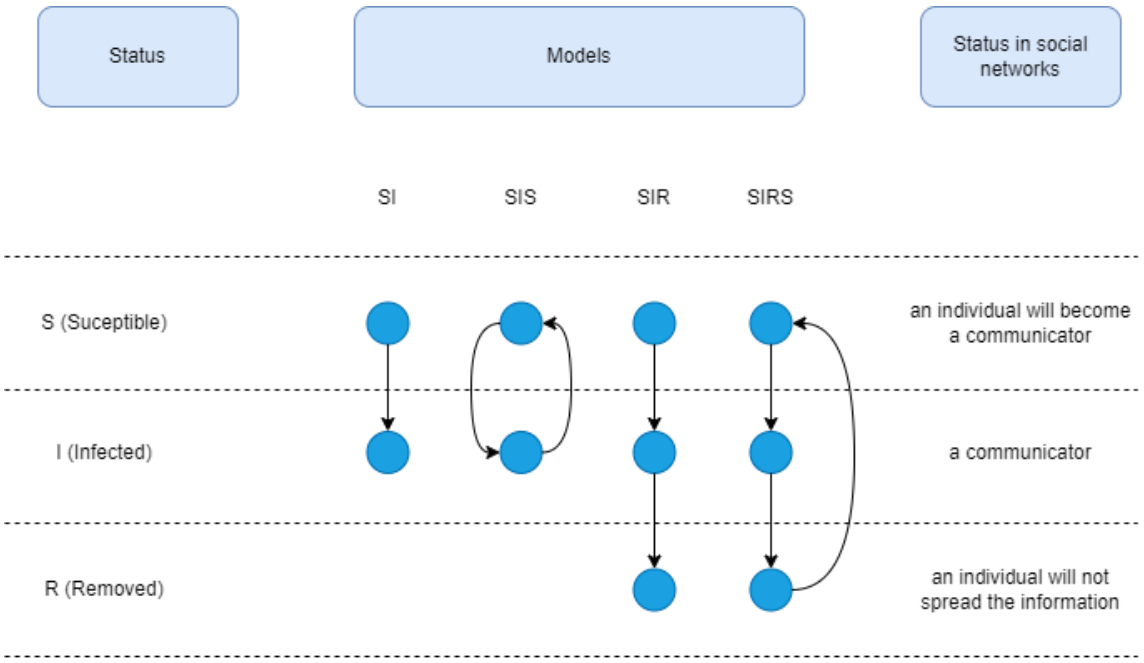
Το μοντέλο αυτό πιστεύει ότι ένας θεραπευμένος χρήστης μπορεί να γίνει ευαίσθητος S χρήστης με πιθανότητα α. Το μοντέλο **SIRS** ορίζεται από τις παρακάτω εξισώσεις:

$$\frac{ds}{dt} = -\lambda si + \alpha r, \qquad (8)$$

$$\frac{di}{dt} = \lambda i(1 - i) - \mu i, \qquad (9)$$

$$\frac{dr}{dt} = \mu i - \alpha r, \qquad (10)$$

Στο Σχήμα 8 παρουσιάζεται η σύγκριση των τεσσάρων βασικών μοντέλων επιδημίας που προαναφέρθηκαν. Δείχνει τη διαδικασία διάχυσης ενός ιού σε μια επιδημία καθώς επίσης παραλληλίζει και την κατάσταση των χρηστών στα κοινωνικά δίκτυα. Έτσι αντιλαμβανόμαστε καλύτερα πως τα μοντέλα επιδημίας μπορούν να χρησιμοποιηθούν στην έρευνα διάχυσης πληροφοριών.



Σχήμα 8: Σύγκριση των τεσσάρων βασικών μοντέλων επιδημίας (βλ. [13])

Η διάχυση πληροφοριών στα κοινωνικά δίκτυα μπορεί να έχει κοινά χαρακτηριστικά με την εξάπλωση μια πανδημίας έχει όμως και άλλα επιπλέον χαρακτηριστικά όπως ο χρόνος, το πόσο δυνατή είναι η σχέση μεταξύ των οντοτήτων, το περιεχόμενο της πληροφορίας κ.α. τα οποία δημιουργούν διαφορές με το μοντέλο της πανδημίας. Έτσι πολλοί επιστήμονες έχοντας σαν βάση τα μοντέλα αυτά, ενισχύοντας και κάνοντας βελτιώσεις ανέπτυξαν νέα μοντέλα όπως τα **SEIR** (Suceptible Exposur Infected Removed), **S-SEIR** (Single layer-SEIR), **SCIR** (Suceptible Contacted Infected Removed), **irSIR** (Infection Recovery SIR), **FSIR** (Fractional SIR) και **ESIS** (Emotional Suceptible Infected Suceptible) για τα οποία όμως δε θα γίνει περαιτέρω ανάλυση.

Μοντέλα πρόβλεψης

Στα κοινωνικά δίκτυα οι σημαντικές πληροφορίες διαδίδονται γρηγορότερα σε σχέση με άλλες. Είναι πολλές οι περιπτώσεις που αρκετοί κλάδοι θα ήθελαν να προβλέψουν τον τρόπο και την ταχύτητα διάδοσης μιας πληροφορίας ώστε να επωφεληθούν από αυτό στο μέλλον. Έτσι έχουν αναπτυχθεί προγνωστικά μοντέλα όπου βάση κάποιων παραγόντων χρησιμοποιούνται για την πρόβλεψη της μελλοντικής διαδικασίας διάχυσης πληροφοριών στα κοινωνικά δίκτυα. Θα αναφέρουμε τρία από αυτά, το ανεξάρτητο μοντέλο διαδοχικών πληροφοριών (Independent Cascade Model), το γραμμικό μοντέλο κατωφλίου (Linear Threshold Model), και το μοντέλο θεωρίας παιγνίων (Game Theory Model).

Independent Cascade Model (ICM)

Στο μοντέλο αυτό θεωρείτε ότι η διαδικασία διάχυσης γίνεται με διακριτά βήματα στο χρόνο. Οποιαδήποτε χρονική στιγμή t ένας κόμβος μπορεί να είναι ενεργός είτε ανενεργός. Οι ενεργοί κόμβοι ενεργοποιούν τους γειτονικούς κόμβους αλλά αυτό μπορεί να γίνει μόνο μια φορά στα βήματα του χρόνου καθώς επίσης και αν ένας κόμβος ενεργοποιηθεί δεν μπορεί να απενεργοποιηθεί σε μεταγενέστερα βήματα, γίνεται μόνο μία φορά. Αρχικά η διαδικασία ξεκινάει με κάποιους ενεργούς κόμβους. Έτσι ένας κόμβος v που ενεργοποιείται τη στιγμή t επιχειρεί να ενεργοποιήσει τους ανενεργούς γειτονικούς κόμβους του τη στιγμή $t+1$. Η πιθανότητα επιτυχίας είναι p_{vu}. Η διαδικασία εκτελείται μέχρι το χρονικό βήμα όπου δεν ενεργοποιούνται άλλοι κόμβοι. [15]

Linear Threshold Model (LTM)

Στο συγκεκριμένο μοντέλο κάθε ενεργός κόμβος τη χρονική στιγμή t έχει ένα κατώφλι ενεργοποίησης. Όλοι οι γειτονικοί κόμβοι προσπαθούν να ενεργοποιήσουν τους γειτονικούς τους. Όλοι οι κόμβοι έχουν έναν βαθμό επιρροής. Όταν ο βαθμός επιρροής των ενεργών κόμβων ξεπεράσει το κάτω όριο ενεργοποίησης του ανενεργού κόμβου, ο κόμβος αυτός θα ενεργοποιηθεί την επόμενη χρονική στιγμή $t + 1$. Τέλος, οι ενεργοί γειτονικοί κόμβοι μπορούν να ενεργοποιήσουν τους γείτονες κόμβους πάνω από μια φορές.

Game Theory Model (GTM)

Η θεωρία παιγνίων είναι μια στρατηγική που βασίζεται στη μεγιστοποίηση του κέρδους. Απευθύνεται κυρίως σε μεγάλο πλήθος ατόμων η συγκεκριμένων ομάδων. Το κόστος είναι ο κύριος παράγοντας επίδρασης στη μετάδοση μιας πληροφορίας. Δηλαδή το πόσο κοστίζει σε κάποιον και τι όφελος προκύπτει από τη μετάδοση της πληροφορίας είναι αυτό που θα καθορίσει εάν στο τέλος θα διαδοθεί η όχι και σε τι βαθμό. Ένας ακόμα παράγοντας που προστέθηκε είναι η σχέση μεταξύ των οντοτήτων. Όσο πιο στενή είναι η σχέση που υπάρχει μεταξύ τους, τόσο πιο εύκολα διαδίδεται η πληροφορία.

Όπως έχουμε προαναφέρει ένα κοινωνικό δίκτυο είναι δυναμικό. Εάν συγκρίνουμε αυτά τα τρία μοντέλα πρόβλεψης προκύπτει ότι το ανεξάρτητο μοντέλο διαδοχικών πληροφοριών (ICM) λειτουργεί με βάση τους αποστολείς της πληροφορίας. Το γραμμικό μοντέλο κατωφλίου (LTM) με βάση τους δέκτες. Ενώ το μοντέλο θεωρίας παιγνίων (GTM) λαμβάνει υπόψιν του το κόστος ή το κέρδος ολόκληρου του δικτύου δίνοντας καλύτερα αποτελέσματα για την πρόβλεψη σε δυναμικά δίκτυα.

2 Το Πρόβλημα Πρόγνωσης Ακμών σε Κοινωνικά Δίκτυα

Ως μέρος της πρόσφατης αύξησης της έρευνας για μεγάλα, πολύπλοκα δίκτυα και τις ιδιότητές τους, έχει δοθεί μεγάλη προσοχή στην υπολογιστική ανάλυση των κοινωνικών δικτύων. Δομές των οποίων οι κόμβοι αντιπροσωπεύουν άτομα ή άλλες οντότητες ενσωματωμένες σε ένα κοινωνικό πλαίσιο και των οποίων οι ακμές αντιπροσωπεύουν την αλληλεπίδραση, τη συνεργασία ή την επιρροή μεταξύ οντοτήτων. Φυσικά παραδείγματα κοινωνικών δικτύων περιλαμβάνουν το σύνολο όλων των επιστημόνων σε έναν συγκεκριμένο κλάδο, με τις άκρες να ενώνουν ζευγάρια που έχουν συγγράψει μαζί εργασίες. Το σύνολο όλων των εργαζομένων σε μια μεγάλη εταιρεία, με άκρες που ενώνουν ζευγάρια εργασίας σε ένα κοινό έργο, ή μια συλλογή ηγετών επιχειρήσεων, με τις άκρες να ενώνουν ζευγάρια που έχουν υπηρετήσει μαζί σε ένα εταιρικό διοικητικό συμβούλιο. Η διαθεσιμότητα μεγάλων, λεπτομερών συνόλων δεδομένων που κωδικοποιούν τέτοια δίκτυα έχει ενθαρρύνει την εκτενή μελέτη των βασικών ιδιοτήτων τους και τον εντοπισμό επαναλαμβανόμενων δομικών χαρακτηριστικών.

Τα κοινωνικά δίκτυα είναι πολύ δυναμικά αντικείμενα. Αυξάνονται και αλλάζουν γρήγορα με την πάροδο του χρόνου, μέσω της προσθήκης νέων άκρων, που σηματοδοτούν την εμφάνιση νέων αλληλεπιδράσεων στην υποκείμενη κοινωνική δομή. Η κατανόηση των μηχανισμών με τους οποίους εξελίσσονται είναι ένα θεμελιώδες ερώτημα που δεν είναι ακόμα καλά κατανοητό και αποτελεί το κίνητρο για την εργασία μας εδώ. Ορίζουμε και μελετάμε ένα βασικό υπολογιστικό πρόβλημα που βασίζεται στην εξέλιξη των κοινωνικών δικτύων, [16] το πρόβλημα της πρόγνωσης ακμών:

Δεδομένου ενός στιγμιότυπου ενός κοινωνικού δικτύου τη στιγμή t, επιδιώκουμε να προβλέψουμε με ακρίβεια τις ακμές που θα προστεθούν στο δίκτυο κατά το διάστημα από τη στιγμή t έως μια δεδομένη μελλοντική στιγμή t΄.

Στην πραγματικότητα, το πρόβλημα πρόγνωσης ακμών θέτει το ερώτημα: σε ποιο βαθμό μπορεί να μοντελοποιηθεί η εξέλιξη ενός κοινωνικού δικτύου χρησιμοποιώντας χαρακτηριστικά εγγενή στο ίδιο το δίκτυο; Σκεφτείτε ένα δίκτυο συν-συγγραφέων μεταξύ επιστημόνων. Υπάρχουν πολλοί λόγοι, εξωγενείς για το δίκτυο, για τους οποίους δύο επιστήμονες που δεν έχουν γράψει ποτέ μια εργασία μαζί θα το κάνουν τα επόμενα χρόνια: για παράδειγμα, μπορεί να τύχει να έρθουν γεωγραφικά κοντά όταν ένας από αυτούς αλλάξει θεσμό. Τέτοιες συνεργασίες μπορεί να είναι δύσκολο να προβλεφθούν.

Αλλά αισθάνεται κανείς επίσης ότι ένας μεγάλος αριθμός νέων συνεργασιών υπονοείται από την τοπολογία του δικτύου: δύο επιστήμονες που είναι «κοντά» στο δίκτυο θα έχουν κοινούς συναδέλφους και θα ταξιδεύουν σε παρόμοιους κύκλους. Αυτό υποδηλώνει ότι οι ίδιοι είναι πιο πιθανό να συνεργαστούν στο εγγύς μέλλον. Στόχος μας είναι να κάνουμε αυτή τη διαισθητική ιδέα ακριβή και να κατανοήσουμε ποια μέτρα «εγγύτητας» σε ένα δίκτυο οδηγούν στις πιο ακριβείς προβλέψεις συνδέσεων.

Βρίσκουμε ότι ορισμένα μέτρα εγγύτητας οδηγούν σε προβλέψεις που ξεπερνούν την πιθανότητα κατά παράγοντες από 40 έως 50, υποδεικνύοντας ότι η τοπολογία του δικτύου περιέχει πράγματι λανθάνουσα πληροφορία από την οποία μπορεί να συναχθεί το συμπέρασμα για μελλοντικές αλληλεπιδράσεις. Επιπλέον, ορισμένες σχετικά λεπτές μετρήσεις που περιλαμβάνουν άπειρα αθροίσματα σε μονοπάτια στο δίκτυο συχνά ξεπερνούν τις πιο άμεσες μετρήσεις, όπως οι αποστάσεις και οι αριθμοί της συντομότερης διαδρομής των κοινών γειτόνων.

Εκτός από τον ρόλο του ως βασικό ερώτημα στην εξέλιξη των κοινωνικών δικτύων, το πρόβλημα πρόβλεψης συνδέσμων θα μπορούσε να σχετίζεται με μια σειρά από ενδιαφέρουσες τρέχουσες εφαρμογές των κοινωνικών δικτύων.

Όλο και περισσότερο, για παράδειγμα, οι ερευνητές στην τεχνητή νοημοσύνη και την εξόρυξη δεδομένων υποστήριξαν ότι ένας μεγάλος οργανισμός, όπως μια εταιρεία, μπορεί να επωφεληθεί από τις αλληλεπιδράσεις εντός του άτυπου κοινωνικού δικτύου μεταξύ των μελών του. Αποτελεσματικές μέθοδοι για την πρόβλεψη συνδέσμων θα μπορούσαν να χρησιμοποιηθούν για την ανάλυση ενός τέτοιου κοινωνικού δικτύου και να προτείνουν πολλά υποσχόμενες αλληλεπιδράσεις ή συνεργασίες που δεν έχουν ακόμη χρησιμοποιηθεί στον οργανισμό.

Με διαφορετικό τρόπο, η έρευνα στον τομέα της ασφάλειας άρχισε πρόσφατα να δίνει έμφαση στο ρόλο της ανάλυσης των κοινωνικών δικτύων, με κίνητρο σε μεγάλο βαθμό το πρόβλημα της παρακολούθησης των τρομοκρατικών δικτύων. Η πρόβλεψη σύνδεσης σε αυτό το πλαίσιο επιτρέπει σε κάποιον να υποθέσει ότι συγκεκριμένα άτομα συνεργάζονται ακόμη κι αν η αλληλεπίδρασή τους δεν έχει παρατηρηθεί άμεσα.

3 Νευρωνικά Δίκτυα σε Γράφους

Ένα από τα σημαντικότερα επιτεύγματα της ανάπτυξης της Τεχνητής Νοημοσύνης όπου η διαδικασία μάθησης είναι ζωτικής σημασίας είναι τα **Νευρωνικά Δίκτυα** (Neural Networks). Αντιγράφοντας χαρακτηριστικά από το βιολογικό νευρικό σύστημα και ειδικότερα του ανθρώπινου εγκεφάλου, έχουν αναπτύξει δυνατότητες όπως την αναπαράσταση εξαρτήσεων καθώς και την πρόβλεψη της κλάσης άγνωστων παρατηρήσεων. Με αποτέλεσμα να εφαρμόζονται κατά κόρον σε πολλές επιστήμες όπως η οικονομία, η ιατρική, η πληροφορική, η χημεία κλπ. Τα δεδομένα είναι αυτά που καθοδηγούν ένα Νευρωνικό Δίκτυο, άρα και ένα μοντέλο ενός νευρωνικού δικτύου είναι αποτέλεσμα της επεξεργασίας των δεδομένων έτσι ώστε να μην προβαίνει σε αυθαίρετες υποθέσεις. [17]

Οι Γράφοι - γραφήματα είναι ένα είδος δομής δεδομένων που μοντελοποιεί ένα σύνολο αντικειμένων και τις σχέσεις τους. Πρόσφατα, οι έρευνες για την ανάλυση γραφημάτων με μηχανική μάθηση τυγχάνουν ολοένα και μεγαλύτερης προσοχής λόγω της μεγάλης εκφραστικής τους δύναμης, δηλαδή τα γραφήματα μπορούν να χρησιμοποιηθούν ως ένδειξη μεγάλου αριθμού συστημάτων σε διάφορους τομείς, συμπεριλαμβανομένων των κοινωνικών επιστημών, των κοινωνικών δικτύων κ.α. Ως μοναδική μη Ευκλείδεια δομή δεδομένων για μηχανική μάθηση, η ανάλυση γραφημάτων εστιάζει σε εργασίες όπως η ταξινόμηση κόμβων, η πρόβλεψη ακμών και η ομαδοποίηση. Τα νευρωνικά δίκτυα γραφημάτων (GNN) είναι μέθοδοι βασισμένες σε βαθιά μάθηση (deep learning). Λόγω της πειστικής τους απόδοσης, έχουν γίνει μια ευρέως εφαρμοσμένη μέθοδος ανάλυσης γραφημάτων και χρησιμοποιείται ολοένα και περισσότερο.

3.1 Γενικός σχεδιασμός Νευρωνικών Δικτύων Γράφων

Η σχεδίαση ενός μοντέλου Νευρωνικών Δικτύων Γράφων (GNN) [18] για την εφαρμογή σε μια εργασία και έναν συγκεκριμένο τύπο γραφήματος σε γενικές γραμμές αποτελείται από τέσσερα βήματα:

1. Εύρεση δομής γραφήματος
2. Καθορισμός του τύπου και της κλίμακας γραφήματος
3. Σχεδιασμός της συνάρτησης απώλειας
4. Κατασκευή μοντέλου με χρήση υπολογιστικών μονάδων

Εύρεση Δομής Γραφήματος

Αρχικά, πρέπει να μάθουμε τη δομή του γραφήματος στην εφαρμογή. Υπάρχουν συνήθως δύο σενάρια: δομικά σενάρια και μη δομικά σενάρια. Σε δομικά σενάρια, η δομή του γραφήματος είναι σαφής στις εφαρμογές, όπως εφαρμογές σε μόρια, φυσικά συστήματα, γραφήματα γνώσης κλπ. Σε μη δομικά σενάρια, τα γραφήματα είναι σιωπηρά, οπότε πρέπει πρώτα να δημιουργήσουμε το γράφημα από την εργασία, όπως η κατασκευή ενός πλήρως συνδεδεμένου γραφήματος «λέξεων» για κείμενο ή η κατασκευή ενός γραφήματος σκηνής για μια εικόνα. Αφού λάβουμε το γράφημα, η μεταγενέστερη διαδικασία σχεδιασμού επιχειρεί να βρει ένα βέλτιστο μοντέλο GNN σε αυτό το συγκεκριμένο γράφημα.

Καθορισμός του τύπου και της κλίμακας γραφήματος

Μετά το πρώτο βήμα και εφόσον έχουμε το γράφημα, το επόμενο βήμα είναι να μάθουμε τον τύπο του γραφήματος και την κλίμακα του. Γραφήματα με σύνθετους τύπους θα μπορούσαν να παρέχουν περισσότερες πληροφορίες για τους κόμβους και τις συνδέσεις τους. Συνηθισμένες κατηγορίες γραφημάτων είναι οι παρακάτω:

Κατευθυνόμενα - Μη κατευθυνόμενα γραφήματα

Η κύρια διαφορά μεταξύ κατευθυνόμενου και μη κατευθυνόμενου γραφήματος είναι ότι ένα κατευθυνόμενο γράφημα περιέχει ένα διατεταγμένο ζεύγος κόμβων ενώ ένα μη κατευθυνόμενο γράφημα περιέχει ένα μη διατεταγμένο ζεύγος κόμβων. Οι ακμές στα κατευθυνόμενα γραφήματα κατευθύνονται όλες από τον έναν κόμβο στον άλλο, και παρέχουν περισσότερες πληροφορίες από τα μη κατευθυνόμενα γραφήματα. Κάθε ακμή στα μη κατευθυνόμενα γραφήματα μπορεί επίσης να θεωρηθεί ως δύο κατευθυνόμενες ακμές.

Ομοιογενή - Ετερογενή Γραφήματα

Στα ομοιογενή γραφήματα οι κόμβοι και οι ακμές περιέχουν τον ίδιο τύπο πληροφορίας, αντίθετα με τα ετερογενή γραφήματα. Αυτός μας κάνει πιο προσεκτικούς καθώς οι τύποι κόμβων και ακμών παίζουν σημαντικό ρόλο σε ένα ετερογενές γράφημα και πρέπει να δίνουμε και την απαραίτητη σημασία.

Στατικά - Δυναμικά Γραφήματα

Όταν τα χαρακτηριστικά εισόδου ή η τοπολογία του γραφήματος ποικίλλουν με το χρόνο, το γράφημα θεωρείται ως δυναμικό γράφημα. Οι πληροφορίες χρόνου πρέπει να λαμβάνονται προσεκτικά υπόψη σε δυναμικά γραφήματα.

Σχεδιασμός της Συνάρτησης Απώλειας

Η συνάρτηση απώλειας σε ένα νευρωνικό δίκτυο ποσοτικοποιεί τη διαφορά μεταξύ του αναμενόμενου αποτελέσματος και του αποτελέσματος που παράγεται από το μοντέλο μηχανικής μάθησης. Από τη συνάρτηση απώλειας, μπορούμε να εξαγάγουμε τις διαβαθμίσεις που χρησιμοποιούνται για την ενημέρωση των βαρών. Ο μέσος όρος όλων των απωλειών αποτελεί το κόστος. Για να σχεδιάσουμε την συνάρτηση απώλειας λαμβάνουμε υπόψη τον τύπο της εργασίας πρόβλεψης και την ρύθμιση εκπαίδευσης. Για την εκμάθηση γραφημάτων, υπάρχουν συνήθως τρία είδη εργασιών ανάλυσης:

- **(Node-level)** Οι εργασίες σε επίπεδο κόμβου εστιάζονται σε κόμβους, οι οποίοι περιλαμβάνουν ταξινόμηση κόμβων, παλινδρόμηση κόμβων, ομαδοποίηση κόμβων. Η ταξινόμηση κόμβων προσπαθεί να κατηγοριοποιήσει τους κόμβους σε πολλές κλάσεις και η παλινδρόμηση κόμβων προβλέπει μια συνεχή τιμή για κάθε κόμβο. Η ομαδοποίηση κόμβων στοχεύει να χωρίσει τους κόμβους σε πολλές χωριστές ομάδες, όπου παρόμοιοι κόμβοι θα πρέπει να βρίσκονται στην ίδια ομάδα.

- **(Edge-level)** Οι εργασίες σε επίπεδο ακμής είναι η ταξινόμηση ακμών και η πρόβλεψη σύνδεσης, οι οποίες απαιτούν από το μοντέλο να ταξινομεί τύπους ακμών ή να προβλέψει εάν υπάρχει ακμή μεταξύ δύο δεδομένων κόμβων.

- **(Graph-level)** Οι εργασίες σε επίπεδο γραφήματος περιλαμβάνουν ταξινόμηση γραφήματος, παλινδρόμηση γραφήματος και αντιστοίχιση γραφημάτων, τα οποία χρειάζονται το μοντέλο για να μάθουν αναπαραστάσεις γραφημάτων.

Υπό την οπτική της επιβλεπόμενης μάθησης, οι εργασίες εκπαίδευσης γραφημάτων μπορούν να κατηγοριοποιηθούν σε τρεις διαφορετικές ρυθμίσεις εκπαίδευσης:

- **(Supervised setting)** Η εποπτευόμενη ρύθμιση παρέχει δεδομένα με ετικέτα για εκπαίδευση.

- **(Semi-supervised setting)** Η ημι-εποπτευόμενη ρύθμιση δίνει μια μικρή ποσότητα κόμβων με ετικέτα και μια μεγάλη ποσότητα κόμβων χωρίς ετικέτα για εκπαίδευση. Στη φάση της δοκιμής, η ρύθμιση μεταγωγής απαιτεί από το μοντέλο να προβλέψει τις ετικέτες των δεδομένων μη επισημασμένων κόμβων, ενώ η επαγωγική ρύθμιση παρέχει νέους μη επισημασμένους κόμβους από την ίδια κατανομή για να συμπεράνουμε. Οι περισσότερες εργασίες ταξινόμησης κόμβων και ακμών είναι ημι-εποπτευόμενες.

- **(Unsupervised setting)** Η ρύθμιση χωρίς επίβλεψη προσφέρει μόνο δεδομένα χωρίς ετικέτα και αφήνει το μοντέλο να βρει μοτίβα. Η ομαδοποίηση κόμβων είναι μια τυπική εργασία μάθησης χωρίς επίβλεψη.

Κατασκευή μοντέλου με χρήση υπολογιστικών μονάδων

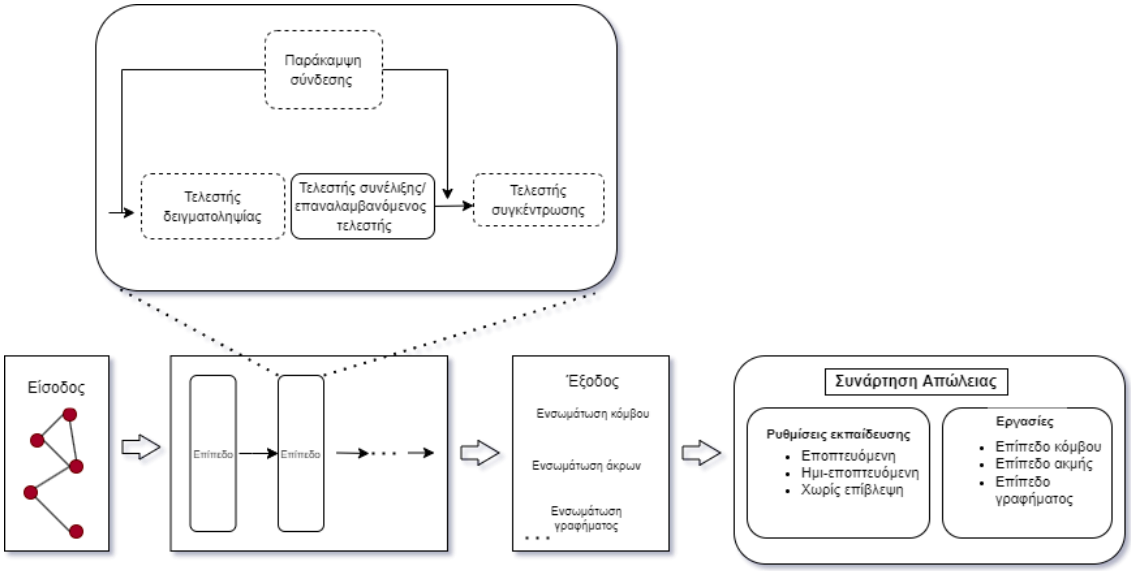
Σε αυτό το στάδιο ξεκινάμε την κατασκευή του μοντέλου χρησιμοποιώντας τις υπολογιστικές μονάδες. Συνηθισμένες υπολογιστικές μονάδες είναι:

- **(Propagation Module)** Η μονάδα διάδοσης χρησιμοποιείται για τη διάδοση πληροφοριών μεταξύ κόμβων. Με αυτόν τον τρόπο ανακαλύπτονται τόσο χαρακτηριστικές όσο και τοπολογικές πληροφορίες. Στις μονάδες διάδοσης, ο τελεστής συνέλιξης (convolution operator) και ο επαναλαμβανόμενος τελεστής (recurrent operator) χρησιμοποιούνται συνήθως για τη συγκέντρωση πληροφοριών από γείτονες, ενώ η λειτουργία παράκαμψης σύνδεσης (skip connection) χρησιμοποιείται για τη συλλογή πληροφοριών από ιστορικές αναπαραστάσεις κόμβων και τον μετριασμό του προβλήματος της υπερβολικής εξομάλυνσης.

- **(Sampling Module)** Η μονάδα δειγματοληψίας χρησιμοποιείται όταν τα γραφήματα είναι μεγάλα και συνήθως συνδυάζεται με τη μονάδα διάδοσης.

- **(Pooling Module)** Η μονάδα συγκέντρωσης χρησιμοποιείται όταν για την εξαγωγή πληροφοριών από κόμβους χρειάζεται να αξιοποιήσουμε αναπαραστάσεις υπογράφων ή γραφημάτων υψηλού επιπέδου.

Ένα μοντέλο νευρωνικών δικτύων γράφων συνήθως δημιουργείται με τις παραπάνω υπολογιστικές μονάδες. Μια κοινή αρχιτεκτονική του μοντέλου απεικονίζεται στο μεσαίο τμήμα του Σχήματος 9 όπου ο τελεστής συνέλιξης ο επαναλαμβανόμενος τελεστής, ο τελεστής δειγματοληψίας και η παράκαμψη σύνδεσης χρησιμοποιούνται για τη διάδοση πληροφοριών σε κάθε επίπεδο και στη συνέχεια ο τελεστής συγκέντρωσης προστίθεται για εξαγωγή υψηλού επιπέδου πληροφορίας. Αυτά τα επίπεδα συνήθως στοιβάζονται για να ληφθούν καλύτερες αναπαραστάσεις.



Σχήμα 9: Ο γενικός σχεδιασμός για ένα μοντέλο GNN (βλ. [18])

3.2 Περιγραφή υπολογιστικών μονάδων

Όπως προαναφέραμε για τον σχεδιασμό και την κατασκευή του μοντέλου χρησιμοποιούμε υπολογιστικές μονάδες. Οι τρεις που έχουμε ήδη αναφέρει είναι η μονάδα δειγματοληψίας, μονάδα συγκέντρωσης και μονάδα διάδοσης. Θα γίνει μια πιο εκτενής περιγραφή καθώς και για την τελευταία θα παρουσιάσουμε και τρία συστατικά: τον τελεστή συνέλιξης, τον επαναλαμβανόμενο τελεστή και την παράβλεψη σύνδεσης.

3.2.1 Μονάδες Διάδοσης

- **Τελεστής Συνέλιξης**

Η βασική ιδέα πίσω από τους τελεστές συνέλιξης που αποτελούν κύριο συστατικό των μονάδων διάδοσης είναι ότι γενικεύουν τις συνελίξεις από κάποιον τομέα στον τομέα του γραφήματος. Κατηγοριοποιούνται σε φασματικές προσεγγίσεις και χωρικές προσεγγίσεις.

Φασματικές προσεγγίσεις

Οι φασματικές προσεγγίσεις λειτουργούν με μια φασματική αναπαράσταση των γραφημάτων. Αυτές οι μέθοδοι βασίζονται θεωρητικά στην επεξεργασία σήματος γραφήματος και ορίζουν τον τελεστή συνέλιξης στο φασματικό πεδίο.

Στις φασματικές μεθόδους, ένα σήμα γραφήματος x μετασχηματίζεται αρχικά στη φασματική περιοχή από τον μετασχηματισμό γραφήματος Fourier $\mathcal{F}$ και στη συνέχεια διεξάγεται η λειτουργία συνέλιξης. Μετά τη συνέλιξη, το σήμα που προκύπτει μετατρέπεται ξανά χρησιμοποιώντας τον αντίστροφο γράφημα μετασχηματισμού Fourier $\mathcal{F}^{-1}$. Αυτοί οι μετασχηματισμοί ορίζονται ως:

$$\mathcal{F}(\mathbf{x}) = \mathbf{U}^T\mathbf{x} \qquad \mathcal{F}^{-1}(\mathbf{x}) = \mathbf{U}\mathbf{x} \tag{1}$$

όπου $\mathbf{U}$ είναι ο πίνακας των ιδιοδιανυσμάτων του κανονικοποιημένου γραφήματος Laplace:

$$\mathbf{L} = \mathbf{I}_N - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$$

εδώ το $\mathbf{D}$ είναι ο πίνακας βαθμών και $\mathbf{A}$ είναι ο πίνακας γειτνίασης του γραφήματος. Το κανονικοποιημένο γράφημα Laplace είναι συμμετρικό και θετικά ημιορισμένο, επομένως μπορεί να παραγοντοποιηθεί ως:

$$\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

όπου $\mathbf{\Lambda}$ είναι ένας διαγώνιος πίνακας των ιδιοτιμών. Με βάση το θεώρημα συνέλιξης (Mallat, 1999), η πράξη συνέλιξης ορίζεται ως:

$$\begin{aligned}\mathbf{g} \star \mathbf{x} &= \mathcal{F}^{-1}(\mathcal{F}(\mathbf{g}) \odot \mathcal{F}(\mathbf{x})) \\ &= \mathbf{U}(\mathbf{U}^T\mathbf{g} \odot \mathbf{U}^T\mathbf{x})\end{aligned}, \qquad (2)$$

όπου το $\mathbf{U}^T\mathbf{g}$ είναι το φίλτρο στο φασματικό πεδίο. Αν απλοποιήσουμε το φίλτρο χρησιμοποιώντας έναν διαγώνιο πίνακα εκμάθησης g_w τότε έχουμε τη βασική συνάρτηση των φασματικών μεθόδων:

$$\mathbf{g}_w \star \mathbf{x} = \mathbf{U}\mathbf{g}_w\mathbf{U}^T\mathbf{x}, \qquad (3)$$

Κάποιες από τις φασματικές μεθόδους παρουσιάζονται παρακάτω.

- **ChebNet**

Σύμφωνα με αυτή τη μέθοδο το g_w μπορεί να προσεγγιστεί με μια περικομμένη επέκταση ως προς τα πολυώνυμα Chebyshev $T_k(x)$ μέχρι την τάξη K^{th}. Το παραπάνω ορίζεται ως:

$$\mathbf{g}_w \star \mathbf{x} \approx \sum_{k=0}^{K} w_k \mathbf{T}_k(\tilde{\mathbf{L}})\mathbf{x}, \qquad (4)$$

όπου $\tilde{\mathbf{L}} = \frac{2}{\lambda_{max}}\boldsymbol{L} - \boldsymbol{I}_N$, και το λ_{max} υποδηλώνει τη μεγαλύτερη ιδιοτιμή του $\boldsymbol{L}$. Το εύρος των ιδιοτιμών του $\tilde{\mathbf{L}}$ είναι [-1,1]. $w \in \mathbb{R}^K$ είναι τώρα ένα διάνυσμα των συντελεστών Chebyshev. Τα πολυώνυμα Chebyshev ορίζονται ως: $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, με $T_0(x) = 1$ και $T_1(x) = x$. Παρατηρούμε ότι η πράξη είναι K-τοπική αφού είναι πολυώνυμο K^{th} τάξης στο γράφημα Laplace. Το παραπάνω χρησιμοποιείται για να οριστεί ένα συνελικτικό νευρωνικό δίκτυο χωρίς την ανάγκη να υπολογιστούν τα ιδιοδιανύσματα του γραφήματος Laplace.

- **Συνελικτικά Δίκτυα Γράφων (GCN)**

Για να μετριαστεί το πρόβλημα της υπερπροσαρμογής οι Kipf και Welling (2017) απλοποίησαν την λειτουργία της συνέλιξης στην Εξ. (4) με $K = 1$. Επιπροσθέτως θεώρησαν $\lambda_{max} \approx 2$ και η εξίσωση μετά από αυτό γράφεται:

$$\mathbf{g}_w \star \mathbf{x} \approx w_0\mathbf{x} + w_1(\mathbf{L} - \mathbf{I}_N)\mathbf{x} = w_0\mathbf{x} - w_1\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{x}, \qquad (5)$$

με δύο ελεύθερες παραμέτρους w_0 και w_1. Με τον περιορισμό παραμέτρου $w = w_0 = -w_1$ λαμβάνουμε το εξής:

$$\mathbf{g}_w \star \mathbf{x} \approx w\left(\mathbf{I}_N + \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\right)\mathbf{x}, \qquad (6)$$

Το GCN εισάγει περαιτέρω ένα τέχνασμα επανακανονικοποίησης για την επίλυση του προβλήματος της κλίσης έκρηξης/εξαφάνισης στην Εξ. (6):

$$\mathbf{I}_N + \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} \rightarrow \widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}$$

με $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ και

$$\widetilde{\boldsymbol{D}}_{ii} = \sum_j \widetilde{\mathbf{A}}_{ij}$$

Τέλος, η συμπαγής μορφή του GCN ορίζεται ως:

$$\boldsymbol{H} = \widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W}, \qquad (7)$$

όπου $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{F}}$ είναι ο πίνακας εισόδου, $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{F}\times\boldsymbol{F}'}$ είναι η παράμετρος και $\boldsymbol{H} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{F}'}$ είναι ο συνελικτικός πίνακας. F και F' είναι οι διαστάσεις της εισόδου και εξόδου, αντίστοιχα.

- **Προσαρμοστικά Συνελικτικά Δίκτυα Γράφων (AGCN)**

Τα μοντέλα αυτά χρησιμοποιούν την αρχική δομή γραφήματος για να υποδηλώσουν σχέσεις μεταξύ κόμβων. Ωστόσο, μπορεί να υπάρχουν σιωπηρές σχέσεις μεταξύ διαφορετικών κόμβων. Έτσι το συγκεκριμένο μοντέλο προτείνεται για να μάθει τις υποκείμενες σχέσεις των κόμβων. Το AGCN μαθαίνει ένα «υπολειπόμενο» γράφημα Laplace και το προσθέτει στον αρχικό πίνακα Laplace. Αποδεικνύεται ότι είναι αποτελεσματικό σε πολλά σύνολα δεδομένων δομημένων με γραφήματα.

- **Συνελικτικό Δίκτυο Διπλού Γραφήματος (DGCN)**

Με το συγκεκριμένο μοντέλο εξετάζεται από κοινού η συνέπεια των γραφημάτων σε τοπικό και συνολικό επίπεδο. Χρησιμοποιεί δύο συνελικτικά δίκτυα για να συλλάβει την τοπική και συνολική συνέπεια και υιοθετεί μια μη επιβλεπόμενη συνάρτηση απώλειας για να τα συνθέσει. Το πρώτο συνελικτικό δίκτυο είναι το ίδιο με την Εξ. (7) και το δεύτερο δίκτυο αντικαθιστά τον πίνακα γειτνίασης με τον πίνακα θετικών σημειακών κοινών πληροφοριών (positive pointwise mutual information) (PPMI). Έτσι προκύπτει:

$$\mathbf{H}' = \rho\left(\mathbf{D}_P^{-\frac{1}{2}}\mathbf{A}_P\mathbf{D}_P^{-\frac{1}{2}}\mathbf{H}\mathbf{W}\right), \qquad (8)$$

Όπου $\mathbf{A}_P$ είναι ο πίνακας PPMI και $\mathbf{D}_P$ είναι ο διαγώνιος πίνακας βαθμών του $\mathbf{A}_P$.

Οι φασματικές προσεγγίσεις λοιπόν βασίζονται καλά στη θεωρία και υπάρχουν επίσης και άλλες θεωρητικές αναλύσεις που προτείνονται γενικότερα στη βιβλιογραφία. Ωστόσο, σχεδόν σε όλες τις φασματικές προσεγγίσεις που αναφέρθηκαν παραπάνω, τα φίλτρα εκμάθησης εξαρτώνται από τη δομή του γραφήματος. Δηλαδή, τα φίλτρα δεν μπορούν να

εφαρμοστούν σε ένα γράφημα με διαφορετική δομή και αυτά τα μοντέλα μπορούν να εφαρμοστούν μόνο κάτω από τη μεταδοτική ρύθμιση των εργασιών ενός γραφήματος.

Χωρικές προσεγγίσεις

Οι χωρικές προσεγγίσεις ορίζουν τις συνελίξεις απευθείας στο γράφημα με βάση την τοπολογία του γραφήματος. Η κύρια πρόκληση των χωρικών προσεγγίσεων είναι ο καθορισμός της λειτουργίας συνέλιξης σε γειτονιές κόμβων διαφορετικού μεγέθους καθώς και να διατηρηθεί η τοπική πληροφορία θέσης.

- **Neural FPs (Neural graph Fingerprints)**

Τα Neural FPs (Duvenaud et al., 2015) χρησιμοποιούν διαφορετικούς πίνακες βάρους για κόμβους με διαφορετικούς βαθμούς:

$$\begin{aligned} \mathbf{t} &= \mathbf{h}_v^t + \sum_{u \in \mathcal{N}_v} \mathbf{h}_u^t \\ \mathbf{h}_v^{t+1} &= \sigma\left(\mathbf{t}\mathbf{W}_{|\mathcal{N}_v|}^{t+1}\right) \end{aligned}, \qquad (9)$$

όπου $\mathbf{W}_{|\mathcal{N}_v|}^{t+1}$ είναι ο πίνακας βάρους για κόμβους με βαθμό $|\mathcal{N}_v|$ στο επίπεδο $t+1$. Ένα από τα κύρια μειονεκτήματα της μεθόδου είναι ότι δεν μπορεί να εφαρμοστεί σε γραφήματα μεγάλης κλίμακας που περιέχουν περισσότερους βαθμούς κόμβων.

- **Συνελικτικό Νευρωνικό Δίκτυο Διάχυσης - Diffusion Convolutional Neural Network (DCNN)**

Το συνελικτικό νευρωνικό δίκτυο διάχυσης (DCNN) (Atwood και Towsley, 2016) χρησιμοποιεί πίνακες μετάβασης για να ορίσει τη γειτονιά για τους κόμβους. Για την ταξινόμηση κόμβων, οι αναπαραστάσεις διάχυσης κάθε κόμβου στο γράφημα μπορούν να εκφραστούν ως:

$$\mathbf{H} = f(\mathbf{W}_c \odot \mathbf{P}^*\mathbf{X}) \in \mathbb{R}^{N \times K \times F}, \qquad (10)$$

όπου $\mathbf{X} \in \mathbb{R}^{N \times F}$ είναι ο πίνακας των χαρακτηριστικών εισόδου (F είναι η διάσταση). Το $\mathbf{P}^*$ είναι ένας τανυστής $N \times K \times N$ που περιέχει τη σειρά ισχύος $\{P, P^2, \dots, P^K\}$ του πίνακα $\mathbf{P}$. Το $\mathbf{P}$ είναι ο κανονικοποιημένος κατά βαθμό πίνακας μετάβασης από τον πίνακα γειτνίασης γραφημάτων $\boldsymbol{A}$. Κάθε οντότητα μετατρέπεται σε μια συνελικτική αναπαράσταση διάχυσης που είναι ένας πίνακας $K \times F$ που ορίζεται από K βήματα της διάχυσης γραφήματος στα χαρακτηριστικά F. Και ορίζεται από έναν πίνακα βάρους $K \times F$ και από μια μη γραμμική συνάρτηση ενεργοποίησης f.

Χωρικές προσεγγίσεις με βάση την προσοχή

Συγκριτικά με τους τελεστές που αναφέρθηκαν προηγουμένως οι τελεστές που βασίζονται στην προσοχή εκχωρούν διαφορετικά βάρη στους γειτονικούς κόμβους, ώστε να μπορούν να μετριάσουν το «θόρυβο» και να επιτύχουν καλύτερα αποτελέσματα.

- **Δίκτυο Προσοχής - Graph Attention Network (GAT)**

Το συγκεκριμένο ενσωματώνει τον μηχανισμό προσοχής στο βήμα διάδοσης. Ακολουθώντας μια στρατηγική αυτό-προσοχής υπολογίζει την «κρυμμένη» κατάσταση κάθε κόμβου παρακολουθώντας τους γείτονές του. Λαμβάνουμε την κρυφή κατάσταση του κόμβου v από:

$$\mathbf{h}_v^{t+1} = \rho\left(\sum_{u\in\mathcal{N}_v} \alpha_{vu}\mathbf{W}\mathbf{h}_u^t\right)$$
$$\alpha_{vu} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_v \parallel \mathbf{W}\mathbf{h}_u]))}{\sum_{k\in\mathcal{N}_v}\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_v \parallel \mathbf{W}\mathbf{h}_k]))}, \quad (11)$$

όπου **W** είναι ο πίνακας βάρους που σχετίζεται με τον γραμμικό μετασχηματισμό που εφαρμόζεται σε κάθε κόμβο, και **a** είναι το διάνυσμα βάρους ενός MLP ενός επιπέδου.

Επιπλέον, το GAT χρησιμοποιεί την προσοχή πολλαπλών κεφαλών [19] (multi-head attention) ή παράλληλα στρώματα προσοχής, για τη σταθεροποίηση της μαθησιακής διαδικασίας. Η προσοχή πολλαπλών κεφαλών επιτρέπει στο μοντέλο να παρακολουθεί από κοινού πληροφορίες από διαφορετικούς υποχώρους αναπαράστασης σε διαφορετικές θέσεις. Με ένα μόνο κεφάλι προσοχής, ο μέσος όρος το εμποδίζει αυτό. Εφαρμόζει δηλαδή K ανεξάρτητους πίνακες κεφαλής προσοχής για τον υπολογισμό των κρυφών καταστάσεων και στη συνέχεια υπολογίζει τον μέσο όρο, με αποτέλεσμα τις ακόλουθες δύο αναπαραστάσεις εξόδου:

$$\mathbf{h}_v^{t+1} = \|_{k=1}^K \sigma\left(\sum_{u\in\mathcal{N}_v} \alpha_{vu}^k \mathbf{W}_k \mathbf{h}_u^t\right)$$
$$\mathbf{h}_v^{t+1} = \sigma\left(\frac{1}{K}\sum_{k=1}^K \sum_{u\in\mathcal{N}_v} \alpha_{vu}^k \mathbf{W}_k \mathbf{h}_u^t\right), \quad (12)$$

Εδώ το α_{ij}^k είναι ο κανονικοποιημένος συντελεστής προσοχής που υπολογίζεται από την k^{th} κεφαλή προσοχής. Η αρχιτεκτονική της προσοχής έχει διάφορες ιδιότητες. Αρχικά ο υπολογισμός των ζευγών κόμβου με τους γειτονικούς μπορεί να παραλληλιστεί, επομένως η λειτουργία είναι πιο αποδοτική. Επίσης μπορεί να εφαρμοστεί σε κόμβους γραφημάτων με διαφορετικούς βαθμούς, καθορίζοντας αυθαίρετα βάρη στους γείτονες. Τέλος, μπορεί να εφαρμοστεί εύκολα στα επαγωγικά μαθησιακά προβλήματα.

Συνδυαστικές Δομές Χωρικών προσεγγίσεων

Πέρα από τις διαφορετικές παραλλαγές χωρικών προσεγγίσεων, έχουν προταθεί και άλλες γενικότερες δομές που έχουν στόχο το συνδυασμό διαφορετικών μοντέλων.

- **Νευρωνικά Δίκτυα Ανταλλαγής Μηνυμάτων - Message Passing Neural Network (MPNN)**

Το συγκεκριμένο εξάγει τα γενικά χαρακτηριστικά μεταξύ πολλών κλασικών μοντέλων. Το μοντέλο περιλαμβάνει δύο φάσεις, μια φάση μετάδοσης μηνύματος και μια φάση ανάγνωσης. Στη φάση μετάδοσης μηνυμάτων, το μοντέλο χρησιμοποιεί πρώτα τη συνάρτηση μηνύματος M_t για να συγκεντρώσει το «μήνυμα» m_v^t από τους γείτονες και στη συνέχεια χρησιμοποιεί τη συνάρτηση ενημέρωσης U_t για να ενημερώσει την κρυφή κατάσταση h_v^t:

$$\mathbf{m}_v^{t+1} = \sum_{u\in\mathcal{N}_v} M_t(\mathbf{h}_v^t, \mathbf{h}_u^t, \mathbf{e}_{vu})$$
$$\mathbf{h}_v^{t+1} = U_t(\mathbf{h}_v^t, \mathbf{m}_v^{t+1}), \quad (13)$$

Το $\mathbf{e}_{vu}$ αντιπροσωπεύει τα χαρακτηριστικά μιας μη κατευθυνόμενης ακμής (v, u). Η φάση ανάγνωσης υπολογίζει ένα διάνυσμα χαρακτηριστικών ολόκληρου του γραφήματος χρησιμοποιώντας τη συνάρτηση ανάγνωσης R:

$$\hat{y} = R(\{h_v^T | v \in G\}), \qquad (14)$$

όπου T υποδηλώνει τα συνολικά χρονικά βήματα. Η συνάρτηση μηνύματος M_t , η συνάρτηση ενημέρωσης κορυφής U_t και η συνάρτηση ανάγνωσης R ενδέχεται να έχουν διαφορετικές ρυθμίσεις. Με αποτέλεσμα το MPNN θα μπορούσε να δημιουργήσει πολλά διαφορετικά μοντέλα μέσω διαφορετικών ρυθμίσεων λειτουργίας. [20]

- **Μη Τοπικό Νευρωνικό Δίκτυο - Non-Local Neural Network (NLNN)**

Το μη τοπικό νευρωνικό δίκτυο γενικεύει και επεκτείνει την κλασική μη τοπική μέση λειτουργία (Buades et al., 2005) [21] στην υπολογιστική όραση. Η μη τοπική πράξη υπολογίζει την κρυφή κατάσταση σε μια θέση ως σταθμισμένο άθροισμα χαρακτηριστικών σε όλες τις πιθανές θέσεις. Οι πιθανές θέσεις μπορεί να είναι στον χώρο, τον χρόνο ή τον χωροχρόνο. Έτσι, το NLNN μπορεί να θεωρηθεί ως ένας συνδυασμός διαφορετικών μεθόδων προσοχής. Ορίζεται ως εξής:

$$\mathbf{h}_v^{t+1} = \frac{1}{\mathcal{C}(\mathbf{h}^t)} \sum_{\forall u} f(\mathbf{h}_v^t, \mathbf{h}_u^t) g(\mathbf{h}_u^t), \qquad (15)$$

όπου u είναι ο δείκτης όλων των πιθανών θέσεων για τη θέση v, $f(\mathbf{h}_v^t, \mathbf{h}_u^t)$ υπολογίζει ένα βαθμωτό μέγεθος μεταξύ v και u που αντιπροσωπεύει τη σχέση μεταξύ τους, το $g(\mathbf{h}_u^t)$ υποδηλώνει έναν μετασχηματισμό της εισόδου $\mathbf{h}_u^t$ και το $\mathcal{C}(\mathbf{h}^t)$ είναι ένας παράγοντας κανονικοποίησης.

- **Δίκτυο γραφημάτων - Graph Network (GN)**

Το δίκτυο γραφημάτων είναι μια πιο γενική δομή σε σύγκριση με άλλες, με την εκμάθηση αναπαραστάσεων σε επίπεδο κόμβου, σε επίπεδο ακμής και σε επίπεδο γραφήματος. Μπορεί να συνδυάσει πολλές παραλλαγές όπως οι παραπάνω MPNN, NLNN κ.α. Η κεντρική μονάδα υπολογισμού του δικτύου γραφημάτων ονομάζεται μπλοκ GN (GN block). Ένα μπλοκ GN ορίζει τρεις συναρτήσεις ενημέρωσης και τρεις συναρτήσεις συγκέντρωσης:

$$\begin{aligned}
\mathbf{e}_k^{t+1} &= \varphi^e\left(\mathbf{e}_k^t, \mathbf{h}_{r_k}^t, \mathbf{h}_{s_k}^t, \mathbf{u}^t\right), \bar{\mathbf{e}}_v^{t+1} = \rho^{e\to h}(\mathbf{E}_v^{t+1}) \\
\mathbf{h}_v^{t+1} &= \varphi^h(\bar{\mathbf{e}}_v^{t+1}, \mathbf{h}_v^t, \mathbf{u}^t), \bar{\mathbf{e}}^{t+1} = \rho^{e\to u}(\mathbf{E}^{t+1}) \quad , \qquad (16) \\
\mathbf{u}^{t+1} &= \varphi^u\left(\bar{\mathbf{e}}^{t+1}, \bar{\mathbf{h}}^{t+1}, \mathbf{u}^t\right), \bar{\mathbf{h}}^{t+1} = \rho^{h\to u}(\mathbf{H}^{t+1})
\end{aligned}$$

όπου r_k είναι ο κόμβος δέκτη και το s_k είναι ο κόμβος αποστολέα της ακμής k. Τα $\mathbf{E}^{t+1}$ και $\mathbf{H}^{t+1}$ είναι οι πίνακες των στοιβαγμένων διανυσμάτων ακμών και των διανυσμάτων κόμβου στο βήμα $t+1$, αντίστοιχα. Το $\mathbf{E}_v^{t+1}$ συλλέγει διανύσματα ακμών με δέκτη τον κόμβο v. Το $\mathbf{u}$ είναι το καθολικό χαρακτηριστικό για την αναπαράσταση γραφήματος. Οι συναρτήσεις φ και ρ μπορούν να έχουν διάφορες ρυθμίσεις και οι συναρτήσεις ρ πρέπει να είναι αμετάβλητες ως προς τις εντολές εισόδου και να λαμβάνουν μεταβλητά μήκη ορισμάτων.

- **Επαναλαμβανόμενος Τελεστής**

Οι μέθοδοι που χρησιμοποιούν τον επαναλαμβανόμενο τελεστή τείνουν να κατέχουν από τις πρώτες θέσεις στην έρευνα των Νευρωνικών Δικτύων Γράφων. Η διαφορά των επαναλαμβανόμενων τελεστών και των τελεστών συνέλιξης που προαναφέραμε είναι ότι τα επίπεδα στους τελεστές συνέλιξης χρησιμοποιούν διαφορετικά βάρη ενώ τα επίπεδα στους επαναλαμβανόμενους τελεστές μοιράζονται τα ίδια βάρη.

Μέθοδοι που βασίζονται στη σύγκλιση

Σε ένα γράφημα, ο κάθε κόμβος ορίζεται από τα χαρακτηριστικά του και τους κόμβους με τους οποίους είναι συνδεδεμένος. Σκοπός ενός νευρωνικού δικτύου γράφου είναι να μάθει μια κατάσταση ενσωμάτωσης $\boldsymbol{h}_v \in \mathbb{R}^s$ που περιέχει τις πληροφορίες της γειτονιάς και της ίδιας της κατάστασης για κάθε κόμβο. Η κατάσταση ενσωμάτωσης $\boldsymbol{h}_v$ είναι ένα διάνυσμα διάστασης s του κόμβου v και μπορεί να χρησιμοποιηθεί για την παραγωγή μιας εξόδου $\boldsymbol{o}_v$ όπως η κατανομή της προβλεπόμενης ετικέτας κόμβου. Ο υπολογισμός των $\boldsymbol{h}_v$ και $\boldsymbol{o}_v$ ορίζονται ως εξής:

$$\begin{aligned} \mathbf{h}_v &= f\left(\mathbf{x}_v, \mathbf{x}_{co[v]}, \mathbf{h}_{\mathcal{N}_v}, \mathbf{x}_{\mathcal{N}_v}\right) \\ \mathbf{o}_v &= g(\mathbf{h}_v, \mathbf{x}_v) \end{aligned}, \qquad (17)$$

όπου $\mathbf{x}_v$ είναι τα χαρακτηριστικά του v, $\mathbf{x}_{co[v]}$ είναι τα χαρακτηριστικά των ακμών του, $\mathbf{h}_{\mathcal{N}_v}$ είναι οι καταστάσεις και $\mathbf{x}_{\mathcal{N}_v}$ τα χαρακτηριστικά των κόμβων στη γειτονιά του v. Η f είναι μια παραμετρική συνάρτηση που ονομάζεται συνάρτηση τοπικής μετάβασης. Τη μοιράζονται μεταξύ τους όλοι οι κόμβοι και ενημερώνει την κατάσταση του κόμβου σύμφωνα με τη γειτονιά εισόδου. Η g είναι η τοπική συνάρτηση εξόδου που περιγράφει πώς παράγεται η έξοδος.

Έστω $\boldsymbol{H}$, $\boldsymbol{O}$, $\boldsymbol{X}$ και $\boldsymbol{X}_N$ οι πίνακες που κατασκευάζονται στοιβάζοντας όλες τις καταστάσεις, όλες τις εξόδους, όλα τα χαρακτηριστικά και όλα τα χαρακτηριστικά του κόμβου, αντίστοιχα. Τότε έχουμε μια συμπαγή μορφή που ορίζεται ως εξής:

$$\begin{aligned} \mathbf{H} &= F(\mathbf{H}, \mathbf{X}) \\ \mathbf{O} &= G(\mathbf{H}, \mathbf{X}_N)' \end{aligned} \qquad (18)$$

όπου F είναι η συνάρτηση καθολικής μετάβασης, και G είναι η καθολική συνάρτηση εξόδου, να είναι οι στοιβαγμένες εκδόσεις των f και g αντίστοιχα για όλους τους κόμβους σε ένα γράφημα. Η τιμή του $\mathbf{H}$ είναι το σταθερό σημείο της εξ. (18) και ορίζεται μοναδικά με την υπόθεση ότι το F είναι μια αντιστοίχιση συστολής.

Με την πρόταση του θεωρήματος σταθερού σημείου του Banach, το GNN χρησιμοποιεί το ακόλουθο κλασικό επαναληπτικό σχήμα για να υπολογίσει την κατάσταση:

$$\boldsymbol{H}^{t+1} = F(\boldsymbol{H}^t, \boldsymbol{X}), \qquad (19)$$

όπου $\boldsymbol{H}^t$ δηλώνει την t-ιοστή επανάληψη του $\boldsymbol{H}$. Το δυναμικό σύστημα Εξ.(19) συγκλίνει εκθετικά γρήγορα στη λύση για οποιαδήποτε αρχική τιμή.

Διάδοση Lagrange - Lagrangian Propagation GNN (LP-GNN)

Η διάδοση Lagrange [22] κανονικοποιεί τη διαδικασία εκμάθησης ως πρόβλημα βελτιστοποίησης περιορισμών στο μοντέλο Lagrange και αποφεύγει τους επαναληπτικούς υπολογισμούς για το σταθερό σημείο. Η διαδικασία σύγκλισης εκφράζεται σιωπηρά από έναν μηχανισμό ικανοποίησης περιορισμών.

- **Παράκαμψη Σύνδεσης**

Πολλές από τις μεθόδους που χρησιμοποιούνται στα νευρωνικά δίκτυα γραφημάτων προσπαθούν να στοιβάξουν σε όσα περισσότερα επίπεδα το γράφημα με στόχο τη λήψη καλύτερων αποτελεσμάτων. Διότι τα πολλά επίπεδα κάνουν κάθε κόμβο να συγκεντρώνει περισσότερες πληροφορίες από τους γείτονές του. Μέσα από δοκιμές όμως, έχει παρατηρηθεί ότι και τα βαθιά αυτά μοντέλα δεν βελτιώνουν, κάποιες φορές κιόλας χειροτερεύουν την απόδοση. Αυτό συμβαίνει διότι κάθε επίπεδο περιέχει «θόρυβο» δηλαδή πληροφορίες που δεν μας ενδιαφέρουν, τον οποίο και μεταδίδει. Προκαλείται επίσης το πρόβλημα της υπερβολικής εξομάλυνσης επειδή οι κόμβοι τείνουν να έχουν παρόμοιες αναπαραστάσεις μετά τη λειτουργία συνάθροισης όταν τα μοντέλα προχωρούν βαθύτερα. Δημιουργήθηκαν λοιπόν μέθοδοι οι οποίες προσθέτουν «παρακάμψεις συνδέσεων» για να κάνουν τα μοντέλα GNN να αποδίδουν σε μεγαλύτερο βάθος, δηλαδή σε περισσότερα επίπεδα.

Highway GCN

Οι Rahimi et al. (2018) προτείνουν ένα Highway GCN που χρησιμοποιεί πύλες κατά επίπεδα παρόμοιες με τα δίκτυα αυτοκινητοδρόμων. Η έξοδος ενός επιπέδου αθροίζεται με την είσοδο του με βάρη πύλης:

$$\mathbf{T}(\mathbf{h}^t) = \sigma(\mathbf{W}_t\mathbf{h}^t + \mathbf{b}_t)$$
$$\mathbf{h}^{t+1} = \mathbf{h}^{t+1} \odot \mathbf{T}(\mathbf{h}^t) + \mathbf{h}^t \odot \left(1 - \mathbf{T}(\mathbf{h}^t)\right)' \qquad (20)$$

Προσθέτοντας τις πύλες του αυτοκινητόδρομου, η απόδοση κορυφώνεται σε 4 επίπεδα σε ένα συγκεκριμένο πρόβλημα που συζητήθηκε στο [23]

Jump Knowledge Network (JKN)

Οι Xu et al. [24] μελέτησαν τις ιδιότητες και τους περιορισμούς των σχημάτων συγκέντρωσης γειτονιών. Προτείνουν το δίκτυο γνώσης άλματος (JKN) που θα μπορούσε να μάθει προσαρμοστικές και δομικές αναπαραστάσεις. Το δίκτυο αυτό επιλέγει από όλες τις ενδιάμεσες αναπαραστάσεις για κάθε κόμβο στο τελευταίο επίπεδο, γεγονός που κάνει το μοντέλο να προσαρμόζει το πραγματικό μέγεθος γειτονιάς για κάθε κόμβο όπως χρειάζεται. Το JKN αποδίδει καλά στα πειράματα στα κοινωνικά δίκτυα, τη βιοπληροφορική και τα δίκτυα αναφορών. Μπορεί επίσης να συνδυαστεί με μοντέλα όπως GCN, GraphSAGE και GAT για τη βελτίωση της απόδοσής τους.

3.2.2 Μονάδες Δειγματοληψίας

Τα μοντέλα νευρωνικών δικτύων γραφημάτων συγκεντρώνουν πληροφορίες για κάθε κόμβο από τους γειτονικούς κόμβους στο προηγούμενο επίπεδο. Έτσι παρατηρούμε ότι όταν έχουμε πολλά επίπεδα σε ένα GNN, το μέγεθος των γειτόνων αυξάνεται εκθετικά με το βάθος. Για να περιορίσουμε αυτό το πρόβλημα ένας αποτελεσματικός και αποδοτικός τρόπος είναι η δειγματοληψία. Επίσης όταν έχουμε να κάνουμε με μεγάλα γραφήματα δεν είναι πάντα δυνατό να αποθηκεύουμε και να επεξεργαζόμαστε όλες τις πληροφορίες της γειτονιάς του κάθε κόμβου, επομένως η μονάδα δειγματοληψίας χρειάζεται για τη διεξαγωγή της διάδοσης. Κάποια από τα είδη δειγματοληψίας γραφημάτων είναι τα εξής: δειγματοληψία κόμβων, δειγματοληψία επιπέδου και δειγματοληψία υπογραφήματος.

Δειγματοληψία κόμβων

Η επιλογή ενός υποσυνόλου από τη γειτονιά κάθε κόμβου είναι ένας τρόπος για να μειωθεί το μέγεθος των γειτονικών κόμβων. Το GraphSAGE (Hamilton et al., 2017) λαμβάνει δείγματα μικρού αριθμού γειτόνων, εξασφαλίζοντας ένα μέγεθος γειτονιάς από 2 έως 50 για κάθε κόμβο.

Για τη μείωση της διακύμανσης δειγματοληψίας, οι Chen et al. (2018) εισάγουν έναν αλγόριθμο

στοχαστικής προσέγγισης με βάση τη μεταβλητή ελέγχου για το GCN χρησιμοποιώντας τις ιστορικές ενεργοποιήσεις κόμβων ως παραλλαγή ελέγχου. Αυτή η μέθοδος περιορίζει το πεδίο υποδοχής στη γειτονιά κατά 1-hop και χρησιμοποιεί την ιστορική κρυφή κατάσταση ως προσιτή προσέγγιση.

Μια άλλη μέθοδος δειγματοληψίας είναι αυτή με βάση τη σημασία του κόμβου. Προσομοιώνοντας τυχαίους περιπάτους και ξεκινώντας από κόμβους που είναι στόχοι, αυτή η προσέγγιση επιλέγει τους κορυφαίους T κόμβους με τον υψηλότερο κανονικοποιημένο αριθμό επισκέψεων.

Δειγματοληψία επιπέδου

Η δειγματοληψία επιπέδου διατηρεί ένα μικρό σύνολο κόμβων για συνάθροιση σε κάθε επίπεδο για τον έλεγχο του συντελεστή επέκτασης. Το FastGCN (Chen et al., 2018b) λαμβάνει απευθείας δείγματα του δεκτικού πεδίου για κάθε στρώμα. Χρησιμοποιεί δειγματοληψία σπουδαιότητας, όπου οι σημαντικοί κόμβοι είναι πιο πιθανό να χρησιμοποιηθούν σας δείγματα.

Μια διαφορετική μέθοδος που προτείνανε οι Huang et al. (2018) είναι να εισάγουν έναν δειγματολήπτη που να δέχεται παραμετροποιήσεις και να εκπαιδεύεται καθώς και να εξαρτάται από το προηγούμενο επίπεδο για την εκτέλεση της δειγματοληψίας. Επιπλέον, αυτός ο προσαρμοστικός δειγματολήπτης θα μπορούσε να βελτιστοποιήσει τη σημασία της δειγματοληψίας και να μειώσει τη διακύμανση ταυτόχρονα.

Δειγματοληψία Υπογραφήματος

Έναντι της δειγματοληψίας κόμβων και ακμών που βασίζονται στο πλήρες γράφημα, ένας ουσιαστικά διαφορετικός τρόπος είναι να δειγματιστούν πολλαπλά υπογραφήματα και να περιοριστεί η αναζήτηση γειτονιάς μέσα σε αυτά. Το ClusterGCN (Chiang et al., 2019) λαμβάνει δείγματα υπογραφημάτων με αλγόριθμους ομαδοποίησης γραφημάτων, ενώ το GraphSAINT (Zeng et al., 2020) λαμβάνει απευθείας δείγματα κόμβων ή ακμών για να δημιουργήσει ένα υπογράφημα.

3.2.3 Μονάδες Συγκέντρωσης

Ένα συνελικτικό επίπεδο ακολουθείται συνήθως από ένα επίπεδο συγκέντρωσης για να ληφθούν πιο γενικά χαρακτηριστικά. Σε πολύπλοκα και μεγάλα γραφήματα συνήθως υπάρχουν πλούσιες ιεραρχικές δομές που έχουν μεγάλη σημασία για εργασίες ταξινόμησης σε επίπεδο κόμβου και σε επίπεδο γραφήματος. Παρόμοια με αυτά τα επίπεδα συγκέντρωσης, πολλή δουλειά επικεντρώνεται στο σχεδιασμό ιεραρχικών επιπέδων συγκέντρωσης σε γραφήματα. Παρακάτω αναφέρονται δύο είδη μονάδων συγκέντρωσης: μονάδες άμεσης συγκέντρωσης και μονάδες ιεραρχικής συγκέντρωσης.

Μονάδες άμεσης συγκέντρωσης

Οι μονάδες άμεσης συγκέντρωσης χρησιμοποιώντας διαφορετικές στρατηγικές επιλογής κόμβων έχουν σαν αποτέλεσμα να μαθαίνουν αναπαραστάσεις σε επίπεδο γραφήματος απευθείας από τους κόμβους.

- **Απλή συγκέντρωση κόμβων**

Η συγκεκριμένη μέθοδος χρησιμοποιείται από αρκετά μοντέλα. Οι λειτουργίες μεγίστου, μέσου όρου, αθροίσματος και προσοχής κόμβων εφαρμόζονται σε χαρακτηριστικά κόμβου για να ληφθεί μια συνολική αναπαράσταση γραφήματος.

- **SortPooling**

Η μέθοδος αυτή ταξινομεί πρώτα τις ενσωματώσεις κόμβων σύμφωνα με τους δομικούς ρόλους των κόμβων και στη συνέχεια οι ταξινομημένες ενσωματώσεις τροφοδοτούνται σε ένα συνελικτικό νευρωνικό δίκτυο για να λάβουν την αναπαράσταση.

Μονάδες ιεραρχικής συγκέντρωσης

Αυτές οι μέθοδοι διερευνούν την ιεραρχική ιδιότητα της δομής του γραφήματος. Ακολουθούν δηλαδή ένα μοτίβο συγκέντρωσης και μαθαίνουν αναπαραστάσεις γραφημάτων ανά επίπεδα.

- **Τραχύτητα γραφήματος - Graph Coarsening**

Οι μέθοδοι φασματικής ομαδοποίησης έχουν το μειονέκτημα του σταδίου ιδιοδιάσπασης (eigendecomposition) με αποτέλεσμα να είναι αναποτελεσματικοί. Οι Dhillon et al., 2007 προτείναν το Graclus το οποίο παρέχει έναν ταχύτερο τρόπο για την ομαδοποίηση κόμβων και εφαρμόζεται ως μονάδα συγκέντρωσης. Το ChebNet για παράδειγμα χρησιμοποιεί το Graclus για να συγχωνεύσει ζεύγη κόμβων και να προσθέσει επιπλέον κόμβους. Με σκοπό να βεβαιωθεί ότι η διαδικασία συγκέντρωσης σχηματίζει ένα ισορροπημένο δυαδικό δέντρο.

- **EigenPooling**

Οι Ma et al., 2019 εισήγαγαν έναν τελεστή συγκέντρωσης EigenPooling που βασίζεται στον μετασχηματισμό γραφήματος Fourier, ο οποίος μπορεί να χρησιμοποιήσει τα χαρακτηριστικά του κόμβου και τις τοπικές δομές κατά τη διαδικασία συγκέντρωσης για την εξαγωγή πληροφοριών υπογράφων. [25]

- **Self-Attention Graph Pooling – SAGPool**

Το SAGPool προτείνεται επίσης να χρησιμοποιήσει από κοινού χαρακτηριστικά και τοπολογία για να μάθει αναπαραστάσεις γραφημάτων. Χρησιμοποιεί μια μέθοδο που βασίζεται στην αυτοπροσοχή με λογική πολυπλοκότητα χρόνου και χώρου. [26]

3.3 Εφαρμογές Νευρωνικών Δικτύων Γράφων

Τα τελευταία χρόνια, τα γραφήματα έχουν αποκτήσει τεράστια δημοτικότητα λόγω της ικανότητάς τους να αναπαριστούν προβλήματα του πραγματικού κόσμου με συνδεδεμένους τρόπους. Οι εφαρμογές είναι δομημένα δεδομένα. Τα δεδομένα χρησιμοποιούνται ως αδόμητη μορφή δεδομένων, όπως σε δοκιμές, οι εικόνες μοντελοποιούνται ως γραφήματα για ανάλυση σε κοινωνικά δίκτυα, μοριακές δομές, δεδομένα συνδέσμων στο διαδίκτυο κ.λπ. Τα GNN έχουν πολλές εφαρμογές σε διάφορες δραστηριότητες και περιοχές. Περιγράφουμε διάφορες εφαρμογές με βάση τα ακόλουθα πεδία έρευνας. [27]

Επεξεργασία Φυσικής Γλώσσας - Natural Language Processing

Η βαθιά μάθηση έχει γίνει η κυρίαρχη προσέγγιση στην αντιμετώπιση διαφόρων εργασιών στην Επεξεργασία Φυσικής Γλώσσας σήμερα, ειδικά όταν λειτουργεί σε σώματα κειμένου μεγάλης κλίμακας. Με την πρόσφατη επιτυχία των τεχνικών ενσωμάτωσης λέξεων, οι προτάσεις αναπαρίστανται συνήθως ως μια ακολουθία διακριτικών σε εργασίες NLP. Ως εκ τούτου, δημοφιλείς τεχνικές βαθιάς μάθησης, όπως τα επαναλαμβανόμενα νευρωνικά δίκτυα και τα συνελικτικά νευρωνικά δίκτυα, έχουν εφαρμοστεί ευρέως για τη μοντελοποίηση ακολουθίας κειμένου. Ωστόσο, υπάρχει μια πλούσια ποικιλία προβλημάτων NLP που μπορούν να εκφραστούν καλύτερα με μια δομή γραφήματος. Για παράδειγμα, οι δομικές πληροφορίες της πρότασης σε μια ακολουθία κειμένου δηλαδή δέντρα συντακτικής ανάλυσης όπως δέντρα ανάλυσης εξάρτησης και συνιστώσας μπορούν να αξιοποιηθούν για να αυξηθούν τα δεδομένα αρχικής ακολουθίας ενσωματώνοντας τη γνώση που σχετίζεται με την εργασία. Ομοίως, οι σημασιολογικές πληροφορίες σε δεδομένα ακολουθίας δηλαδή γραφήματα σημασιολογικής ανάλυσης όπως γραφήματα αναπαράστασης αφηρημένου νοήματος και γραφήματα εξαγωγής πληροφοριών μπορούν επίσης να αξιοποιηθούν για τη βελτίωση των αρχικών δεδομένων ακολουθίας. Επομένως, αυτά τα γραφικά δομημένα δεδομένα μπορούν να κωδικοποιήσουν περίπλοκες σχέσεις ανά ζεύγη μεταξύ των διακριτικών οντοτήτων για την εκμάθηση πιο ενημερωτικών αναπαραστάσεων. [27][28]

Traffic

Σε ένα σύστημα μεταφορών, η ακριβής πρόβλεψη της ταχύτητας κυκλοφορίας, του θορύβου του δρόμου ή της πυκνότητας στα δίκτυα κυκλοφορίας είναι απαραίτητη. Πολλά έργα ερευνητών χρησιμοποιούν χωροχρονικά GNN για την κατασκευή μοντέλων για την αντιμετώπιση διαφόρων προβλημάτων δικτύου κυκλοφορίας. Σε τέτοια έργα, οι συγγραφείς παίρνουν το δίκτυο κυκλοφορίας ως χωροχρονικό γράφημα, τους αισθητήρες που είναι εγκατεστημένοι στους δρόμους ως κόμβους και την απόσταση μεταξύ των ζευγών αισθητήρων ως ακμές. Κάθε κόμβος είναι ένα χαρακτηριστικό δυναμικής εισαγωγής με μέση ταχύτητα κυκλοφορίας κατά τη διάρκεια ενός στιγμιότυπου.

Ταξινόμηση εικόνων - Image Classification

Οι πρώτοι αλγόριθμοι ταξινόμησης εικόνων εγγράφων χρησιμοποιούσαν την οπτική αναγνώριση χαρακτήρων για να αντλήσουν πληροφορίες περιεχομένου. Πολλές προηγμένες τεχνικές, όπως χαρακτηριστικά εικόνας, χαρακτηριστικά κειμένου και πληροφορίες διάταξης εγγράφων για ταξινόμηση εικόνων εγγράφων, εμφανίστηκαν με επιτυχία τις τελευταίες δεκαετίες. Τα Deep Convolutional Neural Networks (DCNN) είναι ένα από τα επιτυχημένα μοντέλα που παρέχουν νέα εργαλεία για την κατηγοριοποίηση εικόνων εγγράφων, καθώς μπορούν να εξάγουν εμφανείς και ιεραρχικές αναπαραστάσεις οπτικών χαρακτηριστικών. Το DCNN μπορεί να αντικατοπτρίζει εν μέρει την ιεραρχική δομή της διάταξης του εγγράφου.

Κατηγοριοποίηση κειμένου

Τα GNN έχουν κερδίσει μεγάλη προσοχή πρόσφατα και γίνονται πιο δημοφιλή στην κατηγοριοποίηση κειμένων. Η μεγαλύτερη πρόκληση είναι να γεφυρωθεί το σημασιολογικό και συντακτικό χάσμα μεταξύ των γλωσσών. Η πλειονότητα των διαθέσιμων τεχνικών αναζητά σημασιολογικές ομοιότητες μεταξύ των γλωσσών και μαθαίνει μια ανεξαρτήτου γλώσσας αναπαράσταση για έγγραφα γραμμένα σε πολλές διαφορετικές γλώσσες.

Χημεία

Οι ερευνητές χρησιμοποιούν επίσης GNN στη χημεία για να χαρακτηρίσουν ενώσεις ή μόρια όπου τα άτομα αναπαρίστανται ως κόμβοι σε ένα γράφημα ενώσεων ή μορίων και οι χημικοί δεσμοί μεταξύ τους αντιπροσωπεύονται ως ακμές. Ο κύριος τομέας των ερευνητών στα γραφήματα ενώσεων ή μορίων είναι η ταξινόμηση κόμβων, η δημιουργία γραφημάτων και η ταξινόμηση γραφημάτων. Βοηθάει σε εργασίες, όπως η εκμάθηση μοριακών δακτυλικών αποτυπωμάτων, η πρόβλεψη μοριακών ιδιοτήτων και η σύνθεση χημικών ενώσεων.

Συστήματα συστάσεων - Recommendation systems

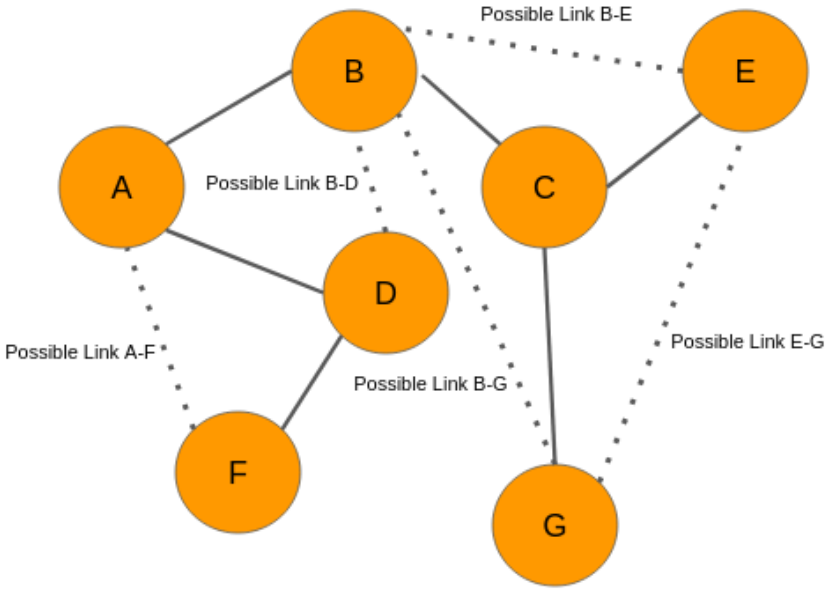
Τα νευρωνικά δίκτυα γραφημάτων χρησιμοποιούνται με επιτυχία σε προβλήματα συστημάτων συστάσεων. Τα συστήματα συστάσεων που βασίζονται σε γραφήματα αντιμετωπίζουν τα πράγματα και τους χρήστες ως κόμβους. Τα συστήματα συστάσεων που βασίζονται σε γραφήματα μπορούν να κάνουν προτάσεις υψηλής ποιότητας αξιοποιώντας τις σχέσεις μεταξύ χρηστών, στοιχείων, χρηστών και στοιχείων και των πληροφοριών περιεχομένου. Το σημαντικό για τη σύσταση σε ένα σύστημα είναι να προσδιοριστεί εάν ένα στοιχείο είναι σημαντικό για τον χρήστη. Μπορεί επομένως να αποτυπωθεί ως ζήτημα με την πρόβλεψη των σχέσεων.

4 Μεθοδολογίες Πρόγνωσης Ακμών

Σε αυτήν την ενότητα, εξετάζουμε μια σειρά μεθόδων για την πρόγνωση ακμών. [16][29] Όλες οι μέθοδοι εκχωρούν ένα βάρος σύνδεσης $score(x,y)$ σε ζεύγη κόμβων $\{x,y\}$ με βάση το γράφημα εισόδου G και στη συνέχεια, παράγουν μια λίστα κατάταξης με φθίνουσα σειρά βαθμολογίας $score(x,y)$. Έτσι, υπολογίζουν ένα μέτρο ομοιότητας μεταξύ των κόμβων x και y, σε σχέση με την τοπολογία του δικτύου. Γενικά, οι μέθοδοι προσαρμόζονται από τεχνικές που χρησιμοποιούνται στη θεωρία γραφημάτων και στην ανάλυση κοινωνικών δικτύων. Σε ορισμένες περιπτώσεις, αυτές οι τεχνικές δεν σχεδιάστηκαν για τη μέτρηση της ομοιότητας κόμβου σε κόμβο και επομένως πρέπει να τροποποιηθούν. Να σημειωθεί ότι κάποιες από τις παρακάτω μεθοδολογίες έχουν σχεδιαστεί μόνο για συνδεδεμένα γραφήματα.

4.1 Μεθοδολογίες με βάση την τοπολογία

Όταν το διάνυσμα από τα χαρακτηριστικά των χρηστών δεν είναι διαθέσιμο, χρησιμοποιούνται μεθοδολογίες βασισμένες στην τοπολογία του δικτύου. Οι μεθοδολογίες βασισμένες στην τοπολογία, υπολογίζουν μια τιμή για κάθε μη υπαρκτή ακμή του δικτύου.



Σχήμα 10: Πιθανές Ακμές Μεταξύ των Κόμβων

Στις μεθοδολογίες που θα αναφερθούν παρακάτω, με $\Gamma(x)$ και $\Gamma(y)$ συμβολίζεται το σύνολο των γειτόνων των κόμβων x και y αντίστοιχα (με τους οποίους συνδέονται απευθείας με ακμή).

4.1.1 Μεθοδολογίες που βασίζονται σε γειτονιές κόμβων

Ορισμένες προσεγγίσεις βασίζονται στην ιδέα ότι δύο κόμβοι x και y είναι πιο πιθανό να σχηματίσουν έναν σύνδεσμο στο μέλλον εάν τα σύνολα των γειτόνων τους $\Gamma(x)$ και $\Gamma(y)$ έχουν μεγάλη επικάλυψη. Ένα παράδειγμα τέτοιων κόμβων x και y είναι όταν αντιπροσωπεύουν συγγραφείς με πολλούς κοινούς συναδέλφους, άρα είναι και πιο πιθανό να έρθουν σε επαφή οι ίδιοι μεταξύ τους.

Κοινοί γείτονες (Common Neighbors)

Η πιο άμεση εφαρμογή αυτής της ιδέας για την πρόβλεψη συνδέσμων είναι να καθοριστεί το

$$score(x, y) \coloneqq |\Gamma(x) \cap \Gamma(y)|$$

Δηλαδή ο αριθμός των γειτόνων που έχουν κοινό τα x και y. Ο υπολογισμός αυτής της ποσότητας έγινε στο πλαίσιο των δικτύων συνεργασίας, επαληθεύοντας μια συσχέτιση μεταξύ του αριθμού των κοινών γειτόνων των x και y τη χρονική στιγμή t και της πιθανότητας να συνεργαστούν στο μέλλον.

Συντελεστής Jaccard (Jaccard Coefficient)

Ο συντελεστής Jaccard είναι μια μετρική ομοιότητας που χρησιμοποιείται συνήθως στην ανάκτηση πληροφοριών. Υπολογίζει την πιθανότητα και το x και το y να έχουν ένα χαρακτηριστικό f, για ένα τυχαία επιλεγμένο χαρακτηριστικό f που έχει είτε το x είτε το y. Αν πάρουμε ότι τα χαρακτηριστικά (features) f είναι γείτονες αυτό οδηγεί στο παρακάτω:

$$score(x, y) \coloneqq \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Λαμβάνει τιμές στο διάστημα $[0,1]$ και όσο μεγαλύτερη είναι η τιμή του, τόσο πιθανότερο είναι στο μέλλον οι δύο συγκεκριμένοι κόμβοι να συνδεθούν με κάποια ακμή.

Συντελεστής Adamic-Adar (Adamic-Adar Coefficient)

Οι Adamic και Adar εξετάζουν ένα σχετικό μέτρο, στο πλαίσιο της απόφασης πότε δύο προσωπικές αρχικές ιστοσελίδες συνδέονται στενά. Για να γίνει αυτό, υπολογίζουν τα χαρακτηριστικά των ιστοσελίδων και καθορίζουν την ομοιότητα μεταξύ δύο σελίδων σύμφωνα με το παρακάτω

$$\sum_{z\,:\,feature\ shared\ by\ x,y} \frac{1}{\log\left(frequency(z)\right)}$$

Αυτό βελτιώνει την απλή καταμέτρηση των κοινών χαρακτηριστικών, σταθμίζοντας περισσότερο τα σπάνια χαρακτηριστικά. Αυτό υπολογίζει το

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

Προτιμώμενη προσκόλληση (Preferential Attachment)

Δηλώνει ότι οι πιο δημοφιλείς κόμβοι, αυτοί με τους περισσότερους γείτονες, έχουν μεγαλύτερες πιθανότητες να αποκτήσουν νέες ακμές.

$$score(x, y) \coloneqq |\Gamma(x)| \cdot |\Gamma(y)|$$

Δείκτης Sorensen (Sorensen Index)

Η συγκεκριμένη τεχνική εξετάζει το πλήθος των κοινών γειτόνων και επισημαίνει ότι κόμβοι με χαμηλότερο βαθμό έχουν μεγαλύτερη πιθανότητα να συνδεθούν με άλλους κόμβους.

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

Ομοιότητα συνημιτόνου Salton

Είναι ακόμα μια κοινή μέθοδος συνημιτόνου η παραλλαγή της οποίας μετρά το πόσο όμοιοι είναι δύο κόμβοι μεταξύ τους.

$$score(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)| * |\Gamma(y)|}}$$

Δείκτης Hub Promoted

Ο δείκτης Hub Promoted ορίζει την τοπολογική επικάλυψη των κόμβων x και y. Παρατηρούμε ότι ο μικρότερος βαθμός κόμβων καθορίζει και την τιμή του δείκτη.

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

Δείκτης Hub Depressed

Ένας άλλος προτεινόμενος δείκτης παρόμοιος με τον προηγούμενο είναι ο δείκτης Hub Depressed. Η διαφορά είναι ότι η ομοιότητα των κόμβων υπολογίζεται σύμφωνα με το μέγιστο βαθμό των κόμβων x, y. Με αποτέλεσμα η βαθμολογία ομοιότητας που δίνει να είναι μικρότερη από τον Hub Promoted.

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

Συντελεστής Leicht-Holme-Nerman

Ο συγκεκριμένος συντελεστής μετρά σύμφωνα με την υψηλή ομοιότητα των ζευγών κόμβων που έχουν πολλούς κοινούς γείτονες συγκριτικά με τον αναμενόμενο αριθμό τέτοιων γειτόνων

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

Συντελεστής Parameter - Dependent

Για τη βελτίωση της πρόβλεψης δημοφιλών και όχι ακμών, οι (Zhu et al.) προτείνανε τη συγκεκριμένη μέτρηση:

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

όπου λ είναι μια ελεύθερη παράμετρος. Όταν το $\lambda = 0$ προκύπτει ο τύπος των κοινών γειτόνων (Common Neighbors). Όταν το $\lambda = 0.5$ προκύπτει ο τύπος ομοιότητας συνημιτόνου Salton και όταν $\lambda = 1$ ο τύπος συντελεστή Leicht-Holme-Nerman.

Κατανομή πόρων - Resource Allocation

Η μέτρηση της κατανομής πόρων είναι παρόμοια με του συντελεστή *Adamic-Adar.* Και οι δύο μετρήσεις επικεντρώνονται σε στοιχεία που μοιράζονται μεταξύ ενός μικρού αριθμού κόμβων. Η διαφορά με την AA, είναι ότι η RA «τιμωρεί» τους υψηλόβαθμους κοινούς γείτονες πιο βαριά. Έτσι το AA και η RA για δίκτυα με μικρούς μέσους βαθμούς έχουν παρόμοια αποτελέσματα, ενώ η RA δίνει καλύτερα αποτελέσματα για δίκτυα με μεγάλους μέσους βαθμούς. Επιπροσθέτως οι AA και RA επικεντρώνονται στους γείτονες των γειτόνων των κόμβων που εξετάζονται και ορίζεται ως εξής:

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

4.1.2 Μεθοδολογίες που βασίζονται στο σύνολο μονοπατιών

Μεθοδολογίες οι οποίες προσδιορίζουν την έννοια της απόστασης της συντομότερης διαδρομής εξετάζοντας το σύνολο όλων των μονοπατιών μεταξύ δύο κόμβων. Δίνοντας μια γενικότερη και πιο εγγύς λύση στο πρόβλημα της πρόβλεψης ακμών, καθώς εξερευνούν όλο τον Γράφο και δεν περιορίζονται μόνο στους γειτονικούς κόμβους.

Δείκτης Katz

Ορίζει ένα μέτρο που αθροίζει άμεσα τη συλλογή μονοπατιών, με εκθετική απόσβεση του μήκους για να μετράει με μεγαλύτερη ακρίβεια τα σύντομα μονοπάτια.

$$score(x, y) \coloneqq \sum_{l=1}^{\infty} \beta^l \cdot \left| paths_{x,y}^{l} \right|$$

Όπου το $paths_{x,y}^{l}$ είναι το σύνολο όλων των μονοπατιών μήκους από το x στο y. Ένα πολύ μικρό β αποδίδει προβλέψεις όπως οι κοινοί γείτονες, καθώς μονοπάτια μήκους τριών ή περισσότερων συμβάλλουν ελάχιστα στην άθροιση. Κάποιος μπορεί να επαληθεύσει ότι ο πίνακας των βαθμολογιών δίνεται από $(I - \beta M)^{-1} - I$, όπου M είναι ο πίνακας γειτνίασης του γραφήματος.

Θεωρούμε δύο παραλλαγές του δείκτη Katz: 1) **χωρίς βάρη**, όπου $paths_{x,y}^{1} = 1$ αν οι κόμβοι x και y έχουν συσχέτιση και 0 διαφορετικά, και 2) **με βάρη**, όπου $paths_{x,y}^{1}$ είναι ο αριθμός των φορών που οι κόμβοι x και y έχουν συσχετιστεί.

Δείκτης Local Path

Ο συγκεκριμένος δείκτης αξιοποιεί τις πληροφορίες των τοπικών μονοπατιών με μήκος 2 και μήκος 3. Αντίθετα με άλλους δείκτες που χρησιμοποιούν μόνο τις πληροφορίες των πλησιέστερων γειτόνων, εκμεταλλεύεται ορισμένες πρόσθετες πληροφορίες των γειτόνων σε αποστάσεις μήκους 3 από τον τρέχοντα κόμβο. Προφανώς, οι διαδρομές μήκους 2 είναι πιο σχετικές από τις διαδρομές μήκους 3, επομένως υπάρχει ένας παράγοντας προσαρμογής α που εφαρμόζεται στο μέτρο. θα πρέπει να είναι ένας μικρός αριθμός κοντά στο 0. Η μέτρηση ορίζεται ως εξής:

$$\mathrm{LP} = A^2 + \alpha A^3$$

Όπου A^2 είναι ο πίνακας γειτνίασης για τους κόμβους με απόσταση μήκους 2 και A^3 ο πίνακας γειτνίασης για τους κόμβους με απόσταση μήκους 3. Επομένως, ο LP είναι επίσης ένας πίνακας γειτνίασης που περιγράφει τα ζεύγη κόμβων με αποστάσεις μήκους 2 και 3.

Ομοιότητα δύναμης σχέσης - Relation Strength Similarity (RSS)

Η παραπάνω μέτρηση είναι ασύμμετρη και χρησιμοποιείται στα σταθμισμένα κοινωνικά δίκτυα. Υπολογίζεται με βάση την δύναμη της σχέσης ομοιότητας $R(x,y)$, μια κανονικοποιημένη βαθμολογία των βαρών των συνδέσμων που καθορίζει το σχετικό βαθμό ομοιότητας μεταξύ γειτονικών κόμβων. Υποθέτοντας ότι υπάρχουν L μονοπάτια $p_1, p_2, \dots, p_L$ μικρότερα από το r μεταξύ των x και y και το μονοπάτι p_L σχηματίζεται από τους K κόμβους $z_1, z_2, \dots, z_{k-1}$ και z_k. Τότε η ομοιότητα δύναμης σχέσης από το x στο y ορίζεται ως εξής:

$$\text{RSS}\ (x,y) = \sum_{l=1}^{L} R^*_{p_l}(x,y)$$

$$R^*_{p_l}(x,y) = \begin{cases} \prod_{k=1}^{K} R(z_k, z_{k+1}) & K \leqslant r \\ 0 & \text{αλλιώς} \end{cases}$$

Δείκτης FriendLink

Ο δείκτης FriendLink εξετάζει την ομοιότητα δύο κόμβων διασχίζοντας όλες τις διαδρομές σε μονοπάτια περιορισμένου μήκους. Υποθέτει ότι τα άτομα σε ένα κοινωνικό δίκτυο μπορούν να χρησιμοποιήσουν όλες τις διαδρομές μεταξύ τους, ανάλογα με τα μήκη διαδρομής. Η ομοιότητα μεταξύ x και y ορίζεται ως το πλήθος των μονοπατιών μεταβαλλόμενου μήκους l από x έως y:

$$score(x,y) = \sum_{i=1}^{l} \frac{1}{i-1} \cdot \frac{|\ paths^{\ i}_{x,y}\ |}{\prod_{j=2}^{i}\ (n-j)}$$

όπου n είναι ο αριθμός των κορυφών στο δίκτυο, l είναι το μήκος μιας διαδρομής μεταξύ x και y και $paths^i_{x,y}$ είναι το σύνολο όλων των διαδρομών μήκους - i από το x στο y. Επιπλέον, μεγάλο l δε σημαίνει και μεγάλη ακρίβεια στις μετρήσεις.

4.1.3 Μεθοδολογίες βάσει τυχαίων περιπάτων

Στα κοινωνικά δίκτυα οι αλληλεπιδράσεις μεταξύ των χρηστών δηλαδή των κόμβων ενός Γράφου μπορούν να μοντελοποιηθούν με την διαδικασία του τυχαίου περιπάτου. Η διαδικασία αυτή χρησιμοποιεί τις πιθανότητες για την διαγραφή της πορείας ενός τυχαίου περιπατητή από τον κόμβο στο οποίο βρίσκεται προς τους γειτονικούς κόμβους. Κάποιοι δείκτες πρόβλεψης ακμών μπορούν και υπολογίζουν την ομοιότητα μεταξύ των κόμβων αξιοποιώντας τη διαδικασία του τυχαίου περιπάτου.

Δείκτης Hitting Time & PageRank

Ο δείκτης Hitting Time $H_{x,y}$ από το x στο y είναι ο αναμενόμενος αριθμός βημάτων που απαιτούνται για έναν τυχαίο περίπατο που ξεκινά από το x για να φτάσει στο y.

Δεδομένου ότι ο δείκτης Hitting Time δεν είναι γενικά συμμετρικός, είναι επίσης φυσικό να λαμβάνεται υπόψη ο χρόνος μετακίνησης, *commute time* $C_{x,y} \coloneqq H_{x,y} + H_{y,x}$. Και οι δύο αυτές μετρήσεις χρησιμεύουν ως μέτρα φυσικής εγγύτητας και ως εκ τούτου (αρνητικά) μπορούν να χρησιμοποιηθούν ως βαθμολογία ομοιότητας score(x,y).

Ένα μειονέκτημα του δείκτη Hitting Time ως μέτρο εγγύτητας είναι ότι το $H_{x,y}$ είναι αρκετά μικρό όταν το y είναι ένας κόμβος με μεγάλη στατική πιθανότητα π_y, ανεξάρτητα από την ταυτότητα του x. Για να αντισταθμίσουμε αυτό το φαινόμενο, εξετάζουμε επίσης κανονικοποιημένες εκδοχές των χρόνων χτυπήματος και μετακίνησης, ορίζοντας:

$$score(x,y) \coloneqq -H_{x,y} \cdot \pi_y \quad or \quad score(x,y) \coloneqq -\left(H_{x,y} \cdot \pi_y + H_{y.x} \cdot \pi_x\right)$$

Μια άλλη δυσκολία με αυτά τα μέτρα είναι η ευαίσθητη εξάρτησή τους από μέρη του γραφήματος μακριά από τα x και y, ακόμη και όταν τα x και y συνδέονται με πολύ σύντομα μονοπάτια. Ένας τρόπος να αντιμετωπιστεί αυτό είναι να επιτραπεί η επανεκκίνηση της τυχαίας βάδισης από το x στο y, επιστρέφοντας στο x με μια πιθανότητα α σε κάθε βήμα. Με αυτόν τον τρόπο, μακρινά μέρη του γραφήματος δεν θα διερευνηθούν σχεδόν ποτέ.

Οι τυχαίες επανεκκινήσεις αποτελούν τη βάση του δείκτη **PageRank** για ιστοσελίδες και μπορούμε να το προσαρμόσουμε για την πρόβλεψη συνδέσμων ως εξής: Ορίζουμε ότι είναι η στατική πιθανότητα να βρεθούμε στον κόμβο y αν ξεκινήσουμε από το κόμβο x και σε κάθε βήμα του τυχαίου περιπάτου, υπάρχει πιθανότητα α να επιστρέψουμε στον κόμβο x και να ξεκινήσουμε από την αρχή και πιθανότητα $1 - \alpha$ να προχωρήσουμε στον επόμενο τυχαίο γειτονικό κόμβο.

Δείκτης Commute Time

Εφόσον ο δείκτης Hitting Time δεν είναι συμμετρικός, ο δείκτης Commute Time χρησιμοποιείται για τη μέτρηση των αναμενόμενων βημάτων τόσο από το x στο y όσο και από το y στο x. Ορίζεται ως εξής:

$$score(x,y) = \mathrm{HT}(x,y) + \mathrm{HT}(y,x) = m\left(L^{\dagger}_{x,x} + L^{\dagger}_{y,y} - 2L^{\dagger}_{x,y}\right)$$

όπου $L^{\dagger}$ είναι το ψευδο-αντίστροφο του πίνακα $L = D_A - A$, και m είναι ο αριθμός των ακμών σε ένα κοινωνικό δίκτυο.

Δείκτης Cosine Similarity Time

Η μέτρηση του συγκεκριμένου δείκτη βασίζεται στο $L^{\dagger}$ υπολογίζοντας την ομοιότητα δύο διανυσμάτων και μπορεί να οριστεί ως εξής:

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

Δείκτης SimRank

Ο δείκτης SimRank ορίζεται αναδρομικά: Δυο κόμβοι έχουν ομοιότητα αν συνδέονται με παρόμοιους γειτονικούς κόμβους. Αριθμητικά, αυτό προσδιορίζεται με τον ορισμό:

$$similarity(x,x) \coloneqq 1$$

και

$$similarity(x,y) \coloneqq \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} similarity(a,b)}{|\Gamma(x)| \cdot |\Gamma(y)|} \ , \gamma\iota\alpha\ \gamma\ \in [0,1]$$

Ο δείκτης SimRank μπορεί επίσης να ερμηνευθεί ως τυχαίος περίπατος σε ένα Γράφο. Είναι η αναμενόμενη τιμή του γ^l , όπου το l είναι μια τυχαία μεταβλητή που δίνει το χρόνο κατά τον οποίο οι τυχαίοι περίπατοι ξεκίνησαν από το x και το y, και συναντώνται για πρώτη φορά στον ίδιο κόμβο, περιδιαβαίνοντας τυχαίες ακμές του γράφου.

Rooted PageRank

Ο Rooted PageRank είναι μια τροποποίηση του δείκτη PageRank, ο οποίος είναι ο βασικός αλγόριθμος που χρησιμοποιείται από τη μηχανή αναζήτησης για την κατάταξη των αποτελεσμάτων αναζήτησης. Η κατάταξη ενός κόμβου στο γράφημα είναι ανάλογη με την πιθανότητα να ανακαλυφθεί ο κόμβος μέσω ενός τυχαίου περιπάτου στο γράφημα. Επιπλέον, υπάρχει ένας παράγοντας ϵ που καθορίζει πόσο πιθανό είναι ο αλγόριθμος να επισκεφθεί τους γείτονες του κόμβου παρά να ξεκινήσει από την αρχή. Έστω D ένας διαγώνιος πίνακας με

$$D_{i,i} = \sum_j A_{i,j}$$

Το μέτρο ορίζεται ως εξής:

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

PropFlow

Ο δείκτης PropFlow είναι παρόμοιος με τον Rooted PageRank. Η διαφορά τους είναι ότι η πιθανότητα ένας τυχαίος περίπατος από τον κόμβο x στον y δεν μπορεί να ξεπερνά τα l βήματα. Ο περιορισμένος περίπατος επιλέγει συνδέσμους βάσει βαρών και τερματίζει όταν φτάσει στο y ή επισκέπτεται ξανά οποιονδήποτε κόμβο για δεύτερη φορά. Εάν τα x και y συνδέονται άμεσα, ο δείκτης υπολογίζεται ως εξής:

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

όπου k είναι ο γείτονας του x του οποίου το βάθος είναι μεγαλύτερο από το βάθος του x από τον αρχικό κόμβο, το w_{xy} υποδηλώνει το βάρος της σύνδεσης μεταξύ των κόμβων x και y και a είναι ο προηγούμενος κόμβος του x σε μια τυχαία διαδρομή περιπάτου. Αν x είναι ο αρχικός κόμβος τότε $\mathrm{PF}(a,x) = 1$. Εάν τα x και y συνδέονται έμμεσα, το $\mathrm{PF}(x,y)$ είναι το άθροισμα του δείκτη PropFlow σε όλες τις συντομότερες διαδρομές από το x στο y.

Σε αντίθεση με τον δείκτη Rooted PageRank, ο υπολογισμός του PropFlow δεν απαιτεί επανεκκίνηση του τυχαίου περιπάτου, απλώς χρησιμοποιεί μια τροποποιημένη πρώτη αναζήτηση πλάτους που περιορίζεται στο ύψος l. Με αποτέλεσμα να είναι μια ταχύτερη μέτρηση από τον Rooted PageRank και τον SimRank.

5 Συμπεράσματα

Τα νευρωνικά δίκτυα γραφημάτων τα τελευταία χρόνια είναι από τις πρώτες επιλογές στον τομέα της μηχανικής μάθησης και παίζουν σημαντικό ρόλο στην έρευνα ανάλυσης κοινωνικών δικτύων. Στόχος αυτής της διπλωματικής εργασίας είναι να γίνει μια γνωριμία με το ερευνητικό πεδίο της ανάλυσης κοινωνικών δικτύων, κυρίως μέσω νευρωνικών δικτύων σε γράφους και να αναδειχθεί το πρόβλημα της πρόγνωσης ακμών σε κοινωνικά δίκτυα. Στο τέλος παρουσιάζονται οι πιο διαδεδομένες τεχνικές πρόβλεψης που εφαρμόζονται σε γραφήματα.

Διαπιστώνουμε ότι σε γενικές γραμμές το πρόβλημα της πρόβλεψης ακμών προσεγγίζεται από τους επιστήμονες είτε από τη δομή του δικτύου είτε από τα χαρακτηριστικά των κόμβων και των συνδέσμων. Η δομή αναφέρεται στον τρόπο με τον οποίο συνδέονται οι κόμβοι που συνθέτουν το δίκτυο καθώς και αντικατοπτρίζει τις πληροφορίες σχετικά με την τοπολογία του δικτύου.

Αν και οι μέθοδοι πρόβλεψης είναι αρκετές δεν έχει υπάρξει ακόμα κάποια σημαντική μέθοδος που να αποδίδει ικανοποιητικά. Ακόμα και νέες προσεγγίσεις, είτε βελτιώσεις των υπάρχων μοντέλων υπό περιορισμούς αποδίδουν καλά στα δεδομένα τα οποία παρουσιάζονται στην εκάστοτε εργασία αλλά δεν έχουν εξίσου καλή απόδοση εάν δοκιμαστούν σε κάποιο άλλο δίκτυο δεδομένων. Γενικότερα η δυσκολία του προβλήματος της πρόβλεψης ακμών είναι τα αραιά δεδομένα που υπάρχουν σε ένα δίκτυο. Όπως προαναφέραμε η δομή του δικτύου αποτελεί κύριο προσεγγιστικό παράγοντα των μοντέλων πρόβλεψης. Τα αραιά δεδομένα λοιπόν έχουν σαν αποτέλεσμα μια δομή δικτύου προβληματική που δυσκολεύει τις μεθόδους πρόβλεψης.

Σύμφωνα με τη βιβλιογραφία η πιο αποδοτική τεχνική είναι ο δείκτης Katz ενώ ακολουθεί η Προτιμώμενη προσκόλληση (Preferential Attachment). Η μέθοδος με τη χειρότερη απόδοση είναι αυτή των Κοινών γειτόνων (Common Neighbors). [30]

Βιβλιογραφία

[1] https://el.wikipedia.org/wiki/Social_media [Accessed 26 July 2022]

[2] Stanley Wasserman and Katherine Faust, “Social network analysis: Methods and applications”, vol. 8, Cambridge university press, 1994.

[3] R. Diestel, "Graph Theory", Springer, 3rd edition, 2005.

[4] Andrea Landherr, Bettina Friedl and Julia Heidemann, “A critical Review of Centrality Measures in Social Networks” [online], *Business & Information Systems Engineering*: Vol. 2: Issue 6, 371-385 (2010). Available from: https://aisel.aisnet.org/bise/vol2/iss6/5/ [Accessed 26 July 2022]

[5] Florian Boudin “A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction” [online], LINA - UMR CNRS 6241, Universite de Nantes, France, International Joint Conference on Natural Language Processing, pages 834–838, Nagoya, Japan, 14-18 October 2013. Available from: https://aclanthology.org/I13-1102/ [Accessed 26 July 2022]

[6] Justin Zhan, Sweta Gurung and Sai Phani Krishna Parsa, “Identifcation of top-K nodes in large networks using Katz centrality” [online], *J Big Data* **4,** 16 (2017). Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0076-5 [Accessed 26 July 2022]

[7] Saumya Chandrashekhar Suvarna, Mashrin Srivastava, Prof. B Jaganathan, Dr. Pankaj Shukla, “PageRank Algorithm using Eigenvector Centrality- New Approach” [online], v3 (2022), Available from: arXiv:2201.05469 [Accessed 26 July 2022]

[8] Rumi Ghosh, Kristina Lerman, “A Parameterized Centrality Metric for Network Analysis” [online], USC Information Sciences Institute (2010), Available from: arXiv:1010.4247 [Accessed 26 July 2022]

[9] Naoki Masuda, Michiko Sakaki, Takahiro Ezaki, Takamitsu Watanabe, “Clustering coefficients for correlation networks” [online], Institute of Cognitive Neuroscience, University College London (June 28, 2018). Available from: arXiv:1806.10228 [Accessed 26 July 2022]

[10] Santo Fortunato, & Darko Hric, “Community detection in networks: A user guide” [online]. Physics Reports 659, 1-44 (2016). Available from: arXiv:1608.00163 [Accessed 26 July 2022]

[11] Santo Fortunato, “Community detection in graphs” [online], Torino: Physics Reports 486, 75-174 (2010). Available from: arXiv:0906.0612 [Accessed 26 July 2022]

[12] Ulrike von Luxburg, “A Tutorial on Spectral Clustering” [online], Statistics and Computing 17(4), (2007). Available from: arXiv:0711.0189 [Accessed 26 July 2022]

[13] Li, Mei & Wang, Xiang & Gao, Kai & Zhang, Shanshan. (2017). “A Survey on Information Diffusion in Online Social Networks: Models and Methods” [online] Information (Switzerland). 8. 10.3390/info8040118. Available from: https://www.researchgate.net/publication/320458975_A_Survey_on_Information_Diffusion_in_Online_Social_Networks_Models_and_Methods [Accessed 26 July 2022]

[14] Pawan Kumar, Adwitiya Sinha, (2021) “Information diffusion modeling and analysis for socially interacting networks” [online] Available from: https://link.springer.com/article/10.1007/s13278-020-00719-7 [Accessed 26 July 2022]

[15] Furkan Gursoy, Ahmet Onur Durahim. (2018). “Predicting Diffusion Reach Probabilities via Representation Learning on Social Networks” [online] Proceedings of the 5th International Management Information Systems Conference. Available from: arXiv:1901.03829 [Accessed 26 July 2022]

[16] David Liben-Nowell and Jon Kleinberg, “The Link Prediction Problem for Social Networks” [online], in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM ’03, pp. 556–559, New York, NY, USA, 2003, ACM. Available from: https://www.cs.cornell.edu/home/kleinber/link-pred.pdf [Accessed 26 July 2022]

[17] Kyrkos, E. (2015). “Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων” [Undergraduate textbook] [online]. Kallipos, Open Academic Editions. Available from: http://hdl.handle.net/11419/1226 [Accessed 26 July 2022]

[18] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, Maosong Sun, “Graph neural networks: A review of methods and applications” [online] (2021). Available from: arXiv:1812.08434 [Accessed 26 July 2022]

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, (2017) “Attention is all you need” [online]. In: Proceeding of NIPS, pp. 5998–6008. Available from: arXiv:1706.03762 [Accessed 26 July 2022]

[20] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, George E. Dahl, (2017) “Neural Message Passing for Quantum Chemistry” [online] Proceedings of ICML 1263–1272. Available from: arXiv:1704.01212 [Accessed 26 July 2022]

[21] Antoni Buades, Bartomeu Coll, Jean-Michel Morel, 2005. “A non-local algorithm for image denoising” [online]. Proceedings of CVPR, 2. IEEE, pp. 60–65. Available from: https://ieeexplore.ieee.org/document/1467423 [Accessed 26 July 2022]

[22] Matteo Tiezzi, Giuseppe Marra, Stefano Melacci, Marco Maggini, (2020) “Deep Constraint-based Propagation in Graph Neural Networks” [online] Available from: arXiv:2005.02392 [Accessed 26 July 2022]

[23] Afshin Rahimi, Trevor Cohn, Timothy Baldwin, (2018) “Semi-supervised User Geolocation via Graph Convolutional Networks” [online] Available from: arXiv:1804.08049 [Accessed 08 August 2022]

[24] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, Stefanie Jegelka, (2018) “Representation Learning on Graphs with Jumping Knowledge Networks” [online] Available from: arXiv:1806.03536 [Accessed 08 August 2022]

[25] Yao Ma, Suhang Wang, Charu C. Aggarwal, Jiliang Tang, (2019) “Graph Convolutional Networks with EigenPooling” [online] Available from: arXiv:1904.13107 [Accessed 08 August 2022]

[26] Junhyun Lee, Inyeop Lee, Jaewoo Kang, (2019) “Self-Attention Graph Pooling” [online] Available from: arXiv:1904.08082 [Accessed 08 August 2022]

[27] Lilapati Waikhom, Ripon Patgiri, (2021) “Graph Neural Networks: Methods, Applications, and Opportunities” [online] Available from: arXiv:2108.10733 [Accessed 08 August 2022]

[28] Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, Bo Long, (2021) “Graph Neural Networks for Natural Language Processing: A Survey” [online] Available from: arXiv:2106.06090 [Accessed 08 August 2022]

[29] Peng Wang, Baowen Xu, Yurong Wu, Xiaoyu Zhou, “Link Prediction in Social Networks: the State-of-the-Art” (2014) [online], Available from: arXiv:1411.5118 [Accessed 26 July 2022]

[30] Fei Gao, Katarzyna Musial, Colin Cooper, Sophia Tsoka, "Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics", [online] Scientific Programming, vol. 2015, Article ID 172879, 13 pages, 2015. Available from: https://doi.org/10.1155/2015/172879 [Accessed 08 August 2022]