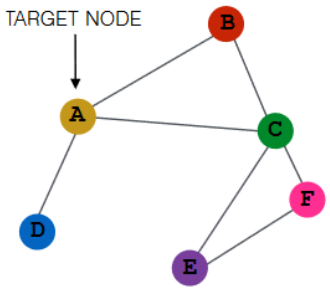
6 - Επαύξηση και εκπαίδευση

Δημήτριος Κοσμόπουλος

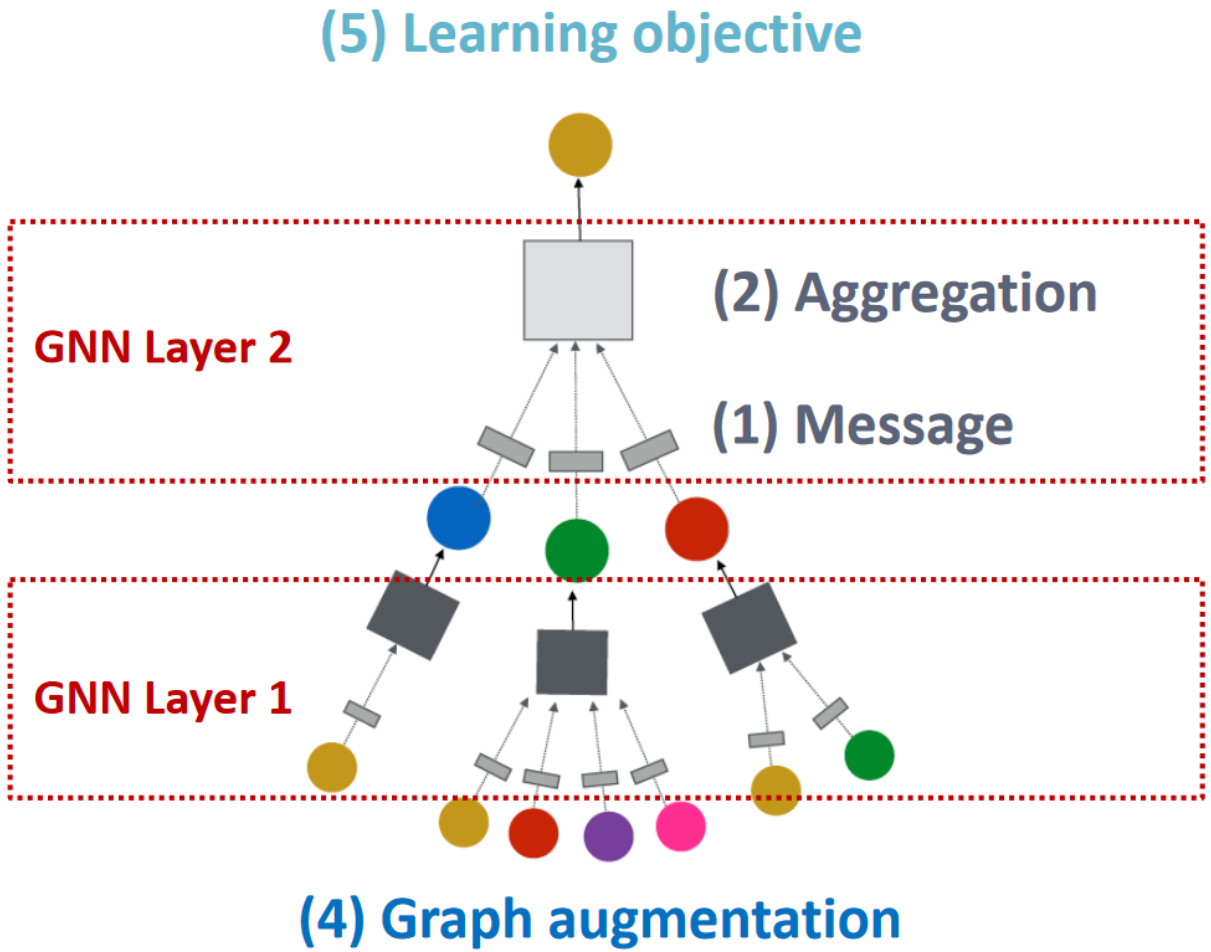
Πανεπιστήμιο Πατρών
Τμήμα Μηχανικών ΗΥ κ Πληροφορικής

12 Ιανουαρίου 2024

Γενικό πλαίσιο GNN



INPUT GRAPH



Επαύξηση γραφήματος

- Χαρακτηριστικά
  - Το γράφημα δεν έχει αρκετά χρήσιμα χαρακτηριστικά → επαύξηση χαρακτηριστικών
- Δομή γραφήματος
  - Το γράφημα είναι αραιό → πρόσθεσε εικονικούς κόμβους
  - Το γράφημα είναι πολύ πυκνό → κάνε δειγματοληψία στους γείτονες κατά τη μετάδοση μηνυμάτων
  - Το γράφημα είναι πολύ μεγάλο → κάνε δειγματοληψία υπογραφημάτων για τον υπολογισμό των ενσωματώσεων

Επαύξηση χαρακτηριστικών

Γιατί;

Ορισμένες δομές δεν μπορούν να προβλεφθούν

- Βαθμός
- Συντελεστής συσταδοποίησης
- *PageRank*
- κεντρικότητα
- ...

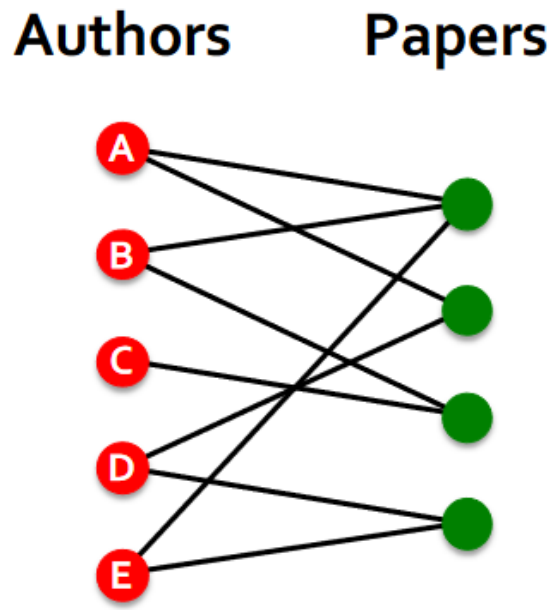
Κάθε πιθανό χαρακτηριστικό μπορεί να χρησιμοποιηθεί για επαύξηση.

Επαύξηση γραφήματος με ακμές

Πρόβλημα: τα μηνύματα δεν διαδίδονται αρκετά γρήγορα σε αραιά γραφήματα.

Λύση: Πρόσθεση εικονικών ακμών.

- Συνήθης τακτική: σύνδεσε γείτονες με βήμα 2 με εικονικές ακμές
- Αντί χρήση ως πίνακα γειτνίασης του πίνακα $\mathbf{A}$ χρήση του $\mathbf{A} + \mathbf{A}^2$
- Περίπτωση: Διμερή γραφήματα
  - Ερευνητές προς δημοσιεύσεις
  - Δημιουργείται γράφημα που συνδέει συνεργαζόμενους ερευνητές

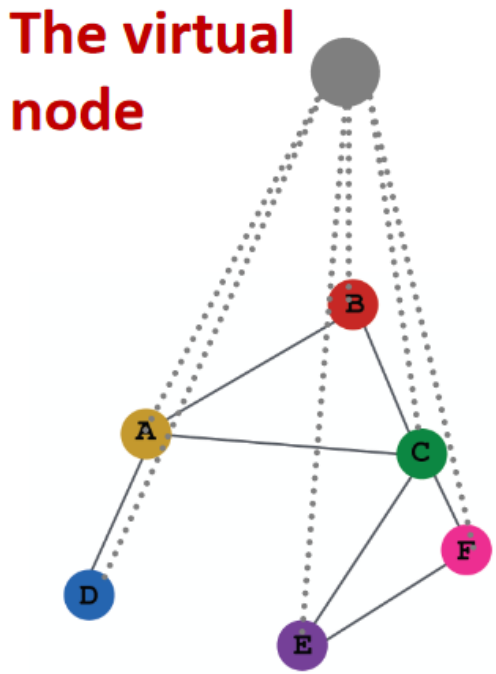


Επαύξηση γραφήματος με κόμβους

Πρόβλημα: τα μηνύματα δεν διαδίδονται αρκετά γρήγορα σε αραιά γραφήματα.

Λύση: Πρόσθεση εικονικών κόμβων.

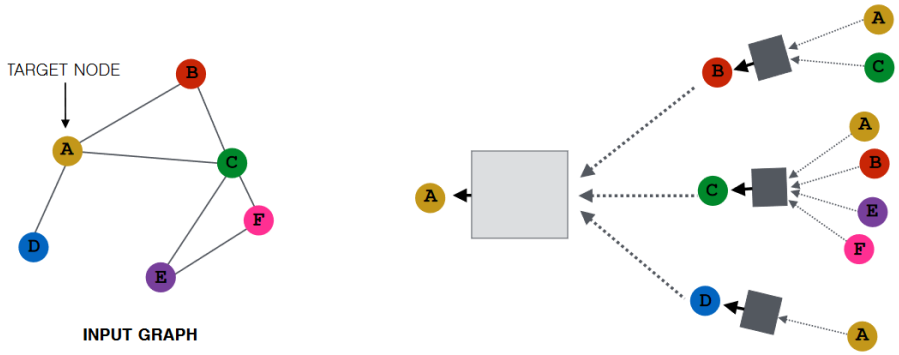
- Ο εικονικός κόμβος συνδέεται με όλους τους κόμβους του γραφήματος
  - Υποθέστε αραιό γράφημα όπου 2 κόμβοι έχουν απόσταση συντομότερου μονοπατιού 100
  - μετά την πρόσθεση του εικονικού κόμβου όλοι οι κόμβοι έχουν απόσταση 2
- βελτιώνεται σημαντικά ο μηχανισμός μηνυμάτων



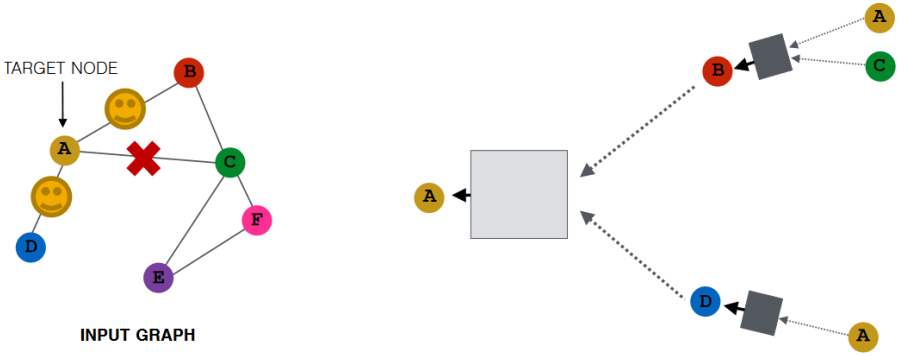
Δειγματοληψία

Πρόβλημα: οι πολλοί κόμβοι επιβραδύνουν την επεξεργασία.
Λύση: δειγματοληψία κόμβων.

- Αρχικά όλοι οι κόμβοι μεταδίδουν μηνύματα



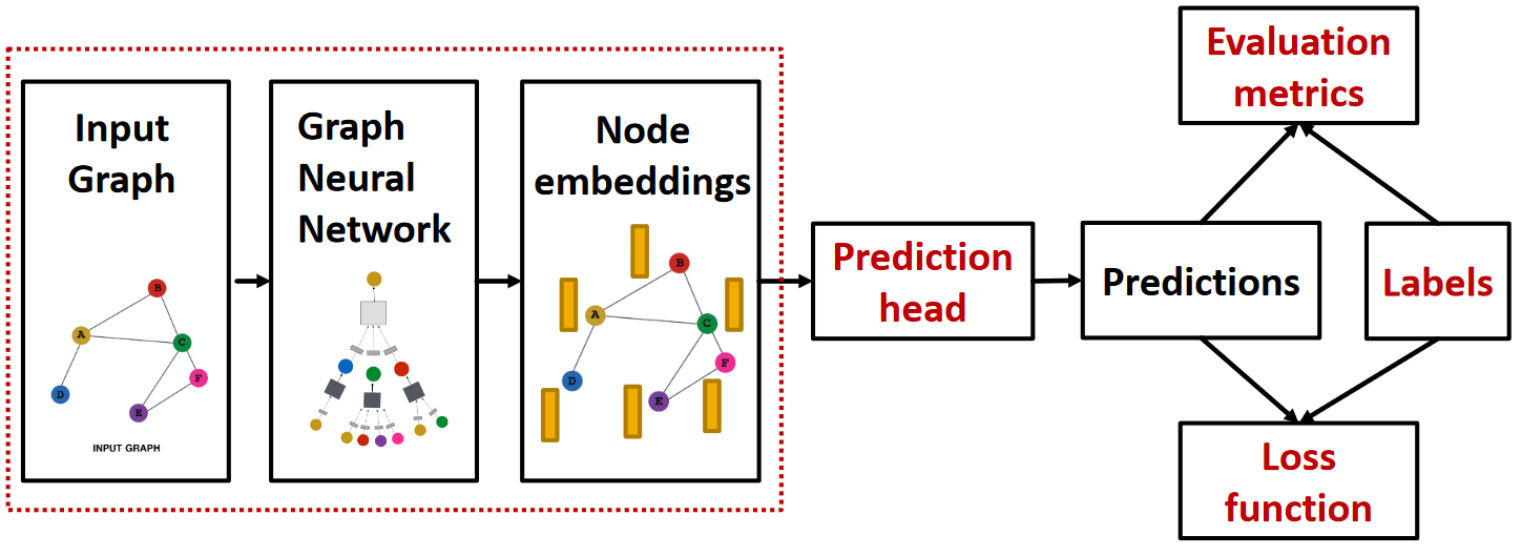
- Μετά τη δειγματοληψία κάποιες συνδέσεις καταργούνται



- Σε επόμενες επαναλήψεις θα γίνει νέα δειγματοληψία
- Το τελικό αποτέλεσμα δεν αλλάζει σημαντικά

Διαδικασία εκπαίδευσης

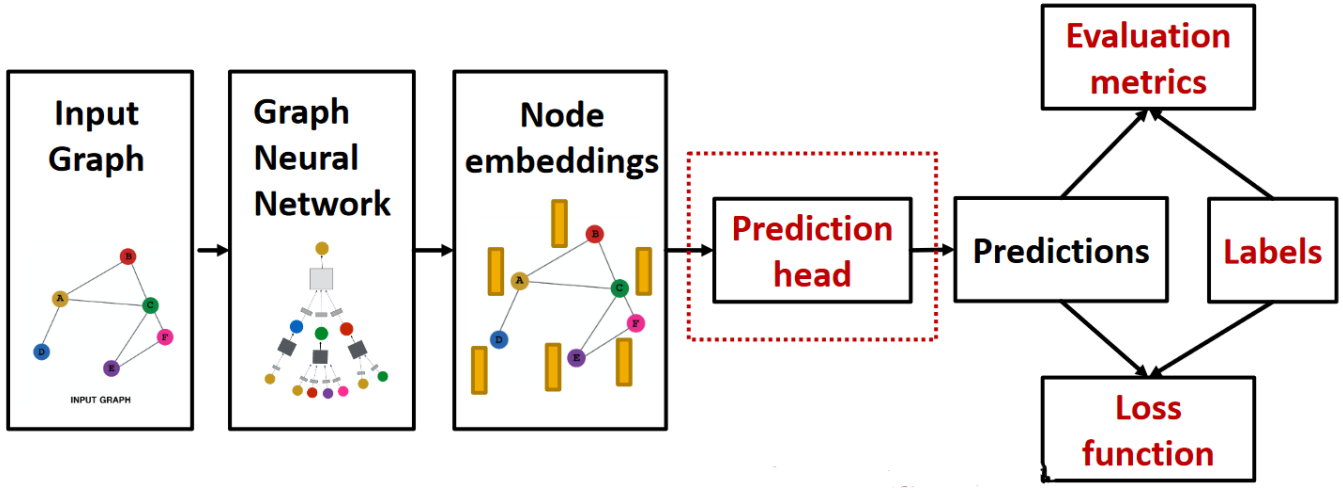
Εντός του κόκκινου διακεκομμένου πλαισίου ό,τι έχουμε καλύψει μέχρι στιγμής:



Αποτέλεσμα το σύνολο των ενσωματώσεων $\{\mathbf{h}_v^L, \forall v \in \mathcal{G}\}$

Διαδικασία εκπαίδευσης

Εντός του κόκκινου διακεκομμένου πλαισίου ό,τι θα καλύψουμε στη συνέχεια:



Εναλλακτικές προβλέψεις σε επίπεδο:

- κόμβων
- ακμών
- γραφημάτων

Πρόβλεψη σε επίπεδο κόμβων

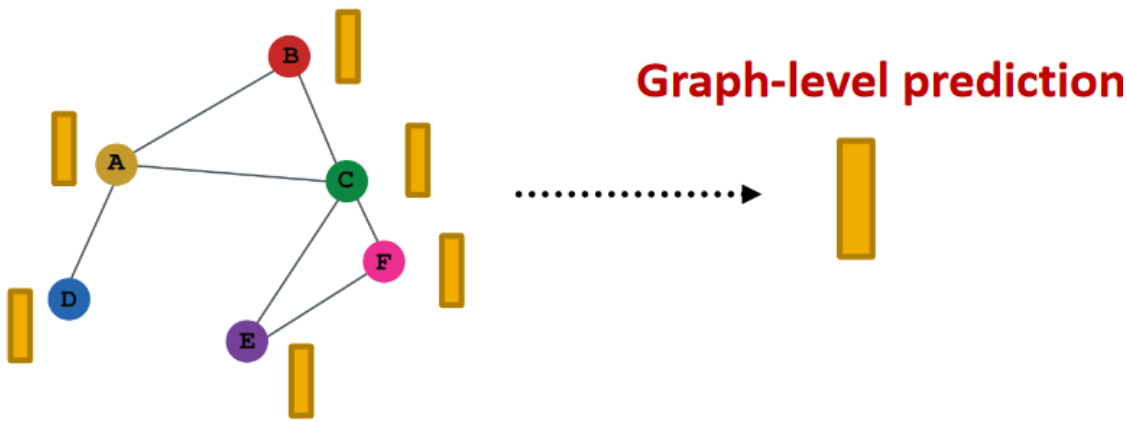
- Χρησιμοποιούμε απευθείας τα διανύσματα των ενσωματώσεων
- Μετά τον υπολογισμό του *GNN* έχουμε το σύνολο $\{\mathbf{h}_v^L \in \mathcal{R}^d, \forall v \in \mathcal{G}\}$
- Μπορούμε να κάνουμε ταξινόμηση σε k-κατηγορίες ή παλινδρόμηση σε k τιμές - στόχους
- $\hat{\mathbf{y}}_v = Head_{node}(\mathbf{h}_v^L) = \mathbf{W}^H \cdot \mathbf{h}_v^L$
  - $\mathbf{W}^H \in \mathcal{R}^{k \times d}$
  - απεικονίζουμε τα διανύσματα των ενσωματώσεων $\mathbf{h}_v^L \in \mathcal{R}^d$ στα $\hat{\mathbf{y}}_v \in \mathcal{R}^d$ ώστε να μπορούμε να υπολογίσουμε την συνάρτηση απώλειας

Πρόβλεψη σε επίπεδο ακμών

1. Συνένωση και γραμμικός μετασχηματισμός
   - $\hat{\mathbf{y}}_{uv} = Linear(Concat(\mathbf{h}_u^L, \mathbf{h}_v^L))$
   - το $Linear(.)$ ισοδυναμεί με πολλαπλασιασμό με πίνακα $\mathbf{W}^H \in \mathcal{R}^{k \times 2d}$
2. Εσωτερικό γινόμενο
   - $\hat{\mathbf{y}}_{uv} = (\mathbf{h}_u^L)^T \cdot \mathbf{h}_v^L$
   - ικανό για επίλυση δυαδικού προβλήματος π.χ. ύπαρξη ακμής ή όχι

Πρόβλεψη σε επίπεδο γραφήματος

- χρησιμοποιούμε όλα τα διανύσματα ενσωματώσεων στο γράφημα
- $\hat{\mathbf{y}}_G = Head_{graph}(\mathbf{h}_v^L \in \mathcal{R}^d, \forall v \in \mathcal{G})$

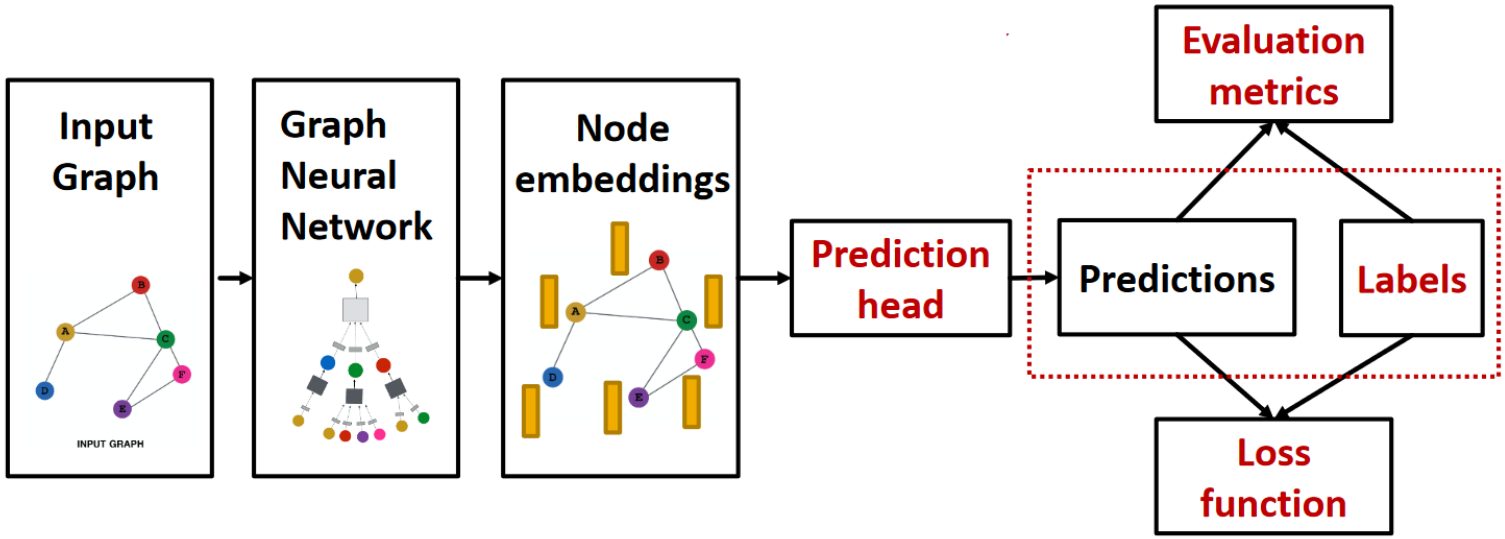


- παρόμοιο με το συγκερασμό $AGG(.)$ σε ένα επίπεδο GNN
  - mean pool
  - max pool
  - sum

Διαδικασία εκπαίδευσης

Από που προέρχονται οι ετικέτες ground truth;

- ετικέτες από επίβλεψη
- μη επιβλεπόμενα σήματα



Επιβλεπόμενη και Μη επιβλεπόμενη μάθηση

- Επιβλεπόμενη μάθηση
  - οι ετικέτες προέρχονται από εξωτερικές πηγές
  - π.χ. αν ένα μόριο είναι τοξικό
- Μη Επιβλεπόμενη μάθηση
  - οι ετικέτες προέρχονται από το ίδιο το γράφημα
  - π.χ. πρόβλεψη ακμής
- Κάποιες φορές οι διαφορές είναι δυσδιάκριτες
  - Έχουμε επίβλεψη σε μη επιβλεπόμενη μάθηση
  - π.χ. εκπαίδευση για πρόβλεψη συντελεστή συσταδοποίησης
  - εναλλακτικά υπάρχει ο όρος αυτοεπιβλεπόμενη μάθηση που αποδίδει ίσως καλύτερα την έννοια

Επιβλεπόμενες ετικέτες

Οι ετικέτες προέρχονται από την εφαρμογή π.χ.:

- ετικέτες κόμβων σε ένα δίκτυο ετεροαναφορών που περιγράφουν τη θεματική περιοχή
- ετικέτες ακμών σε ένα δίκτυο συναλλαγών για το αν μια συναλλαγή συνδέεται με απάτη
- ετικέτες γραφημάτων σε γραφήματα μορίων σε σχέση με την τοξικότητα του μορίου

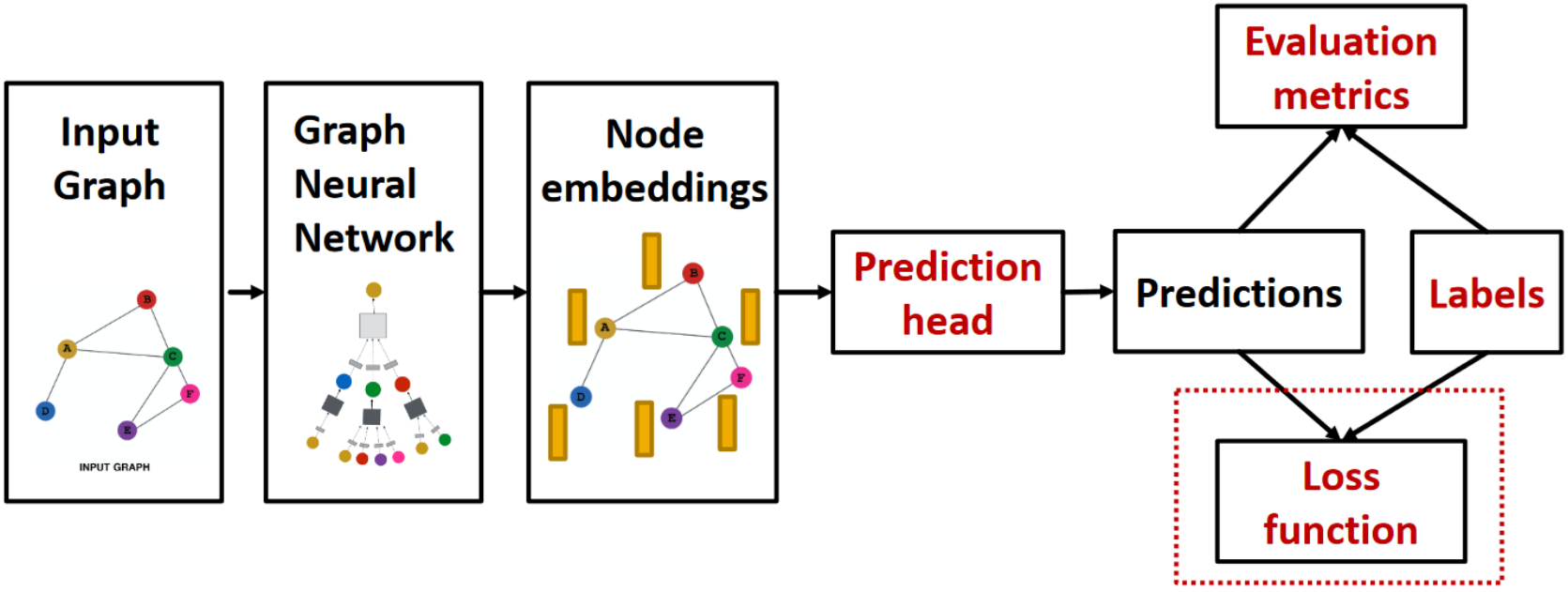
Προσπαθούμε πάντα οι ετικέτες να ανήκουν σε μια από τις παραπάνω κατηγορίες που είναι εύκολα διαχειρίσιμες

- π.χ. αν κάποιοι κόμβοι ξέρουμε ότι ανήκουν σε μια ομάδα μπορούμε να ορίσουμε ως ετικέτα την ομάδα

Μη επιβλεπόμενα σήματα

- Πολλές φορές το μόνο που έχουμε στη διάθεσή μας είναι το γράφημα χωρίς καμία ετικέτα
- Βρίσκουμε επιβλεπόμενα σήματα μέσα στο γράφημα
  - Σε επίπεδο κόμβων $\mathbf{y}_v$: μαθαίνουμε στατιστικά όπως συντελεστής συσταδοποίησης, *PageRank*, κλπ
  - Σε επίπεδο ακμών $\mathbf{y}_{uv}$: απαλείφουμε ακμές και μετά προσπαθούμε να τις προβλέψουμε
  - Σε επίπεδο γραφήματος $\mathbf{y}_g$: στατιστικές π.χ. για να δούμε αν δυο γραφήματα είναι ισομορφικά

Διαδικασία εκπαίδευσης



Πώς υπολογίζουμε τη συνάρτηση απώλειας:

- Απώλεια ταξινόμησης
- Απώλεια παλινδρόμησης

Εκπαίδευση *GNN*

- Έχουμε N δείγματα
- Κάθε δείγμα i μπορεί να είναι κόμβος / ακμή / γράφημα
  - επίπεδο κόμβων: πρόβλεψη $\hat{\mathbf{y}}_v^i$, ετικέτα $\mathbf{y}_v^i$
  - επίπεδο ακμών: πρόβλεψη $\hat{\mathbf{y}}_{uv}^i$, ετικέτα $\mathbf{y}_{uv}^i$
  - επίπεδο γραφήματος: πρόβλεψη $\hat{\mathbf{y}}_g^i$, ετικέτα $\mathbf{y}_g^i$
- Όλα τα παραπάνω θα τα αναπαριστούμε για ευκολία στη συνέχεια ενιαία ως: πρόβλεψη $\hat{\mathbf{y}}^i$, ετικέτα $\mathbf{y}^i$

Συνάρτηση απώλειας για Ταξινόμηση

Συνάρτηση $CrossEntropy - CE$ για ταξινόμηση σε K κλάσεις

- $CE(\mathbf{y}^i, \hat{\mathbf{y}}^i) = -\sum_{j=1}^{K} y_j^i log(\hat{y}_j^i)$
  όπου i είναι το δείγμα και j η κλάσση
- π.χ.
  $\mathbf{y}^i = [0, 0, 1, 0, 0]$ $(one - hot\ encoding)$
  $\hat{\mathbf{y}}_j^i = [0.1, 0.2, 0.5, 0.1, 0.1]$
- Συνολική απώλεια: $Loss = \sum_{i=1}^{N} CE(\mathbf{y}^i, \hat{\mathbf{y}}^i)$

Συνάρτηση απώλειας για Παλινδρόμηση

Χρησιμοποιούμε το μέσο τετραγωνικό σφάλμα (MSE)

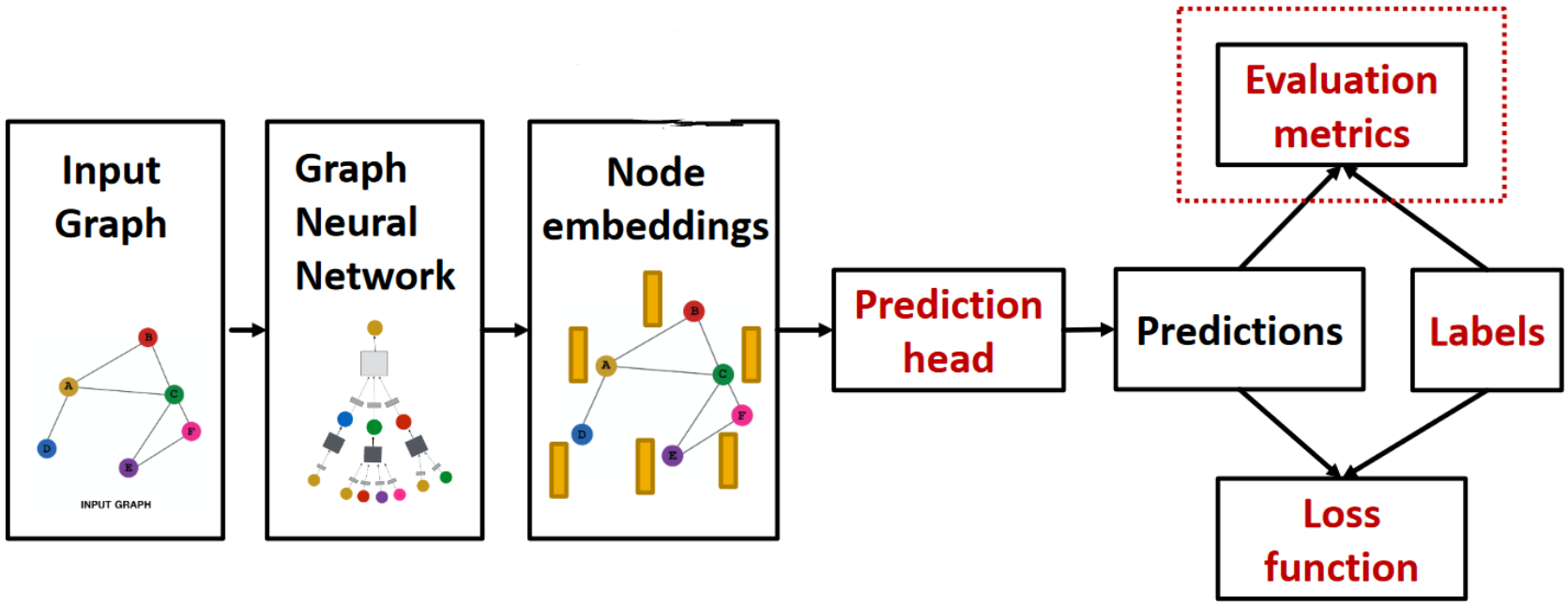
- $MSE(\mathbf{y}^i, \hat{\mathbf{y}}^i) = \sum_{j=1}^{K} (y_j^i - \hat{y}_j^i)^2$
  όπου i είναι το δείγμα και j η διάσταση του διανύσματος που εξετάζουμε
- π.χ.
  $\mathbf{y}^i = [1.5, 2.3, 4.5, 5.6, 0.2]$ διάνυσμα στόχος
  $\hat{\mathbf{y}}^i = [0.5, 2.1, 3.5, 4.6, 0.1]$ διάνυσμα πρόβλεψης

- Συνολική απώλεια: $Loss = \sum_{i=1}^{N} MSE(\mathbf{y}^i, \hat{\mathbf{y}}^i)$

Διαδικασία εκπαίδευσης

Πώς μετράμε την απόδοση ενός *GNN*;

- Ακρίβεια
- *ROC AUC*



Μετρικές

- Παλινδρόμηση:
  - $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}$
  - $\text{MAE} = \sum_{i=1}^{n}|\mathbf{y}_i - \hat{\mathbf{y}}_i|/n$
- Ταξινόμηση
  - ακρίβεια (*accuracy*)
  - precision, recall, F1
  - καμπύλη *ROC*, εμβαδό κάτω από την καμπύλη

Περισσότερες πληροφορίες:
https : *//en.wikipedia.org/wiki/Receiver_operating_characteristic*

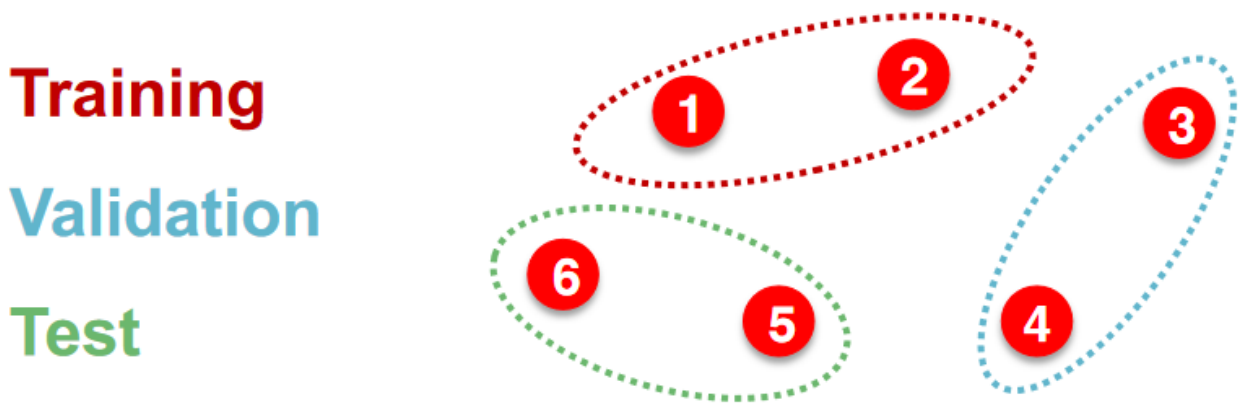
Διαχωρισμός δεδομένων

Πώς χωρίζουμε τα δεδομένα σε εκπαίδευσης/επικύρωσης/τεστ ;

- Σταθερός διαχωρισμός
  - εκπαίδευση: για βελτιστοποίηση παραμέτρων του *GNN*
  - επικύρωση: για διακοπή μάθησης / έλεγχο υπερπαραμέτρων
  - τεστ: κρατείται χωριστό για να αξιολογήσουμε το τελικό αποτέλεσμα
  - Πρόβλημα: δεν μπορούμε κάποιες φορές να εξασφαλίσουμε ότι το σύνολο για τεστ θα είναι ανεξάρτητο από τα προηγούμενα.
- Τυχαίος διαχωρισμός
  - Επαναλαμβάνουμε πολλές φορές και στο τέλος αναφέρουμε τη μέση τιμή από πολλά τρεξίματα

Ιδιαιτερότητες διαχωρισμού δεδομένων σε γραφήματα

- Έστω ότι θέλουμε να διαχωρίσουμε δεδομένα εικόνων
- Τα δείγματα είναι ανεξάρτητα π.χ. η εικόνα 5 δεν επηρεάζει την πρόβλεψη στην εικόνα 1



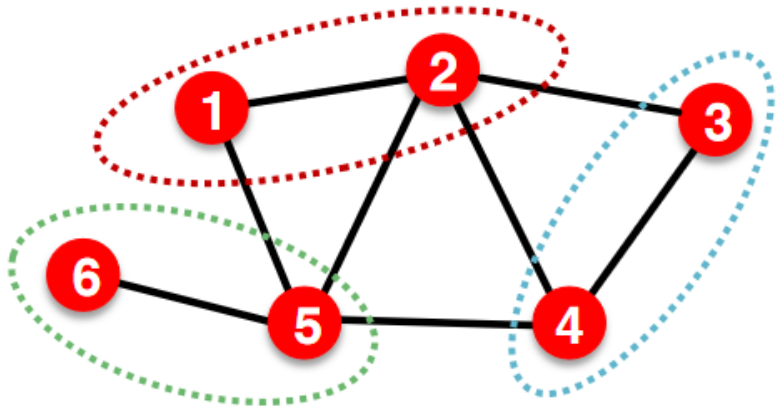
Ιδιαιτερότητες διαχωρισμού δεδομένων σε γραφήματα

- Ταξινόμηση κόμβων: Κάθε δείγμα είναι ένας κόμβος
- Τα δείγματα ΔΕΝ είναι ανεξάρτητα !!!
- Ο κόμβος 5 θα επηρεάσει την πρόβλεψη στον 1 μέσω της συμμετοχής του στο μηχανισμό μηνυμάτων, που καθορίζει την ενσωμάτωση του 1

Training

Validation

Test



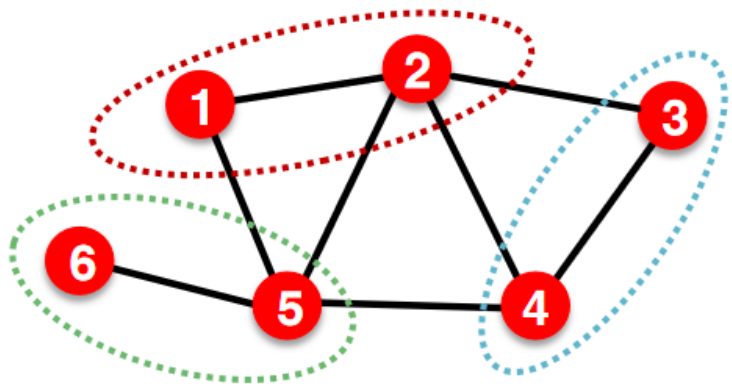
Λύση 1: μεταδοτική (transductive)

- Το γράφημα εμφανίζεται σε όλα τα σύνολα (εκπαίδευσης/επικύρωσης/τεστ)
- Διαχωρίζουμε μόνο τις ετικέτες των κόμβων
- Εκπαιδεύουμε χρησιμοποιώντας τις ενσωματώσεις όλων των κόμβων του γραφήματος και τις ετικέτες των 1,2
- Κάνουμε επικύρωση χρησιμοποιώντας τις ενσωματώσεις όλων των κόμβων του γραφήματος και τις ετικέτες των 3,4
- Κάνουμε τεστ χρησιμοποιώντας τις ενσωματώσεις όλων των κόμβων του γραφήματος και τις ετικέτες των 5,6

Training

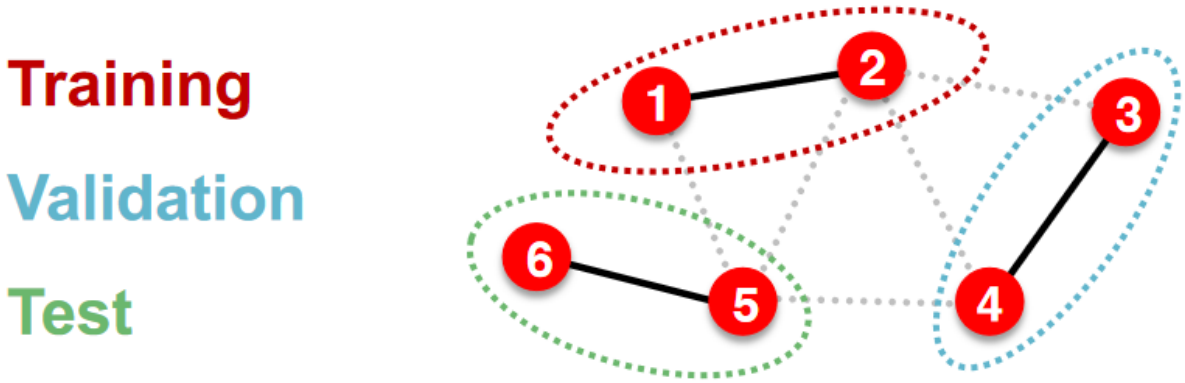
Validation

Test



Λύση 2: επαγωγική (inductive)

- Χωρίζουμε σε επιμέρους γραφήματα (εκπαίδευσης/επικύρωσης/τεστ)
- Τα γραφήματα είναι ανεξάρτητα π.χ. ο 5 δεν επηρεάζει τον 1 πλέον
- Εκπαιδεύουμε υπολογίζοντας τις ενσωματώσεις των κόμβων 1,2 του γραφήματος εκπαίδευσης και τις ετικέτες των 1,2
- Κάνουμε επικύρωση υπολογίζοντας τις ενσωματώσεις των κόμβων 3, 4 του γραφήματος και αξιολογούμε στις ετικέτες των 3,4
- Κάνουμε τεστ χρησιμοποιώντας τις ενσωματώσεις των κόμβων 5,6 και αξιολογούμε στις ετικέτες των 5,6



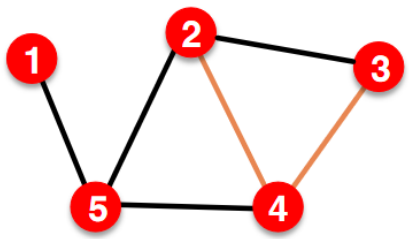
Εφαρμοσιμότητα

1. Μεταδοτική μέθοδος
   - Τα σύνολα στο ίδιο γράφημα
   - Διαχωρίζουμε μόνο τις ετικέτες
   - Εφαρμόσιμο σε προβλέψεις σε επίπεδο κόμβων/ακμών
2. Επαγωγική μέθοδος
   - Τα σύνολα σε χωριστά γραφήματα
   - Ζητούμενο η γενίκευση σε άλλα γραφήματα
   - Εφαρμόσιμο σε προβλέψεις σε επίπεδο κόμβων/ακμών/γραφημάτων

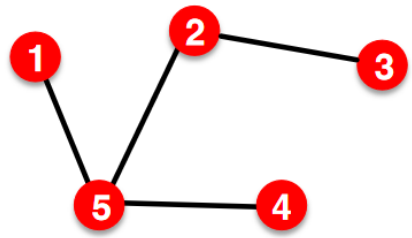
Πρόβλεψη ακμών

Στόχος η πρόβλεψη ακμών που λείπουν

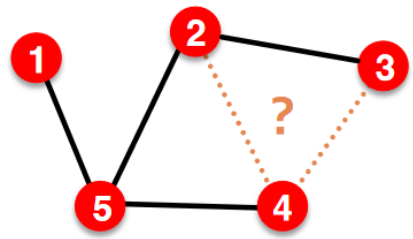
- Μη επιβλεπόμενη / αυτοεπιβλεπόμενη διαδικασία
- Δημιουργούμε τις ετικέτες και κάνουμε το διαχωρισμό μόνοι μας
- Καταργούμε κάποιες ακμές και προσπαθούμε να προβλέψουμε την ύπαρξή τους



Original graph



Input graph to GNN



Predictions made by GNN

Πρόβλεψη ακμών

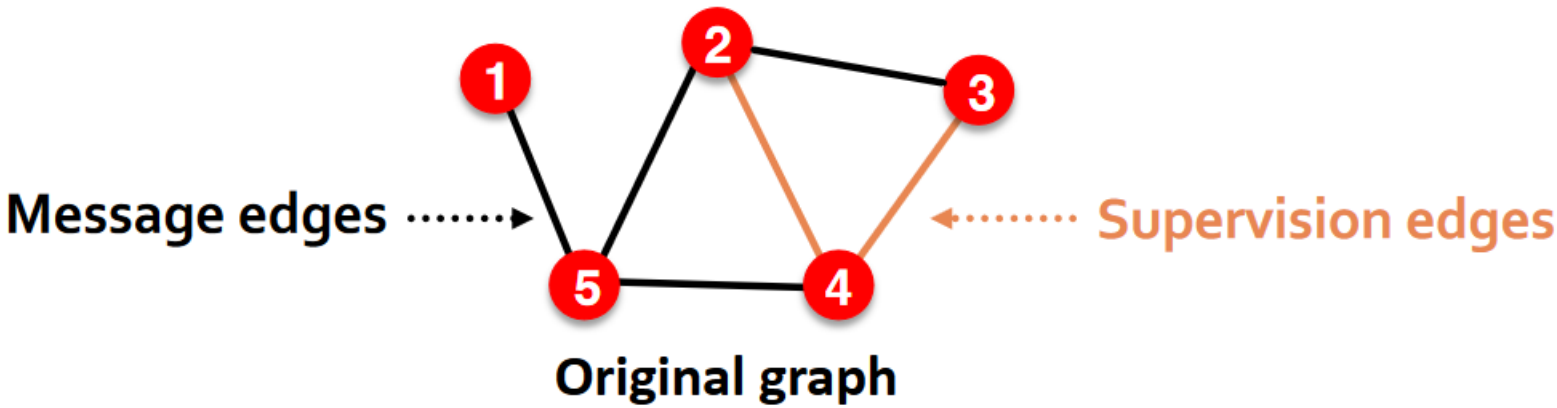
Κάνουμε διαχωρισμό σε 2 βήματα

Βήμα 1: ορίζουμε δύο τύπους ακμών

- Ακμές μηνυμάτων για υπολογισμό ενσωματώσεων
- Ακμές επίβλεψης για βελτιστοποίηση των συναρτήσεων απώλειας

Μετά το Βήμα 1:

- Αφαιρούμε τις ακμές επίβλεψης, οι οποίες δεν χρησιμοποιούνται για τον υπολογισμό των ενσωματώσεων του *GNN*
- Μόνο οι ακμές μηνυμάτων παραμένουν

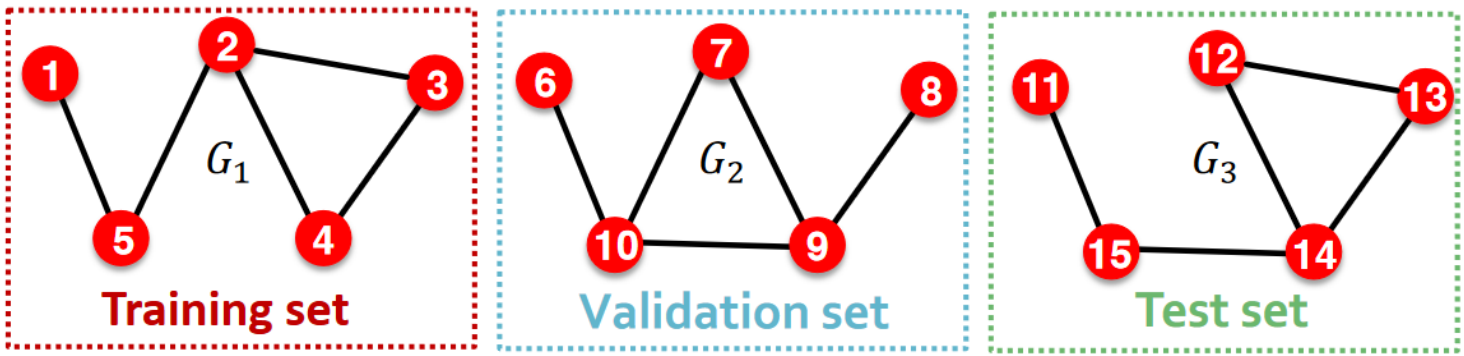


Πρόβλεψη ακμών

Βήμα 2: χωρίζουμε τις ακμές σε σύνολα εκπαίδευσης/επικύρωσης/τεστ

Επιλογή 1: Επαγωγικός διαχωρισμός

- Έστω έχουμε σύνολο από 3 γραφήματα
- Κάθε επαγωγικός διαχωρισμός θα περιέχει ένα ανεξάρτητο γράφημα

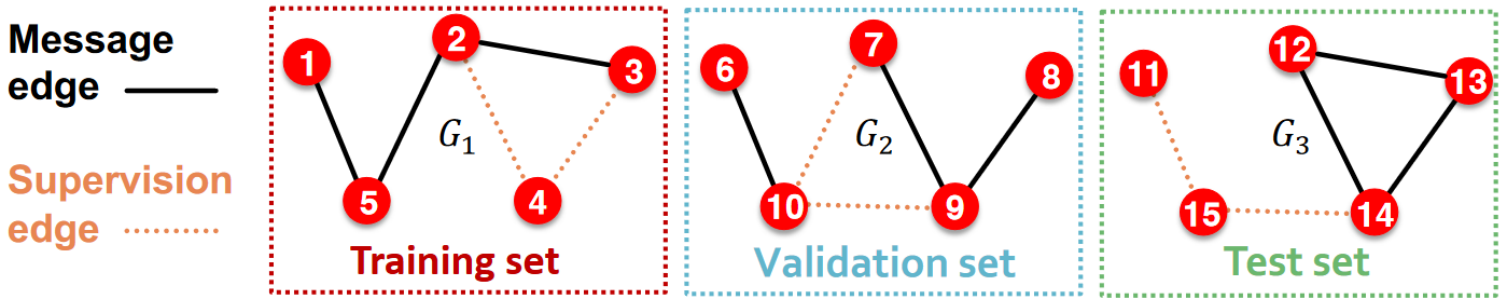


Πρόβλεψη ακμών

Βήμα 2: χωρίζουμε τις ακμές σε σύνολα εκπαίδευσης/επικύρωσης/τεστ

Επιλογή 1: Επαγωγικός διαχωρισμός

- Έστω έχουμε σύνολο από 3 γραφήματα
- Κάθε επαγωγικός διαχωρισμός θα περιέχει ένα ανεξάρτητο γράφημα
- Σε κάθε ένα από τα σύνολα εκπαίδευσης/επικύρωσης/τεστ κάθε γράφημα θα έχει 2 τύπους ακμών: μηνυμάτων και επίβλεψης
- μόνο οι ακμές μηνυμάτων χρησιμοποιούνται για τον υπολογισμό των ενσωματώσεων



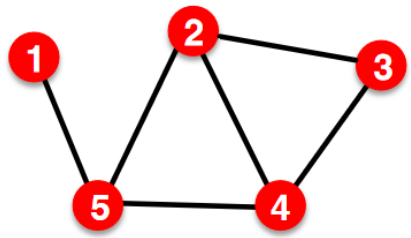
Πρόβλεψη ακμών

Επιλογή 2: Μεταδοτικός διαχωρισμός

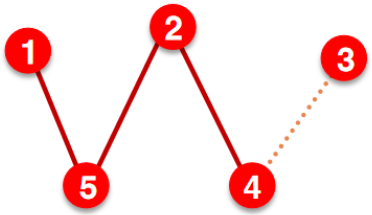
- Η βασική επιλογή για πρόβλεψη ακμών
- Έστω το σύνολο δεδομένων είναι ένα γράφημα
- Εξορισμού όλο το γράφημα είναι ορατό από κάθε ένα από τα σύνολα εκπαίδευσης/επικύρωσης/τεστ
- Θα πρέπει να διαχωρίσουμε από το σύνολο εκπαίδευσης τις ακμές επικύρωσης/τεστ
- Επίσης θα πρέπει να διαχωρίσουμε από το σύνολο εκπαίδευσης τις ακμές επικύρωσης κατά τον υπολογισμό των ενσωματώσεων

Πρόβλεψη ακμών

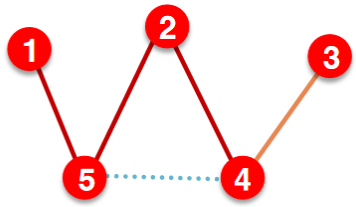
Επιλογή 2: Μεταδοτικός διαχωρισμός



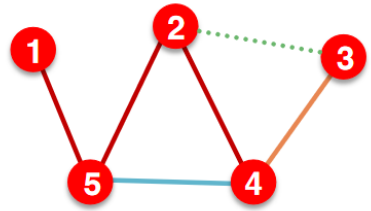
αρχικό γράφημα



Εκπαίδευση: χρησιμοποίησε τις ακμές μηνυμάτων εκπαίδευσης για να προβλέψεις τις ακμές επίβλεψης εκπαίδευσης



Επικύρωση: χρησιμοποίησε τις ακμές μηνυμάτων εκπαίδευσης και τις ακμές επίβλεψης εκπαίδευσης για να προβλέψεις τις ακμές επικύρωσης



Τεστ: χρησιμοποίησε τις ακμές μηνυμάτων εκπαίδευσης και τις ακμές επίβλεψης εκπαίδευσης και τις ακμές επικύρωσης για να προβλέψεις τις ακμές τεστ

Πρόβλεψη ακμών

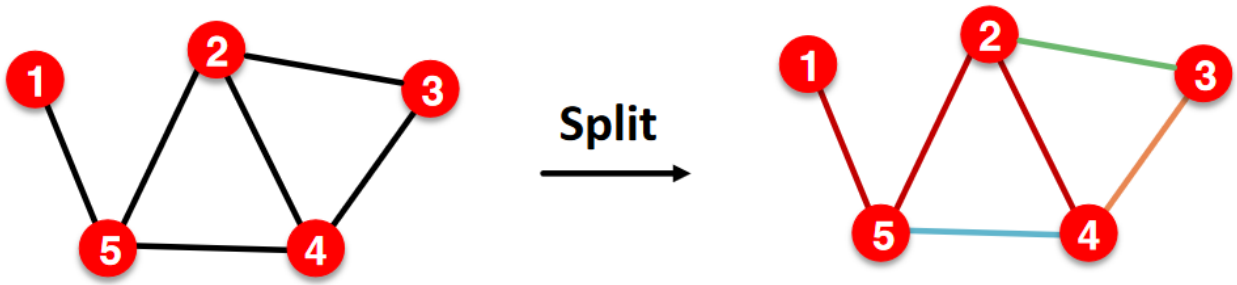
Επιλογή 2: Μεταδοτικός διαχωρισμός

- Γιατί χρησιμοποιούμε αυξανόμενο αριθμό ακμών·
- Μετά την εκπαίδευση οι ακμές επίβλεψης της εκπαίδευσης είναι γνωστές στο *GNN*
- Για το λόγο αυτό ένα ιδεατό μοντέλο θα πρέπει να χρησιμοποιεί ακμές επίβλεψης στον υπολογισμό των ενσωματώσεων στην επικύρωση
- Ομοίως για το τεστ

Πρόβλεψη ακμών

Επιλογή 2: Μεταδοτικός διαχωρισμός

Σύνοψη:

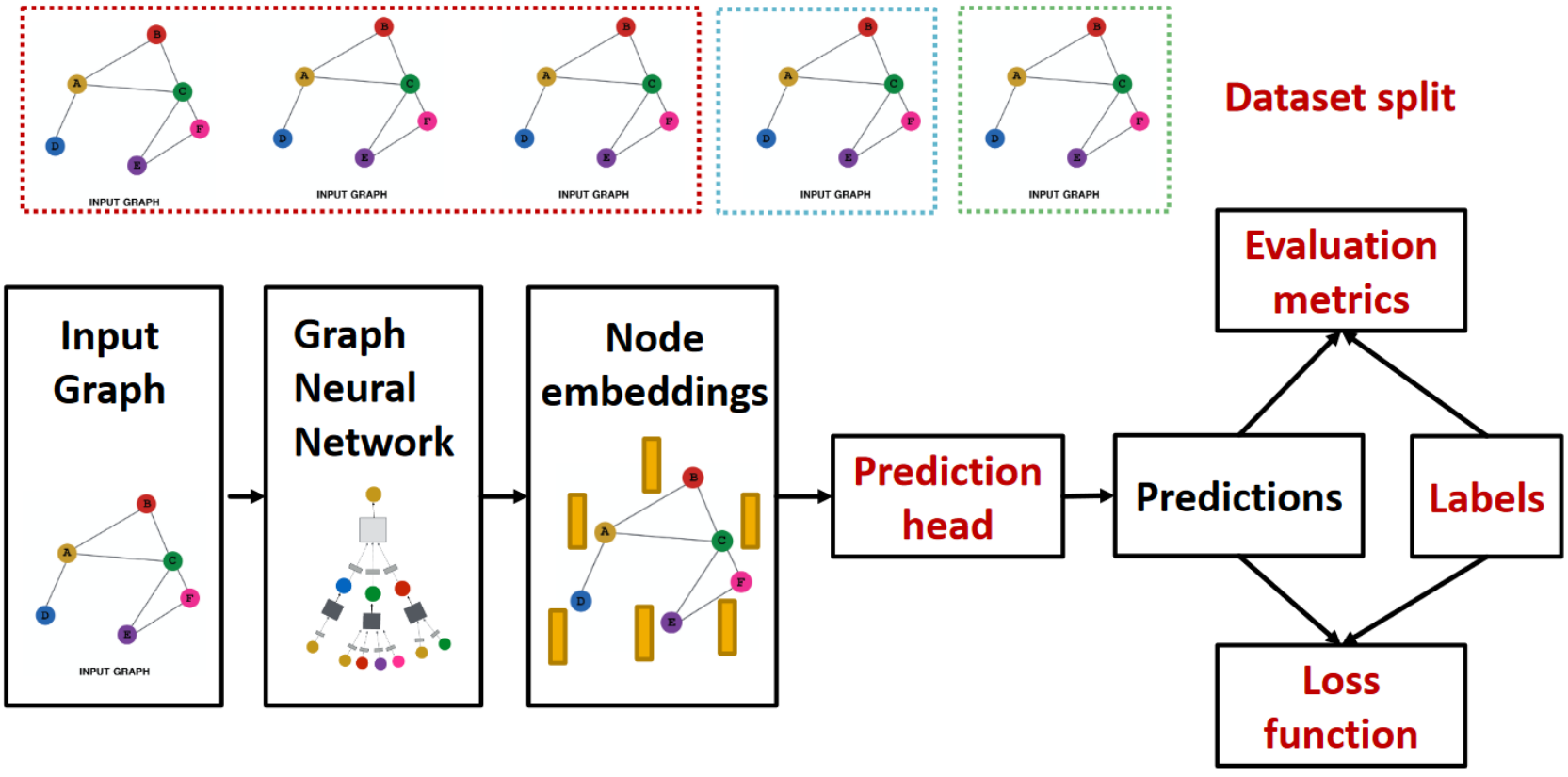


Διαχωρισμός σε 4 είδη ακμών

- ακμές μηνυμάτων εκπαίδευσης
- ακμές επίβλεψης εκπαίδευσης
- ακμές επικύρωσης
- ακμές τεστ

Πρόκειται για μία σύνθετη διαδικασία η οποία όμως υποστηρίζεται από την *PyG*

Διαδικασία εκπαίδευσης - επικύρωσης - τεστ



Υποστήριξη από το περιβάλλον *GraphGym*
https : //github.com/snap − stanford/GraphGym

Βιβλιογραφία

1. W.L. Hamilton, Graph Representation Learning, McGill University, 2020
2. J. Leskovec, Machine Learning with Graphs, Stanford University, Fall 2023