Τεχνολογίες Γραφημάτων και Εφαρμογές
Εργασία-Πρόβλεψη Συνδέσμων (Link Prediction)

Διδάσκων: Δημήτρης Μιχαήλ
2025-2026

Νικόλαος Δούρος,ΑΜ:2022127
Στυλιανός Ορφανίδης, ΑΜ:2022079

Περιεχόμενα :

1) Εισαγωγή:

Η πρόβλεψη συνδέσμων (link prediction) αποτελεί ένα από τα βασικά προβλήματα στην ανάλυση και εξόρυξη γραφημάτων. Δεδομένου ενός γράφου, στόχος του link prediction είναι η εκτίμηση της πιθανότητας ύπαρξης μιας ακμής μεταξύ δύο κόμβων που δεν συνδέονται άμεσα στο παρατηρούμενο γράφημα. Το πρόβλημα αυτό εμφανίζεται σε πληθώρα εφαρμογών, όπως κοινωνικά δίκτυα, δίκτυα αναφορών επιστημονικών άρθρων, συστήματα συστάσεων και βιολογικά δίκτυα.

Στην παρούσα εργασία μελετάται το πρόβλημα της πρόβλεψης συνδέσμων στο γράφημα αναφορών **Cora**, το οποίο αποτελεί κλασικό benchmark στη βιβλιογραφία των γραφημάτων και των Graph Neural Networks. Κάθε κόμβος του γραφήματος αντιστοιχεί σε ένα επιστημονικό άρθρο, ενώ οι ακμές αναπαριστούν σχέσεις αναφοράς μεταξύ άρθρων.

Σκοπός της εργασίας είναι η συγκριτική αξιολόγηση τριών διαφορετικών προσεγγίσεων για την πρόβλεψη συνδέσμων:

1. **Κλασικές ευριστικές μέθοδοι**, οι οποίες βασίζονται αποκλειστικά στην τοπική δομή του γραφήματος και τη γειτονιά των κόμβων.
2. **Ρηχές ενσωματώσεις κόμβων (shallow embeddings)**, με χρήση του αλγορίθμου Node2Vec για την εκμάθηση αναπαραστάσεων κόμβων και ενός επιβλεπόμενου ταξινομητή για την πρόβλεψη ακμών.
3. **End-to-end Graph Neural Networks (GNNs)**, όπου η εκμάθηση των ενσωματώσεων και η πρόβλεψη συνδέσμων πραγματοποιούνται ταυτόχρονα, αξιοποιώντας τόσο τη δομή του γραφήματος όσο και τα χαρακτηριστικά των κόμβων.

Για τη δίκαιη σύγκριση όλων των μεθόδων, ακολουθείται κοινό πειραματικό πλαίσιο αξιολόγησης με διαχωρισμό των ακμών σε σύνολα εκπαίδευσης και δοκιμής, καθώς και ισορροπημένη δειγματοληψία αρνητικών παραδειγμάτων. Ως μετρική αξιολόγησης χρησιμοποιείται η **Area Under the ROC Curve (AUC)**, η οποία μετρά την ικανότητα κάθε μεθόδου να κατατάσσει υψηλότερα τις πραγματικές ακμές σε σχέση με τις μη υπάρχουσες.

Η εργασία ολοκληρώνεται με συγκριτική ανάλυση των αποτελεσμάτων των τριών προσεγγίσεων και συζήτηση των πλεονεκτημάτων και περιορισμών κάθε μεθόδου στο συγκεκριμένο πρόβλημα.

2) Δεδομένα & Πλαίσιο Αξιολόγησης:

2.1 Dataset

Χρησιμοποιήθηκε το Cora citation network μέσω της βιβλιοθήκης DGL (CoraGraphDataset).

Κάθε κόμβος αντιστοιχεί σε επιστημονικό άρθρο.

Κάθε ακμή αναπαριστά σχέση αναφοράς μεταξύ άρθρων.

Οι κόμβοι διαθέτουν διανυσματικά χαρακτηριστικά (bag-of-words).

2.2 Προετοιμασία και Προεπεξεργασία Δεδομένων

Αφαίρεση self-loops.

Μετατροπή του γραφήματος σε απλό γράφημα (χωρίς πολλαπλές ακμές).

Έλεγχος συνεκτικότητας του γραφήματος.

Διατήρηση μόνο της Largest Connected Component (LCC) για:

- αποφυγή απομονωμένων κόμβων,
- ορθή διάδοση πληροφορίας στα GNN μοντέλα.

2.3 Διαχωρισμός Train / Test

Ο διαχωρισμός πραγματοποιείται σε επίπεδο ακμών.

Το 10% των ακμών αφαιρείται για τη δημιουργία του test set (θετικά παραδείγματα).

Για τη διατήρηση της συνεκτικότητας του training graph:

- υπολογίζεται ένα Minimum Spanning Tree (MST),
- όλες οι ακμές του MST παραμένουν στο training set.

Οι υπόλοιπες ακμές επιλέγονται τυχαία για το test set.

2.4 Negative Sampling

Για κάθε θετική ακμή του test set δημιουργείται ένα αρνητικό δείγμα.

Τα αρνητικά δείγματα είναι:

- ζεύγη κόμβων χωρίς ακμή στο πλήρες γράφημα,
- χωρίς self-loops.

Αναλογία θετικών/αρνητικών: 1:1.

2.5 Αναπαραγωγιμότητα (Seed)

Χρησιμοποιείται σταθερό seed (seed = 42) για:

- Python random,
- NumPy,
- PyTorch.

Με τον τρόπο αυτό εξασφαλίζεται η επαναληψιμότητα των πειραμάτων.

2.6 Μετρική Αξιολόγησης

Χρησιμοποιείται η Area Under the ROC Curve (AUC).

Η AUC μετρά την ικανότητα του μοντέλου να:

- κατατάσσει υψηλότερα τις πραγματικές ακμές,
- σε σχέση με τα αρνητικά παραδείγματα,
- ανεξάρτητα από συγκεκριμένο κατώφλι απόφασης.

3) Μέρος Α — Ευριστικές Μέθοδοι

Στόχος

Υπολογισμός “score” για κάθε ζεύγος κόμβων (u, v) στο test set (θετικά + αρνητικά).

Η πρόβλεψη βασίζεται μόνο στη δομή του training graph (χωρίς εκπαίδευση μοντέλου).

3.1 Υλοποιημένες Ευριστικές

Για γειτονιές N(u), N(v):

Common Neighbors (CN)

$$s(u,v) = |N(u) \cap N(v)|$$

Ερμηνεία: όσο περισσότερους κοινούς γείτονες έχουν δύο κόμβοι, τόσο πιθανότερο να συνδέονται.

Jaccard Coefficient

$$s(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

Ερμηνεία: κανονικοποιεί τους κοινούς γείτονες με βάση το μέγεθος της ένωσης.

Adamic–Adar (AA)

$$s(u,v) = \sum_{w \in N(u) \cap N(v)} \frac{1}{\log(|N(w)|)}$$

Ερμηνεία: δίνει μεγαλύτερο βάρος σε “σπάνιους” κοινούς γείτονες (με μικρό βαθμό).

3.2 Διαδικασία Αξιολόγησης

Για κάθε μέθοδο:

Υπολογίζονται scores για:

- test_pos_edges (θετικά)
- test_neg_edges (αρνητικά)

Δημιουργείται διάνυσμα ετικετών:

- y=1 για θετικά, y=0για αρνητικά.

Υπολογίζεται AUC με βάση τα y και τα scores.

Τα αποτελέσματα αποθηκεύονται για συγκριτικό ROC plot.

3.3 Αποτελέσματα & Σχολιασμός

Αποτελέσματα (AUC)

Jaccard: 0.7662

Adamic–Adar: 0.7669

Common Neighbors: 0.7665

Οι ευριστικές επιτυγχάνουν AUC ~0.77, δείχνοντας ότι η τοπική επικάλυψη γειτονιών περιέχει χρήσιμη πληροφορία στο Cora.

Ωστόσο, επειδή βασίζονται μόνο σε τοπική δομή και όχι σε learned representations, αναμένονται χαμηλότερες επιδόσεις από Node2Vec/GNN.

4) Μέρος Β — Node2Vec + MLP

4.1 Στόχος

Εκμάθηση shallow embeddings για τους κόμβους με Node2Vec (unsupervised).

Χρήση των embeddings για link prediction μέσω:

- edge representations (Hadamard product)
- απλού MLP ταξινομητή (supervised).

4.2 Εκμάθηση Node Embeddings

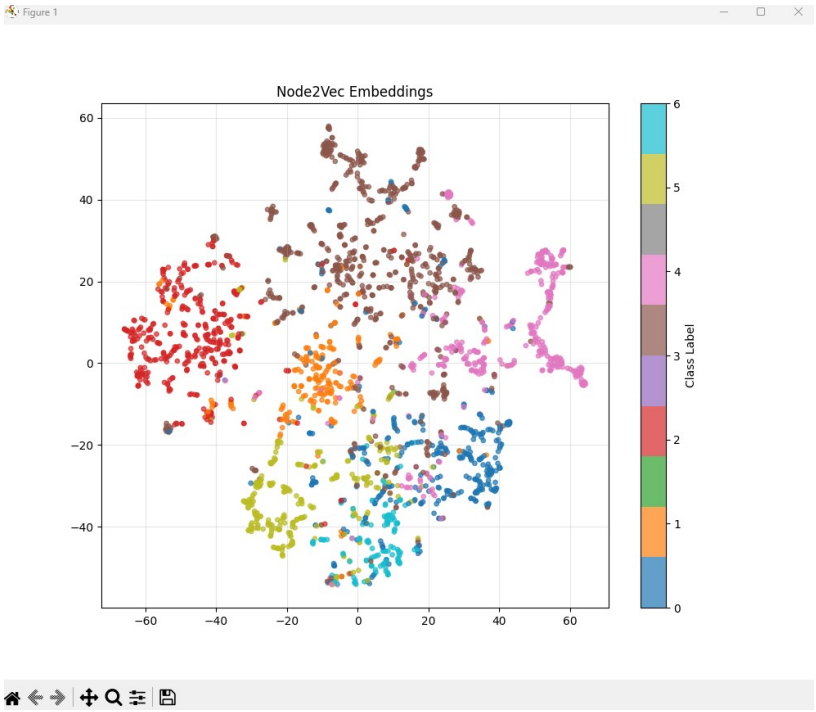
Ο Node2Vec εκπαιδεύεται στο πλήρες γράφημα (full graph), καθώς είναι μη-επιβλεπόμενη διαδικασία.

Παράγονται embeddings z ε R^ 64 για κάθε κόμβο u.

Ρυθμίσεις Node2Vec :

- Embedding dimension: 64
- Walk length: 10
- Number of walks per node: 20
- Window size (Skip-gram): 10
- min_count: 1

t-SNE visualization of Node2Vec node embeddings on the Cora dataset.



4.3 Edge Representation (Hadamard Product)

Για κάθε υποψήφια ακμή (u, v) δημιουργείται ενσωμάτωση ακμής:

$$h_{uv} = z_u \odot z_v$$

όπου dot είναι το element-wise (Hadamard) product.

Το huv χρησιμοποιείται ως feature vector εισόδου στον ταξινομητή.

4.4 Εκπαίδευση MLP Ταξινομητή

Θετικά training δείγματα: όλες οι ακμές του training graph.

Αρνητικά training δείγματα: ίσο πλήθος non-edges από το training graph (negative sampling).

Labels:

- 1 για θετικές ακμές
- 0 για αρνητικά ζεύγη

Αρχιτεκτονική MLP

Input: 64-dim edge feature

2 επίπεδα:

- Linear(64 → 32) + ReLU
- Linear(32 → 1) + Sigmoid

Ρυθμίσεις εκπαίδευσης

Optimizer: Adam

Learning rate: 0.01

Loss: Binary Cross-Entropy (BCELoss)

Epochs: 50

4.5 Αξιολόγηση

Το μοντέλο αξιολογείται στο test set:

- test positive edges
- test negative edges (1:1)

Μετρική: AUC

4.6 Αποτελέσματα & Σχολιασμός

Node2Vec + MLP AUC: 0.9905

Η πολύ υψηλή AUC δείχνει ότι οι embeddings του Node2Vec συλλαμβάνουν αποτελεσματικά τη δομή του γραφήματος.

Σε σχέση με τις ευριστικές, η προσέγγιση αξιοποιεί global structural information μέσω των random walks και βελτιώνει σημαντικά τη διακριτική ικανότητα.

5) Μέρος Γ — End-to-End GNN

5.1 Στόχος

Εκμάθηση node embeddings end-to-end

Χρήση τόσο:

- της δομής του training graph,
- όσο και των χαρακτηριστικών κόμβων (features).

5.2 GNN Encoder (GCN)

Επιλέγεται Graph Convolutional Network (GCN) ως encoder.

Ο encoder χαρτογραφεί κάθε κόμβο u σε embedding z

Αρχιτεκτονική

2 layers GraphConv:

- GraphConv(F → 16) + ReLU
- GraphConv(16 → 16)

Είσοδος: node features (1433 χαρακτηριστικά στο Cora).

Έξοδος: embedding διάστασης 16 για κάθε κόμβο.

5.3 Link Predictor (Decoder)

Χρησιμοποιείται dot-product decoder.

Για ένα ζεύγος κόμβων (u, v):

$$s(u, v) = z_u^\top z_v$$

Η πιθανότητα ακμής δίνεται από:

$$\hat{y}_{uv} = \sigma(s(u, v))$$

όπου σ sigmoid.

5.4 Εκπαίδευση

Εκπαίδευση μόνο στο training graph (χωρίς test edges).

Θετικά παραδείγματα: οι ακμές του training graph.

Αρνητικά παραδείγματα (online):

- σε κάθε epoch γίνεται negative sampling,
- παράγεται ίδιο πλήθος αρνητικών ζευγών με τις θετικές ακμές.

Loss Function

Binary Cross-Entropy με logits:

- logits = [pos_scores, neg_scores]
- labels = [1 για pos, 0 για neg]

Ρυθμίσεις εκπαίδευσης

Optimizer: Adam

Learning rate: 0.01

Epochs: 101

Self-loops προστίθενται στο training graph κατά το forward pass (GCN πρακτική).

5.5 Αξιολόγηση

Υπολογίζονται scores στο test set για:

- test positive edges
- test negative edges

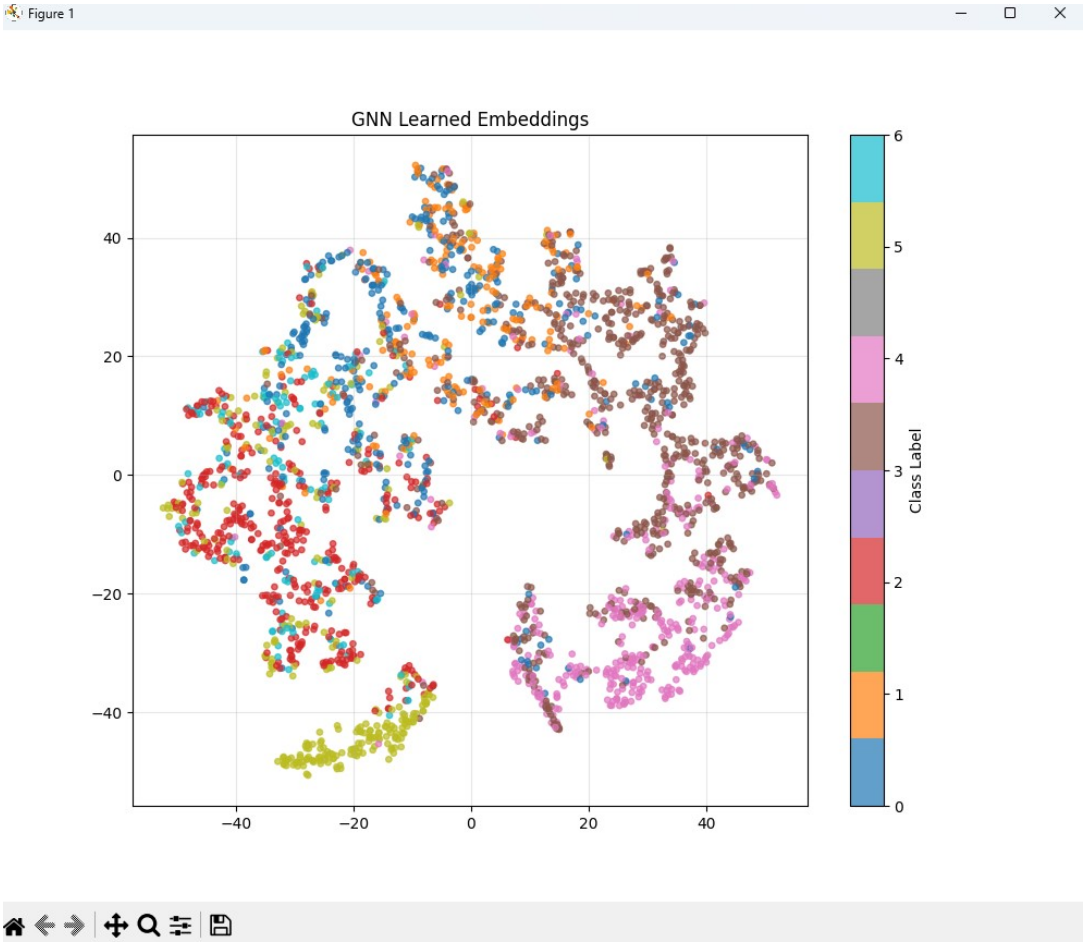
Μετατροπή σε πιθανότητες με sigmoid.

Μετρική: AUC (Area Under ROC Curve).

5.6 Αποτελέσματα & Σχολιασμός

GNN (GCN, end-to-end) AUC: 0.9007

t-SNE visualization of node embeddings learned by the GCN model.



Το GNN έχει σαφώς καλύτερη επίδοση από τις ευριστικές, καθώς μαθαίνει embeddings προσαρμοσμένα στο task.

Παρότι αξιοποιεί node features, έχει χαμηλότερη AUC από το Node2Vec+MLP στο συγκεκριμένο setup, καθώς:

- εκπαιδεύεται μόνο στο training graph,
- πρέπει να μάθει ταυτόχρονα embeddings και classifier μέσω του dot-product decoder.

6) Σύγκριση

6.1 Πίνακας Σύγκρισης Αποτελεσμάτων (AUC)

Μέθοδος	AUC
Common Neighbors	0.7665
Jaccard Coefficient	0.7662
Adamic–Adar Index	0.7669
Node2Vec + MLP	0.9905
GNN (GCN, end-to-end)	0.9007

6.2 Συγκριτική Ανάλυση

Οι ευριστικές μέθοδοι παρουσιάζουν παρόμοια απόδοση (AUC ≈ 0.77), γεγονός που δείχνει ότι:

- η τοπική επικάλυψη γειτονιών αποτελεί χρήσιμο σήμα στο Cora,
- ωστόσο η πληροφορία που αξιοποιούν είναι περιορισμένη (μόνο τοπική δομή).

Η προσέγγιση Node2Vec + MLP επιτυγχάνει τη υψηλότερη απόδοση:

- ο Node2Vec εκπαιδεύεται στο πλήρες γράφημα και μαθαίνει πλούσιες, παγκόσμιες αναπαραστάσεις κόμβων,
- ο MLP αξιοποιεί αυτές τις αναπαραστάσεις επιβλεπόμενα, οδηγώντας σε πολύ υψηλή διακριτική ικανότητα (AUC ≈ 0.99).

Το end-to-end GNN (GCN) παρουσιάζει σημαντική βελτίωση σε σχέση με τις ευριστικές:

- αξιοποιεί τόσο τη δομή του γραφήματος όσο και τα χαρακτηριστικά των κόμβων,
- μαθαίνει embeddings προσαρμοσμένα στο task της πρόβλεψης συνδέσμων.

Παρότι το GNN είναι πιο εκφραστικό μοντέλο, υστερεί σε σχέση με το Node2Vec + MLP στο συγκεκριμένο πείραμα, επειδή:

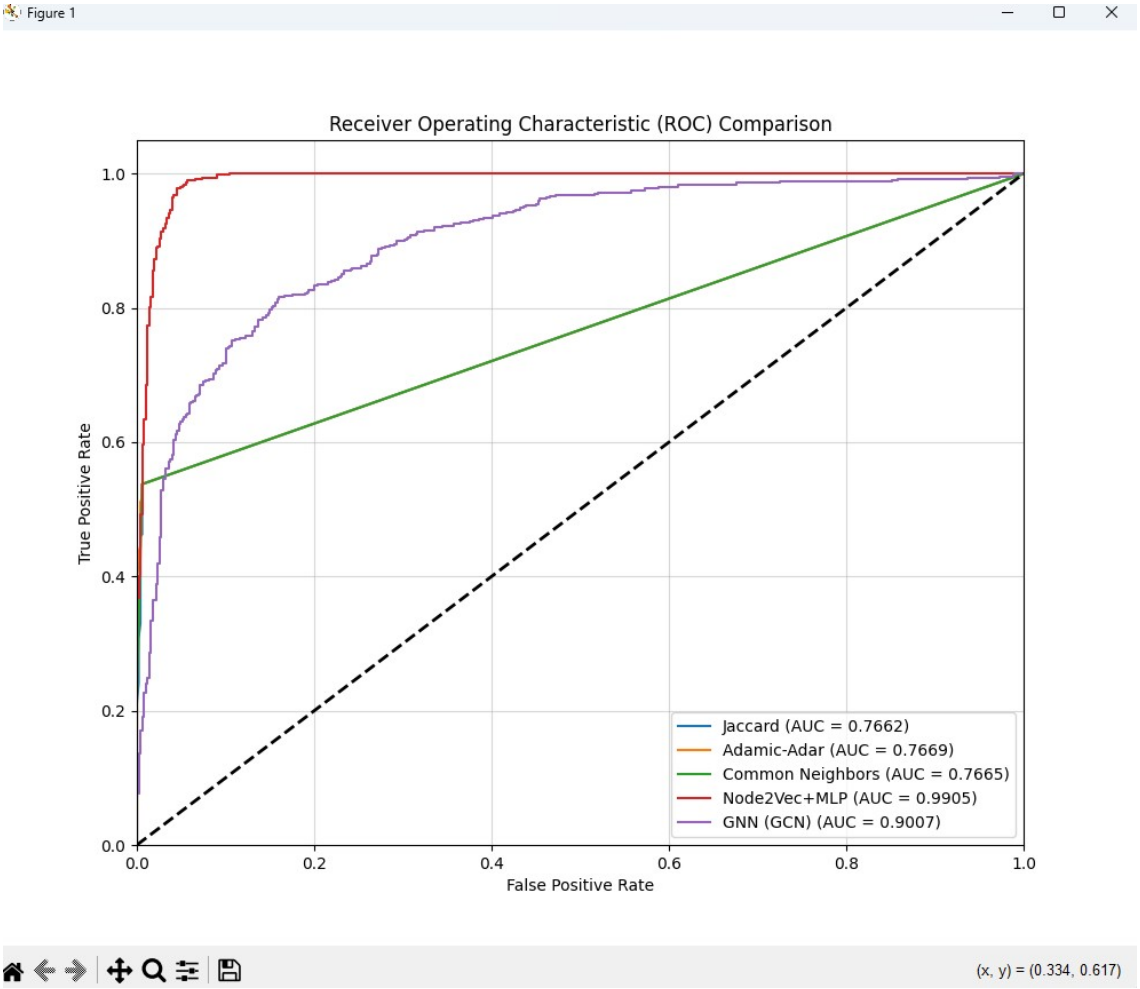
- εκπαιδεύεται μόνο στο training graph,
- πρέπει να μάθει ταυτόχρονα τις αναπαραστάσεις και τον προβλεπτή μέσω dot-product decoder.

6.3 Οπτικοποίηση Αποτελεσμάτων

Η σύγκριση των μεθόδων απεικονίζεται μέσω ROC curves, όπου φαίνεται καθαρά η υπεροχή των learned methods έναντι των ευριστικών.

Οι οπτικοποιήσεις t-SNE των embeddings (Node2Vec και GNN) επιβεβαιώνουν ποιοτικά την ικανότητα των μοντέλων να ομαδοποιούν κόμβους με παρόμοια χαρακτηριστικά.

ROC curves for all evaluated link prediction methods on the Cora dataset.



Κύρια Συμπεράσματα από τη Σύγκριση

Η μετάβαση από ευριστικές σε learned representations οδηγεί σε σημαντική αύξηση της απόδοσης.

Οι shallow embeddings (Node2Vec) παραμένουν ιδιαίτερα ισχυρές για link prediction όταν υπάρχει πρόσβαση στο πλήρες γράφημα.

Τα end-to-end GNNs προσφέρουν μεγαλύτερη ευελιξία και δυνατότητα αξιοποίησης features, αλλά απαιτούν προσεκτική ρύθμιση και περισσότερο training data.

7) Συμπεράσματα

- Στην εργασία αυτή μελετήθηκε το πρόβλημα της πρόβλεψης συνδέσμων στο γράφημα Cora, χρησιμοποιώντας τρεις διαφορετικές κατηγορίες μεθόδων: ευριστικές, shallow embeddings και end-to-end Graph Neural Networks.
- Οι κλασικές ευριστικές μέθοδοι (Common Neighbors, Jaccard, Adamic–Adar) πέτυχαν μέτρια απόδοση (AUC ≈ 0.77), αποδεικνύοντας ότι η τοπική δομή του γραφήματος περιέχει χρήσιμη αλλά περιορισμένη πληροφορία.
- Η προσέγγιση Node2Vec + MLP παρουσίασε την καλύτερη απόδοση (AUC ≈ 0.99), αξιοποιώντας αποτελεσματικά την παγκόσμια δομή του γραφήματος μέσω μη-επιβλεπόμενων embeddings.
- Το end-to-end GNN (GCN) πέτυχε υψηλή απόδοση (AUC ≈ 0.90), ξεπερνώντας τις ευριστικές μεθόδους και αποδεικνύοντας τη δύναμη των GNNs στην εκμάθηση task-specific αναπαραστάσεων.
- Η σύγκριση των μεθόδων ανέδειξε τα πλεονεκτήματα και τους περιορισμούς κάθε προσέγγισης, ανάλογα με τη διαθεσιμότητα δεδομένων και το επίπεδο εποπτείας.
- Μελλοντικές επεκτάσεις θα μπορούσαν να περιλαμβάνουν χρήση πιο σύνθετων decoders (MLP), διαφορετικές αρχιτεκτονικές GNN ή εκτενέστερη ρύθμιση υπερπαραμέτρων.