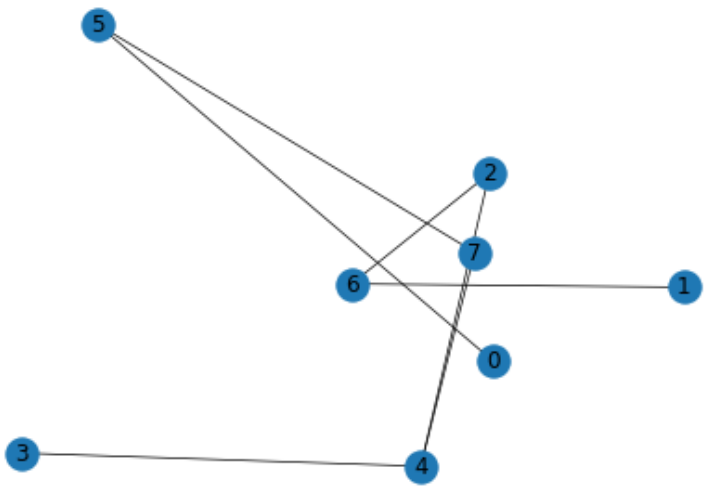
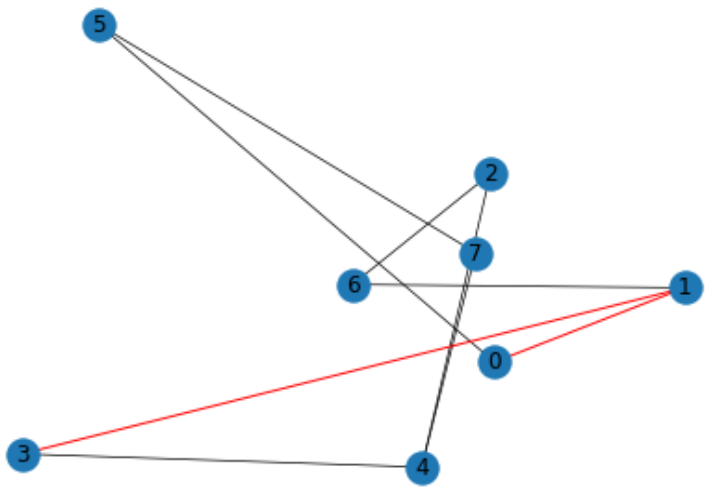
Ανάλυση Κοινωνικών Δικτύων
3η Εργαστηριακή άσκηση
Πρόβλεψη συνδέσμων (Link prediction)

Ορισμός του προβλήματος πρόβλεψης συνδέσμων

- Στατικά δίκτυα: προσδιορισμός των συνδέσμων που δεν είναι εμφανείς στο δίκτυο (missing links).
- Δυναμικά δίκτυα: δοθέντος ενός στιγμιοτύπου του δικτύου την χρονική στιγμή t, πρόβλεψη του σχηματισμού των συνδέσμων στο δίκτυο την χρονική στιγμή $t + k$.



G_t



G_{t+k}

Εξεταζόμενες τεχνικές πρόβλεψης συνδέσμων

- Similarity-based: Για κάθε ζεύγος μη συνδεδεμένων κόμβων στο εξεταζόμενο στιγμιότυπο του δικτύου, υπολογίζεται η ομοιότητά τους, με βάση τα τοπολογικά χαρακτηριστικά τους (common neighbors, degree). Τα ζεύγη με υψηλή τιμή ομοιότητας (που ξεπερνά μια τιμή κατωφλίου), αναπαριστούν τις μελλοντικές συνδέσεις του δικτύου.

  Εξεταζόμενες μετρικές ομοιότητας (similarity metrics) κόμβων:

  - Jaccard Coefficient
  - Preferential Attachment
  - Resource Allocation

- Learning-based: Πρόβλημα δυαδικής κατηγοριοποίησης (binary classification). Στόχος είναι η κατασκευή ενός μοντέλου πρόβλεψης δύο διακριτών κατηγοριών ("ύπαρξη σύνδεσης", "μη-ύπαρξη σύνδεσης").

Similarity metrics

Στο μη κατευθυνόμενο γράφο $G = (V, E)$ για το ζεύγος κόμβων $u, v \in V$ για το οποίο ισχύει $(u, v) \notin E$ και $N(u) = \{w \in V : (u, w) \in E\}$ ορίζονται οι μετρικές ομοιότητας

- **Jaccard Coefficient**

$$JC(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \tag{1}$$

- **Preferential Attachment**
  (Προσοχή! Βασίζεται μόνο στο βαθμό, όχι στους κοινούς γείτονες των κόμβων.)

$$PA(u, v) = |N(u)||N(v)| \tag{2}$$

- **Resource Allocation** (Διαίσθηση: Δύο μη γειτονικοί κόμβοι μοιράζονται resources μέσω των κοινών τους γειτόνων.)

$$RA(u, v) = \sum_{w \in N(u) \cap N(v)} \frac{1}{|N(w)|} \tag{3}$$

Δυναμικό δίκτυο: Κατασκευή στιγμιοτύπων

Συνήθως, τα datasets περιγράφουν ένα μόνο στιγμιότυπο του δικτύου. Για την αξιολόγηση των τεχνικών πρόβλεψης συνδέσμων, χρειαζόμαστε 2 στιγμιότυπα του δικτύου.

Το δοθέν στιγμιότυπο αποτελεί το τελικό δίκτυο $G_{t+k} = (V, E_{t+k})$.
Τότε, το στιγμιότυπο $G_t = (V, E_t)$ (αρχικό δίκτυο) κατασκευάζεται ως εξής:

- Επιλογή των ακμών $(u, v) \in E_{t+k}$ η αφαίρεση των οποίων δεν μεταβάλλει τον αριθμό των συνεκτικών συνιστωσών του δικτύου. Οι ακμές αυτές συνιστούν το σύνολο των **positive samples** PS.
- Το δίκτυο που προκύπτει από την αφαίρεση των ακμών του PS θα είναι το αρχικό δίκτυο $G_t = (V, E_t)$ με $E_t = \{E_{t+k} \setminus PS\}$.

Negative samples $NS = \{(u, v) : (u, v) \notin E_{t+k}\}$

Το σύνολο των ζευγών **μη συνδεδεμένων κόμβων στο αρχικό δίκτυο** G_t είναι $S = PS \cup NS$.

Similarity-based τεχνική πρόβλεψης συνδέσμων

Για το δίκτυο G_t, υπολογίζεται, για κάθε ζεύγος κόμβων που ανήκει στο S, η μετρική ομοιότητας $sim(u, v)$, $sim \in \{JC, PA, RA, ...\}$.

Για τιμή κατωφλίου thr, υπολογίζεται το **σύνολο των predicted συνδέσεων** στο G_t, $P = \{(u, v) \in S : sim(u, v) \geq thr\}$.

Με βάση τα P, S, PS, NS υπολογίζονται τα σύνολα:

- **True Positives** $TP = \{(u, v) : (u, v) \in P \cap PS\}$
- **False Positives** $FP = \{(u, v) : (u, v) \in P \cap NS\}$
- **True Negatives** $TN = \{(u, v) : (u, v) \in \{S \setminus P\} \cap NS\}$
- **False Negatives** $FN = \{(u, v) : (u, v) \in \{S \setminus P\} \cap PS\}$

Ισχύουν οι σχέσεις:

$$|TP| + |FN| = |PS| \tag{4}$$

$$|FP| + |TN| = |NS| \tag{5}$$

Μετρικές αξιολόγησης της πρόβλεψης

- **Precision**

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{6}$$

- **Recall**

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{7}$$

- **Accuracy**

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |FN| + |TN| + |FP|} \tag{8}$$

Learning-based τεχνικές πρόβλεψης συνδέσμων

Πρόβλεψη συνδέσμων ως **πρόβλημα επιβλεπόμενης μάθησης**: δυαδική κατηγοριοποίηση (binary classification) με κατηγορίες "ύπαρξη σύνδεσης", "μη-ύπαρξη σύνδεσης".

- Για το σύνολο των ζευγών μη συνδεδεμένων κόμβων στο αρχικό δίκτυο G_t, S, επιλέγεται ένα σύνολο χαρακτηριστικών (feature set).
- Το S διαμερίζεται σε S_{train} και S_{test}.
- Από το τελικό δίκτυο G_{t+k}, εξασφαλίζεται η γνώση για τα στοιχεία του S_{train} ως προς τις κατηγορίες ταξινόμησης. Για το ζεύγος κόμβων $(u, v) \in S_{train}$ και την αντίστοιχη ετικέτα (label) $l(u, v)$ ισχύει:

$$l(u,v) = \begin{cases} 0, & \text{if } (u,v) \in NS, \\ 1, & \text{if } (u,v) \in PS. \end{cases} \tag{9}$$

Στόχος: Να κατασκευαστεί ένας ταξινομητής ο οποίος, αφού εκπαιδευτεί στο S_{train}, να προβλέπει την κατηγορία στην οποία ανήκουν τα στοιχεία του S_{test}.

Επιλογή feature set

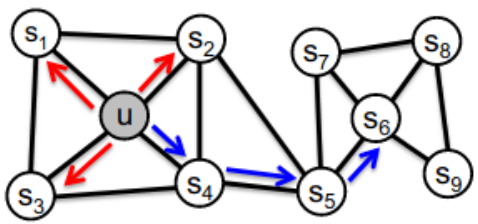
1^{η} μέθοδος επιλογής features: **similarity-based**

Δοθεισών k μετρικών ομοιότητας, για κάθε $(u, v) \in S$ ορίζεται το feature vector $\mathbf{sim}(u, v) = [sim_1(u, v), ..., sim_k(u, v)]$.

2^{η} μέθοδος επιλογής features: **learning-based**

Node2vec: Αλγοριθμικό πλαίσιο μάθησης χαρακτηριστικών των κόμβων (και κατ' επέκταση ζευγών κόμβων, με χρήση δυαδικών τελεστών) ενός δικτύου.

- Κατασκευάζει feature vectors των κόμβων (node embeddings) με βάση τα τοπολογικά χαρακτηριστικά τους (degree, neighborhood), π.χ., κόμβοι που είναι hubs (structural equivalence) ή ανήκουν στην ίδια κοινότητα (homophily) έχουν παρόμοια embeddings.



- Βασίζεται στο Skip-gram model: προβλέπει το context μιας λέξης από ένα σώμα κειμένων (corpus) και κατασκευάζει feature vectors των λέξεων (word embeddings) με βάση το context, μέσω ενός 2-layer νευρωνικού δικτύου.
- Αντιμετωπίζει τον γράφο ως "κείμενο". Εκτελεί τυχαίους περίπατους στον γράφο (second-order biased random walks) για να κατασκευάσει ένα "σώμα προτάσεων" που θα χρησιμοποιηθούν ως είσοδος στο Skip-gram model.

Ζητούμενα άσκησης

1. Κατασκευή δικτύου από το dataset Similarities DBPedia. Επεξεργασία δεδομένων για την εφαρμογή τεχνικών πρόβλεψης συνδέσμων (δημιουργία 2 στιγμιοτύπων του δικτύου, υπολογισμός συνόλων PS,NS,S).
2. **Similarity-based** πρόβλεψη συνδέσμων με χρήση της μετρικής ομοιότητας Jaccard Coefficient για διάφορες τιμές κατωφλίου. Αξιολόγηση της πρόβλεψης, με τις μετρικές Precision, Recall, Accuracy.
3. **Learning-based** πρόβλεψη συνδέσμων: Ταξινόμηση ενός υποσυνόλου των ζευγών μη συνδεδεμένων κόμβων του αρχικού δικτύου στις κατηγορίες "ύπαρξη σύνδεσης", "μη-ύπαρξη σύνδεσης" με τον ταξινομητή Random Forest. Τα feature vectors καθορίζονται από τις μετρικές Jaccard Coefficient, Preferential Attachment, Resource Allocation. Αξιολόγηση της πρόβλεψης με την μετρική Accuracy.
4. **Learning-based** πρόβλεψη συνδέσμων: Ταξινόμηση ενός υποσυνόλου των ζευγών μη συνδεδεμένων κόμβων του αρχικού δικτύου στις κατηγορίες "ύπαρξη σύνδεσης", "μη-ύπαρξη σύνδεσης" με τον ταξινομητή Random Forest. Τα feature vectors καθορίζονται από το αλγοριθμικό πλαίσιο για node embeddings, node2vec. Αξιολόγηση της πρόβλεψης με την μετρική Accuracy.