



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**Πρόγραμμα Μεταπτυχιακών Σπουδών**

**« Προηγμένα Συστήματα Πληροφορική - Ανάπτυξη Λογισμικού και  
Τεχνητής Νοημοσύνης »**

**Μεταπτυχιακή Διατριβή**

Τίτλος Διατριβής	<b>Μεθοδολογίες βαθιάς μάθησης για την πρόβλεψη μεταναστευτικών δικτύων Deep Learning Methodologies for Migration Network Prediction</b>
Όνοματεπώνυμο Φοιτητή	<b>Ασημάκης Αντωνίου</b>
Πατρώνυμο	<b>Βασίλειος</b>
Αριθμός Μητρώου	<b>ΜΠΣΠ/ 2202</b>
Επιβλέπων	<b>Διονύσιος Σωτηρόπουλος, Επίκουρος καθηγητής</b>

Ημερομηνία Παράδοσης **Απρίλιος 2024**

---

Τριμελής Εξεταστική Επιτροπή

Διονύσιος Σωτηρόπουλος  
Επίκουρος Καθηγητής

Γεώργιος Τσιχριντζής  
Καθηγητής

Ευάγγελος Σακκόπουλος  
Αναπληρωτής Καθηγητής

## **Ευχαριστίες**

Θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Σωτηρόπουλο Διονύσιο για όλη τη βοήθεια και συνεχή καθοδήγηση καθόλη τη διάρκεια προετοιμασίας της μεταπτυχιακής μου εργασίας. Τα μαθήματα του αποτέλεσαν έμπνευση για έμένα και έπαιξαν σημαντικό ρόλο στην ανάπτυξη των ενδιαφερόντων μου.

Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου που με στήριξαν στο διάστημα της φοίτησης μου για να μπορώ να επικεντρωθώ στους στόχους μου και στην σχολή μου.

## Περίληψη

Από τα πρώτα χρόνια, η ανθρωπότητα ήταν σε κίνηση. Ορισμένοι άνθρωποι έφευγαν από τον τόπο τους για την αναζήτηση εργασίας, οικονομικών ευκαιριών ή για να σπουδάσουν. Άλλοι μετακινούνται για να ξεφύγουν από συγκρούσεις, διώξεις ή πόλεμους. Αντίθετα, κάποιοι άνθρωποι μεταναστεύουν καθώς έρχονται αντιμέτωποι με τις επιπτώσεις της κλιματικής αλλαγής, των φυσικών καταστροφών ή άλλων περιβαλλοντικών παραγόντων. Η κατανόηση των παγκόσμιων μεταναστευτικών προτύπων έχει σημαντικές συνέπειες για τις οικονομικές πολιτικές, την ανθρωπιστική βοήθεια και την κοινωνική δυναμική. Ωστόσο, η πρόβλεψη αυτών των μεταναστευτικών ροών είναι εγγενώς πολύπλοκη λόγω της πολύπλευρης αλληλεπίδρασης οικονομικών, πολιτικών και περιβαλλοντικών παραγόντων.

Ο σκοπός της διπλωματικής αυτής εργασίας αποσκοπεί στην ανάπτυξη ενός εξελεγκμένου μοντέλου πρόβλεψης με τη χρήση Νευρωνικών Δικτύων Γραφημάτων (GNN) για την ανάλυση και την πρόβλεψη μελλοντικών μεταναστευτικών συνδέσεων. Θεωρώντας τη μετανάστευση ως ένα δυναμικό δίκτυο, το μοντέλο επιδιώκει να αποκαλύψει κρυμμένα μοτίβα και πιθανές μελλοντικές τάσεις στην παγκόσμια ανθρώπινη κινητικότητα.

Οι μέθοδοι που θα χρησιμοποιηθούν αποτελούνται από αλγόριθμους βαθιάς μάθησης. Αρχικά, παρουσιάζουμε την βασική θεωρία στην οποία βασίστηκε η συγκεκριμένη εργασία. Στη συνέχεια, επεξεργαζόμαστε τα δεδομένα μας με τέτοιο τρόπο ώστε να δημιουργήσουμε για κάθε χρονική περίοδο ένα δίκτυο. Έχοντας τα γραφήματα αυτά γίνεται ανάλυση των γραφημάτων με έμφαση την πρώτη και την τελευταία περίοδο 1990-2020. Επιπλέον, δημιουργούνται τα μοντέλα πρόβλεψης μελλοντικών συνδέσεων. Τέλος, ολοκληρώνουμε την εργασία εξετάζοντας την ακρίβεια του εκάστοτε μοντέλου και παρουσιάζουμε τις νέες μεθόδους και το μέλλον που διαφαίνεται στο πεδίο της βαθιάς μηχανικής μάθησης στους γράφους.

## **Abstract**

From the earliest times, humanity was on the move. Some people were leaving their homes in search of work, economic opportunities or to study. Others moved to escape conflict, persecution or war. Instead, some people migrate as they face the effects of climate change, natural disasters or other environmental factors. Understanding global migration patterns has important implications for economic policies, humanitarian assistance and social dynamics. However, forecasting these migration flows is inherently complex due to the multifaceted interaction of economic, political and environmental factors.

The purpose of this thesis aims to develop a sophisticated forecasting model using Graph Neural Networks (GNN) to analyze and predict future migration linkages. By considering migration as a dynamic network, the model seeks to reveal hidden patterns and possible future trends in global human mobility.

The methods to be used consist of deep learning algorithms. First, we present the basic theory on which this thesis is based. Then, we process our data in such a way that we create a network for each time period. Having these graphs, we analyze the graphs with a focus on the first and last period 1990-2020. In addition, we create the models to predict future links(connections). Finally, we conclude the paper by examining the accuracy of each model and present new methods and the future of the field of deep machine learning on graphs.

## Πίνακας Περιεχομένων

Ευχαριστίες .....	3
Περίληψη .....	4
Abstract .....	5
Περιεχόμενα .....	6
Περιεχόμενα Εικόνων .....	7
1 Εισαγωγή.....	9
1.1 Αντικείμενο εργασίας.....	9
1.2 Διαρθρωση εργασίας .....	9
2 Θεωρητικό Υπόβαθρο .....	10
2.1 Νευρωνικά δίκτυα γραφών .....	10
2.2 Graph convolutional neural Layer .....	10
2.3 Αλγόριθμος Node2Vec.....	11
3 Δεδομένα και Προ Επεξεργασία .....	16
3.1 Δεδομένα.....	16
3.2 Επεξεργασία δεδομένων.....	17
4 Ανάλυση γραφημάτων.....	21
4.1 Ορισμοί Θεωρίας γράφων.....	21
4.1.1 Κοινότητες γραφημάτων.....	21
4.1.2 Βαθμός Κεντρικότητας (Degree centrality) .....	21
4.1.3 Κεντρικότητα Εγγύτητας (Closeness Centrality) .....	21
4.1.4 Κεντρικότητα Ενδιαμεσότητας (Betweenness Centrality) .....	22
4.1.5 Ιδιοδιανυσματική Κεντρικότητα (Eigenvector Centrality) .....	22
4.2 Γραφήματα .....	23
4.2.1 Δίκτυο μετανάστευσης 1990 .....	28
4.2.2 Δίκτυο κοινοτήτων 1990 .....	31
4.2.3 Δίκτυο μετανάστευσης 2020 .....	34
4.2.4 Δίκτυο κοινοτήτων 2020 .....	37
5 Πειράματα.....	39
5.1 Υλοποίηση μοντέλου πρόβλεψης συνδέσεων .....	39
5.2 Υλοποίηση μοντέλου πρόβλεψης με χρήση Node2Vec .....	42
6 Πειράματα και Συγκρίσεις Μοντέλων .....	45
7 Επίλογος .....	50
8 Πηγές και Βιβλιογραφία.....	51

## Πίνακας Εικόνων

Εικ.1. Αποτύπωση ενός γράφου στο διανυσματικό χώρο με τη χρήση του αλγόριθμου Node2Vec....	12
Εικ.2. Δομή των αρχικών δεδομένων(Dataset 1).....	16
Εικ.3.Δομή των αρχικών δεδομένων (Dataset 2).....	16
Εικ.4. Δεδομένα μετά την πρώτη επεξεργασία.....	17
Εικ.5.Δομή του τελικού dataset 1.....	18
Εικ.6.Δημιουργία μήτρα γειννίαςης.....	19
Εικ.7. Πίνακας διμερούς μετανάστευσης.....	20
Εικ.8. Ορισμός βαθμού κεντρικότητας.....	21
Εικ.9. Ορισμός κεντρικότητας εγγύτητας.....	22
Εικ.10. Ορισμός κεντρικότητας ενδιαμεσότητας.....	22
Εικ.11. Ορισμούς ιδιοδιανυσματικής κεντρικότητας.....	23
Εικ.12. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 1990.....	23
Εικ.13. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 1995.....	24
Εικ.14. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 2000.....	24
Εικ.15. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 2005.....	25
Εικ.16. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 2010.....	26
Εικ.17. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 2015.....	26
Εικ.18. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 2020.....	27
Εικ.19. Ιστογράμματα πυκνότητας έτος: 1990.....	30
Εικ.20. Κοινότητα I, έτος: 1990.....	31
Εικ.21. Κοινότητα II, έτος: 1990.....	31
Εικ.22. Κοινότητα III, έτος: 1990.....	32
Εικ.23. Κοινότητα IV, έτος: 1990.....	32
Εικ.24. Κοινότητα V, έτος: 1990.....	33
Εικ.25. Ιστογράμματα πυκνότητας έτος: 2020.....	37
Εικ.26. Κοινότητα I, έτος: 2020.....	37
Εικ.27. Κοινότητα II, έτος: 2020.....	38
Εικ.28. Κοινότητα III, έτος: 2020.....	38
Εικ.29.Δημιουργία μοντέλου πρόβλεψης μελλοντικών συνδέσεων.....	40
Εικ.30.Διάγραμμα μοντέλου.....	40
Εικ.31.Εκπαίδευση μοντέλου.....	41
Εικ.32.Μοντέλο πρόβλεψης μελλοντικών.....	42
Εικ.33.Μοντέλο αλγορίθμου Node2Vec.....	43

Εικ.34. Base Model-Loss function.....	45
Εικ.35. Node2Vec Model-Loss function .....	46
Εικ.36. Base Model-True Loss.....	47
Εικ.37. Node2Vec Model-True Loss.....	47
Εικ.38. Base Model Second Loss Function (dataset 2).....	47
Εικ.39. Base Model Second True - Loss Function (dataset 2).....	48
Εικ.40. Base Model Second Loss Function (dataset 2).....	48
Εικ.41. Base Model Second True - Loss Function (dataset 2).....	48



## 1 Εισαγωγή

### 1.1 ΑΝΤΙΚΕΙΜΕΝΟ ΕΡΓΑΣΙΑΣ

Ο βασικός στόχος της εργασίας αυτής είναι να αξιοποίηση των δυνατοτήτων που προσφέρουν τα Νευρωνικά Δίκτυα Γράφων (GNN) και του αλγορίθμου `node2vec` για τον εντοπισμό και την ανάλυση μοτίβων μετανάστευσης σε γραφήματα που αντιπροσωπεύουν πολύπλοκα συστήματα ή δίκτυα. Η μετανάστευση αφορά τη φυσική μετακίνηση ατόμων ή οντοτήτων μεταξύ γεωγραφικών τοποθεσιών. Οι ιδιαιτερότητες αυτών των προτύπων ενσωματώνονται στη δυναμική της δομής και των χαρακτηριστικών των γραφημάτων που αναπαριστούν αυτά τα συστήματα, όπου οι κόμβοι συμβολίζουν τις οντότητες (Χώρες) και οι ακμές υποδηλώνουν τις σχέσεις ή τις αλληλεπιδράσεις μεταξύ τους (πλήθος ανθρώπου που μεταναστεύουν από μια χώρα, σε μια άλλη).

Τα Νευρωνικά Δίκτυα Γραφημάτων παρέχουν ένα ισχυρό πλαίσιο για την εκμάθηση αυτών των μοτίβων, αξιοποιώντας τις πλούσιες σχεσιακές πληροφορίες που ενυπάρχουν στα δεδομένα. Τα GNN υπερέρχουν στην καταγραφή των εξαρτήσεων και της επιρροής των γειτονικών κόμβων, επιτρέποντας στο μοντέλο να κατανοήσει πολύπλοκα δομικά μοτίβα και χαρακτηριστικά κόμβων που είναι ενδεικτικά της συμπεριφοράς μετανάστευσης. Με την επαναληπτική ενημέρωση των αναπαραστάσεων των κόμβων μέσω του παραδείγματος μετάδοσης μηνυμάτων, τα GNNs μπορούν να μάθουν και να προβλέψουν αποτελεσματικά τον τρόπο με τον οποίο οι οντότητες στο δίκτυο θα μπορούσαν να συμπεριφερθούν ή να μετακινηθούν, δεδομένης της τρέχουσας κατάστασής τους και της δομής του τοπικού περιβάλλοντος του γράφου τους.

Συμπληρωματικά, ο αλγόριθμος `node2vec` χρησιμεύει ως κρίσιμο εργαλείο σε αυτή την προσπάθεια, μαθαίνοντας συνεχείς αναπαραστάσεις χαρακτηριστικών για τους κόμβους του γράφου. Η μέθοδος αυτή βοηθάει στην εξερεύνηση ποικίλων γειτονικών κόμβων και την εκμετάλλευση άμεσων κοινοτήτων, αποδίδοντας πλούσιες και κατατοπιστικές για πληροφορίες για τους κόμβους του γραφήματος. Αυτές οι πληροφορίες ενσωματώνουν τόσο τοπικές όσο και παγκόσμιες δομικές πληροφορίες, με τον τρόπο αυτό γίνονται ιδιαίτερα αποτελεσματικές για μεταγενέστερες εργασίες όπως η ομαδοποίηση κόμβων με βάση τα πρότυπα μετανάστευσης ή η πρόβλεψη μελλοντικών κινήσεων.

### 1.2 ΔΙΑΡΘΡΩΣΗ ΕΡΓΑΣΙΑΣ

Ο τρόπος διάρθρωσης της εργασίας, πραγματοποιείται με τον ακόλουθο τρόπο: Αρχικά, γίνεται μια παρουσίαση του απαραίτητου θεωρητικού υποβάθρου για τα νευρωνικά δίκτυα στους γράφους, πως δουλεύει η υλοποίηση Graph Convolution Neural Network καθώς και ο αλγόριθμος `Node2Vec`. Στο τρίτο κεφάλαιο, παρασιάζονται τα αρχικά δεδομένα και πραγματοποιείται μια εξήγηση των βημάτων που έγιναν ώστε να σχηματιστεί το τελικό σύνολο των δεδομένων. Στη συνέχεια τα δεδομένα αυτά μετασχηματίζονται σε γραφήματα ανα χρονική περίοδο και δημιουργούνται οι αντίστοιχες διμερές μήτρες μετανάστευσης. Στο τέταρτο κεφάλαιο γίνεται μια παρουσίαση των θεμελιωδών ορισμών των γράφων σχετικά με τα μέτρα κεντρικότητας και αποτυπώνονται τα γραφήματα για κάθε περίοδο. Επίσης, γίνεται η ανάλυση τις πρώτης και της τελευταίας περιόδου (1990, 2020). Στην ανάλυση αυτή γίνεται ο σχολιασμός των χωρών με τα υψηλότερα μέτρα κεντρικότητας, καθώς και οι διαφορές που υπάρχουν στις κοινότητες αυτών. Στο πέμπτο κεφάλαιο, παρουσιάζεται η υλοποίηση των διαφορετικών μοντέλων και ο αλγόριθμος `Node2Vec`, δείχνοντας αναλυτικά πως δημιουργήθηκαν τα μοντέλα αυτά και πως έγινε η εκπαίδευσή τους. Τέλος, στο έκτο κεφάλαιο παρουσιάζονται τα αποτελέσματα των μοντέλων αυτών κατά την εκπαίδευσή τους αλλά και κατά την επικύρωσή τους σε μελλοντικές προβλέψεις.

## 2 Θεωρητικό Υπόβαθρο

### 2.1 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΓΡΑΦΩΝ

Ένας γράφος ή ένα γράφημα είναι μια αφηρημένη αναπαράσταση ενός συνόλου στοιχείων (κόμβοι), όπου μερικά ζεύγη αυτών συνδέονται μεταξύ τους με δεσμούς(ακμές). Οι κόμβοι ενός γραφήματος μπορούν να αναπαριστούν οντότητες όπως άνθρωποι, αντικείμενα, χώρες ή έννοιες και οι ακμές αναπαριστούν σχέσεις μεταξύ αυτών των οντοτήτων. Για παράδειγμα, στην έρευνα του μεταναστευτικού μπορούμε να αποτυπώσουμε τις χώρες ως κόμβους και οι συνδέσεις μεταξύ αυτών, αποτυπώνουν το πλήθος ανθρώπων που μετακινήθηκαν από την χώρα προέλευσης στην χώρα προορισμού.

Στις μέρες μας η ανάλυση των γραφημάτων με μηχανική μάθηση, λαμβάνει όλο και περισσότερη προσοχή λόγω της μεγάλης εκφραστικής δύναμης που παρέχουν τα γραφήματα. Οι γράφοι παρέχουν τη δυνατότητα αποτύπωσης περίπλοκων καθημερινών δομών δεδομένων και συστημάτων, όπως για παράδειγμα ένα κοινωνικό δίκτυο, ένα σύστημα προτάσεων και η αναζήτηση ενός καινούργιου φαρμάκου. Η ανάλυση των γράφων επικεντρώνεται κυρίως σε εργασίες όπως η κατηγοριοποίηση ενός κόμβου(node classification), την πρόβλεψη συνδέσμων μεταξύ των κόμβων(Link prediction) καθώς και ομαδοποίηση των κόμβων με βάση κάποια κοινά χαρακτηριστικά(Clustering). Τα νευρωνικά δίκτυα γραφημάτων (GNNs) βασίζονται σε μεθόδους βαθιάς προηγμένης μάθησης που λειτουργούν στον τομέα των γραφημάτων.

Ανά τα χρόνια, έχουν παρουσιαστεί αρκετές τεχνικές και μέθοδοι μηχανικής μάθησης για τα νευρωνικά δίκτυα γραφημάτων όπως τα αναδρομικά νευρωνικά δίκτυα (Recursive Neural Networks, RNNs) όπου ερευνητές άρχισαν να διερευνούν τρόπους εφαρμογής των νευρωνικών δικτύων σε δεδομένα υπό τη μορφή γράφου. Ωστόσο, τα νευρωνικά δίκτυα γραφημάτων άρχισαν να κερδίζουν δημοτικότητα μέσα στη δεκαετία αυτή, χάρη στην ανάπτυξη νέων αρχιτεκτονικών και αλγορίθμων. Καταλυτικός παράγοντας στην εύρεση καινούργιων αρχιτεκτονικών σχετικά με τα GNNs ήταν η παρουσίαση των συνελκτικών νευρωνικών δικτύων (Convolutional Neural Network, CCNs) που πέτυχε κορυφαία αποτελέσματα σε διάφορα σύνολα δεδομένων, προκαλώντας νέο ενδιαφέρον για τα GNN. Έκτοτε, έχει αναπτυχθεί ένα ευρύ φάσμα αρχιτεκτονικών και αλγορίθμων GNN, συμπεριλαμβανομένων των Graph Attention Networks(GATs), GraphSage και Graph Isomorphism Networks (GINs). Για την αναγνώριση προτύπων και την πρόβλεψη του μεταναστευτικού, στη συγκεκριμένη εργασία χρησιμοποιήθηκε η αρχιτεκτονική Graph Convolutional Neural Network (GCN).

### 2.2 GRAPH CONVOLUTIONAL NEURAL LAYER

Το στρώμα ενός συνελκτικού νευρωνικού δικτύου (Graph Convolutional Neural Layer) είναι ένας τύπος επιπέδου νευρωνικού δικτύου που έχει σχεδιαστεί ειδικά για την επεξεργασία δεδομένων που είναι δομημένα ως γράφημα. Σε ένα κανονικό νευρωνικό δίκτυο, επεξεργαζόμαστε μεμονωμένα σημεία δεδομένων ανεξάρτητα. Αλλά σε έναν γράφο, ένας κόμβος συχνά ορίζεται από τις σχέσεις του (ακμές) με άλλους κόμβους. Ένα στρώμα GCN λειτουργεί συγκεντρώνοντας πληροφορίες από τους γείτονες ενός κόμβου (άλλους κόμβους με τους οποίους συνδέεται). Για παράδειγμα σε ένα κοινωνικό δίκτυο που ο κάθε κόμβος, είναι ένας χρήστης του δικτύου αυτού, να λαμβάνει ενημερώσεις ή μηνύματα από τους φίλους του (άλλους κόμβους του δικτύου). Η μάθηση ενός τέτοιου μοντέλου επιτυγχάνεται μέσω των χαρακτηριστικών που έχει ένας κόμβος. Κάθε κόμβος σε έναν γράφημα έχει τα χαρακτηριστικά του (όπως στο μεταναστευτικό δίκτυο, το προφίλ της κάθε χώρας μπορεί να έχει χαρακτηριστικά όπως ακαθάριστο εγχώριο προϊόν, τοποθεσία, πολιτεύμα). Κατά τη μάθηση των χαρακτηριστικών του κάθε κόμβου τα συνελκτικά νευρωνικά στρώματα για τους γράφους έχουν την ικανότητα να ενημερώνουν τα χαρακτηριστικά του κάθε κόμβου συνδυάζοντας τα δικά του χαρακτηριστικά με τα χαρακτηριστικά των γειτόνων του. Με αυτόν τον τρόπο, κάθε κόμβος αποκτά μια πιο ολοκληρωμένη αντίληψη που αντικατοπτρίζει τη θέση και το ρόλο του στο ευρύτερο δίκτυο.

Ένα συνελκτικό νευρωνικό στρώμα γραφημάτων έχει εκπαιδευσιμα βάρη. Όταν συνδυάζει τα χαρακτηριστικά ενός κόμβου με τα γειτονικά του, χρησιμοποιεί αυτά τα βάρη για να

δώσει σημασία σε διαφορετικά χαρακτηριστικά. Για παράδειγμα, σε ένα μεταναστευτικό δίκτυο μεταξύ χωρών, μπορεί να μάθει να δίνει μεγαλύτερη σημασία στο πολίτευμα που έχει η κάθε χώρα από ό,τι στον κλίμα που έχει, κατά την ενημέρωση του προφίλ μιας χώρας. Ένα ξεχωριστό χαρακτηριστικό που έχουν είναι η μη γραμμικότητα. Μετά το συνδυασμό και τη στάθμιση των χαρακτηριστικών, εφαρμόζεται συχνά μια μη γραμμική συνάρτηση (όπως η Rectified Linear Unit, ReLU). Αυτό βοηθά το δίκτυο να μάθει σύνθετα μοτίβα και όχι μόνο απλές ευθείες σχέσεις.

Ακριβώς όπως και στα παραδοσιακά νευρωνικά δίκτυα, μπορούν να δημιουργηθούν πολλαπλά επίπεδα συνελκτικών νευρωνικών στρωμάτων στοιβαγμένα το ένα πάνω στο άλλο. Κάθε στρώμα μπορεί να αποτυπώσει πιο σύνθετες σχέσεις στο γράφημα. Με κάθε στρώμα, οι πληροφορίες μπορούν να ταξιδέψουν περισσότερο μέσα στο γράφημα. Το πρώτο στρώμα βλέπει μόνο τους άμεσους γείτονες, αλλά τα επόμενα στρώματα μπορούν να δουν τους γείτονες των γειτόνων και ούτω καθεξής. Συνοπτικά, ένα Graph Convolutional Neural Layer είναι σαν ένας ξύπνιος αλγόριθμος που αναμειγνύει τις πληροφορίες κάθε κόμβου με τις πληροφορίες των γειτόνων του για να κατανοήσει τη συνολική εικόνα. Είναι ιδιαίτερα ισχυρός για σύνθετα δεδομένα όπως ένα γράφημα, όπου οι σχέσεις και οι συνδέσεις μεταξύ των κόμβων είναι το κλειδί για την κατανόηση της συνολικής δομής και των μοτίβων.

Έστω ότι έχουμε ένα μη κατευθυνόμενο γράφημα  $G = (V, E, X)$  όπου  $V$  είναι οι κόμβοι,  $E$  οι ακμές του γραφήματος και  $X$  τα χαρακτηριστικά των κόμβων, τα χαρακτηριστικά αυτά ανήκουν  $X \in \mathbb{R}^{N \times d}$ , όπου  $N$  ο αριθμός των κόμβων και  $d$  ο αριθμός των χαρακτηριστικών. Στη συνέχεια μπορούμε να ορίσουμε τη μήτρα γειτνίασης των γραφήματος ως  $A = [a_{ij}] \in \mathbb{R}^{N \times N}$ . Έχουμε  $d_i = \sum_j a_{ij}$  όπου είναι ο βαθμός του κάθε κόμβου και  $D = \text{diag}(d_1, d_2, \dots, d_n)$  είναι το degree matrix αναφορικά με τη μήτρα γειτνίασης  $A$ . Το  $A' = A + I$ , όπου  $A$  η αρχική μήτρα γειτνίασης και  $I$  ο ταυτοτικός πίνακας και  $d'$  είναι ο αντίστοιχος βαθμός του πίνακα με  $D' = \sum_j a'_{ij}$ . Το σύμβολο  $(\ell)$  χρησιμοποιείται για το  $\ell$ -οστό layer του νευρωνικού και  $H^{(\ell)}$  είναι η είσοδος των χαρακτηριστικών των κόμβων στο  $\ell$ -οστό layer.

Έχοντας έναν μη κατευθυνόμενο γράφημα μπορούμε να περιγράψουμε τα συνελκτικά γραφήματα ως μιας διαδικασίας:

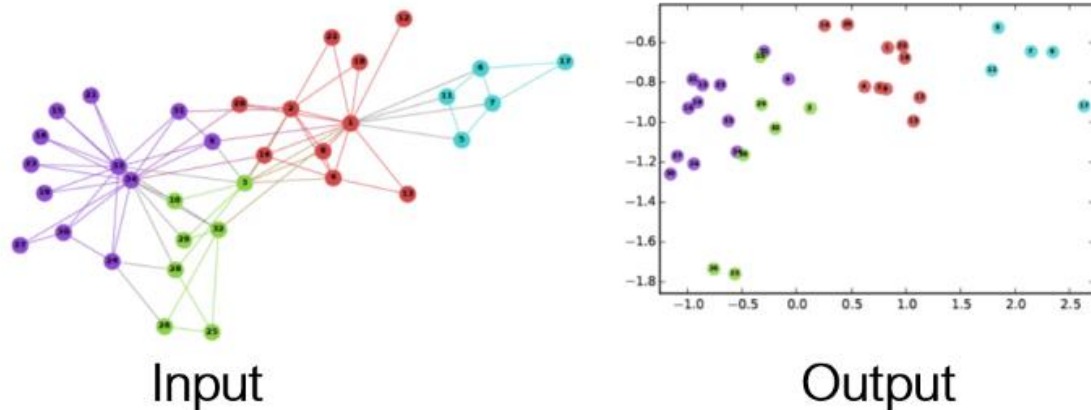
$$H^{(\ell+1)} = \sigma(\tilde{A} H^{(\ell)} W^{(\ell)}) \quad (1)$$

όπου  $\tilde{A} = D'^{-\frac{1}{2}} A' D'^{-\frac{1}{2}}$  ένας πίνακας που προκύπτει από πράξεις των παραπάνω πινάκων και  $W^{(\ell)}$  περιέχει τα βάρη του μοντέλου. Άξιο προς αναφορά είναι πως το  $H^{(\ell+1)}$  είναι η έξοδος του  $\ell$ -οστού επιπέδου και η είσοδος στο επόμενο επίπεδο. Τέλος η παράμετρος  $\sigma$ , είναι η μη γραμμική συνάρτηση ενεργοποίησης, για παράδειγμα RELU και ορίζεται από τον τύπο:

$$\mathcal{L} = - \sum_{v_i \in \mathcal{Y}^{label}} \sum_{c=1}^C Y_{i,c} \log(\tilde{Y}_{i,c}), \quad (2)$$

### 2.3 ΑΛΓΟΡΙΘΜΟΣ NODE2VEC

Ο Node2vec είναι ένας αλγόριθμος που χρησιμοποιείται για την αποτελεσματική αναπαράσταση των κόμβων ενός γράφου σε ένα συνεχή διανυσματικό χώρο. Αυτές οι αναπαραστάσεις μπορούν στη συνέχεια να χρησιμοποιηθούν σε διάφορες εργασίες μηχανικής μάθησης.



Εικ.1. Αποτύπωση ενός γράφου στο διανυσματικό χώρο με τη χρήση του αλγόριθμου Node2Vec

Ακολουθεί μια συνοπτική εξήγηση του τρόπου λειτουργίας του: Όπως στο παράδειγμα της εργασίας με την μετανάστευση ανθρώπων, υπάρχει ένας γράφος με κόμβους και ακμές, όπου οι χώρες των δεδομένων είναι οι κόμβοι και το πλήθος των ανθρώπων που μετακινούνται από μια χώρα σε μια άλλη (migration flows), είναι οι ακμές. Ο στόχος είναι η κατανόηση του κάθε κόμβου (χώρα) εξετάζοντας την τοπική γειτονιά του δικτύου του (άλλες χώρες και migration flows).

Ο Node2vec εκτελεί τυχαίους περιπάτους για να δειγματίσει τη γειτονιά ενός κόμβου. Αυτό είναι σαν να μεταπηδάμε τυχαία από χώρα σε χώρα στο δίκτυο, ξεκινώντας από μια τυχαία χώρα κάθε φορά. Οι διαδρομές που ακολουθούνται κατά τη διάρκεια αυτών των τυχαίων περιπάτων αντιπροσωπεύουν τις διάφορες πιθανές γειτονιές γύρω από έναν κόμβο.

Ο αλγόριθμος μπορεί να ρυθμιστεί ώστε να εξερευνά τη γειτονιά του κάθε με διαφορετικό τρόπο, εστιάζοντας έτσι περισσότερο στις άμεσες χώρες ενός κόμβου (τοπική προβολή) ή περιπλανώμενος πιο μακριά στις χώρες του γραφήματος (παγκόσμια προβολή). Αυτό επιτυγχάνεται μέσω δύο παραμέτρων ( $p$  και  $q$ ) που υπαγορεύουν τη συμπεριφορά του περιπάτου.

Ο αλγόριθμος δημιουργεί μια διανυσματική αναπαράσταση (node embeddings) για κάθε κόμβο που συνοψίζει τη θέση του στο γράφημα. Αυτό το επιτυγχάνει βελτιστοποιώντας τα διανύσματα έτσι ώστε οι κόμβοι που εμφανίζονται συχνά σε παρόμοιες περιπλανήσεις να έχουν παρόμοια διανύσματα.

Ο αλγόριθμος εκπαιδεύεται μέσω μιας τεχνική από την επεξεργασία φυσικής γλώσσας (word2vec αλγόριθμος) που ονομάζεται Skip-Gram, η οποία είναι καλή στην πρόβλεψη του πλαισίου χαρακτηριστικών (στην προκειμένη περίπτωση, γειτονικοί κόμβοι) ενός κόμβου. Ουσιαστικά, αντιμετωπίζει τους κόμβους σαν λέξεις και τους περιπάτους σαν προτάσεις, προσπαθώντας να προβλέψει τους γείτονες κάθε κόμβου.

Οι παράμετροι  $p$  και  $q$  στον αλγόριθμο node2vec είναι κρίσιμες για τον έλεγχο του τρόπου με τον οποίο οι τυχαίοι περίπατοι εξερευνούν το δίκτυο. Ουσιαστικά υπαγορεύουν τη στρατηγική των περιπάτων, εξισορροπώντας μεταξύ της δειγματοληψίας με βάση το εύρος (όπως η εξερεύνηση επίπεδο προς επίπεδο σε ένα δέντρο) και της δειγματοληψίας με βάση το βάθος (όπως η εξερεύνηση ενός μονοπατιού σε βάθος πριν προχωρήσουμε στο επόμενο).

Παράμετρος επιστροφής  $p$ , έχει ως σκοπό τον έλεγχο της πιθανότητας άμεσης επανάληψης ενός κόμβου στον περίπατο. Αυτό έχει σαν αποτέλεσμα, μια υψηλή τιμή της  $p$  καθιστά λιγότερο πιθανό να επιστρέψουμε στον κόμβο από τον οποίο μόλις ήρθαμε, ενθαρρύνοντας την εξερεύνηση μακριά από τον κόμβο προέλευσης. Μια χαμηλότερη τιμή ενθαρρύνει τους στενούς βρόχους, ενισχύοντας την εξερεύνηση της τοπικής γειτονιάς.

Η παράμετρος εισόδου-εξόδου (in-out)  $q$ , έχει ως σκοπό τον έλεγχο της αναζήτησης που θα πραγματοποιηθεί. Μια υψηλή τιμή του  $q$  μεροληπτεί ώστε ο τυχαίος περίπατος να κινείται προς

κόμβους που βρίσκονται κοντά στον προηγούμενο κόμβο (προς τα μέσα), ευνοώντας τα χαρακτηριστικά μιας αναζήτησης σε πλάτος (Breadth-first search, BFS). Μια χαμηλότερη τιμή του  $q$  επιτρέπει στον τυχαίο περίπατο να εξερευνήσει περαιτέρω τμήματα του γράφου, ευνοώντας τα χαρακτηριστικά αναζήτησης σε βάθος (Depth first search, DFS). Αντίθετα, ένα υψηλό  $q$  θα έκανε τον αλγόριθμο να κοιτάξει περισσότερο τις συνδέσεις κοντά στο σημείο όπου έχει βρεθεί προηγουμένως (μένοντας τοπικά).

Μαζί, οι παράμετροι  $p$  και  $q$  επιτρέπουν στο αλγόριθμο να εξερευνά με ευελιξία τη γειτονιά ενός κόμβου. Ρυθμίζοντας αυτές τις παραμέτρους, μπορεί κανείς να κάνει τους τυχαίους περιπάτους να δίνουν προτεραιότητα στην επίσκεψη κόμβων που βρίσκονται κοντά στη δομή του γράφου (εκμετάλλευση) ή να ενθαρρύνουν την επίσκεψη πιο απομακρυσμένων τμημάτων του γράφου (εξερεύνηση). Αυτή η ευελιξία είναι ζωτικής σημασίας, επειδή διαφορετικές εργασίες μπορεί να απαιτούν την κατανόηση διαφορετικών πτυχών της δομής του γράφου. Για παράδειγμα, σε ορισμένες περιπτώσεις, η κατανόηση της άμεσης κοινότητας (τοπική δομή) γύρω από τους κόμβους είναι πιο σημαντική, ενώ σε άλλες περιπτώσεις, η κατανόηση του ρόλου του κόμβου στο ευρύτερο πλαίσιο του γράφου (παγκόσμια δομή) είναι πιο κρίσιμη.

Αναλυτικότερα, πρέπει να μετατραπούν οι κόμβοι ενός γραφήματος σε διανύσματα. Έστω το γράφημα  $G = (V, E)$  όπου  $V$  οι κόμβοι και  $E$  οι ακμές ενός γραφήματος, Έχοντας τη συνάρτηση  $f: V \rightarrow \mathbb{R}^d$  όπου  $d$  τα χαρακτηριστικά των κόμβων προκύπτει η μήτρα με την αναπαράσταση του κάθε κόμβου με μέγεθος  $|V| \times d$ . Ο αρχικός κόμβος  $u \in V$  και  $N_s(u) \subset V$  που είναι η γειτονία του κόμβου  $u$  με βάση τις ακμές του γραφήματος και τη στρατηγική που θα ακολουθήσουμε. Ο αλγόριθμος Node2Vec είναι όλος βασισμένος στους τυχαίους περιπάτους στο γράφημα και την δειγματοληπτική στρατηγική. Ο αλγόριθμος χρησιμοποιεί το μοντέλο Skip-gram και προσπαθεί να βελτιστοποιήσει την παρακάτω συνάρτηση, όπου προσπαθεί να μεγιστοποιήσει τη λογαριθμική πιθανότητα (log-probability) παρατήρησης μιας γειτονιάς  $N_s(u)$  για κάθε κόμβο  $u$  υπό την προϋπόθεση της αναπαράστασης των χαρακτηριστικών του, η οποία δίνεται από τη σχέση  $f$ :

$$\max_f \sum_{u \in V} \log \Pr(N_s(u)|f(u)) \quad (1)$$

Για να γίνει το πρόβλημα βελτιστοποίησης εφικτό, στον αλγόριθμο έγιναν δύο παραδοχές: Ανεξαρτησία υπό όρους, δηλαδή γίνεται η παραγοντοποίηση της πιθανότητα υποθέτοντας ότι η πιθανότητα παρατήρησης ενός γειτονικού κόμβου είναι ανεξάρτητη από την παρατήρηση οποιουδήποτε άλλου γειτονικού κόμβου δεδομένης της αναπαράστασης χαρακτηριστικών της πηγής:

$$\Pr(N_s(u)|f(u)) = \prod_{n_i \in N_s(u)} \Pr(n_i|(f(u))) \quad (2)$$

Συμμετρία στο χώρο χαρακτηριστικών, δηλαδή ένας αρχικός κόμβος και ένας κόμβος  $n_i \in N_s(u)$  έχουν συμμετρική επίδραση ο ένας πάνω στον άλλο στο χώρο χαρακτηριστικών. Κατά συνέπεια, γίνεται η μοντελοποίηση την υπό συνθήκη πιθανότητα κάθε ζεύγους κόμβου πηγής-γειτονιάς ως softmax μονάδα που παραμετροποιείται από το γινόμενο των χαρακτηριστικών τους:

$$\Pr(n_i|f(u)) = \frac{\exp(f(n_i) \cdot f(u))}{\sum_{v \in V} \exp(f(v) \cdot f(u))} \quad (3)$$

Με τις βάση τις παραδοχές που αναφέρθηκαν παραπάνω η βελτιστοποίηση του μοντέλου είναι:

$$\max_f \sum_{u \in V} \left[ -\log Z_u + \sum_{n_i \in N_s(u)} (n_i) \cdot f(u) \right] \quad (4)$$

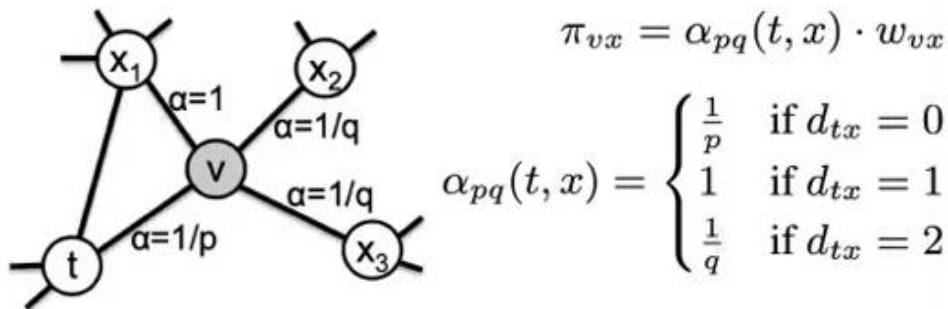
Όπου  $Z_u$  η συνάρτηση διαμερισμού ανά κόμβο είναι:

$$Z_u = \sum_{v \in V} \exp(f(u) \cdot f(v)) \quad (5)$$

Ο υπολογισμός της παραπάνω συνάρτησης είναι πολύ ακριβός, ειδικά σε πολύ μεγάλα δίκτυα. Για τον λόγο αυτό ο υπολογισμός της γίνεται κατά προσέγγιση με τη μέθοδο της αρνητικής δειγματοληψίας. Επίσης, γίνεται η βελτιστοποίηση της εξίσωσης (4) γίνεται με τη χρήση του stochastic gradient ascent [1] πάνω στις παραμέτρους του μοντέλου που ορίζουν τα χαρακτηριστικά  $f$ . Με τον τρόπο γίνεται η εκπαίδευση του μοντέλου χρησιμοποιώντας μερικά θετικά δείγματα και μερικά αρνητικά. Με την παραπάνω θεωρία πραγματοποιείται ένας τυχαίος περίπατος στο δίκτυο. Μπορεί να γίνει μια δειγματοληψία με την Breadth-first στρατηγική ή με την Depth-first στρατηγική και να καταλήξουμε σε τελείως διαφορετικές αναπαραστάσεις των κόμβων. Ο Αλγόριθμος Node2Vec δίνει τη δυνατότητα του ελέγχου στις παραπάνω στρατηγικές και να υπάρξει ένα ενδιάμεσο αποτέλεσμα. Χαρακτηριστικά,  $c_i \rightarrow i$ -οστός κόμβος στον τυχαίο περίπατο και  $c_0 = u$  είναι ο αρχικός κόμβος, τότε μπορούμε να ορίσουμε την πιθανότητα ότι η τρέχουσα κατάσταση  $c_i$  να βρίσκεται στην κατάσταση  $x$ , δεδομένου ότι η προηγούμενη κατάσταση  $c_i - 1$  ήταν στην κατάσταση  $v$ :

$$P(c_i = x \mid c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} \\ 0 \end{cases}$$

Αν  $(v, x) \in E$  τότε ισχύει η  $\frac{\pi_{vx}}{Z}$  είναι η μη κανονικοποιημένη πιθανότητα μετάβασης μεταξύ των κόμβων  $v$  και  $x$ , και  $Z$  είναι η σταθερά κανονικοποίησης. Παρακάτω υπάρχει ένα αναλυτικό παράδειγμα μιας βόλτας στο γράφημα και πως υπολογίζεται η παραπάνω τιμή:



Έστω ότι ο αρχικός μας κόμβος στο παραπάνω γράφημα είναι ο  $t$  και μετακινούμαστε στον κόμβο  $v$ , στη συνέχεια αξιολογούμε αν θα πρέπει να μετακινηθούμε πίσω στον αρχικό κόμβο ή πρέπει να μετακινηθούμε στους κόμβους  $x_1, x_2$  ή  $x_3$ . Με βάση τις δυο υπερ-παραμέτρους που αναφέραμε και προηγουμένως  $p, q$  μπορούμε να έχουμε τον έλεγχο αν ο περίπατος θα κάνει Depth-first, Breadth-first ή κάτι ενδιάμεσο.  $d_{tx}$  είναι η ελάχιστη απόσταση μεταξύ των κόμβων  $t, x$ . Για παράδειγμα:

- $d_{tx}(t, t) = 0$
- $d_{tx}(t, x_1) = 1$
- $d_{tx}(t, x_2) = 2$
- $d_{tx}(t, x_3) = 2$

Με βάση τους παραπάνω αριθμούς διαλέγουμε την πιθανότητα  $\alpha_{pq}(t, x)$ . Παρακάτω εξετάζουμε τις υπερ-παραμέτρους.

$p > q \ \& \ p > 1 \Rightarrow \frac{1}{p} < 1 \ \& \ \frac{1}{p} < \frac{1}{q} \Rightarrow$  έχουμε μικρότερη πιθανότητα να γυρίσουμε στον κόμβο  $t$ . Αντίθετα, αν έχουμε  $p < q \ \& \ p < 1 \Rightarrow \frac{1}{p} > 1 \ \& \ \frac{1}{p} > \frac{1}{q} \Rightarrow$  έχουμε μεγαλύτερη πιθανότητα να επιστρέψουμε στον αρχικό μας κόμβο  $t$ . Σε ένα γενικό πλαίσιο η παράμετρος  $p$  είναι παράμετρος επιστροφής, όσο την αλλάζουμε αλλάζει και η πιθανότητα επιστροφής ή όχι στον αρχικό μας

κόμβο. Αν  $q > 1 \Rightarrow$  ο περίπατος περιστρέφεται κοντα στον αρχικό κόμβο, επομένως έχουμε μια τοπική εξερεύνηση, ενώ αν το  $q < 1 \Rightarrow$  ο περίπατος θα εξερευνήσει περισσότερο κομμάτια του γραφήματος.

Συνοπτικά ο αλγόριθμος `node2vec` δίνει τη δυνατότητα μέσω των χαρακτηριστικών του κάθε κόμβου (`node embeddings`), κατανόησης ομοιοτήτων μεταξύ κόμβων και μπορεί να χρησιμοποιηθεί για την πρόβλεψη συνδέσεων μεταξύ κόμβων ή την ταξινόμηση των κόμβων σε κατηγορίες, μεταξύ άλλων εργασιών, κάτι που μπορεί να είναι χρήσιμο για συστήματα συστάσεων, μελέτες διάχυσης πληροφοριών και πολλές άλλες εφαρμογές σε πολύπλοκα δίκτυα.

### 3 Δεδομένα και Προ Επεξεργασία

#### 3.1 ΔΕΔΟΜΕΝΑ

Τα δεδομένα της εργασίας πάρθηκαν από την ιστοσελίδα [Our World in Data](#). Η πρωταρχική μορφή των δεδομένων παρουσιάζεται στην παρακάτω εικόνα.

A	B	C	D	E	F	G	H
Year	Country	Emigrants from Afghanistan	Immigrants to Afghanistan	Emigrants from Albania	Immigrants to Albania	Emigrants from Algeria	Immigrants to Algeria
1990	Afghanistan	-7679582	57686				
1995	Afghanistan	-4347049	71522				
2000	Afghanistan	-4750677	75917				
2005	Afghanistan	-4116739	87314				
2010	Afghanistan	-5269518	102276				
2015	Afghanistan	-5400916	339432				
2020	Afghanistan	-5853838	144098				
1990	Albania			-180204	66013		
1995	Albania			-501066	71354		
2000	Albania			-824442	76695		
2005	Albania			-966032	64739		
2010	Albania			-1117940	52784		
2015	Albania			-1142844	52031		
2020	Albania			-1250451	48810		
1990	Algeria					-921665	273954
1995	Algeria					-979641	262032
2000	Algeria					-1037985	250110
2005	Algeria					-1599647	197728
2010	Algeria					-1655264	217268
2015	Algeria					-1836502	239473
2020	Algeria					-2022337	250378

Εικ.2. Δομή των αρχικών δεδομένων(Dataset 1)

Στην παραπάνω εικόνα που παρέχετε φαίνονται τα δεδομένα για τις μεταναστευτικές ροές, για τα έτη 1990, 1995, 2000, 2005, 2010, 2015 και 2020. Το σύνολο των δεδομένων περιλαμβάνει στήλες τόσο για τη μετανάστευση (άτομα που εγκαταλείπουν τη χώρα) όσο και για τη μετανάστευση (άτομα που εισέρχονται στη χώρα). Για παράδειγμα για το Αφγανιστάν ο αριθμός των μεταναστών (στήλη C) είναι αρνητικός, γεγονός που μπορεί να υποδηλώνει το πλήθος των ατόμων που έφυγαν από τη συγκεκριμένη χώρα. Ο αριθμός των μεταναστών (στήλη Δ) είναι θετικός, αυξανόμενος με την πάροδο του χρόνου, γεγονός που υποδηλώνει την αύξηση του αριθμού των ανθρώπων που εισέρχονται στο Αφγανιστάν. Από τον παραπάνω πίνακα κρατήσαμε το έτος(στήλη A), την χώρα (στήλη B) και στη συνέχεια όλες τις στήλες που δείχνουν το πλήθος ανθρώπων που μετανάστευσαν στην χώρα της στήλης B. Δηλαδή, τη στήλη που μας δείχνει το πλήθος των ανθρώπων που μετανάστευσαν από την Αλβανία (στήλη E) προς τη χώρα της γραμμής (στήλη B) τη συγκεκριμένη χρονική περίοδο (στήλη A).

Για την αξιολόγηση των μοντέλων που αναπτύχθηκαν στα επόμενα κεφάλαια χρησιμοποιήθηκε επίσης και ένα δεύτερο σύνολο δεδομένων, το οποίο πάρθηκε από την ιστοσελίδα του Οργανισμού Οικονομικής Συνεργασίας και Ανάπτυξης (OECD). Η αρχική μορφή των δεδομένων είναι η παρακάτω:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CO2	Country of birth/nationality	VAR	Variable	GEN	Gender	COU	Country	YEA	Year	Value	Flag Codes	Flags
2	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2000	2000	887		
3	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2001	2001	456		
4	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2002	2002	660		
5	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2003	2003	1015		
6	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2004	2004	1340		
7	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2005	2005	3463		
8	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2006	2006	3465		
9	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2007	2007	2560		
10	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2008	2008	2033		
11	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2009	2009	1696		
12	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2010	2010	1618		
13	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2011	2011	2030		
14	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2012	2012	1623		
15	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2013	2013	2178		
16	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2014	2014	5686		

Εικ.3.Δομή των αρχικών δεδομένων (Dataset 2)



Το σύνολο δεδομένων (Dataset 2) περιέχει πληροφορίες σχετικά με τις εισροές του αλλοδαπού πληθυσμού ανά εθνικότητα, εστιάζοντας ειδικά στα μεταναστευτικά δεδομένα. Περιέχει στήλες όπως είναι η CO2 που αντιπροσωπεύει έναν κωδικό για τη χώρα προέλευσης ή την εθνικότητα των μεταναστών, Country of birth/nationality που αναφέρεται στα ονόματα των χωρών γέννησης ή εθνικότητας των μεταναστών. Η στήλη VAR αποτελεί έναν κωδικό που σχετίζεται με τη στήλη Variable, στην προκειμένη περίπτωση, ο κωδικός "B11", είναι αυτός που θα κρατήσουμε. Η στήλη COU παρέχει και αυτή κωδικούς χωρών και συγκεκριμένα για τη χώρα προορισμού, καθώς η στήλη Country αποτελεί έχει το πλήρες όνομα της χώρας προορισμού στην οποία έφτασαν οι μετανάστες. Η στήλες YEA και Year είναι παρεμφερείς κολόνες και υποδεικνύουν το έτος της εγγραφής δεδομένων. Value είναι ο πραγματικός αριθμός των μεταναστών για τη δεδομένη εθνικότητα, το έτος και τη χώρα προορισμού. Για παράδειγμα, το 2000, 887 άτομα από το Αφγανιστάν μετανάστευσαν στην Αυστραλία. Το dataset αυτό περιέχει μεταναστευτικές ροές, για τα έτη 2000, 2001, 2002 έως το 2020.

### 3.2 ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Από τα αρχικά δεδομένα του dataset 1, πάρθηκαν οι στήλες που αφορούν του μετανάστες που εισήχθησαν στη χώρα (στήλες που ξεκινούν Emigrants from). Οι στήλες στη συνέχεια μετονομάστηκαν και κρατήθηκε μόνο το λεκτικό της κάθε χώρας. Επομένως, μετατράπηκαν σε ένα σύνολο δεδομένων που παρουσιάζεται στη παρακάτω εικόνα:

Year	Country	Afghanistan	Albania	Algeria	American Samoa	Andorra	Angola	Anguilla	Antigua and Barbuda	...	Uzbekistan
1990	Afghanistan	-7679582.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	2027.0
1995	Afghanistan	-4347049.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1836.0
2000	Afghanistan	-4750677.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1646.0
2005	Afghanistan	-4116739.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1422.0
2010	Afghanistan	-5269518.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1199.0
...	...	...	...	...	...	...	...	...	...	...	...
2000	Zimbabwe	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
2005	Zimbabwe	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
2010	Zimbabwe	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
2015	Zimbabwe	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
2020	Zimbabwe	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0

Εικ.4. Δεδομένα μετά την πρώτη επεξεργασία

Οι γραμμές του πίνακα αντιπροσωπεύουν τις χώρες προέλευσης, ενώ οι στήλες αντιπροσωπεύουν τις χώρες προορισμού. Κάθε κελί στο σημείο τομής μιας γραμμής (i) και μιας στήλης (j) δείχνει τον αριθμό των ατόμων που έχουν μεταναστεύσει από τη χώρα i στη χώρα j εντός μιας συγκεκριμένης χρονικής περιόδου t. Η πρώτη στήλη υποδεικνύει τα έτη, παρουσιάζοντας μια χρονοσειρά από το 1990 έως το 2020 ανά διαστήματα (1990, 1995, 2000). Η πρώτη σειρά μετά τους τίτλους των στηλών απαριθμεί διάφορες χώρες, υποδεικνύοντας τους προορισμούς των μεταναστών. Οι αρνητικοί αριθμοί υποδηλώνουν καθαρή εκροή μεταναστών από τη χώρα που αναφέρεται στην αντίστοιχη σειρά, ενώ οι θετικοί αριθμοί υποδηλώνουν καθαρή εισροή. Τα κελιά με μηδενικό αριθμό υποδηλώνουν μηδενική μετανάστευση μεταξύ των χωρών για τη συγκεκριμένη χρονική περίοδο. Η παραπάνω εικόνα είναι ενδεικτική των δεδομένων καθώς υπάρχουν δεδομένα που δεν εμφανίζονται στην τρέχουσα προβολή. Ο πίνακας περιλαμβάνει περισσότερες χώρες και σημεία δεδομένων από όσα είναι ορατά. Ο αρχικός πίνακας έχει 1659 γραμμές × 239 στήλες.

Στη συνέχεια τα δεδομένα μας επεξεργάστηκαν και δημιουργήθηκε ένα καινούργιο dataset (σύνολο δεδομένων), που στη συνέχεια θα χρησιμοποιηθεί για την εκπαίδευση και την πρόβλεψη των μοντέλων. Στην παρακάτω εικόνα παρουσιάζεται ένας δομημένος πίνακας μετανάστευσης, ο οποίος περιγράφει λεπτομερώς τον όγκο της μετανάστευσης μεταξύ ζευγών

χωρών για μια σειρά χρονικών σημείων. Κάθε γραμμή αντιπροσωπεύει μια μοναδική διαδρομή μετανάστευσης από μια χώρα προέλευσης σε μια χώρα-προορισμού. Οι στήλες έχουν ως εξής:

- **source\_country**: Η χώρα από την οποία μεταναστεύουν τα άτομα.
- **target\_country**: Η χώρα στην οποία μεταναστεύουν τα άτομα.
- **1990 έως 2020**: Οι στήλες αυτές αντιπροσωπεύουν τα έτη για τα οποία παρέχονται στοιχεία, ενώ κάθε κελί στις στήλες αυτές δείχνει τον αριθμό των ατόμων που μετανάστευσαν από τη χώρα προέλευσης στη χώρα-στόχο κατά το αντίστοιχο έτος.

	source_country	target_country	1990	1995	2000	2005	2010	2015	2020
0	Afghanistan	Pakistan	8107.0	17225.0	26343.0	37163.0	47984.0	268369.0	102500.0
1	Afghanistan	Tajikistan	40537.0	36733.0	32929.0	28460.0	23991.0	15031.0	4422.0
2	Afghanistan	Uzbekistan	2027.0	1836.0	1646.0	1422.0	1199.0	751.0	219.0
3	Albania	Canada	1151.0	1244.0	1337.0	1128.0	920.0	906.0	849.0
4	Albania	Greece	40087.0	43331.0	46575.0	39315.0	32054.0	31596.0	29643.0
...	...	...	...	...	...	...	...	...	...
12164	Zimbabwe	Malawi	69618.0	67945.0	92989.0	102211.0	103681.0	104247.0	108491.0
12165	Zimbabwe	Mozambique	482855.0	266322.0	197838.0	145402.0	117799.0	118442.0	122923.0
12166	Zimbabwe	South Africa	11426.0	15313.0	17024.0	19010.0	18910.0	19013.0	19731.0
12167	Zimbabwe	United Kingdom	19136.0	18132.0	23747.0	19394.0	12174.0	12243.0	12705.0
12168	Zimbabwe	Zambia	18966.0	22815.0	26337.0	29759.0	30373.0	30538.0	31692.0

12169 rows × 9 columns

Εικ.5. Δομή του τελικού dataset 1

Με βάση το παραπάνω dataset, φτιάχτηκαν οι μήτρες γειννίας (adjacency matrix) για κάθε χρονική στιγμή  $T$ . Ο πίνακας γειννίας είναι ένας τρόπος αναπαράστασης ενός γραφήματος σε μορφή πίνακα. Ο πίνακας αυτός έχει διαστάσεις ίσες με τον αριθμό των κόμβων του γράφου όπου κάθε γραμμή και στήλη του πίνακα αντιστοιχεί σε έναν κόμβο του γράφου. Εάν υπάρχει μια ακμή μεταξύ του κόμβου  $i$  και του κόμβου  $j$ , τότε η τιμή του πίνακα στη γραμμή  $i$  και τη στήλη  $j$  παίρνει την τιμή 1 (ή το βάρος της ακμής). Εάν δεν υπάρχει ακμή μεταξύ των δύο κόμβων τότε η τιμή στην αντίστοιχη εγγραφή τίθεται σε 0. Στις μήτρες γειννίας που αναφέρθηκαν παραπάνω μια ακμή υπάρχει στην περίπτωση που το πλήθος των ατόμων που μεταναστεύουν είναι πάνω από το μηδέν αλλά και πάνω από ένα κατώτατο όριο (threshold). Ανά γράφημα το threshold ορίστηκε από τον πολλαπλασιασμό μιας σταθερής τιμής 0.001 επί τη μέγιστη τιμή μετανάστευσης που υπάρχει στο εκάστοτε γράφημα.

```

181 # Loop through the various time instances.
182 for time,t in zip(time_range_cols,T):
183     print(f'Generating adjacency matrix for T: {t}')
184     # Initialize the current adjacency matrix.
185     mat = np.zeros((nodes_num,nodes_num),dtype=np.float64)
186     # Loop through the records of the dataframe.
187     for record_idx in range(rows):
188         # Get the source country name.
189         source_country_name = migration_df.iloc[record_idx,source_country_idx]
190         # Get the source country index within the current adjacency matrix.
191         source_idx = country_names_ids[source_country_name]
192         # Get the target country name.
193         target_country_name = migration_df.iloc[record_idx,target_country_idx]
194         # Get the target country index within the current adjacency matrix.
195         target_idx = country_names_ids[target_country_name]
196         # Set the migration volume for the current pair of (source,target)
197         # countries.
198         mat[source_idx,target_idx] = migration_df.iloc[record_idx,time]
199     # Get the maximum migration volume for the current time instance.
200     volume_max = mat.max()
201     # Set the percentage of the maximum migration volume above which two countries
202     # will be connected within the corresponding binary adjacency matrix.
203     volume_percentage = 0.001
204     # Append the weighted adjacency matrix for the current time instance.
205     adj_matrices.append(mat)
206     # Check and append the symmetry status.
207     symmetry.append(check_symmetry(mat))
208     # Generate a binary connectivity matrix.
209     bin_mat = mat.copy()
210     bin_mat[bin_mat<volume_percentage*volume_max] = 0
211     bin_mat[bin_mat>=volume_percentage*volume_max] = 1
212     # Append the binary adjacency matrix for the current time instance.
213     binary_matrices.append(bin_mat)
214     # Compute and append the number of edges in each network instance.
215     edges_num.append(bin_mat.sum())
216     # Create and append the graph object.
217     graphs.append(create_graph(bin_mat, t,node_labels))

```

Εικ.6.Δημιουργία μήτρα γειννίαςης

Παρατηρούμε στην γραμμή 203 τη σταθερά που ορίστηκε (threshold) και στη συνέχεια τις γραμμές 210,211 πραγματοποιείται η λογική που αναφέρθηκε παραπάνω. Μια ακμή υπάρχει στην περίπτωση που το πλήθος των ατόμων που μεταναστεύουν είναι πάνω από το μηδέν αλλά και πάνω από το γινόμενο της σταθεράς επί τη μέγιστη τιμή που υπάρχει στο συγκεκριμένο σύνολο δεδομένων.

Στη συνέχεια, έχοντας την μήτρα γειννίαςης δημιουργήθηκαν τα αντίστοιχα γραφήματα για κάθε χρονική στιγμή  $t$  (1990,1995...2020). Με τη βοήθεια των γραφημάτων δημιουργήθηκαν οι παρακάτω πίνακες διμερούς μετανάστευσης(bilateral migration matrix) για κάθε διαφορετική χρονολογία  $t$  ανά χώρα. Ένας πίνακας διμερούς μετανάστευσης είναι ένας τετραγωνικός πίνακας όπου το στοιχείο της  $i$ -οστής γραμμής και της  $j$ -οστής στήλης αντιπροσωπεύει τον αριθμό των ατόμων που μεταναστεύουν από τη χώρα  $i$  στη χώρα  $j$  εντός μιας δεδομένης χρονικής περιόδου  $t$ . Μαθηματικά, μπορεί να παρασταθεί ως εξής:

$$M(t) = \begin{bmatrix} M_{11}(t) & M_{12}(t) & \cdots & M_{1n}(t) \\ M_{21}(t) & M_{22}(t) & \cdots & M_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ M_{N1}(t) & M_{N2}(t) & \cdots & M_{NN}(t) \end{bmatrix}$$

Εικ.7. Πίνακας διμερούς μετανάστευσης

Για κάθε bilateral migration matrix δημιουργήθηκε ένα διαχρονικό γράφημα (temporal graph), που είναι γνωστά ως γραφήματα χρονοσειρών. Τα γραφήματα αυτά, είναι οπτικές αναπαραστάσεις δεδομένων κατά τη διάρκεια μιας χρονικής περιόδου  $t$ . Η δημιουργία των γραφημάτων αυτών έγινε ώστε να παρατηρηθούν τάσεις (αν τυχόν υπάρχουν), κάποιο μοτίβο στην μετανάστευση των ανθρώπων, αλλά και στην παρατήρηση των διακυμάνσεων που μπορεί να έχουν. Κάθε temporal graph αντιστοιχεί σε ένα συγκεκριμένο χρονικό διάστημα, που την περίπτωση αυτή αφορά τις τιμές 1990, 1995...2020.

## 4 Ανάλυση γραφημάτων

### 4.1 ΟΡΙΣΜΟΙ ΘΕΩΡΙΑΣ ΓΡΑΦΩΝ

Για κάθε γράφο που δημιουργήθηκε με βάση το σύνολο δεδομένων που αναφέρθηκε παραπάνω, υπολογίστηκαν τα μέτρα κεντρικότητας (Centrality measures) του γράφου για κάθε χρονική στιγμή  $t$ , καθώς επίσης και οι κοινότητες που υπάρχουν σε κάθε γράφημα (Network communities). Οι μετρικές κεντρικότητας διακρίνονται στις εξής κατηγορίες: κεντρικότητα βαθμού (degree centrality), κεντρικότητα εγγύτητας (closeness centrality), κεντρικότητα ενδιαμεσότητας (betweenness centrality) και ιδιοδιανυσματική κεντρικότητα (eigenvector centrality). Είναι ιδιαίτερα σημαντικές για τη βελτιστοποίηση της λειτουργίας του δικτύου, καθώς χάρη σε αυτές εντοπίζουμε ποιοι κόμβοι είναι σε περίοπτη θέση, ποιοι είναι αυτοί που μεσολαβούν στην επικοινωνία απομακρυσμένων κόμβων, ποιοι συμβάλλουν στην επεκτασιμότητα του δικτύου.

#### 4.1.1 Κοινότητες γραφημάτων

Οι κοινότητες σε ένα γράφημα αντιπροσωπεύουν ομάδες κόμβων που είναι πιο πυκνά συνδεδεμένοι μεταξύ τους απ' ό,τι με άλλους κόμβους στο γράφημα. Ο προσδιορισμός των κοινοτήτων εντός ενός δικτύου είναι ζωτικής σημασίας για την κατανόηση της δομής και της λειτουργίας του δικτύου και μπορεί να αποκαλύψει γνώσεις στις υποκείμενες διαδικασίες που διαμορφώνουν το δίκτυο. Βασικά χαρακτηριστικά περιλαμβάνουν την υψηλή εσωτερική πυκνότητα: Οι κόμβοι εντός μιας κοινότητας συνδέονται με σχετικά υψηλό αριθμό ακμών. Χαμηλή εξωτερική πυκνότητα όπου οι κόμβοι από διαφορετικές κοινότητες συνδέονται με λιγότερες ακμές από ό,τι οι κόμβοι εντός της ίδιας κοινότητας.

#### 4.1.2 Βαθμός Κεντρικότητας (Degree centrality)

Ο βαθμός κεντρικότητας είναι μια απλή αλλά θεμελιώδης έννοια στην ανάλυση δικτύων, η οποία χρησιμοποιείται για τον προσδιορισμό της σημασίας ή της επιρροής ενός κόμβου (μια χώρα, ενός ατόμου ή ενός νευρώνα) σε ένα δίκτυο. Βασίζεται στην υπόθεση ότι οι κόμβοι με περισσότερες συνδέσεις είναι ΠΙΟ ΚΕΝΤΡΙΚΟΙ.

$$C_D(i) = \frac{1}{N-1} \cdot \sum_{j=1}^n M_{ij}$$

Εικ.8. Ορισμός βαθμού κεντρικότητας

Όπου  $C_D(i)$  είναι ο βαθμός εκκεντρικότητας ενός κόμβου  $i$ ,  $M_{ij}$  είναι η τιμή που βρίσκεται στην γραμμή  $i$  και στήλη  $j$  της μήτρας γεινιάσης και  $N$  είναι ο συνολικός αριθμός των κόμβων στο γράφημα.

#### 4.1.3 Κεντρικότητα Εγγύτητας (Closeness Centrality)

Η κεντρικότητα της εγγύτητας είναι ένα μέτρο που χρησιμοποιείται στην ανάλυση δικτύων για τον εντοπισμό των κόμβων που μπορούν να διαδώσουν αποτελεσματικά τις πληροφορίες σε ένα δίκτυο. Μέσω της εγγύτητας είμαστε σε θέση να εκτιμήσουμε πόσο εύκολα ή δύσκολα μπορεί ένας κόμβος του δικτύου να προσεγγίσει άλλους κόμβους μέσα σε αυτό.

$$C_C(v) = (N - 1) \left[ \sum_{j \neq i} d_{SP}(i, j) \right]^{-1}$$

Εικ.9. Ορισμός κεντρικότητας εγγύτητας

Όπου  $C_C(v)$  είναι η κεντρικότητα εγγύτητας ενός κόμβου  $v$ ,  $N$  είναι ο συνολικός αριθμός των κόμβων στο δίκτυο,  $d_{SP}(i, j)$  είναι η πιο σύντομη διαδρομή μεταξύ δυο κόμβων  $i, j$  στο δίκτυο και τέλος, το  $\sum_{j \neq i} d_{SP}(i, j)$  είναι το σύνολο της ελάχιστης απόστασης ενός κόμβου  $i$  προς όλους τους υπόλοιπους κόμβους του δικτύου.

#### 4.1.4 Κεντρικότητα Ενδιαμεσότητας (Betweenness Centrality)

Η κεντρικότητα ενδιαμεσότητας είναι ένα μέτρο που βασίζεται στις συντομότερες διαδρομές. Για κάθε ζεύγος κόμβων σε έναν συνδεδεμένο γράφο, υπάρχουν ενδεχομένως πολλά συντομότερα μονοπάτια μεταξύ τους. Η κεντρικότητα μεταξύ των κόμβων ποσοτικοποιεί τον αριθμό των φορών που ένας κόμβος λειτουργεί ως γέφυρα κατά μήκος της συντομότερης διαδρομής μεταξύ δύο άλλων κόμβων. Επομένως, όσο πιο μεγάλη είναι η ενδιαμεσότητα ενός κόμβου τόσο πιο σημαντικός είναι ο κόμβος αυτός για την πλοήγηση στο δίκτυο. Αφαιρώντας, έναν κόμβο από το δίκτυο με υψηλή ενδιαμεσότητα, υπάρχει ο κίνδυνος να χαλάσει η συνεκτικότητα του δικτύου.

$$C_B(k) = \sum_{i \neq j \neq k} \frac{\sigma_{ij}(k)}{\sigma_{ij}}$$

Εικ.10. Ορισμός κεντρικότητας ενδιαμεσότητας

Όπου  $C_B(k)$  είναι η κεντρικότητα ενδιαμεσότητας,  $\sigma_{ij}$  είναι το πλήθος των σύντομων μονοπατιών μεταξύ δυο κόμβων  $i, j$  και  $\sigma_{ij}(k)$  είναι ο αριθμός αυτών των μονοπατιών που κατά τη διάρκεια της διαδρομής μεταξύ των κόμβων περνάνε από τον κόμβο  $k$ .

#### 4.1.5 Ιδιοδιανυσματική Κεντρικότητα (Eigenvector Centrality)

Η ιδιοδιανυσματική κεντρικότητα αποτελεί μέτρο της σημαντικότητας ενός κόμβου μέσα σε ένα δίκτυο, σύμφωνα με την αρχή ότι η σύνδεση με κόμβους μεγάλης βαθμικής κεντρικότητας προσδίδει περισσότερο κύρος στον υπό εξέταση κόμβο. Το μέτρο αυτό λαμβάνει υπόψη όχι μόνο τον αριθμό των άμεσων συνδέσεων που έχει ένας κόμβος, αλλά και την κεντρικότητα αυτών των συνδέσεων.

$$x_i = \frac{1}{\lambda} \sum_{j=1}^N M_{ij} x_j$$

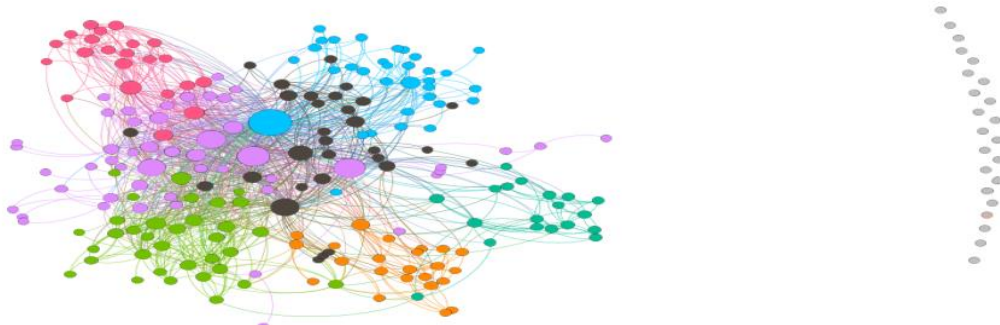
Εικ.11. Ορισμούς ιδιοδιανυσματικής κεντρικότητας

Όπου  $N$  είναι ο συνολικός αριθμός των κόμβων σε ένα δίκτυο,  $\lambda$  είναι η eigenvalue σταθερά.

## 4.2 ΓΡΑΦΗΜΑΤΑ

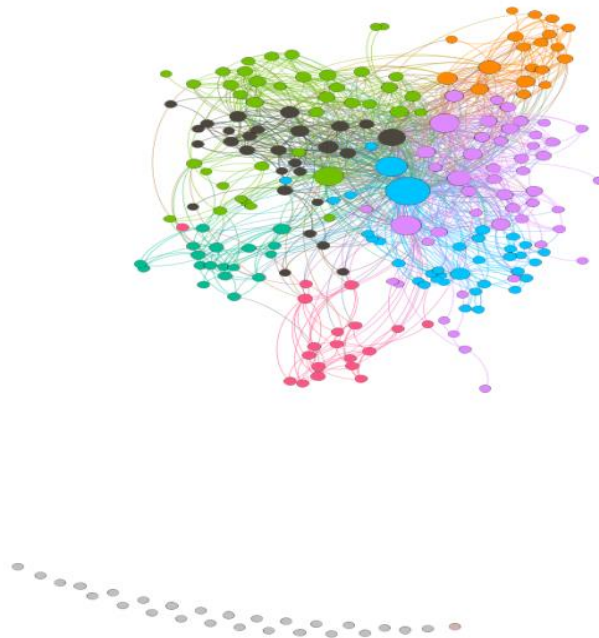
Τα δίκτυα μεταναστευτικών ροών αντιπροσωπεύουν μια συναρπαστική και πολύπλοκη πτυχή της παγκόσμιας δυναμικής, απεικονίζοντας τον τρόπο με τον οποίο οι άνθρωποι μετακινούνται από τη μια τοποθεσία στην άλλη με την πάροδο του χρόνου. Οι μετακινήσεις αυτές καθοδηγούνται από διάφορους παράγοντες, συμπεριλαμβανομένων των οικονομικών ευκαιριών, των πολιτικών καταστάσεων, των περιβαλλοντικών αλλαγών και των κοινωνικών παραγόντων. Η μελέτη των δικτύων μεταναστευτικών ροών μπορεί να προσφέρει ανεκτίμητες γνώσεις σχετικά με την ανθρώπινη συμπεριφορά, τις παγκόσμιες οικονομικές τάσεις και τον αντίκτυπο των διεθνών γεγονότων στις μετακινήσεις πληθυσμών. Για να απεικονίσει κανείς αυτά τα δίκτυα, μπορεί να φανταστεί περίπλοκους ιστούς γραμμών και κόμβων, όπου κάθε κόμβος αντιπροσωπεύει μια γεωγραφική τοποθεσία και κάθε γραμμή υποδηλώνει την κίνηση των ανθρώπων μεταξύ αυτών των τοποθεσιών. Αυτά τα δίκτυα μπορούν να αλλάξουν σημαντικά με την πάροδο του χρόνου, αντανακλώντας τη δυναμική φύση της ανθρώπινης μετανάστευσης. Για να καταδείξουμε αυτή την έννοια, παρακάτω υπάρχει μια σειρά εικόνων που απεικονίζουν το δίκτυο μεταναστευτικών ροών σε συγκεκριμένες χρονικές στιγμές - 1990, 1995, 2000, 2005, 2010, 2015 και 2020. Κάθε εικόνα θα αναπαριστούσε τα παγκόσμια μεταναστευτικά μοτίβα σε αυτές τις χρονικές στιγμές, δείχνοντας πώς ορισμένες διαδρομές κερδίζουν ή χάνουν την προβολή τους, πώς κρίσεις όπως πόλεμοι ή φυσικές καταστροφές επηρεάζουν τη μετανάστευση και πώς οι οικονομικές εξελίξεις σε διάφορες περιοχές προσελκύουν μετανάστες.

Δεδομένων για το 1990:



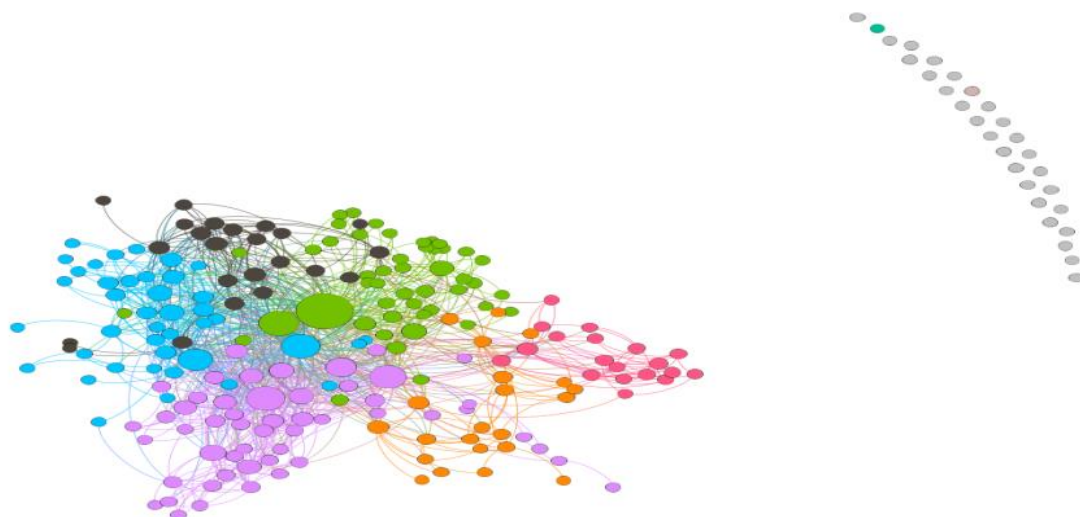
Εικ.12. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 1990

Δεδομένα για το 1995:



Εικ.13. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 1995

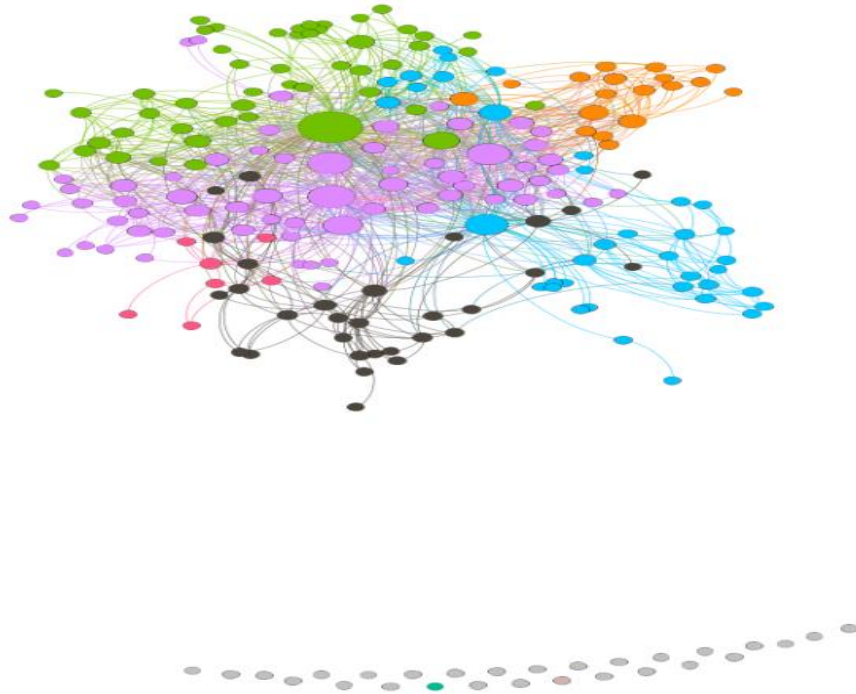
Δεδομένα για το 2000:



Εικ.14. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 2000

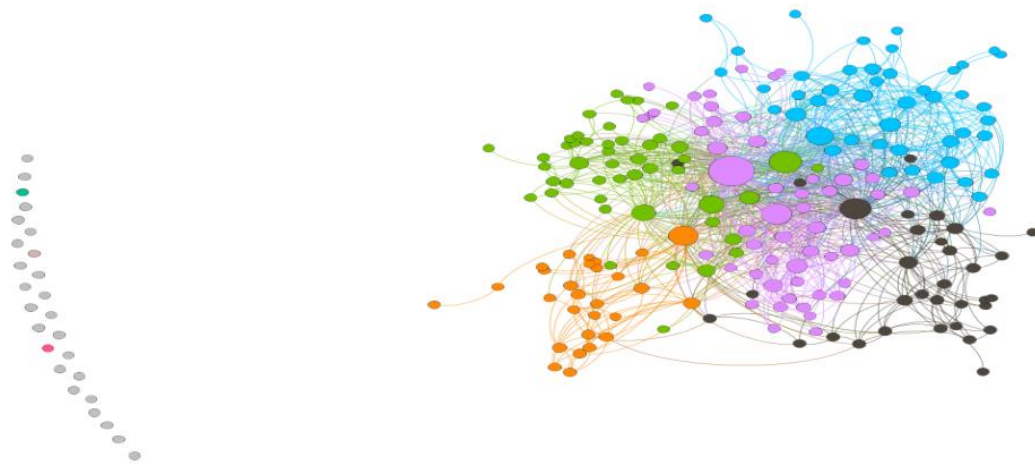


Δεδομένα για το 2005:



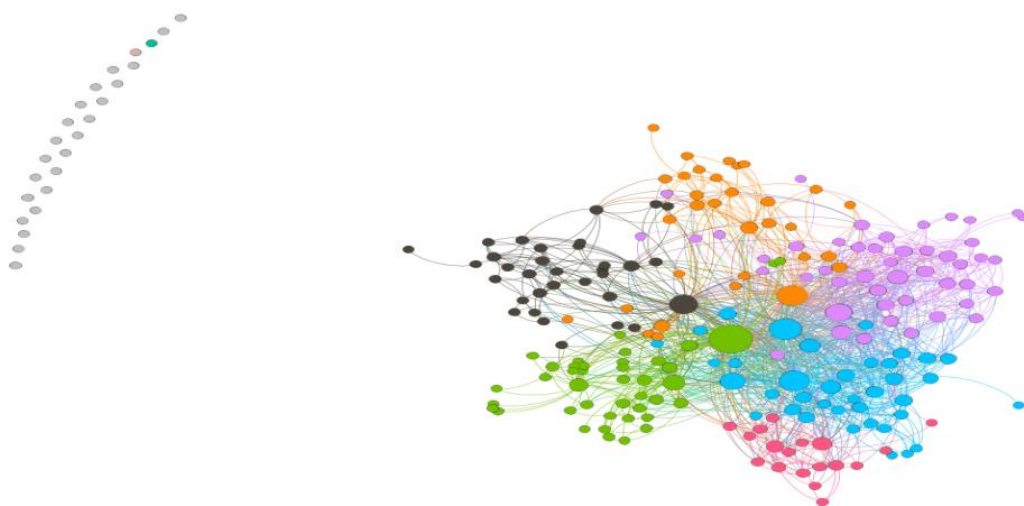
Εικ.15. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 2005

Δεδομένα για το 2010:



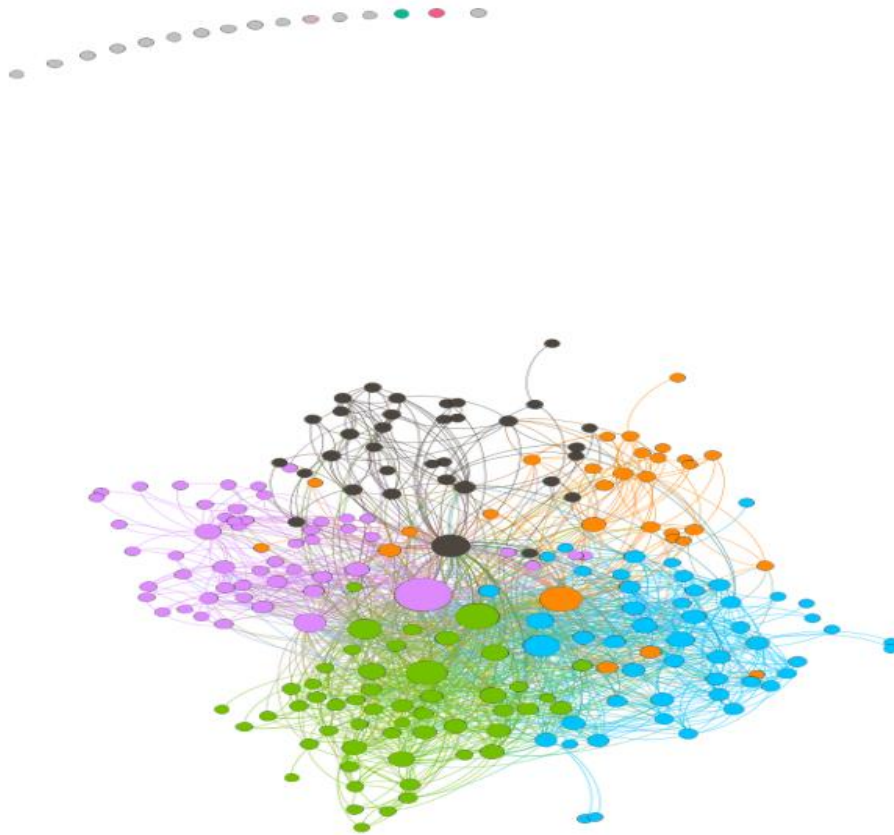
Εικ.16. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 2010

Δεδομένα για το 2015:



Εικ.17. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 2015

Δεδομένα για το 2020:



Εικ.18. Αποτύπωση γραφήματος μέσω του αλγόριθμου ForceAtlas: 2020

Παρακάτω, παρουσιάζονται οι 10 κορυφαίες χώρες με βάση τέσσερα διαφορετικά μέτρα κεντρικότητας (Degree, Eigenvector, Betweenness, Closeness Centralities) σε ένα δίκτυο μεταναστευτικών ροών για το έτος 1990. Όπως αποτυπώθηκε παραπάνω, τα μέτρα κεντρικότητας στην ανάλυση δικτύων είναι ζωτικής σημασίας για τον εντοπισμό των κόμβων με τη μεγαλύτερη επιρροή σε ένα δίκτυο, τα οποία, στο πλαίσιο της μετανάστευσης, βοηθούν στην κατανόηση των χωρών που κατέχουν σημαντικό ρόλο στην παγκόσμια μετακίνηση ανθρώπων.

#### 4.2.1 Δίκτυο μετανάστευσης 1990

Βαθμός κεντρικότητας (Degree Centrality) για τις πρώτες 10 χώρες.

	Country	<b>Degree</b>	Eigenvector	Closeness	Betweenness
1	United States	<b>129</b>	0.298623	0.643777	0.258754
2	Canada	<b>86</b>	0.252418	0.555045	0.068520
3	France	<b>84</b>	0.227926	0.558320	0.166484
4	Germany	<b>74</b>	0.239190	0.537700	0.050363
5	United Kingdom	<b>71</b>	0.215325	0.533156	0.070198
6	Australia	<b>67</b>	0.218259	0.517132	0.042216
7	Italy	<b>54</b>	0.192194	0.503379	0.020851
8	Russia	<b>42</b>	0.146157	0.462764	0.021291
9	Israel	<b>38</b>	0.140388	0.471996	0.011198
10	Switzerland	<b>37</b>	0.160142	0.471996	0.004391

Ιδιοδιανυσματική κεντρικότητα (Eigenvector Centrality) για τις πρώτες 10 χώρες.

	Country	Degree	<b>Eigenvector</b>	Closeness	Betweenness
1	United States	129	<b>0.298623</b>	0.643777	0.258754
2	Canada	86	<b>0.252418</b>	0.555045	0.068520
3	Germany	74	<b>0.239190</b>	0.537700	0.050363
4	France	84	<b>0.227926</b>	0.558320	0.166484
5	Australia	67	<b>0.218259</b>	0.517132	0.042216
6	United Kingdom	71	<b>0.215325</b>	0.533156	0.070198
7	Italy	54	<b>0.192194</b>	0.503379	0.020851
8	Switzerland	37	<b>0.160142</b>	0.471996	0.004391
9	Russia	42	<b>0.146157</b>	0.462764	0.021291
10	Sweden	33	<b>0.141807</b>	0.462764	0.004161

Κεντρικότητα εγγύτητας (Closeness Centrality) για τις πρώτες 10 χώρες.

	Country	Degree	Eigenvector	<b>Closeness</b>	Betweenness
1	United States	129	0.298623	<b>0.643777</b>	0.258754
2	France	84	0.227926	<b>0.558320</b>	0.166484
3	Canada	86	0.252418	<b>0.555045</b>	0.068520
4	Germany	74	0.239190	<b>0.537700</b>	0.050363
5	United Kingdom	71	0.215325	<b>0.533156</b>	0.070198
6	Australia	67	0.218259	<b>0.517132</b>	0.042216
7	Italy	54	0.192194	<b>0.503379</b>	0.020851
8	Switzerland	37	0.160142	<b>0.471996</b>	0.004391
9	Israel	38	0.140388	<b>0.471996</b>	0.011198
10	China	35	0.140581	<b>0.469654</b>	0.012044

Κεντρικότητα Ενδιαμεσότητας (Betweenness Centrality) για τις πρώτες 10 χώρες.

	Country	Degree	Eigenvector	Closeness	<b>Betweenness</b>
1	United States	129	0.298623	0.643777	<b>0.258754</b>
2	France	84	0.227926	0.558320	<b>0.166484</b>
3	United Kingdom	71	0.215325	0.533156	<b>0.070198</b>
4	Canada	86	0.252418	0.555045	<b>0.068520</b>
5	Germany	74	0.239190	0.537700	<b>0.050363</b>
6	Australia	67	0.218259	0.517132	<b>0.042216</b>
7	South Africa	27	0.096002	0.462764	<b>0.034551</b>
8	Caribbean	34	0.085045	0.449573	<b>0.024834</b>
9	Russia	42	0.146157	0.462764	<b>0.021291</b>

10 Italy

54

0.192194

0.503379

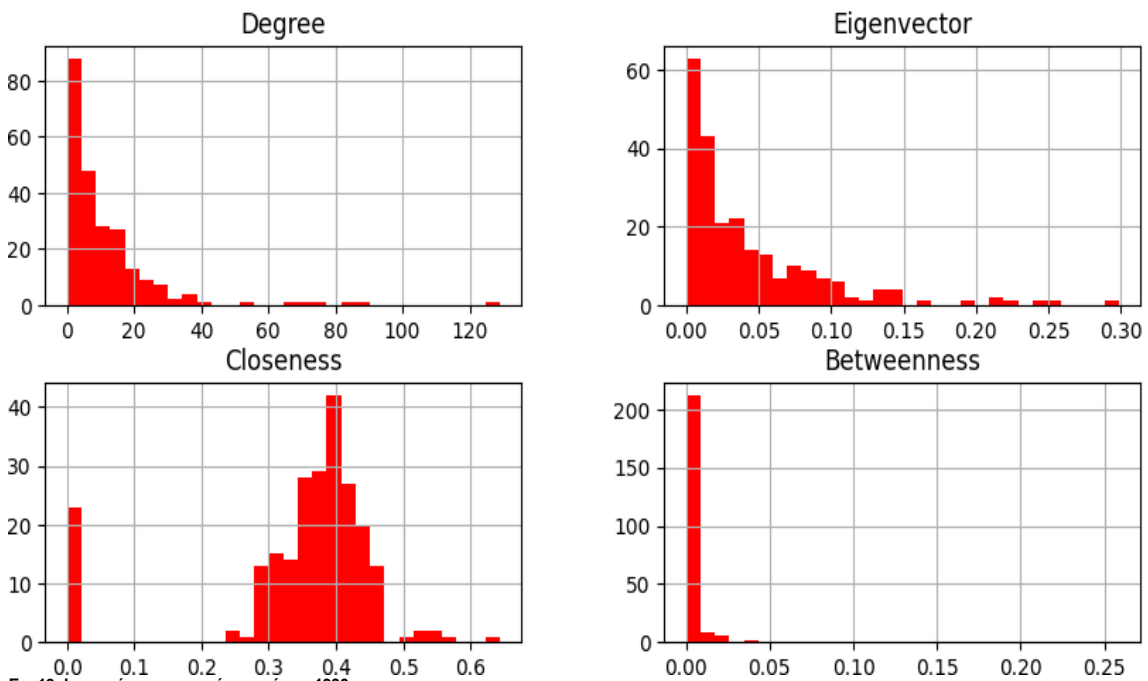
**0.020851**

Από τους παραπάνω πίνακες μπορούμε να πούμε πως, οι Ηνωμένες Πολιτείες της Αμερικής κατατάσσονται σταθερά στην κορυφή σε όλες τις μετρήσεις, γεγονός που υποδηλώνει το ρόλο της χώρας ως σημαντικού κόμβου μετανάστευσης με πολυάριθμες συνδέσεις, που βρίσκεται σε μια γειτονιά με μεγάλη επιρροή, είναι εύκολα προσβάσιμη και συχνά λειτουργεί ως γέφυρα στο δίκτυο. Ο Καναδάς, η Γαλλία και η Γερμανία κατέχουν επίσης εξέχουσα θέση σε όλους τους πίνακες, γεγονός που υποδηλώνει την ισχυρή και σταθερή θέση τους στο δίκτυο. Οι χώρες αυτές έχουν πιθανότατα ποικίλες μεταναστευτικές συνδέσεις, όχι μόνο με τους άμεσους γείτονες αλλά και με πιο απομακρυσμένα έθνη. Το Ηνωμένο Βασίλειο και η Αυστραλία, αν και δεν κατέχουν ηγετική θέση, εξακολουθούν να είναι σημαντικοί παράγοντες, με υψηλή κεντρικότητα εγγύτητας (Closeness centrality) που υποδηλώνει το ρόλο τους ως αποτελεσματικές χώρες διέλευσης στο πλαίσιο της κοινοπολιτείας και άλλων διεθνών μεταναστευτικών συστημάτων.

Η Ιταλία, η Ρωσία και το Ισραήλ παρουσιάζουν αξιοσημείωτη κεντρικότητα ενδιάμεσότητας (Betweenness Centrality), η οποία υποδηλώνει ότι παρά το γεγονός ότι έχουν λιγότερες συνδέσεις, διαδραματίζουν σημαντικό ρόλο στις μεταναστευτικές διαδρομές στις οποίες συμμετέχουν, λειτουργώντας ενδεχομένως ως πύλες εισόδου για τους μετανάστες. Η Ελβετία και η Σουηδία, με την υψηλή ιδιοδιανυσματική κεντρικότητά (Eigenvector Centrality), είναι πιθανό να συνδέονται με άλλες κεντρικές χώρες, γεγονός που είναι ενδεικτικό στρατηγικών μεταναστευτικών συνεργασιών ή συμφωνιών.

Στη συνέχεια, φαίνονται τα ιστογράμματα που εμφανίζουν την κατανομή των μέτρων κεντρικότητας στο δίκτυο μετανάστευσης του 1990. Κάθε ιστόγραμμα αντιπροσωπεύει ένα διαφορετικό μέτρο κεντρικότητας, παρέχοντας εικόνα της δομής και των χαρακτηριστικών του δικτύου.

Density Histograms for 1990



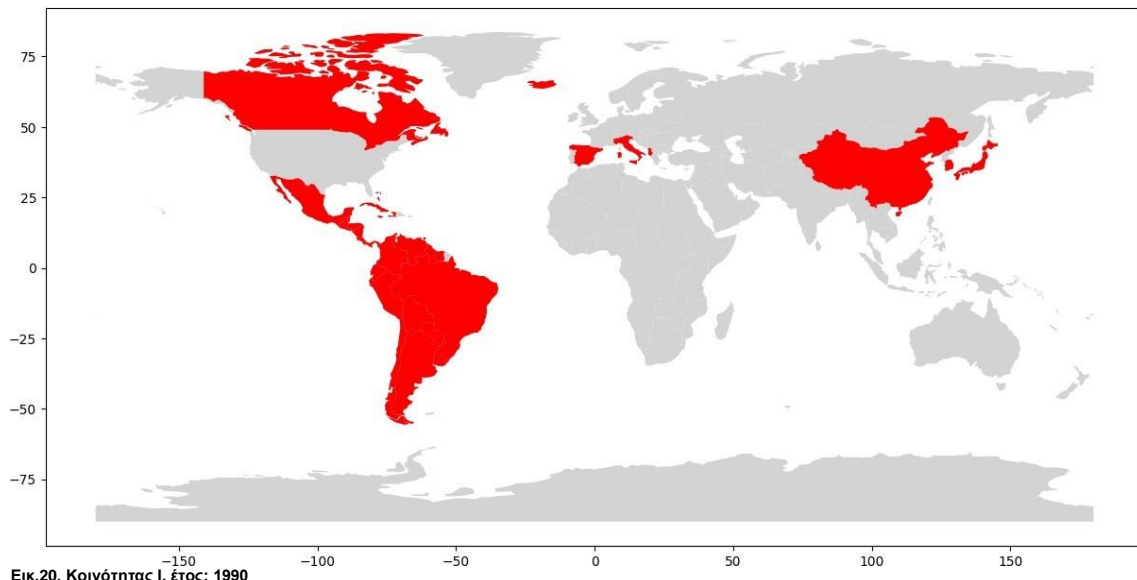
Εικ.19. Ιστογράμματα πυκνότητας έτος: 1990

Συνολικά, αυτά τα ιστογράμματα υποδηλώνουν ένα δίκτυο μετανάστευσης με μικρό αριθμό χωρών με μεγάλη κεντρική θέση και μεγάλο αριθμό χωρών με περιφερειακή σύνδεση. Αυτό είναι χαρακτηριστικό των δικτύων του πραγματικού κόσμου, τα οποία συχνά εμφανίζουν

χαρακτηριστικά χωρίς κλίμακα, όπου λίγοι κόμβοι (χώρες, στην προκειμένη περίπτωση) λειτουργούν ως σημαντικοί κόμβοι, ενώ η πλειονότητα έχει λιγότερες συνδέσεις.

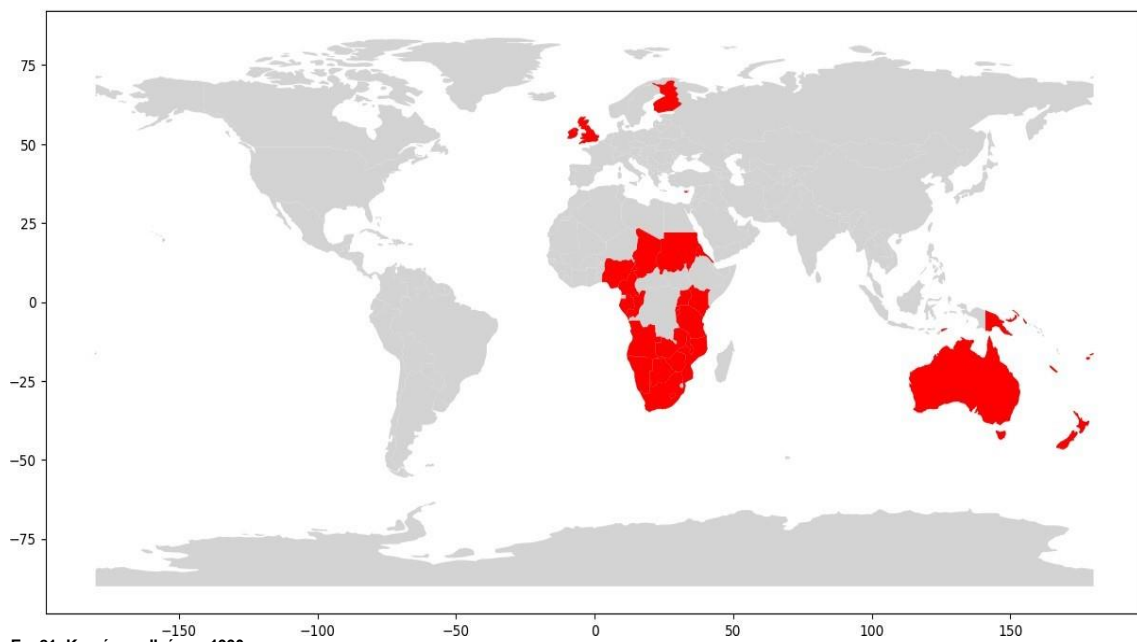
#### 4.2.2 Δίκτυο κοινοτήτων 1990

Migration Network 1990: Community 1 Containing 52 Countries



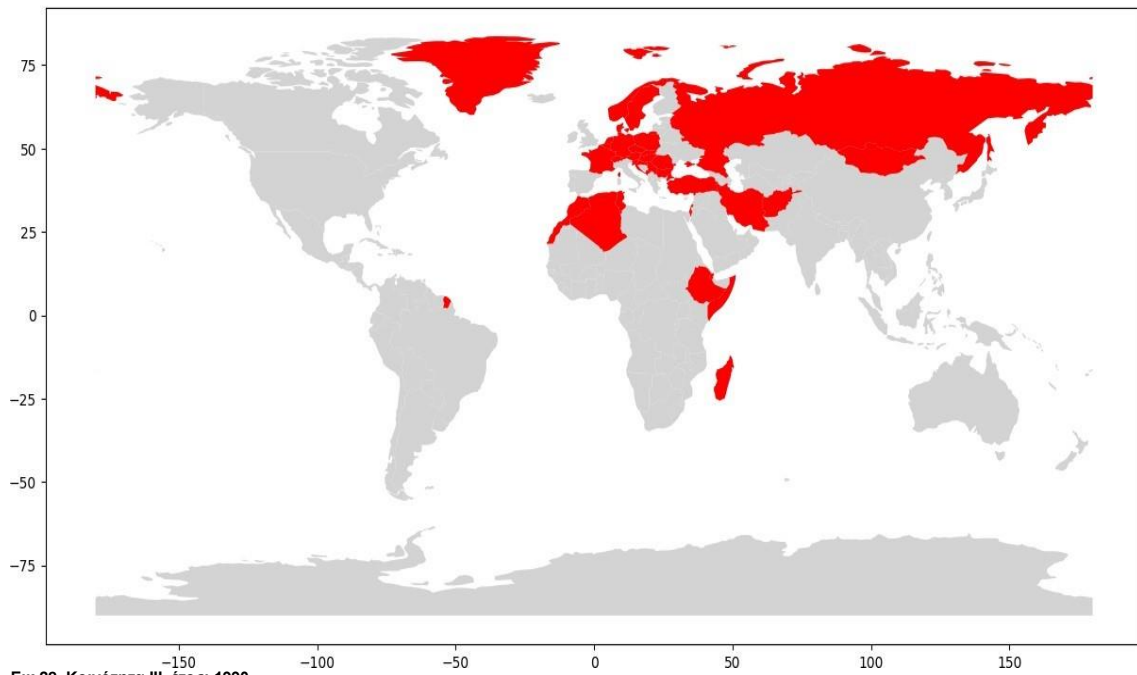
Εικ.20. Κοινότητα I, έτος: 1990

Migration Network 1990: Community 2 Containing 50 Countries



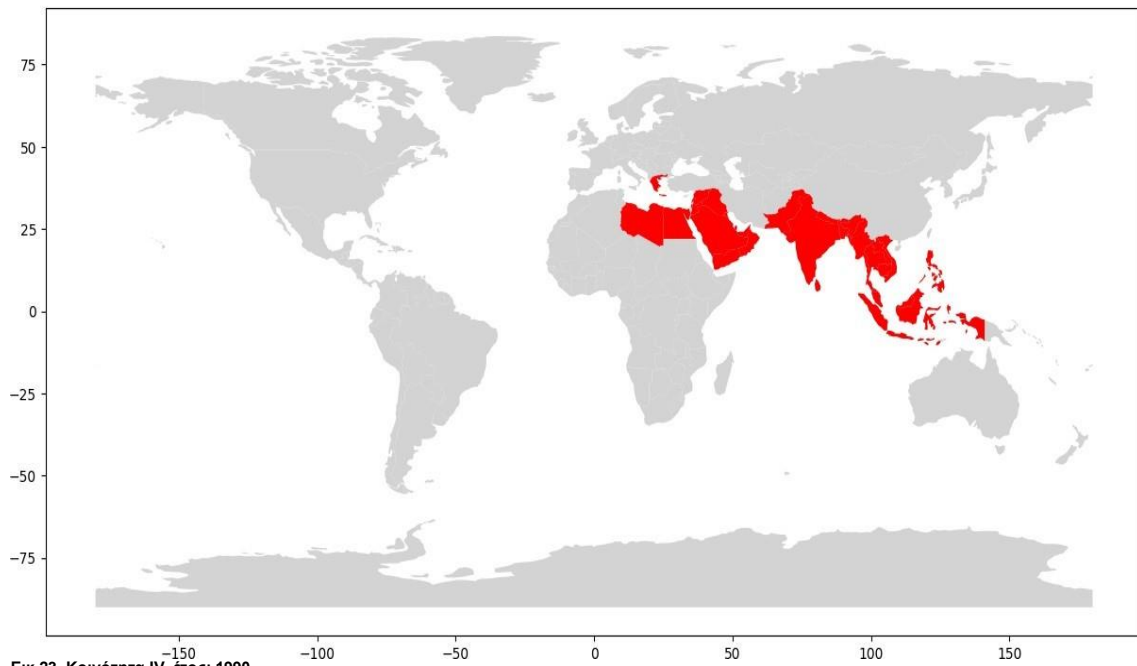
Εικ.21. Κοινότητα II, έτος: 1990

### Migration Network 1990: Community 3 Containing 43 Countries



Εικ.22. Κοινότητα III, έτος: 1990

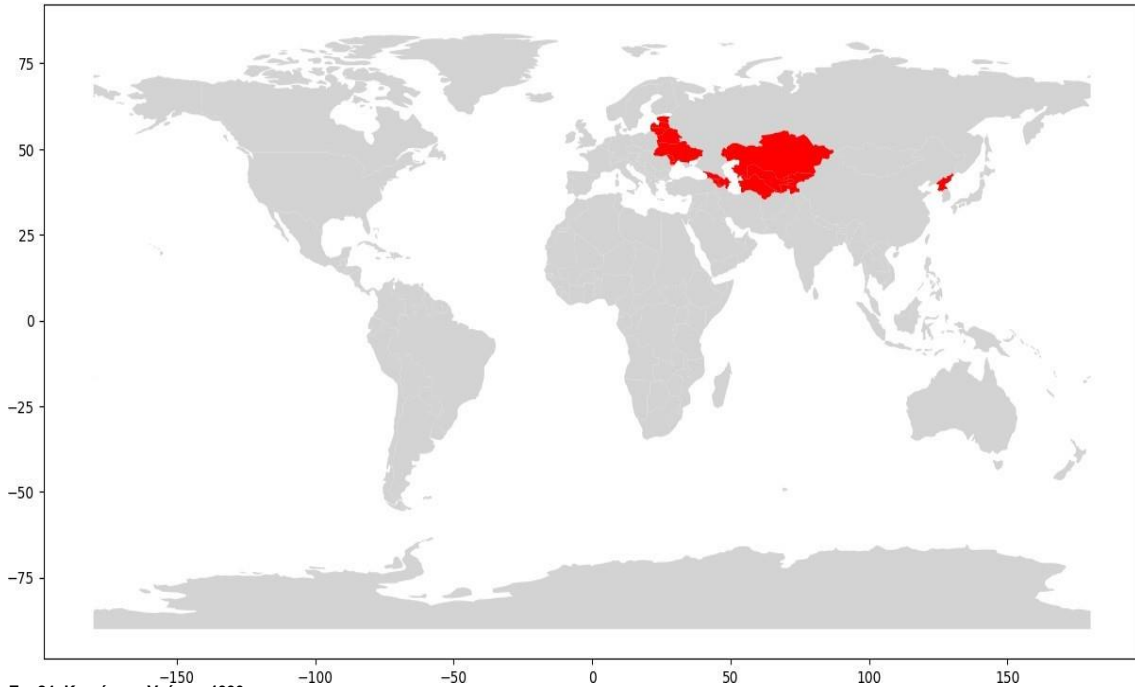
### Migration Network 1990: Community 4 Containing 34 Countries



Εικ.23. Κοινότητα IV, έτος: 1990



## Migration Network 1990: Community 5 Containing 15 Countries



Εικ.24. Κοινότητα V, Έτος: 1990

Σε όλες τις κοινότητες, υπάρχουν διακριτά γεωγραφικά μοτίβα που υποδηλώνουν ότι η περιφερειακή μετανάστευση αποτελεί σημαντικό παράγοντα. Ωστόσο, υπάρχουν επίσης διηπειρωτικές συνδέσεις που αντικατοπτρίζουν τον παγκόσμιο χαρακτήρα της μετανάστευσης κατά τη διάρκεια αυτής της περιόδου. Η ύπαρξη αυτών των κοινοτήτων εντός του δικτύου δείχνει ότι ενώ η μετανάστευση είναι ένα παγκόσμιο φαινόμενο, υπάρχουν ισχυρά περιφερειακά υπο δίκτυα που πιθανώς διαμορφώνονται από ένα συνδυασμό παραγόντων, όπως η εγγύτητα, οι ιστορικοί δεσμοί, οι οικονομικές συνθήκες και οι πολιτικές σχέσεις. Η σύνθεση αυτών των κοινοτήτων και η γεωγραφική τους εξάπλωση παρέχει ένα παράθυρο στο πώς οι παγκόσμιες και περιφερειακές πολιτικές, οι συμφωνίες και τα γεγονότα μπορούν να διαμορφώσουν τη ροή της μετανάστευσης. Για παράδειγμα, η διάλυση της Σοβιετικής Ένωσης γύρω στο 1990 επηρέασε πιθανότατα τα μεταναστευτικά πρότυπα στην Κοινότητα 3. Ομοίως, οικονομικές συμφωνίες μπορεί να επηρεάζουν τα μεταναστευτικά πρότυπα στην Αμερική, γεγονός που ενδεχομένως αντικατοπτρίζεται στην Κοινότητα 1. Κατά την ερμηνεία αυτών των κοινοτήτων, είναι επίσης σημαντικό να ληφθούν υπόψη οι περιορισμοί των δεδομένων και των μεθόδων που χρησιμοποιήθηκαν για τον ορισμό αυτών των κοινοτήτων. Η ανάλυση του δικτύου θα βασιστεί στα διαθέσιμα μεταναστευτικά δεδομένα, τα οποία ενδέχεται να μην καταγράφουν όλες τις μεταναστευτικές ροές, ιδίως τις μη καταγεγραμμένες. Συνολικά, αυτοί οι χάρτες κοινοτήτων είναι ένα πολύτιμο εργαλείο για την κατανόηση του πολύπλοκου ιστού των μεταναστευτικών ροών και μπορούν να βοηθήσουν στον εντοπισμό τόσο των δυνάμεων που οδηγούν τη μετανάστευση όσο και των πιθανών επιπτώσεων στις χώρες εντός κάθε κοινότητας.

### 4.2.3 Δίκτυο μετανάστευσης 2020

Βαθμός κεντρικότητας (Degree Centrality) για τις πρώτες 10 χώρες.

	Country	<b>Degree</b>	Eigenvector	Closeness	Betweenness
1	United States	<b>145</b>	0.283429	0.694497	0.258141
2	Canada	<b>99</b>	0.243889	0.597925	0.066783
3	United Kingdom	<b>96</b>	0.232851	0.592665	0.095241
4	Germany	<b>91</b>	0.233433	0.584100	0.061593
5	France	<b>82</b>	0.192732	0.558284	0.126090
6	Australia	<b>74</b>	0.206803	0.547693	0.030902
7	Italy	<b>67</b>	0.185766	0.536071	0.029013
8	Spain	<b>61</b>	0.166548	0.523572	0.033676
9	China	<b>45</b>	0.144751	0.495340	0.012793
10	Netherlands	<b>42</b>	0.148827	0.502733	0.007124

Ιδιοδιανυσματική κεντρικότητα (Eigenvector Centrality) για τις πρώτες 10 χώρες.

	Country	Degree	<b>Eigenvector</b>	Closeness	Betweenness
1	United States	145	<b>0.283429</b>	0.694497	0.258141
2	Canada	99	<b>0.243889</b>	0.597925	0.066783
3	Germany	91	<b>0.233433</b>	0.584100	0.061593
4	United Kingdom	96	<b>0.232851</b>	0.592665	0.095241
5	Australia	74	<b>0.206803</b>	0.547693	0.030902
6	France	82	<b>0.192732</b>	0.558284	0.126090
7	Italy	67	<b>0.185766</b>	0.536071	0.029013
8	Spain	61	<b>0.166548</b>	0.523572	0.033676
9	Netherlands	42	<b>0.148827</b>	0.502733	0.007124

10	China	45	<b>0.144751</b>	0.495340	0.012793
----	-------	----	-----------------	----------	----------

Κεντρικότητα εγγύτητας (Closeness Centrality) για τις πρώτες 10 χώρες.

	Country	Degree	Eigenvector	<b>Closeness</b>	Betweenness
1	United States	145	0.283429	<b>0.694497</b>	0.258141
2	Canada	99	0.243889	<b>0.597925</b>	0.066783
3	United Kingdom	96	0.232851	<b>0.592665</b>	0.095241
4	Germany	91	0.233433	<b>0.584100</b>	0.061593
5	France	82	0.192732	<b>0.558284</b>	0.126090
6	Australia	74	0.206803	<b>0.547693</b>	0.030902
7	Italy	67	0.185766	<b>0.536071</b>	0.029013
8	Spain	61	0.166548	<b>0.523572</b>	0.033676
9	Netherlands	42	0.148827	<b>0.502733</b>	0.007124
10	China	45	0.144751	<b>0.495340</b>	0.012793

Κεντρικότητα Ενδιαμεσότητας (Betweenness Centrality) για τις πρώτες 10 χώρες.

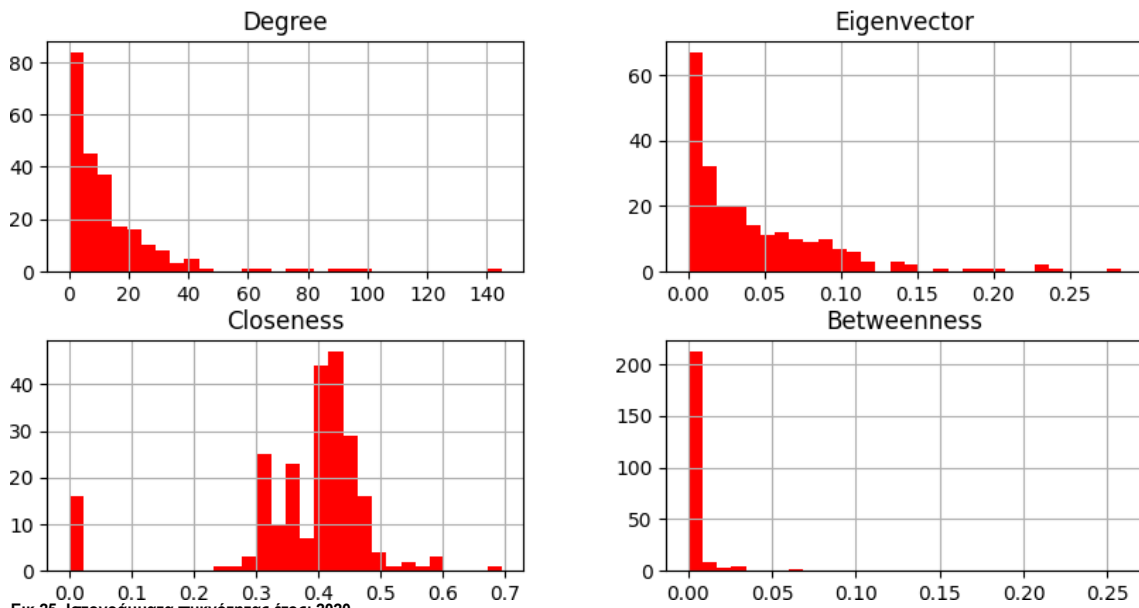
	Country	Degree	Eigenvector	Closeness	<b>Betweenness</b>
1	United States	145	0.283429	0.694497	<b>0.258141</b>
2	France	82	0.192732	0.558284	<b>0.126090</b>
3	United Kingdom	96	0.232851	0.592665	<b>0.095241</b>
4	Canada	99	0.243889	0.597925	<b>0.066783</b>
5	Germany	91	0.233433	0.584100	<b>0.061593</b>
6	Caribbean	39	0.079949	0.472193	<b>0.048362</b>
7	South Africa	32	0.088666	0.467821	<b>0.033954</b>
8	Spain	61	0.166548	0.523572	<b>0.033676</b>

9	Australia	74	0.206803	0.547693	<b>0.030902</b>
10	Italy	67	0.185766	0.536071	<b>0.029013</b>

Το 1990, οι Ηνωμένες Πολιτείες κατείχαν ήδη κυρίαρχη θέση σε όλες τις μετρήσεις κεντρικής θέσης, τάση που θα συνεχιστεί και το 2020. Η παρουσία χωρών όπως η Κίνα και η Ισπανία στην πρώτη δεκάδα για ορισμένες μετρήσεις το 2020, αλλά όχι το 1990, αντανακλά τη μεταβαλλόμενη δυναμική της παγκόσμιας μετανάστευσης, η οποία πιθανότατα επηρεάζεται από την οικονομική ανάπτυξη και τις γεωπολιτικές μετατοπίσεις. Τα μέτρα κεντρικότητας σε όλους τους τομείς είναι υψηλότερα το 2020 σε σύγκριση με το 1990, υποδεικνύοντας ότι το συνολικό δίκτυο μετανάστευσης έχει γίνει πιο διασυνδεδεμένο και πολύπλοκο. Η αύξηση της ενδιάμεσότητας (Betweenness Centrality) της Γαλλίας μέχρι το 2020 είναι αξιοσημείωτη και υποδηλώνει ότι η στρατηγική θέση της χώρας στις μεταναστευτικές διαδρομές έχει γίνει πιο έντονη κατά τη διάρκεια των 30 ετών. Η αύξηση των βαθμών κεντρικότητας (Degree Centrality) από το 1990 έως το 2020 υποδηλώνει ότι οι χώρες διασυνδέονται όλο και περισσότερο, με αύξηση των μεταναστευτικών διαδρομών. Αυτό θα μπορούσε να είναι αποτέλεσμα της παγκοσμιοποίησης, της ευκολίας των διεθνών ταξιδιών και της αυξημένης μετακίνησης ανθρώπων για πολλούς λόγους όπως η εργασία και σπουδές. Οι υψηλές τιμές κεντρικής εγγύτητας (Closeness Centrality) το 2020 υποδηλώνουν ότι πολλές χώρες βρίσκονται σε απόσταση αναπνοής μεταξύ τους όσον αφορά τη μετανάστευση, ενδεχομένως λόγω της βελτίωσης των μεταφορών και των πιο φιλελεύθερων μεταναστευτικών πολιτικών σε ορισμένες περιοχές. Αυτό έρχεται σε αντίθεση με τα δεδομένα του 1990, όπου οι βαθμολογίες ήταν γενικά χαμηλότερες, αντανακλώντας ένα λιγότερο συνδεδεμένο παγκόσμιο μεταναστευτικό δίκτυο εκείνη την εποχή.

Συνοπτικά, η σύγκριση μεταξύ των δεδομένων του 1990 και του 2020 δείχνει μια σημαντική εξέλιξη του παγκόσμιου μεταναστευτικού δικτύου, με αυξημένο αριθμό συνδέσεων και μετατόπιση του ρόλου που διαδραματίζουν συγκεκριμένες χώρες στο δίκτυο αυτό. Η συνολική τάση δείχνει μια παγκοσμιοποίηση της μετανάστευσης με περισσότερες χώρες να συνδέονται μεταξύ τους και ορισμένα έθνη να αναδεικνύονται σε κόμβους με επιρροή λόγω της στρατηγικής τους θέσης ή των πολιτικών τους που προσελκύουν ή διαμετακομίζουν μετανάστες. Τα ιστογράμματα που εμφανίζουν την κατανομή των μέτρων κεντρικότητας στο δίκτυο μετανάστευσης του 2020:

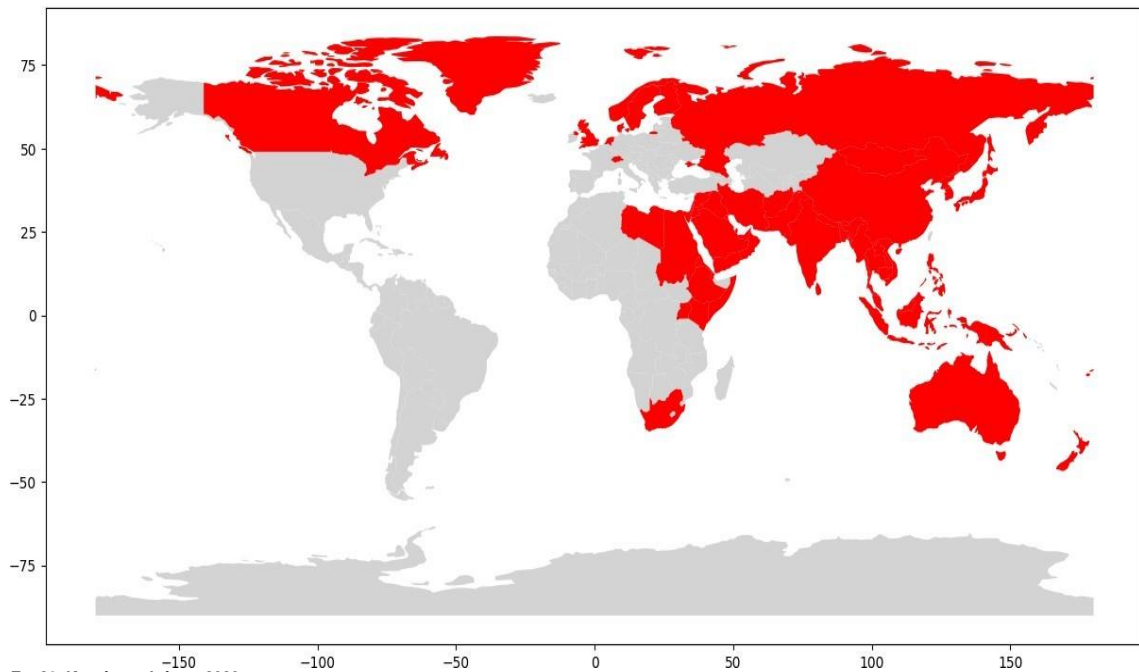
Density Histograms for 2020



Εικ.25. Ιστογράμματα πυκνότητας έτος: 2020

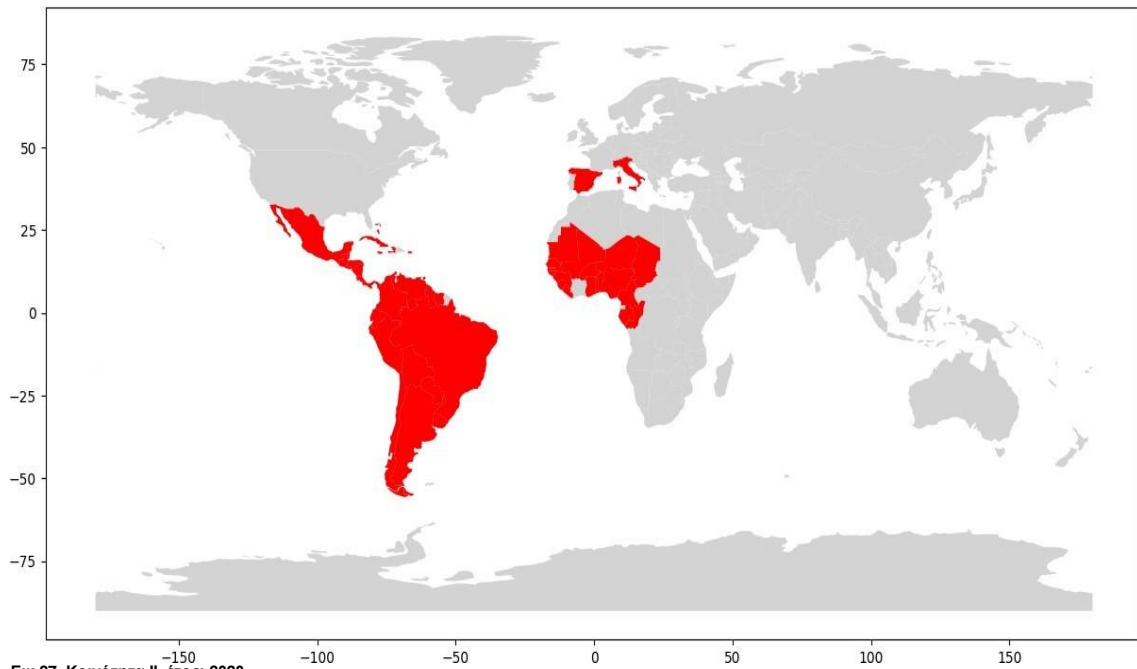
#### 4.2.4 Δίκτυο κοινοτήτων 2020

Migration Network 2020: Community 1 Containing 73 Countries



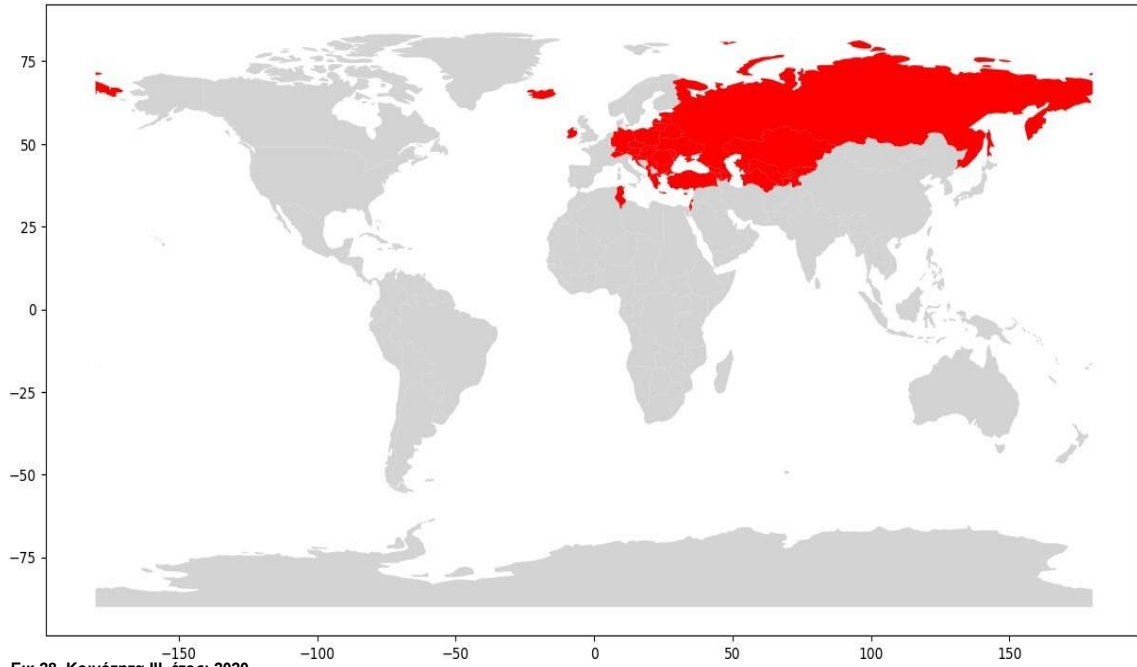
Εικ.26. Κοινότητα I, έτος: 2020

### Migration Network 2020: Community 2 Containing 70 Countries



Εικ.27. Κοινότητα II, έτος: 2020

### Migration Network 2020: Community 3 Containing 39 Countries



Εικ.28. Κοινότητα III, έτος: 2020

## 5 Πειράματα

### 5.1 ΥΛΟΠΟΙΗΣΗ ΜΟΝΤΕΛΟΥ ΠΡΟΒΛΕΨΗΣ ΣΥΝΔΕΣΕΩΝ

Στον τομέα της επιστήμης των δικτύων, η πρόβλεψη μελλοντικών συνδέσεων εντός ενός δικτύου γνωστή και ως πρόβλεψη συνδέσεων (Link Prediction) είναι ένα θεμελιώδες έργο που έχει ευρείες εφαρμογές που κυμαίνονται από την ανάλυση κοινωνικών δικτύων έως τη βιοπληροφορική. Στη συγκεκριμένη εργασία θα εφαρμόσουμε αυτή την στο μεταναστευτικό δίκτυο. Η βασική πρόκληση στην πρόβλεψη συνδέσεων έγκειται στην εκτίμηση της πιθανότητας να δημιουργηθεί στο μέλλον μια σύνδεση μεταξύ δύο κόμβων, με βάση τα παρατηρούμενα δεδομένα και τις δομικές ιδιότητες του δικτύου.

Στο παρόν κεφάλαιο θα γίνει η κατασκευή ενός μοντέλου πρόβλεψης χρονικών συνδέσεων. Το μοντέλο βασίζεται στις αρχές των συνελκτικών δικτύων γράφων (GCN), τα οποία αξιοποιούν την πλούσια σχεσιακή πληροφορία στη δομή του γράφου. Χρησιμοποιώντας μια αρχιτεκτονική GCN δύο επιπέδων, το μοντέλο αποτυπώνει τις πολύπλοκες εξαρτήσεις και υπάρχουν στην εξελισσόμενη τοπολογία του γράφου.

Για την δημιουργία του μοντέλου χρησιμοποιήθηκε η αρχιτεκτονική των συνελκτικών δικτύων με την χρήση δυο Graph Convolution Layers (GCN). Η αρχιτεκτονική αυτή είναι μια καλή λύση στο πρόβλημα της πρόβλεψης μελλοντικών συνδέσεων μεταξύ κόμβων του δικτύου. Το πρώτο στρώμα (conv1) μετασχηματίζει τα χαρακτηριστικά εισόδου των κόμβων σε μια κρυφή αναπαράσταση. Η διαστάσεις των χαρακτηριστικών εισόδου δίνεται από το `input_dim` και το μέγεθος της κρυφής αναπαράστασης ορίζεται από το `hidden_dim`. Το δεύτερο στρώμα (conv2) λαμβάνει την κρυφή αναπαράσταση από το πρώτο στρώμα και την προβάλλει πίσω στις διαστάσεις όπως ήταν κατά την είσοδο στο μοντέλο. Το μοντέλο αυτό θα εκπαιδευτεί σε μια ακολουθία από γραφήματα  $T$  προσπαθώντας να προβλέψει τη μήτρα γειτνίασης του επόμενου γράφου. Η συνάρτηση ενεργοποίησης (activation function) ReLU (Rectified Linear Unit) ακολουθεί το πρώτο στρώμα GCN, εισάγοντας μη γραμμικότητα στο μοντέλο. Αυτή η μη γραμμικότητα είναι ζωτικής σημασίας για το μοντέλο ώστε να καταγράψει πολύπλοκα μοτίβα και αλληλεπιδράσεις μεταξύ των κόμβων, τα οποία δεν είναι δυνατόν να μοντελοποιηθούν με απλά γραμμικούς μετασχηματισμούς. Η τελική συνάρτηση ενεργοποίησης είναι μια σιγμοειδής συνάρτηση, η οποία εφαρμόζεται μετά το δεύτερο στρώμα GCN. Η σιγμοειδής συνάρτηση συμπιέζει την έξοδο μεταξύ 0 και 1, η οποία είναι ιδανική για μια εργασία πρόβλεψης συνδέσεων, όπου η έξοδος μπορεί να ερμηνευτεί ως η πιθανότητα ύπαρξης μιας ακμής μεταξύ δύο κόμβων.

Η ακριβής δομή εμφανίζεται στην εικόνα 29 παρακάτω ενώ διάγραμμα για την ακριβής περιγραφή των εισόδων εξόδων του κάθε layer εμφανίζεται στην εικόνα 30.

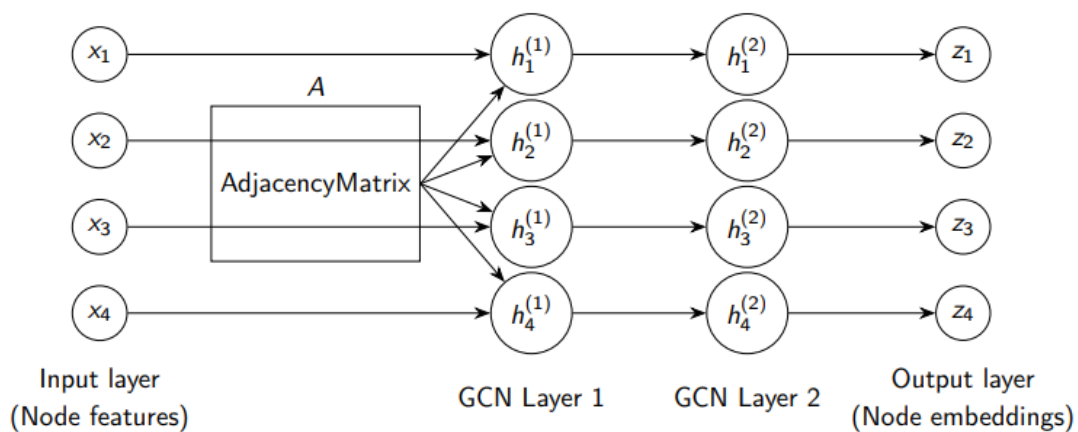
```

class TemporalLinkPredictor(nn.Module):
    def __init__(self, input_dim, hidden_dim):
        super(TemporalLinkPredictor, self).__init__()
        self.conv1 = GCNConv(input_dim, hidden_dim)
        self.conv2 = GCNConv(hidden_dim, input_dim)

    def forward(self, data):
        x, edge_index = data.x, data.edge_index
        x = self.conv1(x, edge_index)
        x = torch.relu(x)
        x = self.conv2(x, edge_index)
        x = torch.sigmoid(x)
        return x

```

Εικ.29. Δημιουργία μοντέλου πρόβλεψης μελλοντικών συνδέσεων



Εικ.30. Διάγραμμα μοντέλου

Τα χαρακτηριστικά του κάθε κόμβου στο παραπάνω πείραμα είναι μία ταυτοτική μήτρα όπου έχει παντού 0, εκτός από την κύρια διαγώνιο που έχει 1. Παράλληλα τροφοδοτείται η μήτρα γειτνίασης στο μοντέλο. Για την εκπαίδευση του μοντέλου φάνηκε να αποδίδει καλύτερα ο adam optimizer με ένα μικρό learning rate ίσο με 0.001 ενώ για Loss Function χρησιμοποιήθηκε η Binary Cross Entropy (BCELoss). Στην παρακάτω εικόνα φαίνεται η εκπαίδευση του μοντέλου στα γραφήματα για τις χρονικές περιόδους  $T-1$ , ώστε να μπορέσει να γίνει στη συνέχεια η αξιολόγηση της πρόβλεψης του μοντέλου στην τελική περίοδο  $T$ .



```
# Initialize list container for storing the cumulative loss during training.
training_loss = []

for epoch in range(EPOCHS):
    # Zero out all the gradient vectors.
    optimizer.zero_grad()
    # Initialize the cumulative loss container.
    cumulative_loss = 0.0
    # Aggregate the total loss on the first T-1 time instances.
    for i in range(T-1):
        # Get the output of the model for the current adjacency matrix.
        output = model(train_data_list[i])
        # Compute the loss between the current output of the model and the next
        # adjacency matrix.
        loss = criterion(output, train_data_list[i+1].x)
        # Accumulate the previously computed loss value.
        # If it's the first iteration, set cumulative_loss to loss
        if i == 0:
            cumulative_loss = loss
        # Add the new loss to cumulative_loss tensor.
        else:
            cumulative_loss += loss
    # Perform the backward pass of the information for training purposes.
    cumulative_loss.backward()
    # Perform the current optimization step.
    optimizer.step()
    # Print information concerning the current optimization step.
    print(f'Performing training epoch: {epoch+1} Loss:{cumulative_loss}')
    training_loss.append(cumulative_loss.item())
```

Εικ.31. Εκπαίδευση μοντέλου

## 5.2 ΥΛΟΠΟΙΗΣΗ ΜΟΝΤΕΛΟΥ ΠΡΟΒΛΕΨΗΣ ΜΕ ΧΡΗΣΗ NODE2VEC

Η υλοποίηση του δεύτερου μοντέλου είναι ίδια με την προηγούμενη με τη μόνη διαφορά ότι πλέον στην είσοδο του μοντέλου δεν δίνουμε τον ταυτοτηκό πίνακα, αλλά τα feature embeddings που δημιουργήθηκαν από τον αλγόριθμο Node2Vec. Στην εικόνα 32 φαίνεται η δομή του μοντέλου, ενώ στην εικόνα 33 το μοντέλο του αλγορίθμου Node2vec.

```
class TemporalGNN(nn.Module):
    def __init__(self, input_dim, hidden_dim):
        super(TemporalGNN, self).__init__()
        self.conv1 = GCNConv(input_dim, hidden_dim)
        self.conv2 = GCNConv(hidden_dim, input_dim)

    def forward(self, x, edge_index):
        x = self.conv1(x, edge_index)
        x = F.relu(x)
        x = self.conv2(x, edge_index)
        x = torch.sigmoid(x)
        return x
```

Εικ.32.Μοντέλο

πρόβλεψης

μελλοντικών

Όπως και το προηγούμενο μοντέλο, αποτελείται από δύο Graph Convolution Layers (GCN) και για την εκπαίδευση του μοντέλου χρησιμοποιήθηκε ο adam optimizer με ένα learning rate ίσο με 0.001 ενώ για Loss Function χρησιμοποιήθηκε η Binary Cross Entropy (BCELoss).

```

# Initialize the Node2Vec model with parameters.
# walk_length: Length of each random walk.
# context_size: Size of the learning context.
# walks_per_node: Number of random walks to start at each node.
# num_negative_samples: Number of negative samples to use.
# p and q: Parameters controlling the random walk behavior.
model = Node2Vec(graph_data.edge_index, embedding_dim=graph_data.x.shape[1], walk_length=20,
                context_size=10, walks_per_node=walks_per_node, num_negative_samples=1, p=p, q=q, sparse=True)

# Create a DataLoader that batches and shuffles the random walks.
loader = model.loader(batch_size=128, shuffle=True)
# Initialize an optimizer for the model parameters.
optimizer = torch.optim.SparseAdam(model.parameters(), lr=0.01)
# Set the model in training mode.
model.train()
# Training loop: Iterate over batches of random walks.
for pos_rw, neg_rw in loader:
    # Reset gradients to zero before starting backpropagation.
    optimizer.zero_grad()
    # Compute the loss based on positive and negative random walks.
    loss = model.loss(pos_rw, neg_rw)
    # Perform backpropagation to compute gradients.
    loss.backward()
    # Update model parameters based on computed gradients.
    optimizer.step()

# Return the learned node embeddings.
return model.embedding.weight.data

```

Εικ.33.Μοντέλο αλγορίθμου Node2Vec

Στην παραπάνω εικόνα παρουσιάζεται το μοντέλο για την παραγωγή των node embeddings. Για τη δημιουργία του μοντέλου χρειάζεται να ορίσουμε τις συνδέσεις των κόμβων, τις διαστάσεις που θέλουμε να έχει ο τελικός πίνακας του μοντέλου που ορίζεται από την μεταβλητή *embedding\_dim*. Οι μεταβλητές *walk\_length*, *context\_size*, *walks\_per\_node* ορίζουν πως θα τρέξει γενικά ο αλγόριθμος, ρυθμίζοντας το μέγεθος της κάθε βόλτας, το μέγεθος πλαισίου (τους γείτονες που υπάρχουν γύρω από τον συγκεκριμένο κόμβο) καθώς επίσης και πόσες βόλτες θα ξεκινήσουν από τον συγκεκριμένο κόμβο. Η μεταβλητή *num\_negative\_samples* είναι η παράμετρος που καθορίζει τον αριθμό των αρνητικών ζευγών node, context που πρέπει να δειγματοληπτηθούν για κάθε θετικό ζεύγος κατά τη διάρκεια της εκπαίδευσης. Για παράδειγμα, εάν η τιμή *num\_negative\_samples* οριστεί σε 5, τότε για κάθε θετικό ζεύγος (node, context), ο αλγόριθμος θα επιλέξει τυχαία 5 άλλους κόμβους (από τον γράφο) που δεν βρίσκονται στο πλαίσιο του κόμβου-στόχου για να χρησιμεύσουν ως αρνητικά παραδείγματα. Οι παράμετροι *p* και *q* όπως αναφέρθηκε και στον δεύτερο κεφάλαιο ελέγχουν την πιθανότητα άμεσης επανάληψης ενός κόμβου στον περίπατο και την πιθανότητα εξερεύνησης απομακρυσμένων τμημάτων του γραφήματος, αντίστοιχα. Στο μοντέλο δημιουργήθηκε ένας *DataLoader* ο οποίος χρησιμοποιείται για την αποτελεσματική φόρτωση των δεδομένων του γραφήματος για την εκπαίδευση. Συγκεντρώνει τις τυχαίες βόλτες και τις ανακατεύει για κάθε επανάληψη, χρησιμοποιώντας ένα μέγεθος δέσμης 128. Γίνεται η εκπαίδευση του μοντέλου και στο τέλος επιστρέφει τα χαρακτηριστικά των κόμβων (node embeddings) υπό μια μορφή συμπίεσης σε ένα χώρο χαμηλότερης διάστασης, διατηρώντας τις δομικές ιδιότητες του γράφου.

Στο παραπάνω μοντέλο, πραγματοποιήθηκε μια επαναληπτική διαδικασία για την εύρεση ορισμένων παραμέτρων όπως το *context\_size*, *walks\_per\_node*, *num\_negative\_samples*, *p* και *q* που θα βοηθούσαν το GCN δίκτυο να ελαχιστοποιήσει τις απώλειες του και να επιτύχει καλύτερες προβλέψεις σε νέα δεδομένα. Βρέθηκε πως το κύριο μοντέλο της εργασίας προβλέπει πιο αξιόπιστα αποτελέσματα όταν ο αλγόριθμος Node2Vec έχει παράξει χαρακτηριστικά για τον κάθε κόμβο με τις παρακάτω παραμέτρους:

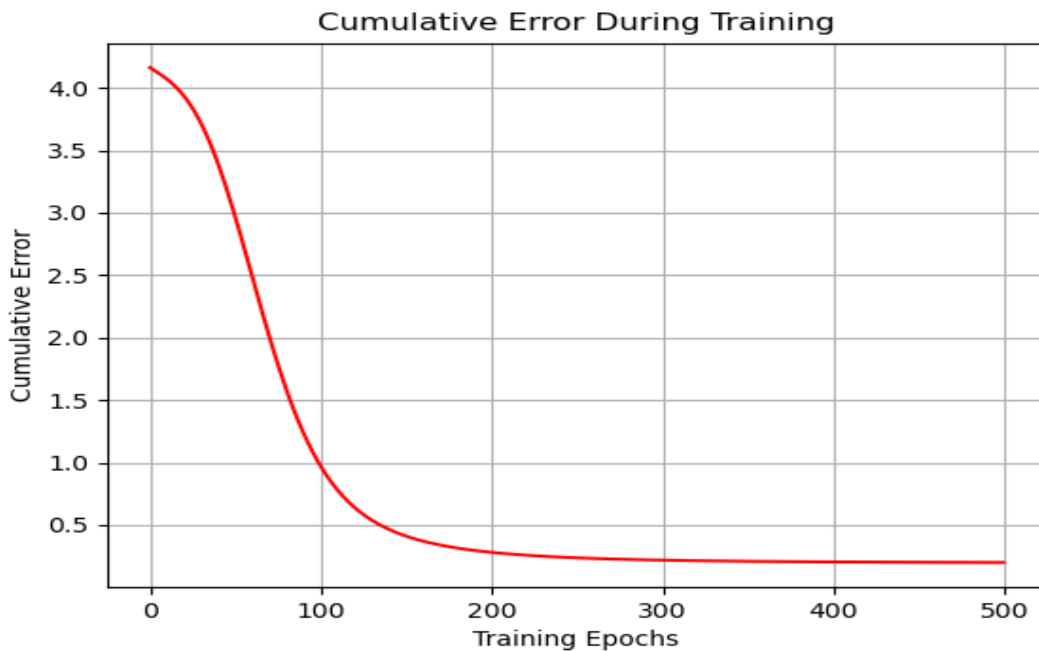
- *context\_size* = 10
- *walks\_per\_node* = 9

- `num_negative_samples = 4`
- `P = 0.5`
- `Q = 1.5`

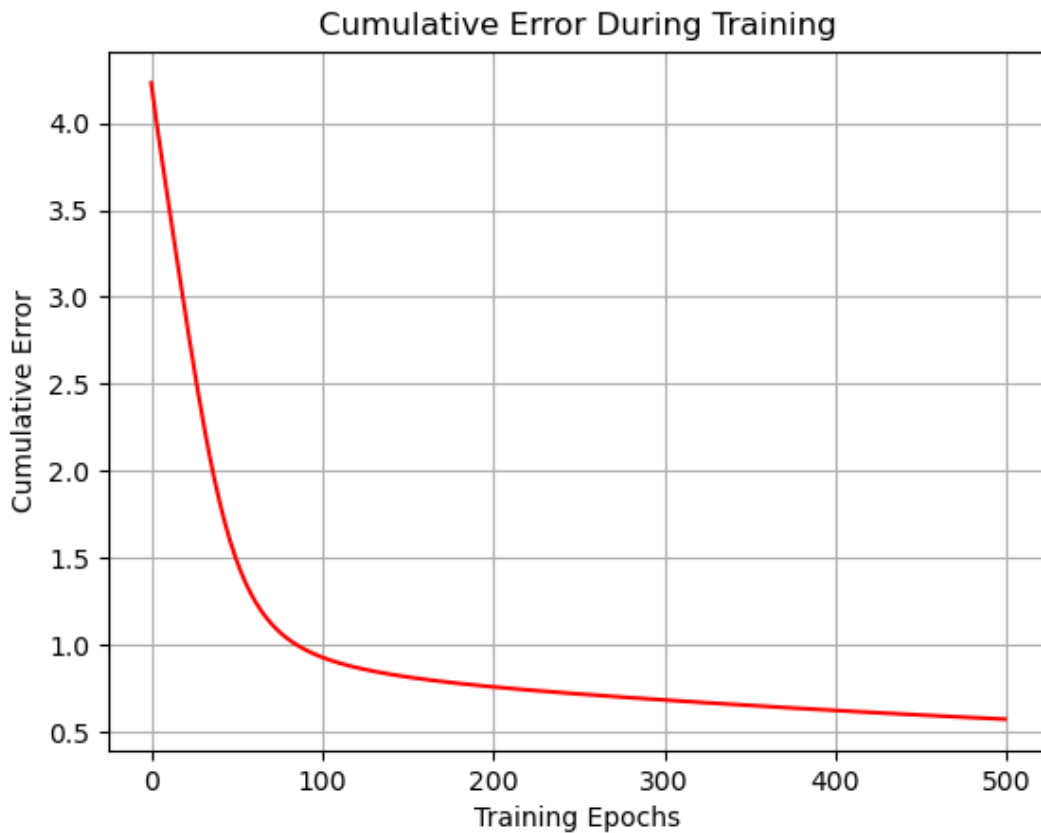
## 6 Πειράματα και Συγκρίσεις Μοντέλων

Για την εκπαίδευση και την πρόβλεψη των δύο μοντέλων χρησιμοποιήθηκε η μετρική Binary Cross Entropy (BCE Loss). Η συγκεκριμένη μετρική είναι γνωστή ως λογαριθμική απώλεια (Log Loss), είναι μια συνάρτηση απώλειας όπως και στο παραπάνω παράδειγμα χρησιμοποιείται συνήθως σε εργασίες δυαδικής ταξινόμησης, όπου ο στόχος είναι να προβλεφθούν αποτελέσματα που είναι δυαδικής φύσης παραδείγματος χάριν αληθές/λάθος, ναι/όχι ή τιμές 1/0. Ποσοτικοποιεί τη διαφορά μεταξύ των προβλεπόμενων πιθανοτήτων και των πραγματικών ετικετών, παρέχοντας ένα μέτρο που δείχνει πόσο καλές είναι οι προβλέψεις πιθανοτήτων του μοντέλου και ταιριάζουν με τις πραγματικές ετικέτες. Στο πλαίσιο της εργασίας τα μοντέλα προσπαθούν να προβλέπουν τη μελλοντική μήτρα γεινίασης, η οποία αποτελείται από 1/0 τιμές.

Το βασικό μοντέλο εκπαιδεύτηκε για 500 επαναλήψεις (epochs) και δόθηκε σε αυτό στην είσοδό του, οι συνδέσεις του γραφήματος στη χρονική στιγμή  $t$ , καθώς επίσης και ο ταυτοτικός πίνακας. Στη συνέχεια, το μοντέλο κλήθηκε να δημιουργήσει μια μελλοντική μήτρα γεινίασης για τη χρονική στιγμή  $t+1$ . Μέσω της λογαριθμικής απώλειας (Binary Cross Entropy) που παίρνει σαν όρισμα την προβλεπόμενη μήτρα γεινίασης του μοντέλου καθώς και την μήτρα γεινίασης τη χρονική στιγμή  $t+1$ , υπολογίζεται το ποσοστό λάθους της πρόβλεψης και έπειτα τα βάρη του μοντέλου. Με τον ίδιο αριθμό επαναλήψεων εκπαιδεύτηκε και το δεύτερο μοντέλο, η διαφορά είναι ότι στην είσοδο του μοντέλου αυτού πλέον δίνεται η μήτρα με τα node embeddings και οι συνδέσεις του γραφήματος. Αντίστοιχα, το δεύτερο μοντέλο χρησιμοποιεί την ίδια μετρική λάθους. Παρακάτω παρουσιάζονται τα γραφήματα της μετρικής λάθους (Loss function). Στην εικόνα 34 είναι η γραφική αναπαράσταση του πρώτου μοντέλου, ενώ στην εικόνα 35 μπορούμε να δούμε το loss function του δεύτερου μοντέλου (node embeddings).



Εικ.34.Base Model-Loss function



Εικ.35.Node2Vec Model-Loss function

Παρατηρώντας τις παραπάνω εικόνες μπορούμε διαπιστώνουμε ότι και τα δύο μοντέλα παρουσιάζουν μείωση των απωλειών, και τα δύο δίκτυα φαίνεται να μαθαίνουν με την πάροδο του χρόνου. Το δίκτυο που χρησιμοποιεί τα node embedding που παράχθηκαν με τον αλγόριθμο Node2Vec αναμένεται να συγκλίνει ταχύτερα επειδή οι ενσωματώσεις Node2Vec έχουν προ-εκπαιδευτεί ώστε να καταγράφουν τη δομή της τοπικής γειτονιάς κάθε κόμβου, δίνοντας ενδεχομένως στο δίκτυο ένα προβάδισμα στη μάθηση.

Η καμπύλη μάθησης του πρώτου μοντέλου που ξεκινά με τον ταυτοτικό πίνακα ως χαρακτηριστικό παρουσιάζει αρχικά μεγαλύτερη διακύμανση, καθώς το δίκτυο πρέπει να μάθει τις αναπαραστάσεις των κόμβων από την αρχή με βάση αποκλειστικά τη δομή του γράφου. Αντίθετα, το μοντέλο που έχει ως χαρακτηριστικά τον πίνακα των χαρακτηριστικών του κάθε κόμβου από το αλγόριθμο Node2Vec χρησιμοποιεί πλουσιότερα αρχικά χαρακτηριστικά που ενσωματώνουν πληροφορίες για τη γειτονιά, γεγονός που οδηγεί σε ομαλότερη και πιο σταθερή διαδικασία μάθησης.

Παρατηρούμε στην αρχή πως και τα δύο μοντέλα η τιμή του συνάρτησης λάθους βρίσκεται σε υψηλές τιμές, η οποία είναι αναμενόμενη, καθώς τα αρχικά βάρη είναι τυχαία και το δίκτυο δεν έχει μάθει ακόμη να προβλέπει αποτελεσματικά. Στη συνέχεια γίνεται μια γρήγορη μείωση αυτής και για τα δύο μοντέλα, γεγονός που υποδεικνύει ότι τα δίκτυα μαθαίνουν τα περισσότερα κατά τη διάρκεια των αρχικών εποχών.

Παρακάτω μπορούμε να δούμε τις τιμές της απώλειας (Loss) όταν τα μοντέλα καλούνται να προβλέψουμε την μελλοντική μήτρα γειτνίασης σε δεδομένα που βλέπουν για πρώτη φορά. Το πρώτο μοντέλο, το οποίο χρησιμοποιεί τον δείκτη ακμών και τον πίνακα ταυτότητας ως είσοδο, έχει ως πραγματική απώλεια περίπου 0,436 (εικόνα 35). Σε αντίθεση, το δεύτερο μοντέλο, το οποίο έχεις τα χαρακτηριστικά των κόμβων που παράγονται από τον αλγόριθμο Node2Vec, έχει πραγματική απώλεια περίπου 0,154 (εικόνα 36).

Συγκρίνοντας αυτές τις δύο τιμές αληθινών απωλειών, είναι προφανές ότι το δεύτερο μοντέλο αποδίδει σημαντικά καλύτερα σε αθέατα δεδομένα, καθώς η απώλειά του είναι σημαντικά χαμηλότερη. Αυτό υποδηλώνει ότι οι προεκπαιδευμένες ενσωματώσεις από τον Node2Vec παρέχουν ένα πιο κατατοπιστικό σύνολο χαρακτηριστικών που επιτρέπει στο μοντέλο να γενικεύει καλύτερα σε νέα δεδομένα.

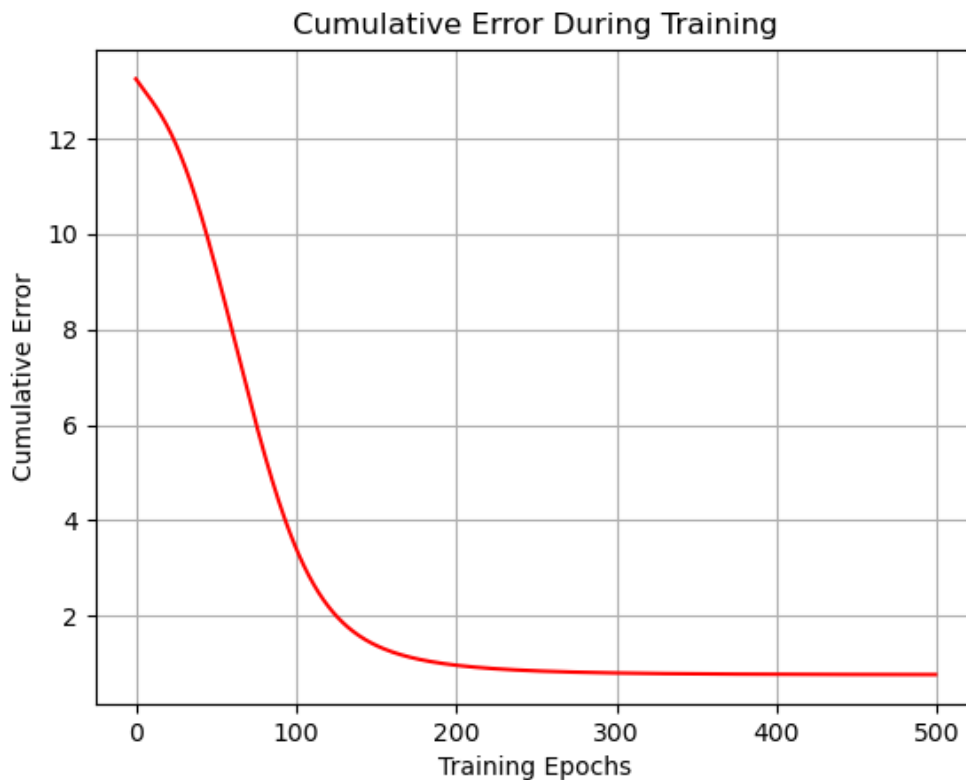
```
Loss with True Matrix: 0.4346179664134979
```

Εικ.36. Base Model-True Loss

```
Loss with True Matrix: 0.15429581701755524
```

Εικ.37. Node2Vec Model-True Loss

Πραγματοποιήθηκε και δεύτερο πείραμα με την χρήση των παραπάνω μοντέλων, αλλά αυτή τη φορά χρησιμοποιήθηκε το δεύτερο σύνολο δεδομένων (dataset 2). Τα δεδομένα αυτά επεξεργάστηκαν με κατάλληλο τρόπο ώστε να έχουν την ίδια ακριβώς δομή όπως το dataset 1. Με αυτόν τον τρόπο χρησιμοποιώντας τα ίδια κομμάτια κώδικα που παρουσιάστηκαν παραπάνω έγινε η εκπαίδευση των μοντέλων.



Εικ.38. Base Model Second Loss Function (dataset 2)

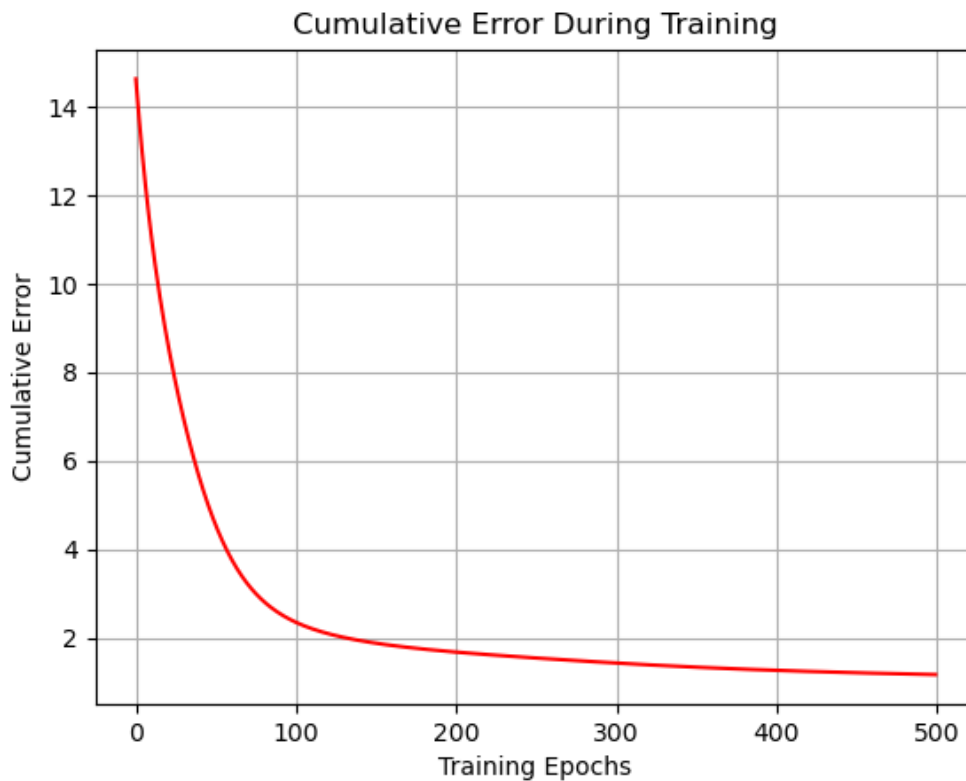
Το μοντέλο παρουσιάζει γρήγορα τη δυναμική της αρχιτεκτονικής του και βλέπουμε ότι η απώλεια του μοντέλου μειώνεται με αξιοσημείωτο ρυθμό. Προς το τέλος των επαναλήψεων η απώλεια του μοντέλου παραμένει σχεδόν η ίδια. Στην παρακάτω εικόνα βλέπουμε τη πρόβλεψη

του μοντέλου ύστερα από την εκπαίδευσή του. Παρατηρούμε πως η πρόβλεψη του μοντέλου, όπως και στο πρώτο πείραμα παρουσιάζει ιδιαίτερη μείωση του loss function.

**Loss with True Matrix: 0.48390093445777893**

Εικ.39. Base Model Second True - Loss Function (dataset 2)

Παρακάτω παρουσιάζεται η απώλεια του δεύτερου μοντέλου που κάνει χρήση του αλγορίθμου Node2Vec. Ξεκινάει από υψηλότερη θέση η τιμή της απώλειας, αλλά παρατηρούμε πως μειώνεται αρκετά γρήγορα. Συμπεραίνουμε πως το μοντέλο μαθαίνει αρκετά πιο γρήγορα τις τοπικές γειτονίες του κάθε κόμβου, αφού το έχουμε τροφοδοτήσει με τα node embeddings. Αντίστοιχα, η απώλεια του δεύτερου μοντέλου σε αθέατα δεδομένα είναι ιδιαίτερα χαμηλή.



Εικ.40. Base Model Second Loss Function (dataset 2)

**Loss with True Matrix: 0.12623924016952515**

Εικ.41. Base Model Second True - Loss Function (dataset 2)

Συμπεραίνουμε ότι ο αλγόριθμος Node2Vec παράγει ενσωματώσεις που αποτυπώνουν το δομικό πλαίσιο των κόμβων σε έναν γράφο, οι οποίες επιτρέπουν στο αντίστοιχο μοντέλο να κάνει ακριβέστερες προβλέψεις σε νέα δεδομένα. Μπορούμε να πούμε σε ένα γενικό πλαίσιο πως το μοντέλο αυτό παρουσιάζει χαμηλότερη απώλεια σε δεδομένα που βλέπει για πρώτη φορά με αποτέλεσμα να είναι ικανό για γενικεύσει καλύτερα από το μοντέλο που εκπαιδεύτηκε μόνο με τις δομικές πληροφορίες. Αντίθετα, το πρώτο μοντέλο ενώ φαίνεται πως κατά την εκπαίδευσή του



μπορεί και παράγει προβλέψεις με μικρότερη απώλεια, κατά την πρόβλεψη σε νέα δεδομένα παρατηρούμε πως έχει υψηλότερη απώλεια. Αυτό μας οδηγεί στο συμπέρασμα πως το μοντέλο αυτό έχει δημιουργήσει μια υπερπροσαρμογή στα αρχικά δεδομένα (Overfitting).

Υπερπροσαρμογή είναι όταν ένα μοντέλο εκπαιδεύεται σε δεδομένα και αυτό μαθαίνει να προβλέπει τέλεια για τα συγκεκριμένα δεδομένα. Το μοντέλο επικεντρώνεται υπερβολικά στις λεπτομέρειες των δεδομένων εκπαίδευσης και χάνει την ικανότητά του να γενικεύει σε νέα δεδομένα και αυτή είναι μια μη επιθυμητή ιδιότητα στα μοντέλα μηχανικής μάθησης. Αντίθετα, η χαμηλότερη απώλεια δοκιμών του δεύτερου μοντέλου υποδεικνύει ότι είναι λιγότερο πιθανό να κάνει υπερβολική προσαρμογή και ότι είναι πιο ικανό να χειριστεί τη μεταβλητότητα των δεδομένων. υποδηλώνοντας ότι το δεύτερο μοντέλο (Node2Vec) έχει μάθει μια πιο ισχυρή αναπαράσταση των δεδομένων.

Συνοπτικά, το μοντέλο που χρησιμοποιεί τις ενσωματώσεις Node2Vec παρουσιάζει καλύτερη απόδοση όσον αφορά τις απώλειες σε αθέατα δεδομένα σε σύγκριση με το μοντέλο που χρησιμοποιεί μόνο τις δομικές πληροφορίες. Αυτό αποδεικνύει την αποτελεσματικότητα της αξιοποίησης των ενσωματώσεων κόμβων από αλγορίθμους όπως ο Node2Vec για την ενίσχυση των προβλεπτικών δυνατοτήτων των συνελικτικών δικτύων γράφων.

## 7 Επίλογος

Από όσα εκθέσαμε παραπάνω, συνάγεται το συμπέρασμα ότι η ενσωμάτωση του αλγορίθμου `node2vec` σε ένα Νευρωνικού Δικτύου Συνελικτικού Γραφήματος (GCN) μπορεί να βελτιώσει σημαντικά την απόδοση εργασιών προηγμένης βαθιάς μάθησης που βασίζονται σε γραφήματα. Έχει σχεδιαστεί για να συλλαμβάνει αποτελεσματικά τις δομικές πληροφορίες ενός γράφου με την εκμάθηση ενσωματώσεων (*features*) για τους κόμβους (*node embeddings*) σε χαμηλές διαστάσεις (*low-dimensional*). Επιτυγχάνεται η διατήρηση της γειτονίας ενός κόμβου, ο αλγόριθμος εξασφαλίζει ότι οι κόμβοι που είναι πιο κοντά στο γράφημα είναι επίσης πιο κοντά και στο χώρο ενσωμάτωσης. Όταν αυτές οι ενσωματώσεις χρησιμοποιούνται ως χαρακτηριστικά εισόδου για ένα συνελικτικό νευρωνικό δίκτυο, παρέχουν στο δίκτυο αυτό μια πλούσια αναπαράσταση των δομικών ρόλων των κόμβων και της τοπικής τοπολογίας του κόμβου, οδηγώντας σε καλύτερες επιδόσεις πρόβλεψης μελλοντικών συνδέσμων.

Με την παραμετροποίηση του αλγορίθμου αυτού δίνεται η δυνατότητα να παράξει ενσωματώσεις που δίνουν έμφαση σε διαφορετικές πτυχές της τοπολογίας του γράφου, γεγονός που μπορεί να είναι ιδιαίτερα επωφελές για το νευρωνικό δίκτυο γράφων που βασίζονται στην ποιότητα των χαρακτηριστικών εισόδου για τα συνελικτικά τους στρώματα.

Επιπλέον ο συνδυασμός του συγκεκριμένου τύπου νευρωνικών δικτύων με τον παραπάνω αλγόριθμο είναι ότι οι ενσωματώσεις που παράγονται από τον `node2vec` μπορούν να βοηθήσουν το μοντέλο να γενικεύσει καλύτερα σε αόρατα δεδομένα. Δεδομένου ότι ο `node2vec` καταγράφει την υποκείμενη δομή του γράφου, οι ενσωματώσεις μπορούν να κάνουν το μοντέλο πιο ανθεκτικό σε μεταβολές της τοπολογίας του γράφου, όπως η προσθήκη ή η αφαίρεση κόμβων/ακμών. Παρόλο που τα ίδια τα συνελικτικά μοντέλα έχουν σχεδιαστεί για να χειρίζονται αποτελεσματικά τα δεδομένα υπό τη μορφή γράφου, η αρχική αναπαράσταση των χαρακτηριστικών των κόμβων διαδραματίζει κρίσιμο ρόλο στη συνολική απόδοση και την επεκτασιμότητα του μοντέλου. Η αποδοτική στρατηγική τυχαίου περιπάτου του αλγορίθμου για τη δειγματοληψία γειτονιών επιτρέπει την κλιμακούμενη δημιουργία ενσωματώσεων, η οποία είναι επωφελής για μεγάλους γράφους.

Ωστόσο, πρέπει να σημειωθεί ότι τα οφέλη από την ενσωμάτωση του αλγορίθμου στα συνελικτικά μοντέλα εξαρτώνται από τη συγκεκριμένη εφαρμογή ή πρόβλημα που χρειάζεται επίλυση καθώς και τα χαρακτηριστικά των δεδομένων που υπάρχουν. Η προσέγγιση αυτή μπορεί να είναι ιδιαίτερα αποτελεσματική, όταν η δομή των πληροφοριών του εκάστοτε γραφήματος είναι κρίσιμη για τη διαδικασία της εκμάθησης και όταν το γράφημα παρουσιάζει πολύπλοκα μοτίβα συνδεσιμότητας τα οποία μπορούν να αποτυπωθούν αποτελεσματικά μέσω της διαδικασίας ενσωμάτωσης του `node2vec`.

Μπορούμε να πούμε πως το μέλλον των Νευρωνικών Δικτύων Συνελικτικού Γραφήματος (GCNs) φαίνεται εξαιρετικά ελπιδοφόρο, με πιθανές εξελίξεις και εφαρμογές που καλύπτουν διάφορους τομείς. Μία από τις κύριες προκλήσεις των δικτύων αυτών είναι η κλιμάκωση τους σε πολύ μεγάλους γράφους, όπως τα κοινωνικά δίκτυα ή οι γράφοι του διαδικτύου, λόγω περιορισμών υπολογιστικών δυνατοτήτων και μνήμης. Στο μέλλον ενδέχεται να αναπτυχθούν πιο αποδοτικοί αλγόριθμοι και αρχιτεκτονικές που μπορούν να χειριστούν πιο αποτελεσματικά δισεκατομμύρια κόμβους και ακμές, αξιοποιώντας ενδεχομένως τις εξελίξεις στο *hardware*, όπως οι μονάδες επεξεργασίας γραφημάτων (GPU), και τεχνικές βελτιστοποίησης όπως η δειγματοληψία γραφημάτων. Οι μελλοντικές εξελίξεις ενδέχεται να φέρουν περισσότερα υβριδικά μοντέλα που θα ενσωματώνουν τα πλεονεκτήματα διαφόρων αρχιτεκτονικών GNN, όπως τα Graph Attention Networks (GATs), τα Graph Isomorphism Networks (GINs) και άλλα. Αυτό θα μπορούσε να οδηγήσει σε πιο ισχυρά και ευέλικτα μοντέλα ικανά να αποτυπώνουν πολύπλοκες δομές και δυναμικές γραφημάτων.

## 8 Πηγές και Βιβλιογραφία

- [1] L. Wu, P. Cui, J. Pei, και L. Zhao, Επιμ., *Graph Neural Networks: Foundations, Frontiers, and Applications*. Singapore: Springer Nature, 2022. doi: [10.1007/978-981-16-6054-2](https://doi.org/10.1007/978-981-16-6054-2).
- [2] G. Panagopoulos, G. Nikolentzos, και M. Vazirgiannis, 'Transfer Graph Neural Networks for Pandemic Forecasting'. arXiv, 12 Απρίλιος 2021. doi: [10.48550/arXiv.2009.08388](https://doi.org/10.48550/arXiv.2009.08388).
- [3] P. Sgrignoli, R. Metulini, S. Schiavo, και M. Riccaboni, 'The Relation Between Global Migration and Trade Networks', *Physica A: Statistical Mechanics and its Applications*, τ. 417, σσ. 245–260, Ιανουαρίου 2015, doi: [10.1016/j.physa.2014.09.037](https://doi.org/10.1016/j.physa.2014.09.037).
- [4] A. Grover και J. Leskovec, 'node2vec: Scalable Feature Learning for Networks'. arXiv, 3 Ιούλιος 2016. doi: [10.48550/arXiv.1607.00653](https://doi.org/10.48550/arXiv.1607.00653).
- [5] K. Z. Khanam, A. Singhal, και V. Mago, 'NODDLE: Node2vec based deep learning model for link prediction', τ. 480, 2023, σσ. 196–212. doi: [10.1007/978-3-031-33614-0\\_14](https://doi.org/10.1007/978-3-031-33614-0_14).
- [6] E. Tranos, M. Gheasi, και P. Nijkamp, 'International Migration: A Global Complex Network', *Environment and Planning B: Planning and Design*, τ. 42, σσ. 4–22, Φεβρουαρίου 2015, doi: [10.1068/b39042](https://doi.org/10.1068/b39042).
- [7] G. Fagiolo και M. Mastrorillo, 'International migration network: Topology and modeling', *Phys. Rev. E*, τ. 88, τχ. 1, σ. 012812, Ιουλίου 2013, doi: [10.1103/PhysRevE.88.012812](https://doi.org/10.1103/PhysRevE.88.012812).
- [8] K. F. Davis, P. D'Odorico, F. Laio, και L. Ridolfi, 'Global Spatio-Temporal Patterns in Human Migration: A Complex Network Perspective', *PLOS ONE*, τ. 8, τχ. 1, σ. e53723, Ιανουαρίου 2013, doi: [10.1371/journal.pone.0053723](https://doi.org/10.1371/journal.pone.0053723).
- [9] J. Zhou κ.ά., 'Graph Neural Networks: A Review of Methods and Applications'. arXiv, 6 Οκτώβριος 2021. doi: [10.48550/arXiv.1812.08434](https://doi.org/10.48550/arXiv.1812.08434).
- [10] Sejan, Mohammad Abrar Shakil κ.ά. 2023. 'Graph Convolutional Network Design for Node Classification Accuracy Improvement'. *Mathematics* 11(17): 3680. doi: [10.3390/math11173680](https://doi.org/10.3390/math11173680).
- [11] Kipf, Thomas N., και Max Welling. 2017. 'Semi-Supervised Classification with Graph Convolutional Networks'. doi: [10.48550/arXiv.1609.02907](https://doi.org/10.48550/arXiv.1609.02907).
- [12] Wang, Junfu κ.ά. 2021. 'Bi-GCN: Binary Graph Convolutional Network'. Στο *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, 1561–70. doi: [10.1109/CVPR46437.2021.00161](https://doi.org/10.1109/CVPR46437.2021.00161).