Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Εντοπισμός εξελικτικών μοντέλων σε πραγματικά και συνθετικά δίκτυα

Διπλωματικη Εργασια

της

ΦΑΙΔΡΑΣ ΑΝΑΣΤΑΣΙΑΣ Ι. ΑΝΘΟΠΟΥΛΟΥ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβριος 2025

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Εντοπισμός εξελικτικών μοντέλων σε πραγματικά και συνθετικά δίκτυα

Διπλωματικη Εργασια

της

ΦΑΙΔΡΑΣ ΑΝΑΣΤΑΣΙΑΣ Ι. ΑΝΘΟΠΟΥΛΟΥ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 2α Οκτωβρίου 2025.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Συμεών Παπαβασιλείου	Ελένη Στάη	Βασίλειος Καρυώτης
Καθηγητής Ε.Μ.Π	Επίκουρη Καθηγήτρια Ε.Μ.Π	Καθηγητής Ι.Π

Αθήνα, Οκτώβριος 2025

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Φαίδρα Αναστασία Ανθοπούλου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

2 Οκτωβρίου 2025

Περίληψη

Τις τελευταίες δύο δεκαετίες, η συζήτηση της επιστημονικής κοινότητας περί του εξελικτικού μοντέλου στο οποίο ταξινομούνται διάφορα είδη πραγματικών δικτύων έχει υπάρξει έντονη, συνοδευόμενη από πληθώρα αντικρουόμενων ερευνών πάνω στο θέμα αυτό. Επιπλέον, πρόσφατα εισήχθη η θεώρηση ότι μεγάλο ποσοστό των πραγματικών δικτύων αποτελούνται από συνιστώσες που χαρακτηρίζονται από διαφορετικά εξελικτικά μοντέλα. Τα γεγονότα αυτά ενέπνευσαν την επιλογή του αντικειμένου της παρούσας εργασίας, η οποία αποσκοπεί στην διερεύνηση της συχνότητας εμφάνισης κάθε εξελικτικού μοντέλου στα πραγματικά δίκτυα, και την ισχύ της προαναφερθείσας θεώρησης. Για την πραγματοποίηση του στόχου αυτού, αξιολογήθηκαν αρχικά 20 μετρικές δικτύων σχετικά με την καταλληλότητά τους για χρήση ως είσοδοι ενός ταξινομητή. Έπειτα, δημιουργήθηκε ένα σύνολο 600 συνθετικών δικτύων, και υπολογίστηκαν οι τιμές των επιλεγμένων μετρικών για αυτά. Ύστερα οι τιμές αυτές χρησιμοποιήθηκαν ως τα δεδομένα εκπαίδευσης και ελέγχου μίας σειράς ταξινομητών, οι οποίοι αξιολογήθηκαν με σκοπό την επιλογή του βέλτιστου. Ο βέλτιστος ταξινομητής χρησιμοποιήθηκε για την ταξινόμηση 17 πραγματικών δικτύων, τα οποία ύστερα τμηματοποιήθηκαν με χρήση επιλεγμένου αλγορίθμου τμηματοποίησης, και οι συνιστώσες που προέκυψαν ταξινομήθηκαν και αυτές. Η ανάλυση των αποτελεσμάτων έδειξε πως πιο συχνά εμφανίζονται τα Δίκτυα Μικρού Κόσμου (Small-World) και τα Μη-Κλιμακούμενα Δίκτυα (Scale-Free). Ακόμη, η παρουσία του μοντέλου Small-World είναι ιδιαίτερα εκτεταμένη στα δίκτυα μεταφορών, ενώ το Scale-Free εμφανίζεται ως επί το πλείστον ως η μεγαλύτερη συνιστώσα σε κοινωνικά δίκτυα. Το μοντέλο Waxman επίσης εμφανίζεται συχνά ως συνιστώσα στα κοινωνικά δίκτυα, ενώ για τα υπόλοιπα μοντέλα δεν παρατηρήθηκε κάποιο μοτίβο. Τέλος, συμπεραίνουμε πως η θεώρηση σχετικά με την ύπαρξη συνιστωσών με διαφορετικά μοντέλα στα δίκτυα είναι σωστή.

Λέξεις Κλειδιά

Σύνθετα Δίκτυα, Εξελικτικά Μοντέλα, Μηχανική Μάθηση, Τμηματοποίηση Δικτύων, Ταξινόμηση Δικτύων

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Σ. Παπαβασιλείου για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου την παρούσα διπλωματική εργασία, καθώς και για την καθοδήγηση και τις πολύτιμες συμβουλές του. Επιπλέον, θα ήθελα να ευχαριστήσω τον Δρ. Κ. Τσιτσεκλή για τις υποδείξεις του και την βοήθειά του σε όλη την διάρκεια της εκπόνησής της.

Αθήνα, Οκτώβριος 2025

Φαίδρα Αναστασία Ανθοπούλου

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Το κεφάλαιο αυτό αποτελεί μία εισαγωγή στην παρούσα διπλωματική. Περιλαμβάνει μια συνοπτική παρουσίαση του αντικειμένου της, την περιγραφή της προσέγγισης που ακολουθήσαμε στα πλαίσια της επίλυσης του προβλήματος που αυτή πραγματεύεται, και πληροφορίες σχετικά με την δομή της.

1.1 Αντικείμενο

Η παρουσία των δικτύων είναι εξαιρετικά εκτεταμένη στην συντριπτική πλειοψηφία των πτυχών της ζωής ενός ανθρώπου. Από τις κοινωνικές συναναστροφές μας, τις συναλλαγές μας, τις μετακινήσεις μας, έως και τα ίδια τα κύτταρά μας, τα δίκτυα απαντώνται παντού. Για τον λόγο αυτό, η μελέτη τους με σκοπό την καλύτερη κατανόηση της δομής και της συμπεριφοράς τους έχει απασχολήσει πολύ την επιστημονική κοινότητα.

Ένα σημαντικό εργαλείο της Επιστήμης Δικτύων για την επιτέλεση αυτού του έργου είναι τα μοντέλα των δικτύων, τα οποία (όπως υποδηλώνει και το όνομά τους) μοντελοποιούν τα δίκτυα και την συμπεριφορά τους [1]. Συγκεκριμένα, τα εξελικτικά μοντέλα, τα οποία αποτελούν το αντικείμενο της παρούσας εργασίας, αφορούν την εξέλιξη ενός δικτύου σε βάθος χρόνου, δηλαδή τον τρόπο με τον οποίο σχηματίζονται νέες συνδέσεις και κατ' επέκταση την υφιστάμενη μορφή του δικτύου βάσει του παρελθόντος. Η ταξινόμηση, λοιπόν, ενός δικτύου στο σωστό εξελικτικό μοντέλο βάσει της παρούσας μορφής του, μπορεί να μας προσφέρει μία πληθώρα πληροφοριών σχετικά με την μελλοντική συμπεριφορά και δομή του [2]. Για παράδειγμα, η γνώση του εξελικτικού μοντέλου (και επομένως της τοπολογίας) ενός ηλεκτρικού δικτύου μας επιτρέπει την προσομοίωσή του μέσω ενός συνθετικού δικτύου που έχει κατασκευαστεί βάσει αυτού του μοντέλου, και διευκολύνει την διεξαγωγή ανάλυσης ευρωστίας πάνω σε αυτό. [3]

Υπάρχουν πολλές διαφορετικές προσεγγίσεις για την διεξαγωγή της ταξινόμησης αυτής. Στην βιβλιογραφία εντοπίζονται μέθοδοι ταξινόμησης βάσει θεωρητικών μετρικών, αλλά και μέσω της χρήσης εργαλείων Τεχνητής Νοημοσύνης όπως τα Μοντέλα Μηχανικής Μάθησης και τα Νευρωνικά Δίκτυα. Επιπλέον, εμφανίζεται η θεώρηση πως αρκετά δίκτυα περιγράφονται καλύτερα από δύο ή περισσότερα μοντέλα που αντιστοιχούν σε διαφορετικές συνιστώσες αυτών, αντί από ένα μοναδικό που αντιστοιχεί σε ολόκληρο το δίκτυο.

1.2 Προσέγγιση

Με έναυσμα τα παραπάνω, αποφασίσαμε και εμείς να ερευνήσουμε το θέμα του εντοπισμού εξελικτικών μοντέλων σε πραγματικά και συνθετικά δίκτυα, στα πλαίσια της παρούσας διπλωματικής. Η δική μας προσέγγιση περιλαμβάνει την επιλογή ενός συνόλου μετρικών δικτύων, κατόπιν αξιολόγησής τους ως προς την καταλληλότητά τους για το πρόβλημά μας. Οι μετρικές αυτές ύστερα υπολογίζονται για το δίκτυο που επιθυμούμε να ταξινομήσουμε, και οι τιμές τους χρησιμοποιούνται ως είσοδοι σε έναν ταξινομητή που έχει αποδειχθεί ο πιο αποτελεσματικός. Η έξοδος του ταξινομητή μας δίνει την απάντηση στο ερώτημα ποιο εξελικτικό μοντέλο περιγράφει καλύτερα το δίκτυο. Επιπλέον, εξετάζουμε και την πεποίθηση ότι τα δίκτυα ενίοτε αποτελούνται από συνιστώσες που χαρακτηρίζονται από διαφορετικά εξελικτικά μοντέλα, διασπώντας τα σε υποδίκτυα με την χρήση κατάλληλης μεθόδου τμηματοποίησης (η οποία έχει επίσης επιλεγεί από εμάς μέσω ελέγχων επίδοσης), και ταξινομώντας ύστερα τα υποδίκτυα αυτά ακολουθώντας την διαδικασία που περιγράφηκε για τα αυτούσια δίκτυα.

1.3 Δομή

Η παρούσα διπλωματική οργανώνεται στην συνέχεια στα εξής κεφάλαια:

- **Κεφάλαιο 2 (Θεωρητικό Υπόβαθρο):** Παρέχει στον αναγνώστη τους ορισμούς και ορισμένες ιδιότητες των εννοιών που χρησιμοποιήθηκαν, με σκοπό την ομαλή εισαγωγή του στο αντικείμενο της διπλωματικής
- **Κεφάλαιο 3 (Σχετική Βιβλιογραφία):** Παρουσιάζει περιληπτικά ένα πλήθος ερευνών που σχετίζονται με το αντικείμενο της παρούσας διπλωματικής, σκιαγραφώντας έτσι το επιστημονικό πλαίσιο στο οποίο αυτή διεξήχθη
- **Κεφάλαιο 4 (Μεθοδολογία):** Απαριθμεί τις μεθόδους και τις τεχνολογίες που χρησιμοποιήθηκαν για την διεκπεραίωση των στόχων της διπλωματικής, όπως και τα κριτήρια που οδήγησαν στην επιλογή τους
- **Κεφάλαιο 5 (Αποτελέσματα):** Περιέχει τα αποτελέσματα του πειραματικού μέρους της διπλωματικής, ορισμένες παρατηρήσεις πάνω σε αυτά, καθώς και τις αποφάσεις που λήφθηκαν υπ' όψιν τους
- **Κεφάλαιο 6 (Κατακλείδα):** Περιέχει μια σύνοψη της μεθοδολογίας και των αποτελεσμάτων, τα συμπεράσματα στα οποία καταλήξαμε βάσει των αποτελεσμάτων αυτών, καθώς και ορισμένες ιδέες για μελλοντικές επεκτάσεις

Μέρος I

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

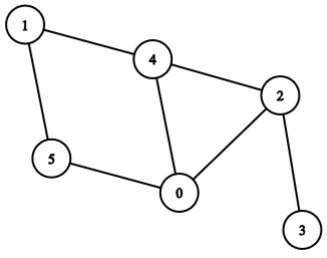
Στο παρόν κεφάλαιο παρέχονται πληροφορίες σχετικά με τις βασικές έννοιες που χρησιμοποιούνται στην εργασία αυτή, όπως και με τις τεχνολογίες που χρειάστηκαν για την υλοποίηση του πειραματικού μέρους της. Συγκεκριμένα, θα αναλυθούν έννοιες και αλγόριθμοι που σχετίζονται με τους Γράφους, τα Σύνθετα Δίκτυα, και την Μηχανική Μάθηση.

2.1 Βασική Θεωρία Γράφων

2.1.1 Ορισμός Γράφου

Ένας γράφος είναι η μαθηματική αναπαράσταση ενός συνόλου στοιχείων, και των σχέσεων μεταξύ αυτών. Τα στοιχεία αυτά αποτελούν τις **κορυφές** V του γράφου, ενώ οι σχέσεις μεταξύ τους αποτυπώνονται στις **ακμές** E του. Έτσι ένας γράφος μπορεί να αναπαρασταθεί μαθηματικά από το διατεταγμένο ζεύγος συνόλων $\{V, E\}$. [4] [5]

Οι γράφοι συχνά αναφέρονται και ως **δίκτυα**, αφού αποτελούν ένα σημαντικότατο εργαλείο στην μελέτη και την αναπαράσταση πραγματικών και τεχνητών δικτύων, σε σημείο κάποιες φορές οι έννοιες να ταυτίζονται. Για παράδειγμα, μία ομάδα συμφοιτητών αποτελεί ένα κοινωνικό δίκτυο που μπορεί να αναπαρασταθεί μέσω ενός γράφου του οποίου οι κόμβοι αντιστοιχούν στους φοιτητές και οι ακμές στις φιλίες μεταξύ τους. [6] [5]



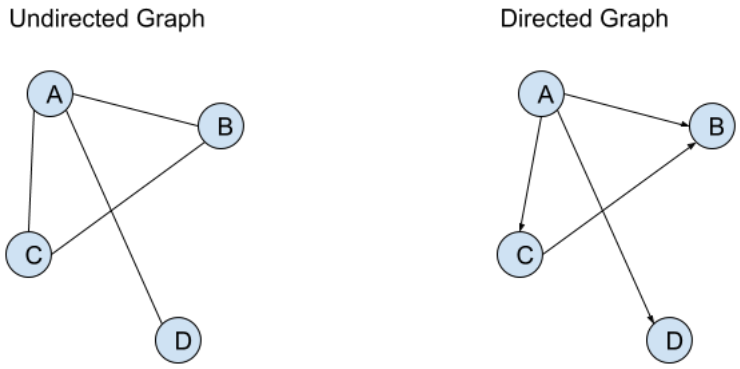
Σχήμα 2.1: *Παράδειγμα γράφου με 6 κορυφές και 7 ακμές*

2.1.2 Ιδιότητες Γράφων

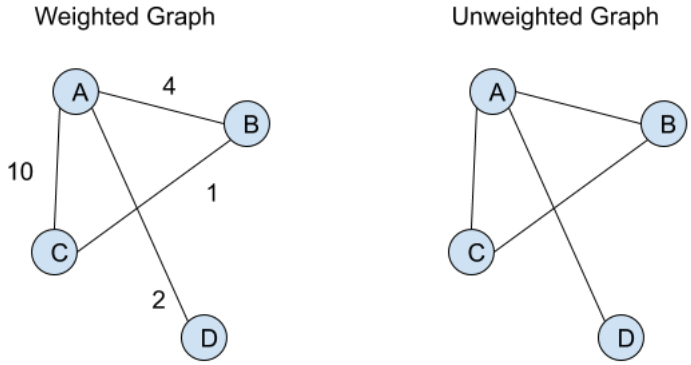
Κατευθυντικότητα

Ένας γράφος ονομάζεται **μη-κατευθυνόμενος** αν κάθε ακμή (v_i, v_j) είναι ισοδύναμη με την ακμή (v_j, v_i), όπου v_i, v_j οι κόμβοι στους οποίους προσπίπτει η ακμή. Σε έναν τέτοιο γράφο, η κάθε ακμή υποδηλώνει μια αμφίδρομη σχέση. Ένα παράδειγμα δικτύου που αναπαριστάται με μη-κατευθυνόμενο γράφο είναι ένα μέσο κοινωνικής δικτύωσης όπου σχηματίζονται συμμετρικές σχέσεις φιλίας μεταξύ των μελών (π.χ. Facebook). [5]

Αντιθέτως, ένας γράφος ονομάζεται **κατευθυνόμενος** εάν οι ακμές (v_i, v_j) και (v_j, v_i) είναι διαφορετικές μεταξύ τους. Σε αυτήν την περίπτωση η κάθε ακμή θεωρείται ότι έχει μια συγκεκριμένη κατεύθυνση, η οποία καταδεικνύεται από την σειρά που αναγράφονται οι κόμβοι πρόσπτωσης στο ζεύγος. Η σχέση που υποδηλώνει κάθε ακμή, λοιπόν, είναι μονόδρομη. Αντιστοίχως, ένα παράδειγμα δικτύου που αναπαριστάται με κατευθυνόμενο γράφο είναι ένα μέσο κοινωνικής δικτύωσης όπου σχηματίζονται μη συμμετρικές σχέσεις ακόλουθου και ακολουθούμενου μεταξύ των μελών (π.χ. X). [5]



Σχήμα 2.2: *Δύο γράφοι: Αριστερά ένας μη-κατευθυνόμενος, Δεξιά ένας κατευθυνόμενος* [7]



Σχήμα 2.3: *Δύο γράφοι: Αριστερά ένας με βάρη, Δεξιά ένας χωρίς βάρη* [7]

Βάρη

Ένας γράφος λέμε ότι έχει βάρη όταν σε κάθε ακμή αντιστοιχίζεται μια αριθμητική τιμή, η οποία αποτελεί το **βάρος** της ακμής, και ότι δεν έχει βάρη στην αντίθετη περίπτωση. Το βάρος μιας ακμής αντιπροσωπεύει την τιμή ενός χαρακτηριστικού της σύνδεσης στην οποία αντιστοιχεί, διαφοροποιώντας έτσι τις συνδέσεις και επιτρέποντας την μοντελοποίηση φαινομένων όπως το κόστος, η απόσταση, η χωρητικότητα, και άλλα. Σε γράφους χωρίς βάρη όλες οι ακμές, και άρα οι συνδέσεις, θεωρούνται ισοδύναμες μεταξύ τους. [5]

Γειτνίαση και Διαδρομές

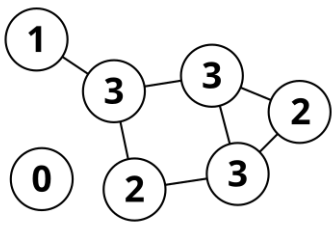
Δύο κόμβοι v_i, v_j ονομάζονται **γείτονες** εφόσον υπάρχει μία ακμή (v_i, v_j) ή (v_j, v_i) που τους συνδέει. Το πλήθος των γειτόνων ενός κόμβου αποτελεί τον **βαθμό** του κόμβου αυτού εφόσον ο γράφος είναι χωρίς βάρη. Εάν είναι με βάρη, τότε ως βαθμός ενός κόμβου ορίζεται το άθροισμα των βαρών όλων των ακμών που τον συνδέουν με τους γείτονές του. [8] Στους κατευθυνόμενους γράφους ορίζονται ο εισερχόμενος και ο εξερχόμενος βαθμός κόμβου, που ισούνται με το πλήθος των ακμών με κατεύθυνση προς ή από τον κόμβο αντίστοιχα.

Εάν υπάρχει μια ακολουθία ακμών (και αντίστοιχων κορυφών) που μπορούμε να ακολουθήσουμε για να φτάσουμε από τον κόμβο v_i στον κόμβο v_j, τότε αυτή αποτελεί μια **διαδρομή** μεταξύ των κόμβων v_i και v_j. Εάν μάλιστα δεν επαναλαμβάνεται καμία ακμή στην διαδρομή τότε αυτή ονομάζεται **μονοπάτι**, ενώ αν επιπλέον δεν επαναλαμβάνεται κανένας κόμβος τότε ονομάζεται **απλό μονοπάτι**. [9]

Συνεκτικότητα και Συνιστώσες

Ένας γράφος ονομάζεται **συνεκτικός** αν υπάρχει τουλάχιστον μία διαδρομή μεταξύ οποιωνδήποτε δύο κόμβων του (αγνοώντας τις τυχούσες κατευθύνσεις των ακμών), και **μη συνεκτικός** στην αντίθετη περίπτωση. Εάν ο γράφος είναι κατευθυνόμενος και συνεκτικός, τότε αν λαμβάνοντας υπ' όψιν και τις κατευθύνσεις των ακμών παραμένει συνεκτικός αυτός ονομάζεται **ισχυρά συνεκτικός** αλλιώς ονομάζεται **ασθενώς συνεκτικός**. [10]

Μια **συνεκτική συνιστώσα** ενός γράφου αποτελείται από ένα υποσύνολο κορυφών και ακμών του γράφου (δηλαδή έναν **υπογράφο**), που είναι συνεκτικό. Κάθε γράφος αποτελείται τουλάχιστον μία συνεκτική συνιστώσα. Αφαιρώντας ακμές από έναν γράφο μπορούμε να τον διασπάσουμε σε περισσότερες συνεκτικές συνιστώσες, τεχνική που χρησιμοποιείται σε πολλούς αλγορίθμους για γράφους (παραδείγματος χάριν αυτούς που αποσκοπούν στην ανίχνευση κοινοτήτων). [10]



Σχήμα 2.4: *Ένας μη-συνεκτικός γράφος με δύο συνεκτικές συνιστώσες* [10]

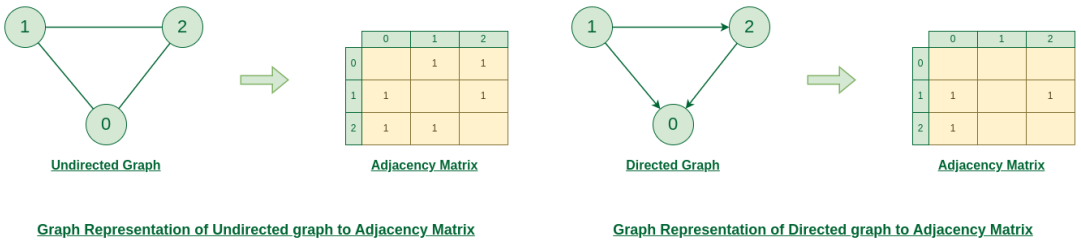
2.1.3 Αναπαράσταση Γράφων

Οι γράφοι συνήθως αναπαρίστανται είτε μέσω ενός **πίνακα γειτνίασης** είτε μέσω μιας **συνδεδεμένης λίστας γειτνίασης**.

Πίνακας Γειτνίασης

Ο πίνακας γειτνίασης ενός γράφου χωρίς βάρη είναι ένας πίνακας μεγέθους $N \times N$, όπου N το πλήθος των κορυφών του γράφου. Εάν ο γράφος είναι μη-κατευθυνόμενος, τότε οι θέσεις με συντεταγμένες (i, j) και (j, i) του πίνακα περιέχουν τον αριθμό 1 εάν υπάρχει ακμή που συνδέει τους κόμβους v_i και v_j και το 0 αλλιώς. Εάν ο γράφος είναι κατευθυνόμενος, τότε η θέση με συντεταγμένες (i, j) περιέχει το 1 μόνο εάν υπάρχει ακμή (v_i, v_j) (δηλαδή από τον κόμβο i προς τον κόμβο j) και το 0 αλλιώς. [11]

Για έναν γράφο με βάρη ισχύουν τα ίδια, εκτός από το γεγονός ότι η ύπαρξη μιας ακμής στον πίνακα γειτνίασης δεν αντιπροσωπεύεται από τον αριθμό 1, αλλά από το βάρος της. [12]



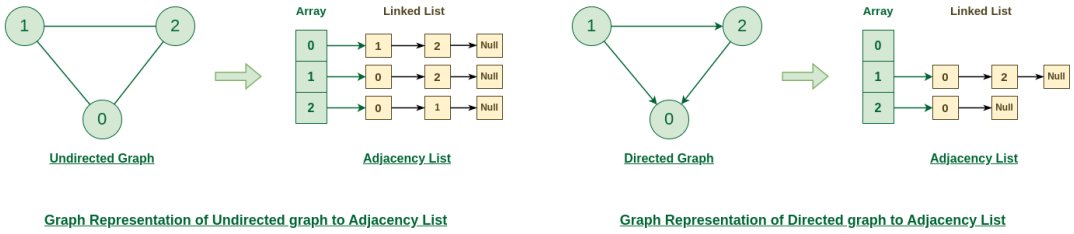
(α') (β')

Σχήμα 2.5: *Παραδείγματα γράφων χωρίς βάρη και των αντίστοιχων πινάκων γειτνίασης: (α') Μη-κατευθυνόμενος γράφος, (β') Κατευθυνόμενος γράφος*
[11]

Συνδεδεμένη λίστα γειτνίασης

Η συνδεδεμένη λίστα γειτνίασης ενός γράφου χωρίς βάρη αποτελείται από έναν πίνακα μεγέθους $N \times 1$, σε κάθε θέση i του οποίου βρίσκεται μια συνδεμένη λίστα που περιέχει όλους τους γείτονες του v_i. Εάν ο γράφος είναι κατευθυνόμενος τότε στην λίστα γειτόνων ενός κόμβου v_i περιλαμβάνονται μόνο οι γείτονές του v_j που συνδέονται μαζί του με ακμή κατεύθυνσης από τον v_i προς τον v_j (δηλαδή ακμή (v_i, v_j)). Εάν είναι μη-κατευθυνόμενος, τότε όλοι οι γείτονες ενός κόμβου συμπεριλαμβάνονται στην αντίστοιχη λίστα, καθώς οι σχέσεις είναι αμφίδρομες. [11]

Για έναν γράφο με βάρη ισχύουν όλα τα παραπάνω, συν το γεγονός ότι μαζί με κάθε γείτονα v_j ενός κόμβου v_i, που καταγράφεται στην συνδεδεμένη λίστα γειτόνων του v_i, καταγράφεται και το βάρος της ακμής (v_i, v_j). Έτσι κάθε στοιχείο της συνδεδεμένης λίστας γειτόνων ενός κόμβου περιέχει δύο τιμές: το αναγνωριστικό ενός γειτονικού κόμβου, και το βάρος της ακμής που τους συνδέει. [12]



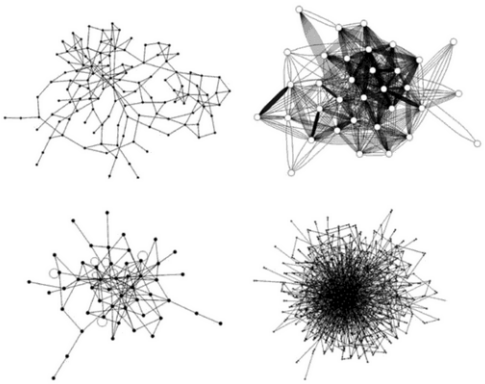
(α΄) (β΄)

Σχήμα 2.6: *Παραδείγματα γράφων και των αντίστοιχων συνδεδεμένων λιστών γειτνίασης: (α΄) Μη-κατευθυνόμενος γράφος, (β΄) Κατευθυνόμενος γράφος*
[11]

2.2 Σύνθετα Δίκτυα

2.2.1 Ορισμός Σύνθετου Δικτύου

Ένα δίκτυο είναι μία αναπαράσταση ενός συστήματος, μέσω των κόμβων του που αναπαριστούν τις οντότητες του συστήματος και των ακμών του που αναπαριστούν τις σχέσεις ή αλληλεπιδράσεις μεταξύ των οντοτήτων αυτών. Ένα **σύνθετο δίκτυο** είναι ένα δίκτυο του οποίου η δομή είναι πολύπλοκη, αναφορικά με τον όγκο πληροφορίας που αναπαριστά και εμπεριέχει. Παραδείγματα σύνθετων δικτύων αποτελούν τα οδικά δίκτυα, τα δίκτυα αλληλεπίδρασης μεταξύ πρωτεϊνών, ο Παγκόσμιος Ιστός, τα εμπορικά δίκτυα, τα δίκτυα συνεργασιών, τα δίκτυα παραπομπών, κι άλλα. [13]



Σχήμα 2.7: *Διαφορετικές κατηγορίες δικτύων: (Πάνω αριστερά) Αναπαράσταση ενός ηλεκτρονικού κυκλώματος, (Πάνω δεξιά) Αναπαράσταση των αλληλεπιδράσεων μίας ομάδας ερευνητών πανεπιστημίου της West Virginia, (Κάτω Αριστερά) Αναπαράσταση των πρωτεϊνικών αλληλεπιδράσεων ενός στελέχους του ιού Έρπη, (Κάτω Δεξιά) Αναπαράσταση μιας συνεκτικής συνιστώσας του μεταβολικού δικτύου του Ελικοβακτηρίου του Πυλωρού μέσω κατευθυνόμενου δικτύου*

[13]

2.2.2 Εξελικτικά Μοντέλα Σύνθετων Δικτύων

Σε αυτήν την υποενότητα θα παρουσιαστούν τα εξελικτικά μοντέλα τα οποία αποσκοπεί η παρούσα εργασία να εντοπίσει σε σύνθετα δίκτυα.

Τυχαίοι Γράφοι (Random Graphs)

Ένας Τυχαίος Γράφος είναι ένας γράφος όπου η πιθανότητα να υπάρχει μια ακμή μεταξύ δύο κορυφών είναι ανεξάρτητη από τις ιδιότητες των κορυφών αυτών και εξαρτάται μόνο από κάποιο μέτρο πιθανότητας. Στην συντριπτική πλειοψηφία των περιπτώσεων, ένας Τυχαίος Γράφος αναπαρίσταται μέσω ενός εκ των επόμενων δύο μοντέλων: του μοντέλου Erdos-Renyi (G(n, m)) ή του μοντέλου Gilbert (G(n, p)). [14]

- **Erdos-Renyi (G(n, m)):** Με το μοντέλο αυτό επιλέγεται ένας από όλους τους δυνατούς γράφους με n στο πλήθος κορυφές και m ακμές, ισοπίθανα με τους υπόλοιπους. [14]
- **Gilbert (G(n, p)):** Με το μοντέλο αυτό ο γράφος αποτελείται από n στο πλήθος κορυφές, ενώ κάθε δυνατή ακμή σχηματίζεται ανεξάρτητα από τις άλλες με πιθανότητα p. [14]

Μερικά παραδείγματα κατηγοριών δικτύων όπου αυτά τα μοντέλα βρίσκουν εφαρμογή είναι τα κοινωνικά δίκτυα, τα βιολογικά δίκτυα, και ο Παγκόσμιος Ιστός. [14]

Χωρικά Δίκτυα (Spatial Networks)

Ένα Χωρικό Δίκτυο είναι ένα δίκτυο (ή γράφος) στον οποίο η πιθανότητα ύπαρξης μιας ακμής καθορίζεται από κάποιο μέτρο απόστασης μεταξύ των κόμβων στους οποίους προσπίπτει. [15] Στο πλαίσιο αυτής της εργασίας, από όλα τα μοντέλα αναπαράστασης Χωρικών Δικτύων επιλέχθηκαν ο Τυχαίος Γεωμετρικός Γράφος ή Random Geometric Graph (G(n, r)) και το μοντέλο Waxman (G(n, q, s)).

- **Random Geometric Graph (G(n, r)):** Με το μοντέλο αυτό επιλέγονται n σημεία σε τυχαίες θέσεις ενός επιπέδου, που θα αποτελέσουν τους κόμβους του γράφου, και όσοι κόμβοι έχουν απόσταση μικρότερη ή ίση με r μεταξύ τους ενώνονται με μία ακμή. [16]
- **Waxman (G(n, q, s)):** Με το μοντέλο αυτό επιλέγονται n σημεία σε τυχαίες θέσεις ενός επιπέδου (συνήθως του μοναδιαίου τετραγώνου), που θα αποτελέσουν τους κόμβους του γράφου, και κάθε δυνατή ακμή μεταξύ δύο κόμβων i, j σχηματίζεται με πιθανότητα $p(d_{ij}) = qe^{-sd_{ij}}$, όπου d_{ij} η απόσταση μεταξύ τους. [17]

Αυτά τα μοντέλα βρίσκουν εφαρμογή σε κατηγορίες δικτύων όπου η γεωγραφική απόσταση είναι σημαντική, όπως τα δίκτυα κινητής τηλεφωνίας, τα οδικά δίκτυα, και τα δίκτυα αλληλεπιδράσεων μεταξύ κάποιων ζώων. [16] [17]

Τέλος, θα πρέπει να σημειωθεί πως εισάγεται τυχαιότητα και στα δύο αυτά μοντέλα, επομένως μπορούν να θεωρηθούν υποκατηγορία των Τυχαίων Γράφων. Είναι δηλαδή συγχρόνως και Τυχαίοι και Χωρικοί γράφοι. [16] [17]

Μη-Κλιμακούμενα Δίκτυα (Scale-Free Networks)

Ένα Scale-Free δίκτυο είναι ένα δίκτυο, ή γράφος, στο οποίο η πιθανότητα ένας τυχαία επιλεγμένος κόμβος να έχει ακριβώς k γείτονες μειώνεται εκθετικά συναρτήσει του k, δηλαδή ακολουθεί την κατανομή Νόμου Δύναμης $P(k) \sim k^{-\gamma}$ (Power Law), όπου γ παράμετρος για την οποία συνήθως ισχύει $2 < \gamma < 3$. [18] [19] Το όνομα «Μη-κλιμακούμενο Δίκτυο» προέρχεται από το γεγονός ότι η εκθετική αυτή κατανομή των βαθμών κόμβων είναι ανεξάρτητη του μεγέθους (κλίμακας) του δικτύου.

Το μοντέλο που χρησιμοποιείται για αυτά τα δίκτυα ονομάζεται **Barabasi-Albert G(n, m)** όπου n το πλήθος των κόμβων και m το πλήθος των ακμών που προστίθενται κάθε φορά που εισέρχεται ένας νέος κόμβος ώστε να τον συνδέσουν με τους προϋπάρχοντες. [5] [20]

Τέτοια δίκτυα, όπου νέοι οι κόμβοι είναι πιθανότερο να συνδεθούν με τους ήδη υπάρχοντες κόμβους με μεγαλύτερο βαθμό (preferential attachment) [5], είναι πάρα πολύ κοινά στην φύση (ειδικά όταν πρόκειται για κοινωνικά δίκτυα). Το μοντέλο αυτό για παράδειγμα, μπορεί να χρησιμοποιηθεί για την αναπαράσταση διαφόρων τύπων δικτύων όπως ο Παγκόσμιος Ιστός, δίκτυα ανθρώπινων σεξουαλικών επαφών, και χημικά δίκτυα κυττάρων. [18]

Δίκτυα Μικρού-Κόσμου (Small-World Networks)

Τα δίκτυα Μικρού-Κόσμου αποτελούν ένα ενδιάμεσο μοντέλο μεταξύ των Κανονικών Γράφων (γράφοι στον οποίο όλοι οι κόμβοι έχουν τον ίδιο βαθμό), και των Τυχαίων Γράφων. Αυτό επιτυγχάνεται μέσω του μοντέλου **Watts-Strogatz G(n, k, p)** ως εξής: Ξεκινώντας από έναν Κανονικό Γράφο - Δαχτυλίδι με n κορυφές όπου η κάθε μια είναι συνδεδεμένη με τους k κοντινότερούς της γείτονες, η κάθε ακμή (v_i, v_j) έχει πιθανότητα p να μετασχηματιστεί στην (v_i, v_w) όπου v_w ένας διαφορετικός τυχαία επιλεγμένος κόμβος του γράφου, εφόσον δεν υπάρχει ήδη η ακμή (v_i, v_w). [21]

Τα δίκτυα αυτά παίρνουν το όνομά τους από το «φαινόμενο του μικρού κόσμου» ή αλλιώς «6 βαθμοί διαχωρισμού» (small-world phenomenon / 6 degrees of separation), αφού η μέση απόσταση μεταξύ δύο κόμβων σε έναν τέτοιο γράφο είναι περίπου 6 ακμές. Κάποια από τα παραδείγματα δικτύων που αναπαριστώνται με αυτό το μοντέλο είναι το νευρωνικό δίκτυο του σκώληκα Caenorhabditis elegans, το ηλεκτρικό δίκτυο των ΗΠΑ, και δίκτυα συνεργασιών μεταξύ ηθοποιών. [21]

2.2.3 Μετρικές Σύνθετων Δικτύων

Σε αυτήν την υποενότητα θα παρουσιαστούν οι μετρικές σύνθετων δικτύων που χρησιμοποιήθηκαν για τους σκοπούς της παρούσης εργασίας.

Κατανομή Βαθμών Κόμβων

Η **κατανομή βαθμών των κόμβων** δείχνει το πλήθος κόμβων που έχουν έναν συγκεκριμένο βαθμό κόμβου, για κάθε υπαρκτό στο δίκτυο βαθμό κόμβου. Είναι πολύ χρήσιμη μετρική καθώς αρκετά από τα μοντέλα που εξετάσαμε έχουν χαρακτηριστική μορφή κατανομής βαθμών κόμβου (παραδείγματος χάριν, η κατανομή βαθμών κόμβων ενός Scale-Free

δικτύου παρουσιάζει μορφή κατανομής Power Law). Ο **μέσος βαθμός κόμβου** αποτελεί τον μέσο όρο όλων των βαθμών των κόμβων του δικτύου, και αποτελεί κι αυτός χρήσιμη πληροφορία για το δίκτυο. [22]

Μέσο Μήκος Μονοπατιού

Το **συντομότερο μονοπάτι** μεταξύ δυο κόμβων είναι το μονοπάτι με το μικρότερο μήκος (δηλαδή με το μικρότερο πλήθος ακμών) που τους ενώνει. Ο όρος **μέσο μήκος μονοπατιού** αναφέρεται στον μέσο όρο του μήκους των συντομότερων μονοπατιών μεταξύ οποιωνδήποτε δύο κόμβων στον γράφο. [23] Μπορεί να διατυπωθεί μαθηματικά ως $l = \frac{1}{N(N-1)} \sum_{i \neq j} d(v_i, v_j)$, όπου $d(v_i, v_j)$ η απόσταση (συντομότερο μονοπάτι) μεταξύ των κόμβων v_i και v_j, και N το πλήθος των κόμβων του γράφου. [5]

Αποτελεί μια χρήσιμη μετρική που μπορεί να δώσει ενδεικτικές πληροφορίες για έναν γράφο. Ένα παράδειγμα αυτού του γεγονότος αποτελεί η ιδιότητα των 6 βαθμών διαχωρισμού, που αφορά το μέσο μήκος μονοπατιού και είναι χαρακτηριστική για τους Γράφους Μικρού-Κόσμου. [23] [5]

Διάμετρος

Η **διάμετρος** ενός γράφου είναι η μέγιστη απόσταση μεταξύ οποιωνδήποτε δύο κόμβων του γράφου. Μπορεί να διατυπωθεί μαθηματικά ως $diam = max_{v_i, v_j \in V}(d(v_i, v_j))$, όπου $d(v_i, v_j)$ η απόσταση μεταξύ των κόμβων v_i και v_j, και V το σύνολο των κόμβων του γράφου.

Η διάμετρος παρέχει σημαντικές πληροφορίες για ένα δίκτυο, ως το ανώτατο όριο των αποστάσεων μέσα σε αυτό. Για παράδειγμα, όταν πρόκειται για ένα δίκτυο επικοινωνιών, η διάμετρος υποδεικνύει τον χρόνο μετάδοσης και την εξασθένηση του σήματος κατά την αποστολή ενός μηνύματος στην χειρότερη περίπτωση. [24]

Κεντρικότητες

Οι κεντρικότητες είναι μετρικές που αποσκοπούν στον εντοπισμό των πιο σημαντικών (ή με την μεγαλύτερη επιρροή) κόμβων, σύμφωνα με διάφορα κριτήρια. [5] Οι κεντρικότητες που χρησιμοποιήθηκαν στην παρούσα εργασία είναι οι εξής:

- **Κεντρικότητα της Εγγύτητας (Closeness Centrality):** Αφορά την εγγύτητα ενός κόμβου με όλους τους άλλους του γράφου στον οποίο ανήκει, κι επομένως είναι αντιστρόφως ανάλογη του αθροίσματος των αποστάσεων από όλους τους υπόλοιπους κόμβους. Διατυπώνεται μαθηματικά ως $C_C(v_i) = \frac{1}{\sum_j d(v_i, v_j)}$. Ξεκινώντας με αφετηρία έναν κόμβο με μεγάλη κεντρικότητα εγγύτητας μπορούμε να φτάσουμε πιο γρήγορα σε άλλα μέρη του δικτύου. Αποτελεί μία από τις δημοφιλέστερες μετρικές για την εξέταση ενός δικτύου. [5]

- **Κεντρικότητα της Πληροφορίας (Information Centrality):** Αφορά την ροή της πληροφορίας μέσα σε ένα δίκτυο, αφού λαμβάνει υπ' όψιν όλα τα δυνατά μονοπάτια μεταξύ δύο κόμβων και αξιολογεί την ποσότητα της πληροφορίας που διέρχεται από αυτά (η οποία είναι αντιστρόφως ανάλογη του μήκους του μονοπατιού). Η κεντρικότητα

της πληροφορίας ενός κόμβου αποτελεί τον αρμονικό μέσο όρο της πληροφορίας που περιέχουν όλα τα μονοπάτια που ξεκινούν από αυτόν τον κόμβο, και διατυπώνεται μαθηματικά ως $C_I(i) = \frac{n}{\sum_{j=1}^{n} \frac{1}{I_{ij}}}$, όπου n το πλήθος των κόμβων του γράφου και I_{ij} η ποσότητα της πληροφορίας στο μονοπάτι μεταξύ των v_i και v_j (πίνακας αν υπάρχουν παραπάνω από ένα μονοπάτια). Σχετίζεται και με την κεντρικότητα της εγγύτητας, αφού ταυτίζεται με την κεντρικότητα της εγγύτητας της τρέχουσας ροής (current flow closeness centrality). [25] [26]

- **Κεντρικότητα των Αρμονικών (Harmonic Centrality):** Στοχεύει στην λύση του προβλήματος που εισάγει η πιθανή απουσία μονοπατιού από έναν κόμβο σε έναν άλλο στον υπολογισμό της κεντρικότητας της εγγύτητας. Αυτό το επιτυγχάνει αντικαθιστώντας την μέση απόσταση με τον αρμονικό μέσο των αποστάσεων, δηλαδή η κεντρικότητα των αρμονικών ενός κόμβου διατυπώνεται μαθηματικά ως $C_h(i) = \sum_{i \neq j} \frac{1}{d(v_i, v_j)}$. Αν και η διαφορά ίσως φαίνεται μικρή με την κεντρικότητα της εγγύτητας, στην πραγματικότητα αυτή η αλλαγή επιφέρει σημαντική διαφορά στα αποτελέσματα και διευκολύνει την ανάλυση ασθενώς συνεκτικών δικτύων. [27]

- **Ενδιαμεσική Κεντρικότητα (Betweenness Centrality):** Η ενδιαμεσική κεντρικότητα ενός κόμβου v ισούται με το πλήθος των συντομότερων μονοπατιών μεταξύ άλλων κόμβων που διέρχονται από αυτόν, και αποτελεί μέτρο της επιρροής ενός κόμβου στην ροή της πληροφορίας μεταξύ άλλων κόμβων. Διατυπώνεται μαθηματικά ως $C_{NB}(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$, όπου σ_{st} το συνολικό πλήθος των συντομότερων μονοπατιών από τον κόμβο s ως τον κόμβο t, και $\sigma_{st}(v)$ το πλήθος των συντομότερων μονοπατιών από τον s ως τον t που διέρχονται από τον v. [28] Ακριβώς αντίστοιχα ορίζεται και η ενδιαμεσική κεντρικότητα μιας ακμής e, $C_{EB}(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}$. [29] [30]

Συνάφεια (Assortativity)

Η **συνάφεια** αποτελεί μέτρο της τάσης των κόμβων ενός γράφου να συνδέονται με άλλους παρόμοιους κόμβους. Υπάρχουν διάφορα κριτήρια για την ομοιότητα των κόμβων, με το συνηθέστερο από αυτά να είναι ο βαθμός τους. Σε αυτήν περίπτωση, η συνάφεια βαθμού κόμβων αποτελεί μία *συσχέτιση Pearson*, και μπορεί να διατυπωθεί μαθηματικά ως $r = \frac{1}{\sigma_q^2}[(\sum_{jk} jke_{j,k}) - \mu_q^2]$, όπου e_jk η από κοινού κατανομή πιθανότητας δύο γειτόνων να έχουν βαθμό j και k αντίστοιχα, ενώ σ_q^2 και mu_q^2 η τυπική απόκλιση και η μέση τιμή της πιθανοτικής κατανομής βαθμού κόμβων του δικτύου. [31]

Η συνάφεια παίρνει τιμές στο διάστημα $[-1, 1]$, με το $r = 1$ να σημαίνει ότι όλοι οι κόμβοι έχουν γείτονες μόνο με τον ίδιο βαθμό με τον δικό τους, το $r = -1$ να σημαίνει ότι κανένας κόμβος δεν έχει γείτονα με τον ίδιο βαθμό με τον δικό τους, και $r = 0$ ότι ο οποιοσδήποτε κόμβος μπορεί να συνδεθεί τυχαία με οποιονδήποτε άλλο. Γενικά, θετικές τιμές της συνάφειας υποδηλώνουν ότι οι κόμβοι τείνουν να συνδέονται με όμοιούς τους, ενώ οι αρνητικές τιμές το αντίθετο. Η τάση αυτή κάποιου δικτύου εξαρτάται από την φύση του: παραδείγματος χάριν ένα δίκτυο φιλιών μεταξύ ανθρώπων μάλλον θα έχει θετική συνάφεια ενώ ένα δίκτυο που αντιπροσωπεύει μια τροφική αλυσίδα μάλλον θα έχει αρνητική συνάφεια. Έτσι, η συνάφεια μπορεί να αποτελέσει ισχυρή ένδειξη για την δομή, και άρα το εξελικτικό

μοντέλο που ακολουθεί ένα δίκτυο. [31]

2.2.4 Μέθοδοι Διαμέρισης Σύνθετων Δικτύων

Σε αυτήν την υποενότητα θα συζητηθούν κάποιοι τρόποι με τους οποίους ένα δίκτυο μπορεί να διαμεριστεί σε υποδίκτυα (ή υπογράφους).

Εντοπισμός Κοινοτήτων (Community Detection)

Μία **κοινότητα** μπορεί να οριστεί ως μια ομάδα κόμβων που αλληλεπιδρούν πιο συχνά και είναι πιο ισχυρά συνδεδεμένοι μεταξύ τους σε σύγκριση με τους άλλους κόμβους του δικτύου. Οι κόμβοι αυτοί μπορεί επίσης να είναι πιο κοντινοί μεταξύ τους, είτε υπό την έννοια της απόστασης είτε υπό την έννοια της ομοιότητας. [32] [33]

Ο **εντοπισμός κοινοτήτων** είναι η διαδικασία που αποσκοπεί στην εύρεση των κοινοτήτων μέσα σε ένα δίκτυο, και μπορεί να υλοποιηθεί με πολλούς διαφορετικούς αλγορίθμους. [33]. Η μετρική που χρησιμοποιείται ως επί των πλείστων για την αξιολόγηση της ποιότητας της διαμέρισης ενός δικτύου σε κοινότητες ονομάζεται **αρθρότητα (modularity)**, η οποία μετρά την πυκνότητα των ακμών στο εσωτερικό των κοινοτήτων σε σύγκριση με την πυκνότητα των ακμών των οποίων τα άκρα ανήκουν σε διαφορετικές κοινότητες. Η αρθρότητα παίρνει τιμές στο διάστημα $[-1, 1]$ και διατυπώνεται μαθηματικά ως $Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j)$, όπου m το πλήθος των ακμών του γράφου, A_{ij} η τιμή στην θέση (i,j) του πίνακα γειτνίασης του γράφου, k_i και k_j οι βαθμοί των κόμβων v_i και v_j αντίστοιχα, c_i καιc_j οι κοινότητες στις οποίες ανήκουν v_i και v_j αντίστοιχα, και δ η συνάρτηση $\delta(u, v)$ που ισούται με το 1 αν $u = v$ και με το 0 αλλιώς. [34] [8]

Ακολουθεί η περιγραφή των αλγορίθμων που χρησιμοποιήθηκαν στα πλαίσια της παρούσης εργασίας:

- **Girvan-Newmann:** Αλγόριθμος που βασίζεται στην ενδιαμεσική κεντρικότητα. Συγκεκριμένα, ο αλγόριθμος αυτός αφαιρεί τις ακμές με την μεγαλύτερη ενδιαμεσική κεντρικότητα, κάνοντας την υπόθεση πως οι ακμές αυτές αποτελούν γέφυρες μεταξύ διαφορετικών κοινοτήτων, και διαμερίζοντας έτσι τον γράφο σε ξεχωριστές κοινότητες. Η αφαίρεση των ακμών σταματάει είτε όταν ο γράφος έχει διαμεριστεί σε έναν συγκεκριμένο αριθμό κοινοτήτων που ορίζει ο εκάστοτε προγραμματιστής, είτε όταν έχουν αφαιρεθεί όλες οι ακμές από τον γράφο. Η ποιότητα της κάθε διαμέρισης που έχει προκύψει στην πορεία μπορεί να αξιολογηθεί μέσω μιας μετρικής (π.χ. του modularity), έτσι ώστε να βρεθεί η βέλτιστη. Ο Girvan-Newmann έχει χρονική πολυπλοκότητα $O(m^2 n)$, όπου n το πλήθος των κόμβων και m το πλήθος των ακμών. Ενώ αρχικά ορίστηκε για μη κατευθυνόμενους γράφους χωρίς βάρη, έχει γενικευτεί ώστε να μπορεί να εφαρμοστεί και σε κατευθυνόμενους γράφους και γράφους με βάρη. [30] [33] [8]

- **Walktrap:** Αλγόριθμος που βασίζεται στο γεγονός ότι οι τυχαίοι περίπατοι σε έναν γράφο τείνουν να «εγκλωβίζονται» σε ισχυρά συνδεδεμένα τμήματα του γράφου, που αντιστοιχούν σε κοινότητες. Οι τυχαίοι περίπατοι είναι διαδρομές σε έναν γράφο, οι οποίες σχηματίζονται ξεκινώντας από έναν κόμβο και επιλέγοντας κάθε φορά τον επόμενο τυχαία από τους γείτονες του τρέχοντα κόμβου. Αν ο τρέχοντας κόμβος είναι ο

v_i, τότε η πιθανότητα μετάβασης στον v_j ισούται με $P_{ij} = \frac{A_{ij}}{d(i)}$, όπου A_{ij} η τιμή της θέσης (i,j) του πίνακα γειτνίασης του γράφου και $d(i)$ ο βαθμός του v_i. Η ακολουθία των κόμβων που σχηματίζει τον περίπατο αποτελεί μια *αλυσίδα Markov*, με καταστάσεις τους κόμβους του γράφου, και πιθανότητα μετάβασης ίση με P_{ij} σε κάθε βήμα. Ο αλγόριθμος αφορά μόνο μη κατευθυνόμενους γράφους, με βάρη ή χωρίς. Έχει χρονική πολυπλοκότητα $O(mn^2)$ και χωρική πολυπλοκότητα $O(n^2)$ στην χειρότερη περίπτωση, ενώ στην μέση περίπτωση έχει χρονική πολυπλοκότητα $O(n^2 log(n))$ και $O(n^2)$ χωρική. [35]

- **Infomap:** Αλγόριθμος που χρησιμοποιεί το ρεύμα πιθανότητας τυχαίων περιπάτων σε έναν γράφο (δηλαδή τις πιθανότητες εισόδου και εξόδου για κάθε κόμβο), ώστε να τον τμηματοποιήσει σύμφωνα με την ροή πληροφορίας σε αυτόν. Μία ομάδα κόμβων μέσα στην οποία η πληροφορία ρέει γρήγορα και εύκολα θεωρείται ισχυρά συνδεδεμένη, και αποτελεί μία κοινότητα. Συγκεκριμένα, ο αλγόριθμος αυτός χρησιμοποιεί μία συνάρτηση χάρτη (map function) που παίρνει ως είσοδο μια διαμέριση του γράφου σε υπογράφους και υπολογίζει το μέσο μήκος ενός τυχαίου περιπάτου σε αυτήν, με σκοπό να επιλεγεί η διαμέριση με το μικρότερο δυνατό μέσο μήκος περιπάτου. Τότε οι υπογράφοι της διαμέρισης αυτής αντιστοιχούν σε κοινότητες, αφού το γεγονός ότι οι τυχαίοι περίπατοι είναι συντομότεροι σε αυτούς σημαίνει ότι (ανά υπογράφο-κοινότητα) οι συνδέσεις μεταξύ των κόμβων είναι πυκνότερες και η πληροφορία ρέει γρηγορότερα. Ο Infomap μπορεί να χρησιμοποιηθεί για κάθε κατηγορία γράφων (κατευθυνόμενων ή μη, και με βάρη ή χωρίς). [36]

- **Leiden:** Αλγόριθμος που αποτελεί μια βελτιωμένη έκδοση ενός άλλου αλγορίθμου εντοπισμού κοινοτήτων, του Louvain. Ο Louvain βασίζεται στην μεγιστοποίηση της μετρικής modularity. Αρχικά κάθε κόμβος ορίζεται ως διαφορετική κοινότητα, και γίνονται αλλεπάλληλες δοκιμές συνένωσης κοινοτήτων ώστε η προκύπτουσα κοινότητα να έχει μεγαλύτερο modularity από τις αρχικές. Όταν το modularity δεν γίνεται να αυξηθεί άλλο, αντικαθιστά κάθε κοινότητα μέ έναν κόμβο και επαναλαμβάνει την διαδικασία ώσπου το modularity να μεγιστοποιηθεί όσο δύναται. Η διαδικασία αυτή ενίοτε έχει ως αποτέλεσμα οι τελικές κοινότητες να μην είναι καλά συνδεδεμένες ή ακόμη και να μην είναι συνεκτικές. Ο Leiden δεν παρουσιάζει το πρόβλημα των μη-συνεκτικών κοινοτήτων, και επιπλέον προσφέρει καλύτερα αποτελέσματα ενώ η εκτέλεσή του είναι γρηγορότερη. [34] [37] Ο αλγόριθμος αυτός αφορά μόνο μη-κατευθυνόμενους γράφους, με βάρη ή χωρίς. [33]

- **Spectral Clustering:** Αλγόριθμος που ομαδοποιεί τους κόμβους σε k κοινότητες βάσει της ομοιότητάς τους, η οποία υπολογίζεται μέσω των k πρώτων ιδιοδιανυσμάτων του Λαπλασιανού πίνακα του γράφου ($L = D - W$, όπου D διαγώνιος πίνακας που περιέχει τον βαθμό του κάθε κόμβου και W ο πίνακας γειτνίασης του γράφου). Εδώ μια ενδιαφέρουσα παρατήρηση είναι πως οι ιδιοτιμές του Λαπλασιανού πίνακα σχετίζονται ποιοτικά με τον γράφο με διάφορους τρόπους. Για παράδειγμα, σε έναν συνδεδεμένο η γράφο η ιδιοτιμή 0 έχει πάντα πολλαπλότητα ίση με 1, ενώ η πρώτη μη-μηδενική ιδιοτιμή αποκαλύπτει το πόσο καλά συνδεδεμένος είναι ο γράφος, με μεγαλύτερες τιμές

να αντιστοιχούν σε περισσότερες συνδέσεις. Αξίζει επίσης να σημειωθεί ότι οι ιδιοτιμές βρίσκονται σε αύξουσα σειρά όσον αφορά τις τιμές τους, με την πρώτη να ισούται πάντα με το 0. Ο αλγόριθμος Spectral Clustering δεν χρησιμοποιείται αποκλειστικά για εύρεση κοινοτήτων σε γράφους, αλλά για ένα μεγάλο εύρος προβλημάτων. Εκτελώντας τον όμως με όρισμα τον πίνακα γειτνίασης ενός γράφου, μπορούμε να λάβουμε μια πολύ καλή διαμέριση του γράφου αυτού. [38] [39] Εφαρμόζεται κυρίως σε μη-κατευθυνόμενους γράφους, αλλά είναι δυνατό να τροποποιηθεί ώστε να εφαρμοστεί και σε κατευθυνόμενους. [40] Επιπλέον, μπορεί να εφαρμοστεί είτε σε γράφους με βάρη είτε χωρίς. [39]

2.3 Μηχανική Μάθηση

2.3.1 Ορισμός Μηχανικής Μάθησης

Η **Μηχανική Μάθηση** αποτελεί ένα πεδίο της επιστήμης υπολογιστών που αφορά αλγορίθμους και τεχνικές για την αυτοματοποίηση της λύσης σύνθετων προβλημάτων που είναι δύσκολο να λυθούν με την χρήση κλασικών προγραμματιστικών τεχνικών. [41]

Συγκεκριμένα, σε αντίθεση με την κλασική προσέγγιση όπου ο προγραμματιστής ορίζει ένα συγκεκριμένο σύνολο εντολών για την επίλυση ενός προβλήματος, οι αλγόριθμοι μηχανικής μάθησης λύνουν το πρόβλημα μόνοι τους, δημιουργώντας ένα σύνολο κανόνων που ονομάζεται **μοντέλο** και το οποίο εκπαιδεύεται από τα δεδομένα ώστε να μπορεί να μπορεί να κάνει προβλέψεις για μία νέα είσοδο. Έτσι διευκολύνεται η επίλυση περίπλοκων προβλημάτων όπως η κατηγοριοποίηση δεδομένων (παραδείγματος χάριν κατηγοριοποίηση των email σε spam και μη), η αναγνώριση κειμένου από εικόνα, η αναγνώριση φωνής, η αυτόματη οδήγηση, και άλλα. [41]

Οι αλγόριθμοι Μηχανικής Μάθησης τείνουν να δίνουν πιο ακριβή αποτελέσματα, καθώς οι κανόνες που δημιουργούν βασίζονται στα αποκλειστικά στα δεδομένα και έτσι είναι πιο αντικειμενικοί από αυτούς που ορίζει ένας πιθανώς προκατειλημμένος προγραμματιστής. Από την άλλη, το γεγονός πως δεν γνωρίζουμε ακριβώς το πως έφτασαν σε μία απόφαση αποτελεί πρόβλημα, καθώς μειώνει την εμπιστοσύνη μας σε αυτή, ειδικά όταν πρόκειται για ένα σημαντικό πρόβλημα όπως μία ιατρική διάγνωση. [41] Το πρόβλημά αυτό (που ονομάζεται Πρόβλημα του Μαύρου Κουτιού ή Black Box Problem) μελετάται και γίνεται προσπάθεια να λυθεί μέσω ενός ερευνητικού πεδίου που ονομάζεται Εξηγήσιμη Τεχνητή Νοημοσύνη (Explainalble AI). [42]

2.3.2 Επιβλεπόμενη και Μη Επιβλεπόμενη Μηχανική Μάθηση

Τα δεδομένα στα οποία εκπαιδεύεται ένας αλγόριθμος Μηχανικής Μάθησης μπορούν να είναι **επισημασμένα (labeled)** ή **μη επισημασμένα (unlabeled)**, ανάλογα με το αν έχουν ετικέτα ή όχι, αντίστοιχα. Η **ετικέτα (label)** είναι ένα χαρακτηριστικό ή μία κατηγορία στην οποία ανήκει ένα δεδομένο, και η οποία ζητείται να προβλεφθεί για μία νέα είσοδο σε ένα πρόβλημα Μηχανικής Μάθησης που βασίζεται σε επισημασμένα δεδομένα. Έτσι, η Μηχανική Μάθηση κατηγοριοποιείται σε **Επιβλεπόμενη (Supervised)** όταν πρόκειται για

αλγορίθμους που προορίζονται για προβλήματα με επισημασμένα δεδομένα, και σε **Μη Επιβλεπόμενη (Unsupervised)** όταν τα δεδομένα είναι μη επισημασμένα. [43]

Η **Επιβλεπόμενη Μάθηση** αφορά προβλήματα **ταξινόμησης (classification)**, τα οποία αναλύονται στην επόμενη υποενότητα, και προβλήματα **παλινδρόμησης (regression)**, δηλαδή προβλήματα όπου βάσει των τιμών των χαρακτηριστικών μίας εισόδου ζητείται να προβλεφθεί μια αριθμητική τιμή ως έξοδος (η οποία θα αποτελέσει την ετικέτα). Παραδείγματα τέτοιων προβλημάτων είναι ο διαχωρισμός των email σε spam και όχι spam (ταξινόμηση), και η πρόβλεψη των μελλοντικών εσόδων μιας επιχείρησης βάσει των παρελθόντων εσόδων (παλινδρόμηση). [43]

Η **Μη Επιβλεπόμενη Μάθηση** αφορά κυρίως προβλήματα **συσταδοποίησης (clustering)**, δηλαδή προβλήματα όπου τα δεδομένα πρέπει να ομαδοποιηθούν βάσει της ομοιότητάς τους, **προβλήματα συσχέτισης (association)**, δηλαδή προβλήματα όπου πρέπει να βρεθούν σχέσεις και συσχετίσεις μεταξύ των δεδομένων βάσει των τιμών των χαρακτηριστικών τους, και προβλήματα **μείωσης της διαστατικότητας (dimensionality reduction)**, δηλαδή προβλήματα όπου το πλήθος των χαρακτηριστικών του προβλήματος πρέπει να μειωθεί χωρίς να επηρεαστεί η ποιότητα της πληροφορίας που παρέχουν. Παραδείγματα τέτοιων προβλημάτων είναι η τμηματοποίηση της αγοράς βάσει των ομοιοτήτων μεταξύ των καταναλωτών (συσταδοποίηση), η υλοποίηση ενός συστήματος συστάσεων για έναν ιστότοπο ηλεκτρονικού εμπορίου (συσχέτιση), και η αφαίρεση θορύβου από δεδομένα εικόνας με σκοπό την βελτίωση της ευκρίνειας (μείωση της διαστατικότητας). [43]

2.3.3 Ταξινομητές

Ορισμός Ταξινομητή

Το πρόβλημα της **ταξινόμησης** αποτελεί την επιλογή της κατηγορίας που ανήκει μια νέα είσοδος, από ένα σύνολο δυνατών κατηγοριών. Ένας **ταξινομητής** είναι το μοντέλο, το οποίο εκπαιδεύεται από ένα σύνολο επισημασμένων δεδομένων ώστε να μπορεί να λύσει το πρόβλημα της ταξινόμησης. Το σύνολο των δυνατών κατηγοριών στις οποίες μπορεί να κατατάξει ένας ταξινομητής μια νέα είσοδο είναι το σύνολο των διαφορετικών ετικετών που παρουσιάζονται στα δεδομένα εκπαίδευσης. Για την κατασκευή ενός ταξινομητή μπορούν να χρησιμοποιηθούν πολλές διαφορετικές τεχνικές μηχανικής μάθησης, όπως τα Νευρωνικά Δίκτυα, τα Τυχαία Δάση, τα Μαρκοβιανά μοντέλα, και άλλα. [44]

Μετρικές επίδοσης ταξινομητών

Η επίδοση των ταξινομητών αξιολογείται ως επί των πλείστων με τις μετρικές **ευστοχία (precision), ανάκληση (recall), ακρίβεια (accuracy)**, και **F1 score**. [45] Εδώ είναι αναγκαίο να οριστούν οι εξής έννοιες για ένα **δυαδικό ταξινομητή** (δηλαδή που επιλέγει μόνο μεταξύ δύο κατηγοριών), αν η μία κατηγορία είναι τα θετικά (positive) και η άλλη τα αρνητικά (negative):

- **Αληθώς Θετικό (True Positive):** Η είσοδος ανήκει στην θετική κατηγορία, και ταξινομείται ως θετική

- **Ψευδώς Θετικό (False Positive):** Η είσοδος ανήκει στην αρνητική κατηγορία, αλλά ταξινομείται ως θετική
- **Αληθώς Αρνητικό (True Negative):** Η είσοδος ανήκει στην αρνητική κατηγορία, και ταξινομείται ως αρνητική
- **Ψευδώς Αρνητικό (False Negative):** Η είσοδος ανήκει στην θετική κατηγορία, και ταξινομείται ως αρνητική

[46] [47]

	Actual Positive Class	**Actual Negative Class**
Predicted Positive Class	True positive (*tp*)	False negative (*fn*)
Predicted Negative Class	False positive (*fp*)	True negative (*tn*)

Σχήμα 2.8: *Πίνακας που αποτελεί οπτικοποίηση του ορισμού των ανωτέρω όρων* [46]

Έτσι μπορούμε τώρα να ορίσουμε τις παραπάνω μετρικές ως εξής:

- **Precision** = $\frac{TP}{TP+FP}$ (οι αληθώς θετικές είσοδοι προς όλες τις εισόδους που έχουν ταξινομηθεί ως θετικές)
- **Recall** = $\frac{TP}{TP+TN}$ (οι αληθώς θετικές είσοδοι ως προς όλες που είναι πραγματικά θετικές)
- **Accuracy** = $\frac{TP+TN}{TP+FP+TN+FN}$ (οι σωστά ταξινομημένες είσοδοι προς όλες τις εισόδους)
- **F1 Score** = $2\frac{precision \cdot recall}{precision+recall}$ (αρμονικός μέσος όρος των precision και recall)

[46] [47]

Όταν πρόκειται για ταξινομητές που κατηγοριοποιούν τις εισόδους σε παραπάνω από δύο κατηγορίες (**multi-class classification**), το precision και το recall υπολογίζονται είτε με την μακροσκοπική προσέγγιση (macro precision και macro recall) είτε με την μικροσκοπική (micro precision και micro recall). Προτιμάται η μακροσκοπική, αφού με την μικροσκοπική το precision καταλήγει να ισούται με το recall, όπως και το accuracy με το F1 score. [47]

Στην **μακροσκοπική προσέγγιση**, υπολογίζεται το precision και το recall για κάθε κατηγορία, και ύστερα υπολογίζεται ο μέσος όρος τους ώστε να βγει το συνολικό αποτέλεσμα για τον ταξινομητή. Οι έννοιες των αληθώς θετικών, ψευδώς θετικών, και ψευδώς αρνητικών γενικεύονται ως εξής: τα **αληθώς θετικά** για μια κατηγορία k είναι όσα όντως ανήκουν σε αυτή και ταξινομήθηκαν σε αυτή, τα **ψευδώς θετικά** είναι όσα ταξινομήθηκαν σε αυτή αλλά ανήκουν σε άλλη κατηγορία, ενώ τα **ψευδώς αρνητικά** είναι όσα ταξινομήθηκαν σε άλλη κατηγορία αλλά ανήκουν στην k. (Σημείωση: Ο πίνακας που περιέχει το πλήθος των αληθώς θετικών, ψευδώς θετικών, και ψευδώς αρνητικών για όλες τις κατηγορίες ονομάζεται **πίνακας σύγχυσης (confusion matrix)**.) Οι τύποι για τον υπολογισμό των ολικών **precision** και **recall** παραμένουν οι ίδιοι, ενώ χρησιμοποιούνται οι γενικευμένες έννοιες που μόλις αναφέρθηκαν. Ομοίως με πριν, το **F1 score** υπολογίζεται ως ο αρμονικός μέσος όρος του ολικού precision και του ολικού recall, ενώ το **accuracy** εξακολουθεί να ισούται με το πλήθος των ορθά ταξινομημένων εισόδων προς το πλήθος όλων των εισόδων. [47]

Παραδείγματα ταξινομητών

Παρακάτω παρουσιάζονται συνοπτικά οι ταξινομητές που χρησιμοποιήθηκαν στα πλαίσια της παρούσης εργασίας:

- **Ταξινομητής Διανυσμάτων Στήριξης (Support Vector Classifier ή SVC) και Γραμμικός Ταξινομητής Διανυσμάτων Στήριξης (Linear SVC):** Ο ταξινομητής αυτός είναι μία **Μηχανή Διανυσμάτων Στήριξης (Support Vector Machine ή SVM)**, εξού και το όνομά του. Τα SVM είναι μοντέλα, των οποίων ο στόχος του αλγορίθμου εκπαίδευσής τους είναι η εύρεση του **βέλτιστου ορίου (optimal hyperplane)** μεταξύ των περιοχών που σχηματίζουν τα μέλη κάθε κατηγορίας σε έναν γεωμετρικό χώρο που ορίζεται με τους άξονές του να αντιστοιχούν στα χαρακτηριστικά των στοιχείων του προβλήματος (δηλαδή αν τα στοιχεία έχουν παραδείγματος χάριν 7 χαρακτηριστικά, θα έχουμε έναν 7-διάστατο χώρο με 7 άξονες που ο καθένας αντιστοιχεί στις τιμές που παίρνει ένα χαρακτηριστικό). Τα μέλη της κάθε κλάσης που βρίσκονται πιο κοντά στο όριο αποτελούν τα **διανύσματα στήριξης (support vectors)** από τα οποία παίρνει το όνομά του ο ταξινομητής. Ως βέλτιστο όριο θεωρείται το όριο εκείνο το οποίο μεγιστοποιεί την απόσταση μεταξύ του ίδιου και των διανυσμάτων στήριξης. Εάν οι κλάσεις είναι γραμμικά διαχωρίσιμες, δηλαδή το όριο μεταξύ τους μπορεί να είναι μια ευθεία, τότε μπορούμε να χρησιμοποιήσουμε ένα γραμμικό SVM (**Linear SVC**) που αναζητά συγκεκριμένα μια ευθεία ως όριο (και όχι μια οποιαδήποτε καμπύλη όπως η γενική περίπτωση του **SVC**). Και στις δύο περιπτώσεις, ο χώρος χωρίζεται σε περιοχές που αντιστοιχούν σε κάθε κατηγορία, και μια νέα είσοδος ταξινομείται βάσει της περιοχής στην οποία τοποθετείται σύμφωνα με τις τιμές των χαρακτηριστικών της. [48]

- **Ταξινομητής *k*-πλησιέστερων γειτόνων (*k*-nearest neighbours classifier):** Ο ταξινομητής αυτός επίσης βασίζεται στην θέση μιας νέας εισόδου σε σχέση με αυτή των δεδομένων εκπαίδευσης σε έναν γεωμετρικό χώρο που ορίζεται ακριβώς όπως περιγράφηκε για τα SVM. Σε αυτήν την περίπτωση, εντοπίζονται οι ***k* κοντινότεροι γείτονες** μιας νέας εισόδου (δηλαδή τα δεδομένα εκπαίδευσης του προβλήματος που έχουν την μικρότερη απόσταση από την νέα είσοδο), και αποφασίζεται σε ποια κατηγορία ανήκει η νέα είσοδος ανάλογα με το ποια είναι η κατηγορία στην οποία ανήκει η πλειοψηφία των k κοντινότερων γειτόνων. Κάποιες φορές, για την λήψη της απόφασης αυτής λαμβάνονται υπ' όψιν και οι ακριβείς αποστάσεις των γειτόνων από την είσοδο ως βάρη, με μικρότερες αποστάσεις να αντιστοιχούν σε μεγαλύτερα βάρη. Η επιλογή του κατάλληλου k έχει πολύ μεγάλη σημασία για την επίδοση του ταξινομητή αυτού, και πρέπει να γίνεται με προσοχή. Υπάρχουν αρκετές μέθοδοι για την επίτευξη του στόχου αυτού, με την απλούστερη να είναι η επανάληψη της ταξινόμησης με διαφορετικές τιμές του k, και η επιλογή του k εκείνου που επιφέρει τα καλύτερα αποτελέσματα σύμφωνα με τις μετρικές επίδοσης του ταξινομητή που έχουν επιλεγεί. [49]

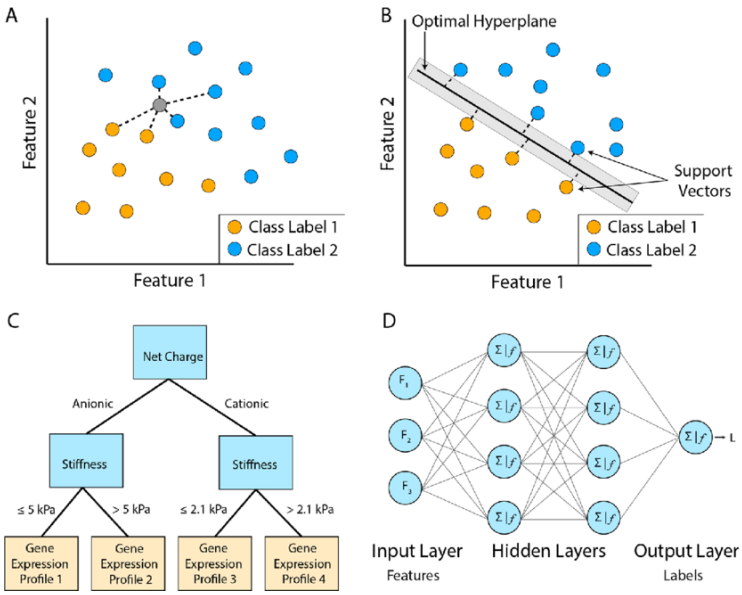
- **Ταξινομητές Συνόλου (Ensemble classifiers):** Οι ταξινομητές συνόλου είναι μία κατηγορία ταξινομητών που αποτελούνται από και χρησιμοποιούν έναν συνδυασμό άλλων, απλούστερων ταξινομητών με σκοπό την επίτευξη καλύτερης συνολικής επίδοσης. [48] Μερικοί από αυτούς είναι οι εξής:

α) AdaBoost: Ο ταξινομητής αυτός βασίζεται στην αντιστοίχιση βαρών σε όλα τα δεδομένα εκπαίδευσης, τα οποία αρχικά είναι όλα ίσα. Μετά την εκπαίδευση κάθε (αδύναμου) ταξινομητή από τον οποίο αποτελείται, υπολογίζεται το σφάλμα για κάθε ένα από τα δεδομένα εκπαίδευσης. Αν υπάρχει σφάλμα για ένα δεδομένο τότε το βάρος του αυξάνεται, ενώ αν δεν υπάρχει σφάλμα μειώνεται. Αυτό αναγκάζει τους ταξινομητές να επικεντρωθούν στα αδύναμα σημεία τους, με αποτέλεσμα να βελτιωθούν. Έτσι προκύπτει τελικά ένας συνολικά ισχυρότερος ταξινομητής. [50]

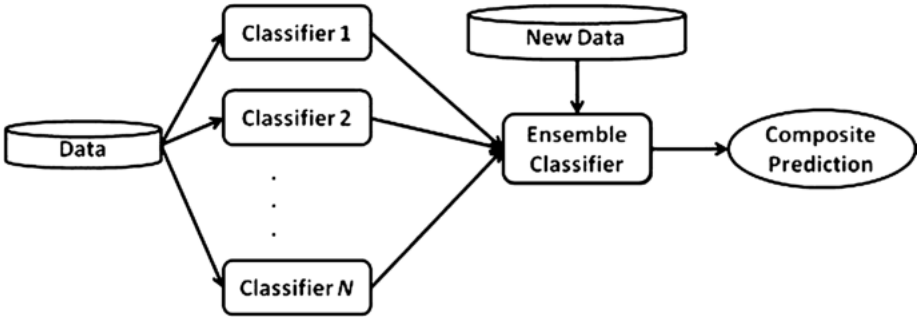
β) Gradient boosting: Ο ταξινομητής αυτός αποτελείται από πολλούς επί μέρους απλούς ταξινομητές, οι οποίοι εκπαιδεύονται διαδοχικά. Όλοι οι επί μέρους ταξινομητές είναι συσχετισμένοι με το αρνητικό της κλίσης της συνάρτησης απώλειας (loss function) του συνολικού ταξινομητή, έτσι ώστε μετά την εκπαίδευση του κάθε επί μέρους ταξινομητή η τιμή της συνάρτησης απώλειας να μειώνεται. Έτσι ο συνολικός ταξινομητής συνεχώς βελτιώνεται. Η συνάρτηση απώλειας μπορεί να είναι οποιαδήποτε, και η επιλογή της εξαρτάται από το εκάστοτε πρόβλημα. [51]

γ) Bagging (Bootstrap Aggregating): Ο ταξινομητής αυτός αποτελείται από πολλούς (συνήθως ασταθείς) ταξινομητές, ο κάθε ένας εκ των οποίων έχει εκπαιδευτεί σε ένα διαφορετικό δείγμα bootstrap των δεδομένων εκπαίδευσης. Ένα δείγμα bootstrap είναι ένα υποσύνολο των δεδομένων εκπαίδευσης που έχει επιλεγεί τυχαία με ομοιόμορφη κατανομή πιθανότητας, ενώ ένα δεδομένο που έχει επιλεγεί για ένα bootstrap δεν αφαιρείται από το σύνολο και μπορεί να επανεπιλεγεί για ένα άλλο bootstrap. Έτσι, κάθε bootstrap περιέχει ένα μείγμα μοναδικών και κοινών δεδομένων εκπαίδευσης. Η τελική απόφαση του συνολικού ταξινομητή λαμβάνεται μέσω ψηφοφορίας (voting), δηλαδή αποφασίζεται ότι η είσοδος ανήκει στην κατηγορία που επιλέχθηκε από την πλειοψηφία των επί μέρους ταξινομητών. Με αυτόν τον τρόπο ο συνολικός ταξινομητής είναι πιο ακριβής από τους επί μέρους, καθώς μειώνονται τα λάθη λόγω overfitting (δηλαδή υπερβολικής προσαρμογής του ταξινομητή στα δεδομένα εκπαίδευσης), και στατιστικής διασποράς. [52]

δ) Τυχαίο Δάσος (Random Forest): Ο ταξινομητής αυτός αποτελείται από πολλά **Δέντρα Απόφασης (Decision Trees)**, τα οποία είναι απλοί ταξινομητές που λειτουργούν χωρίζοντας αναδρομικά τα δεδομένα σε υποσύνολα. Συγκεκριμένα, κατασκευάζεται ένα δέντρο ξεκινώντας από την ρίζα, και ανάλογα με την τιμή κάποιου από τα χαρακτηριστικά των δεδομένων (αποφασίζεται ποιο βάσει κάποιου κριτηρίου κέρδους) ο κάθε κόμβος αποκτά δύο παιδιά, ώστε να εξαντληθούν τα χαρακτηριστικά και να φτάσουμε στα φύλλα του δέντρου. Μια νέα είσοδος ταξινομείται ακολουθώντας την διαδρομή από την ρίζα του δέντρου που ορίζουν οι τιμές των χαρακτηριστικών της. [53] Το Τυχαίο Δάσος εκπαιδεύει τα Δέντρα από τα οποία αποτελείται με διαφορετικά υποσύνολα δεδομένων εκπαίδευσης αλλά και χαρακτηριστικών, τα οποία επιλέγονται από το σύνολο των δεδομένων εκπαίδευσης και των χαρακτηριστικών ακριβώς όπως περιγράφηκε και για τον ταξινομητή Bagging. Στο τέλος το αποτέλεσμα της κατηγοριοποίησης βγαίνει πάλι με ψηφοφορία, δηλαδή επιλέγεται η απόφαση της πλειοψηφίας των Δέντρων. Αυτή η διαδικασία προσφέρει πολύ καλύτερα αποτελέσματα από την χρήση ενός μοναδικού Δέντρου Απόφασης. [54]



Σχήμα 2.9: *Α: Ένας ταξινομητής k-γειτόνων με $k = 4$ σε πρόβλημα ταξινόμησης όπου τα δεδομένα έχουν 2 χαρακτηριστικά, Β: Ένα γραμμικό SVC σε πρόβλημα ταξινόμησης όπου τα δεδομένα έχουν 2 χαρακτηριστικά, C: Ένα Δέντρο Απόφασης σε πρόβλημα ταξινόμησης όπου τα δεδομένα έχουν 2 χαρακτηριστικά, D: Ένα Νευρωνικό Δίκτυο με ένα επίπεδο εισόδου, δύο κρυφά επίπεδα, και ένα επίπεδο εξόδου (δεν εξετάστηκε στην παρούσα ενότητα)*
[55]



Σχήμα 2.10: *Μία γενικευμένη περιγραφή της δομής και του τρόπου λειτουργίας των Ταξινομητών Συνόλου: Ο Ταξινομητής Συνόλου αποτελείται από πολλούς επί μέρους ταξινομητές, οι οποίοι εκπαιδεύονται στα δεδομένα του προβλήματος. Όταν πρόκειται να ταξινομηθεί μια νέα είσοδος, η λειτουργία ή τα αποτελέσματα των επιμέρους ταξινομητών συνδυάζονται κατάλληλα, ώστε ο Ταξινομητής Συνόλου να καταλήξει σε μία συνολική απόφαση.*
[56]

Κεφάλαιο 3

Σχετική Βιβλιογραφία

Στο κεφάλαιο αυτό παρουσιάζεται συνοπτικά μεγάλο μέρος της υπάρχουσας έρευνας και βιβλιογραφίας σχετικά με την ταξινόμηση δικτύων σε εξελικτικά μοντέλα, η οποία είναι αρκετά περιορισμένη. Επιπλέον, παρουσιάζεται το άρθρο σχετικά με την συχνότητα εμφάνισης του Scale-Free μοντέλου σε πραγματικά δίκτυα, το οποίο ενέπνευσε το αντικείμενο έρευνας της παρούσας διπλωματικής, καθώς και κάποια άλλα άρθρα που σχετίζονται με αυτό το θέμα.

3.1 Γενικές αναφορές στην ταξινόμηση δικτύων σε μοντέλα

3.1.1 Ταξινόμηση γράφων μέσω θεωρητικών μετρικών

Οι Linhong Zhu, Wee Keong Ng, και Shuguo Han μελέτησαν την δυνατότητα ταξινόμησης ενός γράφου βάσει ορισμένων χαρακτηριστικών του, και επιχείρησαν να προσδιορίσουν ποια είναι τα βέλτιστα χαρακτηριστικά για την πραγματοποίηση της ταξινόμησης αυτής. Για την επίτευξη του στόχου τους χρησιμοποίησαν τις εξής 12 μετρικές ως τα χαρακτηριστικά που αντιπροσωπεύουν κάθε γράφο: τον Συντελεστή Συσταδοποίησης (Clustering Coefficient), το Μέσο Μήκος Μονοπατιού, την Καθολική Αποδοτικότητα (Global Efficiency), την Κατανομή Βαθμών Κόμβων, την Πυρηνικότητα (Coreness), την Πυκνότητα (Density), την Περιφέρεια (Girth), την Περίμετρο (Circumference), την Συνεκτικότητα, τον Λόγο Συμπίεσης Κορυφών (Vertex Compression Ratio), τον Λόγο Συμπίεσης Ακμών (Edge Compression Ratio), και την Ακυκλικότητα (Acyclicity). Ανέπτυξαν, επιπλέον, έναν αλγόριθμο επιλογής των πιο αντιπροσωπευτικών και διαχωρίσιμων από τα υπόλοιπα χαρακτηριστικών, αντιμετωπίζοντας την επιλογή αυτή ως ένα πρόβλημα βελτιστοποίησης. [57]

Ο αλγόριθμος αυτός επέστρεψε 8 από τα παραπάνω χαρακτηριστικά, τα οποία ύστερα χρησιμοποίησαν για την εκπροσώπηση συνολικά 300 γράφων (πραγματικών, και συνθετικών), αντιστοιχίζοντας ένα διάνυσμα χαρακτηριστικών σε κάθε γράφο. Σε κάθε έναν από αυτούς τους γράφους αντιστοιχίστηκε επίσης μια ετικέτα εκ τριών δυνατών επιλογών (INTERVAL, HOPI, και DUAL) ανάλογα με το ποια από αυτές αποτιμήθηκε ως ο βέλτιστος δείκτης-λύση στο πρόβλημα επιλογής δείκτη προσιτότητας ή reachability index (δηλαδή στο πρόβλημα επιλογής της κατάλληλης δομής δεδομένων για την αποθήκευση πληροφορίας σχετικά με τη δυνατότητα μετάβασης από κάθε κόμβο προς τους υπόλοιπους), μετά από την αξιολόγηση της προσιτότητας του γράφου μέσω του υπολογισμού του συνολικού χρόνου πρόσβασης ενός εκατομμυρίου τυχαίων ερωτημάτων στον γράφο για κάθε μία από τις πιθανές

επιλογές. Η πλειοψηφία των 300 αυτών γράφων χρησιμοποιήθηκε ως τα δεδομένα εκπαίδευσης δύο ταξινομητών, ενός ταξινομητή k-πλησιέστερων γειτόνων (k-NN) κι ενός αντίστροφου ταξινομητή k-πλησιέστερων γειτόνων (reverse k-nearest neighbours classifier ή Rk-NN), ενώ οι υπόλοιποι χρησιμοποιήθηκαν ως τα δεδομένα δοκιμής των παραπάνω ταξινομητών. [57]

Τέλος συγκρίθηκε η επίδοση των δύο ταξινομητών μέσω των μετρικών accuracy και micro precision, καθώς και του χρόνου εκτέλεσής τους και της μνήμης που κατανάλωσαν. Βρέθηκε πως ο k-NN έχει μικρότερο χρόνο εκτέλεσης και καταναλώνει λιγότερη μνήμη, αλλά ο Rk-NN αποδίδει καλύτερα ως προς το accuracy και το micro precision (αν και τα αποτελέσματά του είναι πιο ασταθή). Τα αποτελέσματα έδειξαν πως η ταξινόμηση γράφων βάσει κατάλληλων χαρακτηριστικών τους είναι πράγματι εφικτή. [57]

Η προηγούμενη δημοσίευση δεν αφορούσε άμεσα την ταξινόμηση γράφων σε εξελικτικά μοντέλα, όμως απέδειξε πως η ταξινόμηση γράφων μέσω μετρικών είναι δυνατή, μια τεχνική που χρησιμοποιήθηκε και στην παρούσα εργασία. Όσον αφορά την επιλογή των κατάλληλων μετρικών για το δικό μας πρόβλημα, μπορούμε να συμβουλευτούμε πηγές όπως το άρθρο των Reka Albert και Albert-Laszlo Barabasi [58] ή το βιβλίο των Sergey Dorogovtsev και José Fernando Mendes [59], τα οποία περιγράφουν αναλυτικά αρκετά από τα εξελικτικά μοντέλα που χρησιμοποιήσαμε και προσδιορίζουν τα πιο αντιπροσωπευτικά χαρακτηριστικά τους. Υπάρχει, ακόμη, και η δημοσίευση των Gábor Szárnyas, Zsolt Kővári, Ágnes Salánki, και Dániel Varró [60] στην οποία εξετάζεται μία πληθώρα μετρικών σχετικά με το πόσο χαρακτηριστικές είναι για έναν γράφο, με κριτήρια την ομοιογένεια τους (δηλαδή το κατά πόσο τα δίκτυα που ανήκουν σε μία κατηγορία έχουν παρόμοιες τιμές αυτής της μετρικής) και την διακριτότητά τους (δηλαδή το κατά πόσο τα δίκτυα που ανήκουν σε ξεχωριστές κατηγορίες έχουν διαφορετικές τιμές της μετρικής).

3.1.2 Ταξινόμηση γράφων μέσω Νευρωνικών Δικτύων

Ένα ακόμη σημαντικό εργαλείο για την ταξινόμηση γράφων, πάνω στο οποίο μάλιστα έχει γίνει εκτεταμένη έρευνα, είναι τα Νευρωνικά Δίκτυα, τα οποία προσφέρουν διάφορα πλεονεκτήματα σε περιπτώσεις που η μέθοδος επιλογής θεωρητικών μετρικών για την εκπροσώπηση των γράφων δεν είναι αποτελεσματική. Για παράδειγμα, οι John Boaz Lee, Ryan Rossi και Xiangnan Kong [61] επιχειρούν να λύσουν τα προβλήματα που ενδέχεται να προκύψουν με την προηγούμενη μέθοδο λόγω της παρουσίας θορύβου σε πραγματικά δίκτυα κατασκευάζοντας ένα Επαναλαμβανόμενο Νευρωνικό Δίκτυο (Recurrent Neural Network). Το μοντέλο αυτό αντλεί πληροφορία μόνο από τις γειτονιές συγκεκριμένων χαρακτηριστικών κόμβων του δικτύου που επιλέγονται μέσω περιπάτων που καθοδηγούνται από έναν *μηχανισμό προσοχής* (attention mechanism), γεγονός το οποίο επιτρέπει την αγνόηση του θορύβου και μειώνει τα υπολογιστικά κόστη. Επιπλέον, στα σύνολα δεδομένων που εξετάστηκαν, διαπιστώθηκε πως έχει μεγαλύτερη ακρίβεια από άλλες μεθόδους που λαμβάνουν υπ' όψην ολόκληρο τον γράφο, η οποία μάλιστα αυξάνεται ακόμη περισσότερο με την προσθήκη εξωτερικής μνήμης.

Ένα ακόμη είδος Νευρωνικών Δικτύων που είναι εξαιρετικά χρήσιμα για την ταξινόμηση δικτύων είναι τα Νευρωνικά Δίκτυα Γράφων (Graph Neural Networks ή GNN), τα οποία είναι ειδικά σχεδιασμένα για την επεξεργασία δεδομένων σε μορφή γράφου. Τα GNN εξάγουν αυτόματα και αποτελεσματικά τα αντιπροσωπευτικά χαρακτηριστικά ενός γράφου, γεγονός

που έχει αυξήσει την δημοτικότητά τους και συνεπώς τον αριθμό μοντέλων που υπάρχουν. Οι Federico Errica, Marco Podda, Davide Bacciu και Alessio Micheli [62] έχουν πραγματοποιήσει μια λεπτομερή αξιολόγηση και σύγκριση διαφόρων δημοφιλών μοντέλων GNN για ταξινόμηση γράφων (με το μοντέλο GIN των Xu et al. [63] να φαίνεται πως έχει συνολικά την μεγαλύτερη ακρίβεια από αυτά που μελετήθηκαν), η οποία μπορεί να αποδειχθεί πολύ χρήσιμη για την επιλογή του κατάλληλου μοντέλου για ένα πρόβλημα ταξινόμησης. Τέλος, θα πρέπει να αναφερθεί ότι ενώ τα μοντέλα αυτά είναι για ταξινόμηση γράφων σε οποιεσδήποτε κατηγορίες, μπορούν να χρησιμοποιηθούν και για την ταξινόμηση γράφων σε εξελικτικά μοντέλα, καθιστώντας τα έτσι σχετικά με το αντικείμενο έρευνας της παρούσης εργασίας.

3.1.3 Εντοπισμός εξελικτικών μοντέλων σε πραγματικά δίκτυα

Τεχνολογικά δίκτυα

Οι Bálint Hartmann και Viktória Sugár, με έναυσμα την αμφιλεγόμενη θεώρηση ότι τα δίκτυα ηλεκτρικής ισχύος μπορούν να αντιπροσωπευτούν από το μοντέλο Small-World ή το μοντέλο Scale-Free, αποφάσισαν να μελετήσουν το ηλεκτρικό δίκτυο της Ουγγαρίας χρησιμοποιώντας δεδομένα 70 ετών ώστε να προσδιορίσουν κατά πόσο είναι πράγματι δυνατό να αντιπροσωπευτεί από ένα από αυτά τα δύο μοντέλα και αν αυτό αλλάζει με την πάροδο του χρόνου. Για την επίτευξη του σκοπού αυτού, κατασκεύασαν έναν γράφο για κάθε έτος (δηλαδή συνολικά 70 γράφους), με τους κόμβους να αντιστοιχούν στις γεννήτριες, τους μετασχηματιστές, και τους υποσταθμούς, και τις ακμές στις γραμμές μεταφοράς. Για να αποφανθούν σχετικά με το αν οι γράφοι μπορούν να ταξινομηθούν ως Small-World χρησιμοποίησαν τον συντελεστή μικρού κόσμου (small-world coefficient) σ και την μετρική ω (όπως αυτή προτάθηκε από τους Telesford et al. [64]), ενώ για την ταξινόμηση σε Scale-Free χρησιμοποίησαν την κατανομή βαθμών κόμβων. Υπολόγισαν επιπλέον μερικές άλλες μετρικές με σκοπό την μελέτη της εξέλιξης του ηλεκτρικού δικτύου σε βάθος χρόνου. Τα αποτελέσματα έδειξαν μεγάλες αλλαγές τα πρώτα 20 χρόνια και μια σταθεροποίηση των τιμών των μετρικών τα επόμενα 50, τα οποία συνέδεσαν με ιστορικά γεγονότα όπως την κατασκευή ενός νέου σταθμού παραγωγής ενέργειας και την πρόσθεση νέων γραμμών μεταφοράς. Συμπέραναν πως αυτά τα γεγονότα επηρέασαν και το εξελικτικό μοντέλο του δικτύου, αφού κατέληξαν πως το δίκτυο μπορεί να κατηγοριοποιηθεί ως Small-World τα τελευταία 50 χρόνια της μελέτης τους, αλλά όχι τα πρώτα 20. Η αλλαγή αυτή συνδέθηκε με την προσθήκη διαφόρων επιπέδων τάσεων στο δίκτυο. Επιπλέον, συμπέραναν πως το μοντέλο Scale-Free δεν είναι αντιπροσωπευτικό του δικτύου αυτού σε καμία χρονική περίοδο. [65]

Η παραπάνω μελέτη δεν είναι η μοναδική που αποσκοπεί στον εντοπισμό κάποιου εξελικτικού μοντέλου σε ηλεκτρικά δίκτυα, καθώς η γνώση αυτή χρησιμεύει στην επίλυση σημαντικών προβλημάτων όπως η κατασκευή συνθετικών ηλεκτρικών δικτύων με σκοπό την προσομοίωση πραγματικών, και η ανάλυση ευρωστίας. Από την πληθώρα ερευνών πάνω σε ηλεκτρικά δίκτυα της Ευρώπης, της Αμερικής, και της Ασίας, εμείς επιλέγουμε να παρουσιάσουμε, ακόμη, ενδεικτικά αυτή των Rafael Espejo, Sara Lumbreras, και Andres Ramos πάνω σε 15 Ευρωπαϊκά δίκτυα διανομής ηλεκτρικής ενέργειας. Η έρευνα αυτή είναι ενδελεχής και μελετάει σε βάθος την δομή των 15 αυτών δικτύων, εξετάζοντας τα επίπεδα

τάσης κάθε δικτύου διανομής (220kV και 400kV) ως ξεχωριστά δίκτυα αλλά και μαζί. Χρησιμοποιούν 6 μετρικές για την ανάλυση της τοπολογίας των δικτύων, εκ των οποίων αυτή που σχετίζεται άμεσα με το αντικείμενο της παρούσας διπλωματικής είναι ο συντελεστής μικρού-κόσμου σ, καθώς χρησιμοποιείται για την αξιολόγηση του εάν ένα δίκτυο μπορεί να θεωρηθεί Small-World. Τα αποτελέσματα δείχνουν πως όταν τα δύο επίπεδα τάσης εξετάζονται μαζί (δηλαδή συνδυάζονται σε έναν γράφο), τότε τα δίκτυα διανομής όλων των χωρών μπορούν να θεωρηθούν Small-World. Αντιθέτως όταν τα επίπεδα τάσης εξετάζονται ξεχωριστά (δηλαδή αντιπροσωπεύονται από δύο διαφορετικούς γράφους) μια μειοψηφία αυτών δεν μπορεί να θεωρηθεί Small-World, γεγονός το οποίο οφείλεται στις διαφορές στην δομή του δικτύου κάθε χώρας. Τα αποτελέσματα αυτά φαίνεται να συνάδουν με τα συμπεράσματα των Hartmann και Sugár. [3]

Βιολογικά δίκτυα

Ένα ακόμη πεδίο στο οποίο γίνονται προσπάθειες εντοπισμού εξελικτικών μοντέλων είναι τα βιολογικά δίκτυα, με σκοπό την καλύτερη κατανόηση τους. Δεδομένου ότι διάφορα ήδη βιολογικών δικτύων όπως τα μεταβολικά δίκτυα, τα δίκτυα αλληλεπίδρασης πρωτεϊνών, και τα δίκτυα αλληλεπίδρασης και έκφρασης γονιδίων έχουν παρατηρηθεί να συμπεριφέρονται ως Scale-Free δίκτυα, οι Raya Khanin και Ernst Wit αποφάσισαν να ερευνήσουν εάν όντως τα βιολογικά δίκτυα μπορούν να θεωρηθούν Scale-Free. Για τον σκοπό αυτό, επικεντρώθηκαν σε δίκτυα που είχαν ήδη ταξινομηθεί ως Scale-Free από άλλους ερευνητές, οι οποίοι είχαν χρησιμοποιήσει την μέθοδο της προσαρμογής μιας ευθείας στην γραφική παράσταση της κατανομής βαθμών κόμβων σε λογαριθμική κλίμακα (log-log plot). Αρχικά, εκτίμησαν τον εκθέτη γ (της κατανομής power law) για κάθε ένα από αυτά, χρησιμοποιώντας την στατιστική μέθοδο της μέγιστης πιθανοφάνειας. Έπειτα, για κάθε δίκτυο πραγματοποιήθηκε έλεγχος καλής προσαρμογής, ώστε να διαπιστωθεί εάν πράγματι αυτό προκύπτει από την κατανομή power law που αντιστοιχεί στο γ που υπολογίστηκε. Τα αποτελέσματα του ελέγχου αυτού έδειξαν πως τα δίκτυα που εξετάστηκαν δεν μπορούν τελικά να χαρακτηριστούν Scale-Free, καθώς η κατανομή βαθμών κόμβων τους δεν ακολουθεί κατανομή power law. Τέλος, οι Khanin και Wit επισημαίνουν ότι τα αποτελέσματά τους συνάδουν με αυτά των Stumpf et al. [66], και τονίζουν την ανάγκη περισσότερης έρευνας πάνω στην τοπολογία των βιολογικών δικτύων. Πράγματι, το ερώτημα σχετικά με το εξελικτικό μοντέλο που ακολουθούν τα βιολογικά δίκτυα παραμένει ανοιχτό. [67]

Δίκτυα Μεταφορών

Προσπάθειες εντοπισμού εξελικτικών μοντέλων πραγματοποιούνται και για τα δίκτυα μεταφορών. Οι Bin Jiang και Christophe Claramunt, με έναυσμα την χρησιμότητα της ανάλυσης δικτύων στο πεδίο των Γεωγραφικών Πληροφοριακών Συστημάτων, εξέτασαν ένα σύνολο χαρακτηριστικών ορισμένων δικτύων μεταφορών, συμπεριλαμβανομένου του κατά πόσο αυτά επιδεικνύουν Small-World ή Scale-Free συμπεριφορά. Τα δίκτυα που εξετάστηκαν ήταν τα οδικά δίκτυα του Gävle της Σουηδίας, του Μονάχου της Γερμανίας, και του San Francisco των ΗΠΑ, ενώ ως κριτήριο της παρουσίας Scale-Free συμπεριφοράς χρησιμοποιήθηκε η μορφή της γραφικής παράστασης της συνάρτησης κατανομής πιθανότητας της

συνδεσιμότητας των οδών των πόλεων σε λογαριθμικούς άξονες. Τα αποτελέσματα έδειξαν ότι κανένα από τα οδικά δίκτυα αυτών των πόλεων δεν χαρακτηρίζεται από το μοντέλο Scale-Free. Όσον αφορά τον έλεγχο παρουσίας Small-World συμπεριφοράς, αυτός έγινε μέσω των τιμών του μέσου μήκους μονοπατιού και του συντελεστή συσταδοποίησης των δικτύων, οι οποίες έδειξαν πως πράγματι αυτά μπορούν να χαρακτηριστούν ως Small-World. Επομένως, οι Jiang και Claramunt κατέληξαν στο συμπέρασμα πως τα οδικά δίκτυα πόλεων δεν είναι Scale-Free, αλλά Small-World. [68]

Μία ακόμη έρευνα που αφορά τον εντοπισμό κάποιου εξελικτικού δικτύου σε δίκτυο μεταφορών είναι αυτή των Sen et al., η οποία επικεντρώνεται στην δομική ανάλυση του Ινδικού Σιδηροδρομικού Δικτύου. Στα πλαίσια αυτά, εξετάζουν εάν το δίκτυο αυτό μπορεί να χαρακτηριστεί ως Small-World ή Scale-Free, χρησιμοποιώντας κριτήρια παρόμοια με αυτά των Jiang και Claramunt. Συγκεκριμένα, ως κριτήριο εντοπισμού του Small-World μοντέλου χρησιμοποιούνται πάλι το Μέσο Μήκος Μονοπατιού και ο Συντελεστής Συσταδοποίησης, ενώ για το Scale-Free χρησιμοποιούνται η μορφή της κατανομής βαθμών κόμβων και ο συντελεστής συσταδοποίησης. Βάσει των αποτελεσμάτων, οι Sen et al. συμπεραίνουν ότι το Ινδικό Σιδηροδρομικό δίκτυο ανήκει στην κατηγορία των Small-World δικτύων, αλλά όχι σε αυτή των Scale-Free, και εκφράζουν την άποψη πως τα αποτελέσματα αυτά μπορούν να γενικευτούν για το σιδηροδρομικό δίκτυο οποιασδήποτε χώρας. [69]

3.2 Η συχνότητα εμφάνισης του μοντέλου Scale-Free σε πραγματικά δίκτυα

3.2.1 Τα Scale-free Δίκτυα Είναι Σπάνια

Τα Scale-Free δίκτυα αποτελούν δημοφιλές αντικείμενο μελέτης στον χώρο της επιστήμης δικτύων, καθώς θεωρείται ότι συνιστούν ένα μεγάλο μέρος των πραγματικών δικτύων. Ενώ υπάρχει πληθώρα ερευνών που υποστηρίζει αυτόν τον ισχυρισμό, υπάρχουν ακόμη πολλές άλλες έρευνες που τον καταρρίπτουν, γεγονός που τον καθιστά αμφιλεγόμενο. Επιπλέον, τα κριτήρια που χρησιμοποιούνται για να κατηγοριοποιηθεί ένα δίκτυο ως Scale-Free δεν είναι τα ίδια σε όλες τις έρευνες, αλλά υπάρχουν μερικές διαφορετικές παραλλαγές. Τα γεγονότα αυτά οδήγησαν τους Anna D. Broido και Aaron Clauset στο να διεξάγουν την δική τους έρευνα σχετικά με την συχνότητα εμφάνισης του μοντέλου Scale-Free στα πραγματικά δίκτυα, με τίτλο *Τα Scale-free Δίκτυα Είναι Σπάνια (Scale-free networks are rare)*. [70]

Οι Broido και Clauset εργάστηκαν πάνω σε 928 δίκτυα, τα δεδομένα των οποίων άντλησαν από το Index of Complex Networks (ή ICON). Το 53% αυτών ήταν βιολογικά δίκτυα, το 22% τεχνολογικά δίκτυα, το 16% κοινωνικά δίκτυα, το 7% δίκτυα μεταφορών, και το 2% δίκτυα πληροφορίας, συμπεριλαμβάνοντας έτσι αρκετές κατηγορίες ώστε να μπορεί να βγει ένα πιο καθολικό συμπέρασμα. Τα δίκτυα αυτά ήταν επίσης διαφόρων μεγεθών, με το πλήθος των κόμβων τους να κυμαίνεται από μερικές δεκάδες έως και εκατομμύρια, και διέφεραν ως προς ένα εύρος ιδιοτήτων όπως η κατευθυντικότητα, τα βάρη, η χρονική μεταβλητότητα, το αν είναι πολυεπίπεδα (multiplex), το αν είναι διμερή (bipartate), και το αν είναι πολυγραφήματα (multigraphs). Επειδή η παρουσία αρκετών από αυτές τις ιδιότητες οδηγεί στην ύπαρξη πολλών κατανομών βαθμών κόμβων για τον ίδιο γράφο (παραδείγμα-

τος χάρην για έναν κατευθυνόμενο γράφο υπάρχει η κατανομή in-degree και η κατανομή out-degree), αποφάσισαν να διαχωρίσουν κάθε μη απλό γράφο σε απλούς οι οποίοι θα τον αντιπροσωπεύουν. Για παράδειγμα, ένας πολυεπίπεδος γράφος με n επίπεδα «σπάει» σε n απλούστερους που αντιστοιχούν στα επίπεδά του, και ύστερα ο κάθε επιμέρους γράφος χωρίζεται σε ακόμη απλούστερος ανάλογα με τις υπόλοιπες ιδιότητές του, και ούτω καθεξής, ώσπου όλοι οι επιμέρους γράφοι να είναι απλοί (δηλαδή να μην διαθέτουν καμία από τις ιδιότητες που προαναφέρθηκαν).

Το βασικό κριτήριο το οποίο χρησιμοποίησαν ώστε να αποφασίσουν ποια δίκτυα μπορούν να θεωρηθούν Scale-Free ήταν το κατά πόσο η κατανομή βαθμών κόμβων τους προσομοιάζει την κατανομή power law $P(k) = Ck^{-a}$, $k \geq k_{min} \geq 1$, όπου k ο βαθμός κόμβου, C η σταθερά κανονικοποίησης, a η παράμετρος της κατανομής, και k_{min} ο ελάχιστος βαθμός κόμβου για τον οποίον ζητείται οι βαθμοί κόμβων να ακολουθούν την κατανομή power law. Το γεγονός ότι εξετάζεται μόνο η άνω ουρά της κατανομής, από το k_{min} και ύστερα, καθιστά την επιλογή του k_{min} πολύ σημαντική. Αυτή έγινε μέσω της κλασικής μεθόδου ελαχιστοποίησης Kolmogorov-Smirnov (standard KS-minimization), ενώ η εξίσου σημαντική επιλογή της παραμέτρου a έγινε μέσω του διακριτού εκτιμητή μέγιστης πιθανοφάνειας. Η αξιολόγηση του κατά πόσο η κατανομή power law που προκύπτει από τα $\hat{k_{min}}$ και $\hat{a}$ που υπολογίστηκαν προσομοιάζει την πραγματική κατανομή κόμβων του δικτύου έγινε μέσω ενός κλασικού ελέγχου καλής προσαρμογής, ο οποίος επέστρεφε μια τιμή p ως μέτρο της ομοιότητας των κατανομών αυτών μεταξύ τους. Επιπλέον, η κατανομή power law που αντιστοιχεί στο ζεύγος παραμέτρων $(\hat{k_{min}}, \hat{a})$ συγκρίθηκε με τέσσερεις άλλες κατανομές (εκθετική, log-normal, power law με εκθετική αποκοπή, και Weibull) στο τμήμα $k \geq k_{min}$ μέσω ελέγχου κανονικοποιημένου λόγου πιθανοφάνειας Vuong, σχετικά με το ποια από αυτές προσομοιάζει την πραγματική κατανομή κόμβων καλύτερα. Εδώ θα πρέπει να σημειωθεί ότι η εφαρμογή αυτών των μεθόδων, και άρα η ταξινόμηση, πραγματοποιήθηκε στους απλουστευμένους γράφους και όχι στους αρχικούς.

Ένα ακόμη βασικό κομμάτι του έργου των Broido και Clauset υπήρξε η κατασκευή πέντε κατηγοριών Scale-Free δικτύων, συνοψίζοντας έτσι τα διαφορετικά κριτήρια που εμφανίζονται στην βιβλιογραφία, και ταυτοχρόνως δημιουργώντας μια κλίμακα για την ισχύ της ιδιότητας Scale-Free σε ένα δίκτυο. Οι κατηγορίες αυτές είναι οι εξής:

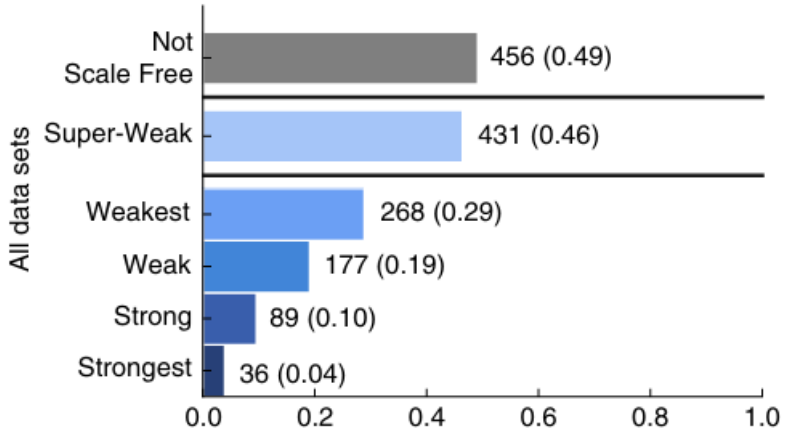
- **Υπερ-Αδύναμα (Super-Weak):** Για τουλάχιστον 50% των απλών γράφων, καμία άλλη κατανομή δεν προτιμάται αντί του power law
- **Αδυναμότατα (Weakest):** Για τουλάχιστον 50% των απλών γράφων, η κατανομή power law δεν μπορεί να απορριφθεί ($p \geq 0.1$)
- **Αδύναμα (Weak):** Ισχύουν οι προϋποθέσεις των Weakest, και η περιοχή που περιγράφεται από την power law κατανομή περιέχει τουλάχιστον 50 κόμβους
- **Δυνατά (Strong):** Ισχύουν οι προϋποθέσεις των Weakest και των Weak, και ταυτόχρονα $2 < a < 3$ για τουλάχιστον 50% των απλών γράφων
- **Δυνατότατα (Strongest):** Ισχύουν οι προϋποθέσεις των Strong για τουλάχιστον 90% των απλών γράφων, και οι προϋποθέσεις των Super-Weak για τουλάχιστον 95% των

απλών γράφων

Η ταξινόμηση ενός δικτύου στην πρώτη κατηγορία (Super Weak) αποτελεί έμμεση ένδειξη της παρουσίας Scale-Free δομής σε αυτό, ενώ η ταξινόμηση σε κάποια από τις επόμενες κατηγορίες αποτελεί άμεση ένδειξη. Θα πρέπει να σημειωθεί επίσης πως οι τέσσερεις τελευταίες κατηγορίες είναι εμφωλευμένες, δηλαδή αν ένα δίκτυο ανήκει σε μία από αυτές, τότε θα ανήκει και σε όλες τις από-πάνω της (εκτός φυσικά από την Super-Weak). Τέλος, για την εκπροσώπηση των μη Scale-Free δικτύων ορίζεται μια επιπλέον κατηγορία:

- **Μη Scale-Free (Not Scale Free):** Δίκτυα που δεν είναι ούτε Super-Weak ούτε Weakest

Τα αποτελέσματα της ταξινόμησης των δικτύων του συνόλου δεδομένων στις παραπάνω κατηγορίες ακολουθώντας την μέθοδο που περιγράφηκε συνοψίζονται στο σχήμα 3.1.



Σχήμα 3.1: *Ποσοστά των δικτύων ανά κατηγορία στην οποία ανήκουν σχετικά με την ιδιότητα Scale-Free. Οι οριζόντιες γραμμές χωρίζουν την κατηγορία Super-Weak από τις εμφωλευμένες, και από την Not Scale Free*

[70]

Όπως βλέπουμε, 49% των δικτύων δεν μπορεί να θεωρηθεί Scale-Free, 46% ανήκει στην κατηγορία Super-Weak, 29% ανήκει στην Weakest, 19% στην Weak, 10% στην Strong, και 4% στην Strongest. Οι Broido και Clauset ανέλυσαν τα αποτελέσματα αυτά και ανά κατηγορία δικτύων. Για τα **βιολογικά δίκτυα** βρήκαν πως 63% από αυτά είναι μη Scale-Free το 33% Super-Weak, το 19% Weakest, και το 6% Strongest. Παρατήρησαν πως το μεγαλύτερο μέρος των μη Scale-Free βιολογικών δικτύων ήταν δίκτυα μυκήτων. Μάλιστα, όλα τα δίκτυα μυκήτων που εξέτασαν ήταν μη Scale-Free, το οποίο ίσως ήταν αναμενόμενο δεδομένης της χωρικά εξαρτώμενης δομής τους. Γι αυτόν τον λόγο αποφάσισαν να εξετάσουν ποια θα ήταν τα ποσοστά κάθε κατηγορίας εάν το σύνολο δεδομένων δεν περιείχε τα δίκτυα μυκήτων. Βρήκαν πως σε αυτήν την περίπτωση ενώ τα ποσοστά των κατηγοριών Super-Weak, Weakest, και Weak θα παρουσίαζαν σημαντική αύξηση, η αύξηση των ποσοστών των Strong και Strongest θα ήταν ελάχιστη, γεγονός που σημαίνει ότι τελικά η συμπερίληψη των δικτύων μυκήτων δεν επηρεάζει τα ποιοτικά συμπεράσματα της έρευνας. Όσον αφορά τα **κοινωνικά δίκτυα** τα αποτελέσματα είχαν αρκετά διαφορετική μορφή από τα συνολικά αποτελέσματα,

καθώς το 50% κατηγοριοποιήθηκε ως μη Scale-Free, το 41% ως Super-Weak, το 48% ως Weakest, και το 31% ως Weak, αλλά κανένα ως Strong ή Strongest. Τα αποτελέσματα ήταν διαφορετικά από τα συνολικά και για τα **τεχνολογικά δίκτυα**, καθώς μόνο 8% αυτών κατατάχθηκαν στα μη Scale-Free, ενώ 90% κατατάχθηκαν στα Super-Weak, 42% στα Weakest, 37% στα Weak, 28% στα Strong, και μόλις 1% στα Strongest. Τέλος, τα **δίκτυα μεταφορών** δεν είναι αρκετά για μια αντίστοιχη ποσοστιαία ανάλυση, αλλά παρατηρήθηκε πως όλα τους άνηκαν στις κατηγορίες Not Scale Free, Super-Weak, ή Weakest, ενώ κανένα στις Strong και Strongest.

Επιπλέον, για να διασφαλιστεί ότι τα αποτελέσματα αυτά δεν επηρεάστηκαν από την ίδια την μέθοδο αξιολόγησης, πραγματοποιήθηκαν τέσσερεις έλεγχοι ευρωστίας. Πιο συγκεκριμένα, ελέγχθηκε το αν επηρεάζονται υπό τις εξής συνθήκες: α) όλα τα δίκτυα που εξετάζονται είναι εκ φύσεως απλά (δηλαδή χωρίς βάρη, μη κατευθυνόμενα, μονοεπίπεδα, και χωρίς πολύ-ακμές (multiedges)), β) η κατανομή power law με εκθετική αποκοπή αφαιρείται από τις εναλλακτικές κατανομές, γ) χαμηλώνονται τα ποσοστιαία κατώφλια των κατηγοριών έτσι ώστε να επιτρέπεται η ταξινόμηση σε αυτές εάν έστω κι ένας επί μέρους απλός γράφος πληροί τα κριτήρια, δ) εξετάζεται η κλιμακωτή συμπεριφορά της πρώτης και της δεύτερης ροπής της κατανομής βαθμών κόμβων. Τα αποτελέσματα των ελέγχων αυτών έδειξαν πως πράγματι οι μέθοδοι που χρησιμοποιήθηκαν δεν υπήρξαν μεροληπτικές.

Τελικά, οι Broido και Clauset συμπέραναν πως τα Scale-Free δίκτυα δεν είναι πανταχού παρόντα, αλλά πολύ πιο σπάνια από ότι υποδεικνύει η σχετική βιβλιογραφία, και πως η χρήση τους ως το αρχικό στάδιο της μοντελοποίησης και της δομικής ανάλυσης πραγματικών δικτύων είναι αβάσιμη. Επισήμαναν επίσης, πως η πιθανότητα εμφάνισης Scale-Free δομής αλλάζει ανάλογα με την κατηγορία των δικτύων. Για παράδειγμα, είναι πιο πιθανό να υπάρχει σε συγκεκριμένους τύπους βιολογικών δικτύων και τεχνολογικών δικτύων, και πολύ λιγότερο πιθανό στα κοινωνικά δίκτυα. Τα μεταφορικά δίκτυα και τα δίκτυα πληροφορίας ήταν πολύ λίγα ώστε να βγει κάποιο συμπέρασμα γι αυτά, αλλά δεν υπήρχε κάποια ισχυρή ένδειξη πως η δομή τους διαφέρει πολύ από των υπολοίπων.

3.2.2 Είναι τα Scale-Free δίκτυα πράγματι σπάνια; Ένα αμφιλεγόμενο ζήτημα

Το άρθρο των Broido και Clauset [70] προκάλεσε αρκετές αντιδράσεις, ακόμη και από τον ίδιο τον Barabasi [71], δημιουργό του μοντέλου Scale-Free, ο οποίος ενέστη για την μεθοδολογία και τα συμπεράσματα του προαναφερθέντος άρθρου. Το άρθρο αυτό σε συνδυασμό με τον εκτεταμένο προβληματισμό της επιστημονικής κοινότητας σχετικά με την συχνότητα εμφάνισης των Scale-Free δικτύων ενέπνευσαν τους Voitalov et al. [72] να διεξάγουν την δική τους έρευνα πάνω σε αυτό το αντικείμενο, δίνοντας και οι ίδιοι έναν αυστηρό ορισμό για τα Scale-Free ως απάντηση στην ασάφεια που υπάρχει στην βιβλιογραφία, και πραγματοποιώντας στατιστική ανάλυση σε ένα σύνολο πραγματικών δικτύων. Τα αποτελέσματα της έρευνας αυτής επιβεβαίωσαν ότι πολλά από τα δίκτυα που είχαν βρεθεί στο παρελθόν να είναι Scale-Free είναι πράγματι Scale-Free (όπως ο Παγκόσμιος Ιστός, δίκτυα αλληλεπίδρασης ανθρώπινων πρωτεϊνών, δίκτυα μελών κοινωνικών ομάδων, και άλλα), σχηματίζοντας μια διαφορετική εικόνα για την συχνότητα εμφάνισης των Scale-Free δικτύων από ότι το άρθρο

των Broido και Clauset. Τα ίδια γεγονότα αποτέλεσαν έναυσμα και για τους Liu et al. [73] να ερευνήσουν την συχνότητα εμφάνισης του μοντέλου Scale-Free, αυτήν την φορά σε εμπορικά δίκτυα. Τα δικά τους αποτελέσματα έδειξαν πως η δομή των εμπορικών δικτύων παρουσιάζει μεγάλη ποικιλομορφία λόγω της διαφορετικής φύσης των εμπορικών προϊόντων που αφορά το κάθε ένα, αλλά η εμφάνιση Scale-Free δομής δεν είναι σπάνια. Οι ίδιοι επισήμαναν, επιπλέον, πως η λάθος διαχείριση των βαρών των ακμών όταν αυτά αντιπροσωπεύουν χωρητικότητες μπορεί να οδηγήσει στην παράβλεψη της πιθανώς Scale-Free δομής ενός δικτύου.

Μία ακόμη αξιοσημείωτη έρευνα σχετικά με την συχνότητα εμφάνισης του μοντέλου Scale-Free σε πραγματικά δίκτυα είναι αυτή των Artico et al. [74], οι οποίοι ισχυρίστηκαν πως οι Broido και Clauset [70] και οι Khanin και Wit [67] έθεσαν υπερβολικά αυστηρά κριτήρια σχετικά τη μορφή της κατανομής βαθμών κόμβων ενός δικτύου ώστε αυτό να μπορεί να θεωρηθεί Scale-Free, και εμπνεύστηκαν από την προσέγγιση των Voitalov et al. [72] αλλά και άλλων ερευνητών ώστε να θέσουν τα δικά τους, λιγότερο αυστηρά κριτήρια. Επιπλέον, οι Artico et al. τροποποιήσαν τους κλασικούς ελέγχους Kolmogorov-Smirnov και καλής προσαρμογής που χρησιμοποίησαν οι Broido και Clauset, και ύστερα χρησιμοποίησαν αυτές τις μεθόδους πάνω στο ίδιο σύνολο δεδομένων με τους τελευταίους ώστε να επαναξιολογήσουν τα συμπεράσματά τους. Τα αποτελέσματα των Artico et al. έδειξαν πως 64% των δικτύων που εξετάστηκαν ήταν Scale-Free, επομένως οι ίδιοι συμπέραναν πως τα Scale-Free δίκτυα τελικά δεν είναι σπάνια, αντικρούοντας έτσι τον ισχυρισμό των Broido και Clauset.

Ο προβληματισμός σχετικά με την συχνότητα εμφάνισης Scale-Free δομής σε πραγματικά δίκτυα δεν εμφανίστηκε βέβαια με την έκδοση του άρθρου των Broido και Clauset [70], αλλά προϋπήρχε σχεδόν από την στιγμή της δημιουργίας του μοντέλου Scale-Free το 1999. Ένα παράδειγμα παλαιότερης έρευνας πάνω σε αυτό το αντικείμενο αποτελεί το άρθρο των Gipsi Lima-Mendez και Jacques van Helden [75] με θέμα την ορθότητα δημοφιλών πεποιθήσεων εκείνης της εποχής (2000-2010) σχετικά με την συχνότητα εμφάνισης της power law κατανομής και του μοντέλου Scale-Free σε βιολογικά δίκτυα. Ύστερα από στατιστική ανάλυση, οι Lima-Mendez και van Helden συμπέραναν ότι οι εξής πέντε πεποιθήσεις είναι μύθοι: α) η κατανομή βαθμών κόμβων των βιολογικών δικτύων είναι η κατανομή power law, β) τα βιολογικά δίκτυα είναι Scale-Free, γ) τα μεταβολικά δίκτυα είναι Small-World, δ) τα δίκτυα Small-World είναι ευάλωτα σε στοχευμένες επιθέσεις αλλά όχι σε τυχαίες διαγραφές, ε) τα βιολογικά δίκτυα μεγαλώνουν μέσω προνομιακής προσκόλλησης (preferential attachment). Μία, ακόμη, παρόμοια έρευνα της εποχής εκείνης είναι αυτή των Khanin και Wit [67] που παρουσιάστηκε στην παράγραφο 3.1.3, η οποία επίσης καταλήγει στο συμπέρασμα πως τα βιολογικά δίκτυα δεν μπορούν εν γένει να χαρακτηριστούν ως Scale-Free.

Την ίδια χρονική περίοδο δημοσιεύτηκε και η έρευνα των Mislove et al. [76] στην οποία εξετάστηκαν τέσσερα διαδικτυακά κοινωνικά δίκτυα (Flickr, LiveJournal, Orkut και YouTube) ως προς διάφορες ιδιότητες, συμπεριλαμβανομένης της Scale-Free συμπεριφοράς. Μέσω των γραφικών παραστάσεων των κατανομών βαθμών κόμβων των δικτύων αυτών σε λογαριθμική κλίμακα, οι Mislove et al. συμπέραναν πως τα τρία από τα τέσσερα (συγκεκριμένα το Flickr, το LiveJournal, και το Orkut) εκδηλώνουν ισχυρή Scale-Free συμπεριφορά. Μελέτησαν επίσης το κατά πόσο η κατανομή βαθμών κόμβων τους προσομοιάζει την

power law κατανομή, πραγματοποιώντας στατιστική ανάλυση αντίστοιχη με αυτή που παρουσιάστηκε σε προαναφερθείσες έρευνες, και κατέληξαν στο ότι και τα τέσσερα αυτά δίκτυα παρουσιάζουν power law κατανομή βαθμών κόμβων.

Οι προαναφερθείσες έρευνες δεν είναι, φυσικά, οι μόνες που έχουν διεξαχθεί στα πλαίσια της μακράς και ζωηρής συζήτησης της επιστημονικής κοινότητας σχετικά με την συχνότητα εμφάνισης του Scale-Free μοντέλου. Πέρα από τα περιεχόμενα της παρούσης εργασίας, μία περίληψη της συζήτησης αυτής μπορεί να βρεθεί στο άρθρο του Petter Holme [77], μαζί με την δική του οπτική πάνω στο συγκεκριμένο ζήτημα.

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 4

Μεθοδολογία

4.1 Επισκόπηση

Στο παρόν κεφάλαιο παρουσιάζεται το σύνολο των μεθόδων που χρησιμοποιήθηκαν στο πλαίσιο της διπλωματικής αυτής με σκοπό τον προσδιορισμό των εξελικτικών μοντέλων που χαρακτηρίζουν ένα σύνολο πραγματικών και τεχνητών δικτύων, αλλά και την τμηματοποιήση των δικτύων σε υποδίκτυα που χαρακτηρίζονται από διαφορετικά εξελικτικά μοντέλα.

Την κεντρική ιδέα όσον αφορά την ταξινόμηση δικτύων σε εξελικτικά μοντέλα αποτέλεσε η εκπαίδευση ενός ταξινομητή, του οποίου η είσοδος θα ήταν οι τιμές ορισμένων χαρακτηριστικών ενός γράφου, και η έξοδος το εξελικτικό μοντέλο που χαρακτηρίζει τον συγκεκριμένο γράφο. Το πρώτο βήμα στην διαδικασία αυτή ήταν η επιλογή των εξελικτικών μοντέλων που θα συνιστούσαν τις πιθανές εξόδους του ταξινομητή. Ακολούθησε η επιλογή των μετρικών που θα χρησιμοποιούνταν ως είσοδοί του, η οποία βασίστηκε κυρίως στο κατά πόσο οι τιμές τους αποκλίνουν για διαφορετικά μοντέλα. Επιπλέον δημιουργήθηκε το σύνολο δεδομένων εκπαίδευσης του ταξινομητή, μέσω της κατασκευής ενός πλήθους τεχνητών δικτύων κάθε είδους, και του υπολογισμού των τιμών των προαναφερθεισών μετρικών για κάθε ένα από αυτά. Το επόμενο βήμα αποτέλεσε ο έλεγχος της επίδοσης διαφόρων ταξινομητών πάνω στο σύνολο δεδομένων αυτό, και η επιλογή εκείνου με την καλύτερη επίδοση. Ο επιλεγμένος ταξινομητής χρησιμοποιήθηκε ύστερα για την ταξινόμηση ορισμένων πραγματικών δικτύων.

Όσον αφορά τον εντοπισμό τμημάτων ενός δικτύου που χαρακτηρίζονται από διαφορετικά εξελικτικά μοντέλα, δοκιμάστηκε ένα πλήθος μεθόδων τμηματοποίησης με σκοπό να βρεθεί η πιο ακριβής. Συγκεκριμένα, η κάθε μια εφαρμόστηκε σε τεχνητά δίκτυα που είχαν δημιουργηθεί μέσω της συνένωσης δύο ή περισσότερων δικτύων διαφορετικών τύπων, και ύστερα χρησιμοποιήθηκε ο επιλεγμένος ταξινομητής ώστε να προσδιοριστεί το εξελικτικό μοντέλο που χαρακτηρίζει το κάθε υποδίκτυο. Η μέθοδος με τα πιο ακριβή αποτελέσματα (δηλαδή αυτά που προσέγγιζαν πιο πολύ την πραγματική φύση των αντίστοιχων δικτύων) χρησιμοποιήθηκε για τον εντοπισμό υποδικτύων στα πραγματικά δίκτυα που ταξινομήθηκαν νωρίτερα, ενώ τα υποδίκτυα που προέκυψαν ταξινομήθηκαν έπειτα σε εξελικτικά μοντέλα με την βοήθεια του επιλεγμένου ταξινομητή.

4.2 Επιλογή Εξελικτικών Μοντέλων

Η επιλογή των εξελικτικών μοντέλων τα οποία έγινε προσπάθεια να εντοπιστούν σε πραγματικά και τεχνητά δίκτυα αποτελεί θεμελιώδες μέρος της παρούσης εργασίας. Αρχικά πραγματοποιήθηκε αναζήτηση στην βιβλιογραφία, όπου βρέθηκαν τα εξής μοντέλα: Κανονικός Γράφος (Regular Graph), Erdos-Renyi, Gilbert, Τυχαίος Γεωμετρικός Γράφος, Watts-Strogatz, Newman-Watts, Barabasi-Albert, Waxman, και Ιεραρχικός Γράφος (Hierarchic Graph). Εξ αυτών τα μοντέλα Erdos-Renyi και Gilbert ανήκουν στην κατηγορία των Τυχαίων Γράφων, τα μοντέλα Watts-Strogatz και Newman-Watts ανήκουν στην κατηγορία Small-World, το μοντέλο Waxman ανήκει στην κατηγορία των Χωρικών Δικτύων, και τέλος ο Τυχαίος Γεωμετρικός Γράφος ανήκει και στην κατηγορία των Τυχαίων Γράφων και σε αυτή των Χωρικών Δικτύων.

Στην συνέχεια ορισμένα από αυτά απορρίφθηκαν, με κριτήρια την συχνότητα εμφάνισής τους σε πραγματικά δίκτυα, και την ομοιότητά τους με άλλα μοντέλα. Συγκεκριμένα, οι Κανονικοί Γράφοι κατέχουν μία πολύ ιδιαίτερη μορφή (όλοι οι κόμβοι τους έχουν τον ίδιο βαθμό), η οποία παρουσιάζεται σε πολύ λίγα είδη πραγματικών δικτύων, παραδείγματος χάριν σε δίκτυα που αναπαριστούν την δομή κρυστάλλων [78]. Επιπλέον, οι Ιεραρχικοί Γράφοι έχουν την μορφή Δέντρου, η οποία επίσης παρουσιάζεται σε πολύ συγκεκριμένα είδη δικτύων, που περιγράφουν ιεραρχικές σχέσεις (παραδείγματος χάριν δίκτυα συγγένειας) [15]. Όσον αφορά την ομοιότητα των δικτύων, αποφασίσαμε να επιτρέψουμε την χρήση δύο ή περισσότερων μοντέλων της ίδια κατηγορίας, με την εξαίρεση των Watts-Strogatz και Newman-Watts αφού το δεύτερο αποτελεί (βελτιωμένη) παραλλαγή του πρώτου [79]. Από αυτά τα δύο αποφασίσαμε να κρατήσουμε το μοντέλο Watts-Strogatz, καθώς εμφανίζεται πολύ πιο συχνά στην βιβλιογραφία και στις υπόλοιπες έρευνες σχετικά με την ταξινόμηση δικτύων σε εξελικτικά μοντέλα.

Έτσι, τα μοντέλα που χρησιμοποιήθηκαν τελικά ήταν τα **Erdos-Renyi**, **Gilbert**, **Τυχαίος Γεωμετρικός Γράφος**, **Watts-Strogatz**, **Barabasi-Albert**, και **Waxman**, ενώ οι Κανονικοί Γράφοι, οι Ιεραρχικοί Γράφοι, και το μοντέλο Newman-Watts απορρίφθηκαν.

4.3 Επιλογή Μετρικών Δικτύων

Η επιλογή των μετρικών δικτύων που χρησιμοποιήθηκαν ως είσοδοι στον ταξινομητή ξεκίνησε πάλι με μια ενδελεχή έρευνα στην βιβλιογραφία. Εκεί βρέθηκαν 20 μετρικές οι οποίες μπορούν να προσφέρουν σημαντικές πληροφορίες για την δομή και την συμπεριφορά ενός δικτύου, κι επομένως να βοηθήσουν στην ταξινόμηση αυτού σε κάποιο εξελικτικό μοντέλο. Οι μετρικές αυτές ήταν οι εξής: Κατανομή Βαθμών Κόμβων, Μέσο Μήκος Μονοπατιού, Διάμετρος, Ακτίνα, Συντελεστής Συσταδοποίησης, Ενδιαμεσική Κεντρικότητα, Κεντρικότητα της Εγγύτητας, Κεντρικότητα της Πληροφορίας, Κεντρικότητα της Δρομολόγησης (Routing Centrality), Κεντρικότητα της Γεφύρωσης (Bridging Centrality), Φασματική Κεντρικότητα (Spectral Centrality), Κεντρικότητα των Ιδιοδιανυσμάτων (Eigenvector Centrality), Κεντρικότητα Katz, Λαπλασιανή Κεντρικότητα (Laplacian Centrality), Κεντρικότητα των Αρμονικών, Κεντρικότητα της Διήθησης (Percolation Centrality), Κεντρικότητα του Φορτίου (Load Centrality), Συντελεστής Πλουσίου Συλλόγου (Rich Club Coefficient), Ζωτικότητα

της Εγγύτητας (Closeness Vitality), και Συνάφεια.

Ύστερα δημιουργήθηκαν 6 τεχνητά δίκτυα, το κάθε ένα εκ των οποίων κατασκευάστηκε έτσι ώστε να χαρακτηρίζεται από ένα από τα 6 εξελικτικά μοντέλα που επιλέχθηκαν πρωτύτερα. Αποφασίστηκε, ακόμη, τα δίκτυα αυτά να κατασκευαστούν όλα με το ίδιο πλήθος κόμβων, έτσι ώστε να διασφαλιστεί ότι οι τιμές των μετρικών δεν θα επηρεάζονταν από το μέγεθος των δικτύων κατά τον υπολογισμό τους πάνω σε αυτά. Επιπλέον έγινε προσπάθεια και οι υπόλοιπες παράμετροί τους να έχουν τις ίδιες τιμές, όσο αυτό ήταν δυνατό καθώς δεν έχουν όλα τα μοντέλα τις ίδιες παραμέτρους (παραδείγματος χάριν κάποια έχουν ως παράμετρο την πιθανότητα σχηματισμού ακμής, ενώ κάποια άλλα το πλήθος νέων συνδέσεων που ξεκινούν από κάθε κόμβο).

Οι 20 προαναφερθείσες μετρικές υπολογίστηκαν και για τους 6 τεχνητούς γράφους, έτσι ώστε να διαπιστωθεί το κατά πόσο η κάθε μία παρουσιάζει διαφορά στις τιμές της για διαφορετικά μοντέλα, και το αν είναι εφικτό να υπολογιστεί για ένα μεγάλο πλήθος δικτύων μέσα σε ένα εύλογο χρονικό διάστημα. Πιο αναλυτικά, για κάθε μετρική υπολογίστηκε η τιμή της για κάθε έναν από τους γράφους, εφόσον αυτή είναι μοναδική ανά γράφο (όπως για την Συνάφεια), ή η κατανομή της, η μέση τιμή της, και η διασπορά της, εφόσον αυτή ορίζεται για όλους τους κόμβους ενός γράφου (παραδείγματος χάριν οι κεντρικότητες). Επιπλέον, για κάθε μετρική υπολογίστηκε η διασπορά της μέσης (ή της μοναδικής) τιμής της, όπως και ο συντελεστής διασποράς της τιμής αυτής, ως πιο αντικειμενικό μέτρο της διακύμανσης των τιμών της μετρικής ανάλογα με το μοντέλο ενός δικτύου (καθώς το μέγεθος της διασποράς εξαρτάται από το μέγεθος της μέσης τιμής, ενώ ο συντελεστής διασποράς δεν εμπεριέχει την εξάρτηση αυτή). Τέλος, για κάθε μετρική υπολογίστηκε ο χρόνος εκτέλεσης του υπολογισμού της ανά δίκτυο, ο συνολικός χρόνος εκτέλεσης και για τα 6 τεχνητά δίκτυα, όπως και ο λόγος του συντελεστή διασποράς της προς τον συνολικό χρόνο εκτέλεσής της, ως μέτρο επίδοσης.

Έπειτα πραγματοποιήθηκε συνολική επισκόπηση επί των αποτελεσμάτων, λαμβάνοντας υπ' όψιν τις τιμές που υπολογίστηκαν, αλλά και τις ποιοτικές διαφορές στις μορφές των κατανομών, για όποιες μετρικές αυτές υπήρχαν. Συγκεκριμένα, για κάθε μετρική αξιολογήθηκαν θετικά οι διαφορές στις κατανομές και στις μέσες ή απλές τιμές ανά δίκτυο, όπως και οι συγκριτικά μεγάλες τιμές της διασποράς της μέσης ή απλής τιμής, του συντελεστή της διασποράς, και του λόγου του συντελεστή της διασποράς προς τον χρόνο εκτέλεσης, ενώ αρνητικά αξιολογήθηκαν οι μεγάλοι χρόνοι εκτέλεσης. Έτσι, αφότου απορρίφθηκαν ορισμένες μετρικές λόγω υπερβολικά μεγάλου χρόνου εκτέλεσης ή άλλων προβλημάτων κατά τον υπολογισμό τους, επιλέχθηκαν οι εξής 7 που κρίθηκε ότι παρουσιάζουν αρκετά μεγάλες διαφορές ανά μοντέλο ώστε να επιτευχθεί μια καλή επίδοση του ταξινομητή: **Κατανομή Βαθμών Κόμβων**, **Μέσο Μήκος Μονοπατιού**, **Διάμετρος**, **Κεντρικότητα της Εγγύτητας**, **Κεντρικότητα της Πληροφορίας**, **Κεντρικότητα των Αρμονικών**, και **Συνάφεια**.

4.4 Δημιουργία Συνόλου Δεδομένων

Το σύνολο δεδομένων για την εκπαίδευση και την αξιολόγηση του ταξινομητή δημιουργήθηκε κατασκευάζοντας 100 γράφους για κάθε εξελικτικό μοντέλο με την βοήθεια των αντίστοιχων συναρτήσεων της βιβλιοθήκης NetworkX της γλώσσας προγραμματισμού Python, η οποία είναι αυτή που χρησιμοποιήθηκε για την υλοποίηση του πειραματικού μέρους της

εργασίας μας. Για κάθε έναν από τους 600 αυτούς γράφους υπολογίστηκαν επιπλέον οι τιμές των μετρικών που επιλέχθηκαν ως είσοδοι του ταξινομητή, έτσι ώστε κάθε γράφος να αντιπροσωπεύεται στο σύνολο δεδομένων από το μοντέλο του (που αποτελεί την ετικέτα του), και τις τιμές των προαναφερθεισών μετρικών υπολογισμένων πάνω σε αυτόν. Το 80% των γράφων του συνόλου δεδομένων (δηλαδή 480 γράφοι) επιλέχθηκαν με τυχαίο τρόπο ώστε να αποτελέσουν το σύνολο δεδομένων εκπαίδευσης του ταξινομητή, ενώ το υπόλοιπο 20% (120 γράφοι) χρησιμοποιήθηκε ως το σύνολο δεδομένων ελέγχου του.

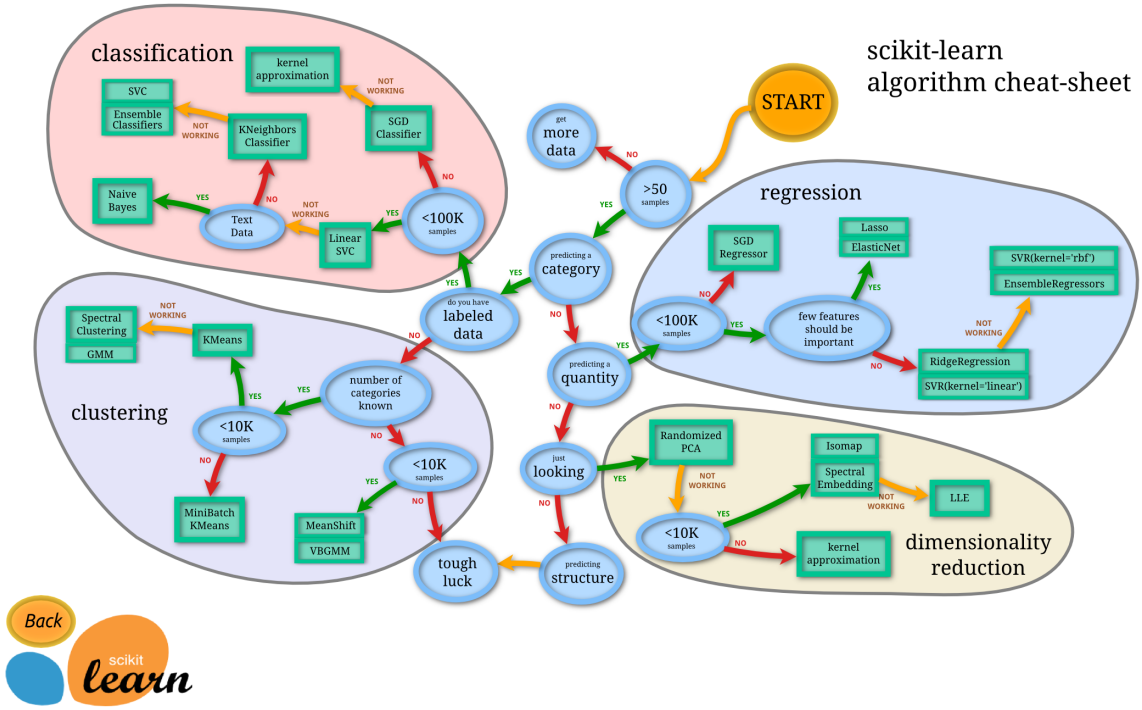
4.5 Επιλογή Ταξινομητή

Η επιλογή του κατάλληλου ταξινομητή έγινε με την βοήθεια του διαγράμματος της εικόνας 4.1, που προέρχεται από τα έγγραφα της βιβλιοθήκης scikit-learn της Python [80]. Συγκεκριμένα, ξεκινώντας από την αρχή (START) του διαγράμματος ακολουθήθηκε η προτεινόμενη διαδρομή που προκύπτει απαντώντας καταφατικά στις ερωτήσεις αν υπάρχουν πάνω από 50 δείγματα, αν προβλέπεται κάποια κατηγορία, αν τα δεδομένα έχουν ετικέτα, και αν τα δείγματα είναι λιγότερα από 100000. Η διαδρομή αυτή οδηγεί στον ταξινομητή Linear SVC, επομένως αυτός αποτέλεσε έναν από τους υποψήφιους ταξινομητές για το παρόν πρόβλημα ταξινόμησης. Με σκοπό να βρεθεί ο βέλτιστος ταξινομητής για το πρόβλημα αυτό, ακολουθήθηκε η διαδρομή που αντιστοιχεί στην κακή λειτουργία του Linear SVC, ύστερα στην απουσία δεδομένων κειμένου, και τέλος στην κακή λειτουργία του ταξινομητή k-Nearest Neighbors (αφού προστέθηκε βέβαια και αυτός στην λίστα με τους υποψηφίους). Η διαδρομή αυτή καταλήγει στον ταξινομητή SVC και στην κατηγορία των Ταξινομητών Συνόλου, από την οποία επιλέχθηκαν οι Ada Boost, Gradient Boosting, Random Forest, και Bagging ως υποψήφιοι.

Έτσι, το σύνολο των υποψήφιων ταξινομητών αποτελούταν από τους Linear SVC, k-Nearest Neighbors, SVC, Ada Boost, Gradient Boosting, Random Forest, και Bagging. Στην συνέχεια αυτοί οι ταξινομητές αξιολογήθηκαν πάνω στο σύνολο δεδομένων του προβλήματος, βάσει των μετρικών precision, recall, accuracy, και F1 score. Επιλέχθηκε αυτός με την μεγαλύτερη τιμή των παραπάνω μετρικών (στην περίπτωσή μας ο **Random Forest**), και υπολογίστηκε ο Πίνακας Σύγχυσής του έτσι ώστε να αποκτηθούν λεπτομέρειες σχετικά με τις αδυναμίες του. Επιπλέον πραγματοποιήθηκε Μελέτη Αφαίρεσης (Ablation Study), δηλαδή επαναξιολόγηση της επίδοσης του ταξινομητή μετά από την αφαίρεση κάθε μίας εκ των μετρικών στην είσοδό του, έτσι ώστε να εξεταστεί το ενδεχόμενο κάποια από τις μετρικές που χρησιμοποιήθηκαν να χειροτερεύει την επίδοσή του.

4.6 Ταξινόμηση Πραγματικών Δικτύων

Τα πραγματικά δίκτυα τα οποία ταξινομήθηκαν ως προς το εξελικτικό μοντέλο τους στα πλαίσια της παρούσης εργασίας επιλέχθηκαν κυρίως από το Index of Complex Networks ή ICON (https://icon.colorado.edu/networks), αλλά και από την Συλλογή Δεδομένων Μεγάλων Δικτύων του πανεπιστημίου Stanford (https://snap.stanford.edu/data/), τα αρχεία του πρότζεκτ KONECT του πανεπιστημίου του Koblenz–Landau (http://konect.cc/), και την πλατφόρμα Kaggle (https://www.kaggle.com/). Τα δίκτυα αυτά παρουσιάζουν ποικιλία στην



Σχήμα 4.1: *Διάγραμμα σχετικά με την επιλογή του σωστού αλγορίθμου μηχανικής μάθησης ανάλογα με το είδος του προβλήματος και τις παραμέτρους του*
[80]

μορφή τους, με το πλήθος των κόμβων και των ακμών τους να κυμαίνεται από εκατοντάδες έως και δεκάδες χιλιάδες, και τις ιδιότητες και τις κατηγορίες τους να διαφέρουν.

Συγκεκριμένα, τα δίκτυα που επιλέχτηκαν ήταν τα εξής:

- **Football:** Κοινωνικό δίκτυο, με 115 κόμβους και 613 ακμές, από το KONECT. Μη κατευθυνόμενο, χωρίς βάρη. Δίκτυο παιχνιδιών αμερικάνικου football μεταξύ κολεγίων το φθινόπωρο του 2000.

- **GOT:** Κοινωνικό δίκτυο, με 114 κόμβους και 396 κόμβους, από το Kaggle. Μη κατευθυνόμενο, χωρίς βάρη. Δίκτυο αλληλεπιδράσεων μεταξύ των χαρακτήρων της 5ης σεζόν της τηλεοπτικής σειράς Game of Thrones.

- **Email-EU:** Κοινωνικό δίκτυο, με 1005 κόμβους και 25571 ακμές, από το Stanford. Κατευθυνόμενο, χωρίς βάρη. Δίκτυο ηλεκτρονικής αλληλογραφίας.

- **AdvogatoTrust:** Κοινωνικό δίκτυο, με 6541 κόμβους και 51127 ακμές, από το ICON. Κατευθυνόμενο, με βάρη. Δίκτυο των σχέσεων εμπιστοσύνης μεταξύ των χρηστών του Advogato.

- **MorenoCrime:** Κοινωνικό δίκτυο, με 1380 κόμβους και 1476 ακμές, από το KONECT. Μη κατευθυνόμενο, χωρίς βάρη, και διμερές. Δίκτυο που καταγράφει την ανάμειξη ορισμένων προσώπων (αριστερό μέρος του γράφου) με εγκλήματα (δεξί μέρος του γράφου).

- **NetScience:** Κοινωνικό δίκτυο, με 1589 κόμβους και 2742 ακμές, από το ICON. Μη κατευθυνόμενο, με βάρη. Δίκτυο συνεργασίας μεταξύ επιστημόνων για την συγγραφή δημοσιεύσεων στο πεδίο της επιστήμης δικτύων.

- **Euroroad:** Δίκτυο μεταφορών, με 1174 κόμβους και 1417 ακμές, από το ICON. Μη κατευθυνόμενο, χωρίς βάρη. Οδικό δίκτυο Ε-δρόμων, κυρίως στην Ευρώπη.

- **ChicagoRegional:** Δίκτυο μεταφορών, με 12982 κόμβους και 39018 ακμές, από το ICON. Κατευθυνόμενο, χωρίς βάρη. Δίκτυο οδικής κυκλοφορίας μεταξύ διασταυρώσεων στο Chicago, Illinois των ΗΠΑ.

- **Flights-FAA:** Δίκτυο μεταφορών, με 1226 κόμβους και 2615 ακμές, από το ICON. Κατευθυνόμενο, χωρίς βάρη. Δίκτυο αεροπορικών διαδρομών, από την βάση δεδομένων προτιμώμενων διαδρομών του Εθνικού Κέντρου Δεδομένων Πτήσεων (NFDC) της Ομοσπονδιακής Διοίκησης Πτήσεων (FAA) των ΗΠΑ.

- **Bristol:** Δίκτυο μεταφορών, με 3731 κόμβους και 3101 ακμές, από το ICON. Μη κατευθυνόμενο, χωρίς βάρη. Δίκτυο διαδρομών λεωφορείων στο Bristol της Αγγλίας.

- **Coach:** Δίκτυο μεταφορών, με 3915 κόμβους και 3228 ακμές, από το ICON. Μη κατευθυνόμενο, χωρίς βάρη. Δίκτυο διαδρομών πούλμαν στο HB.

- **YuYeast:** Βιολογικό δίκτυο, με 2018 κόμβους και 2930 ακμές, από το ICON. Μη κατευθυνόμενο, χωρίς βάρη. Δίκτυο πρωτεϊνικών αλληλεπιδράσεων στον μύκητα S. cerevisiae.

- **BacillusMegaterium:** Βιολογικό δίκτυο, με 4179 κόμβους και 5644 ακμές, από το ICON. Μη κατευθυνόμενο, χωρίς βάρη, και διμερές. Μεταβολικό δίκτυο του βακτηρίου Bacillus megaterium.

- **Malaria:** Βιολογικό δίκτυο, με 307 κόμβους και 1446 ακμές, από το ICON. Μη κατευθυνόμενο, χωρίς βάρη, και αποτελεί ένα από τα επίπεδα ενός μεγαλύτερου, πολυεπίπεδου γράφου. Δίκτυο ανασυνδυασμένων γονιδίων αντιγόνων του ανθρώπινου παράσιτου της ελονοσίας, P. falciparum.

- **Fullerene:** Χημικό δίκτυο, με 3840 κόμβους και 5760 ακμές, από το ICON. Μη κατευθυνόμενο, χωρίς βάρη. Δίκτυο δεσμών μεταξύ των ατόμων άνθρακα σε ένα μόριο φουλερενίου.

- **PowerNet:** Τεχνολογικό δίκτυο, με 4941 κόμβους και 6594 ακμές, από το ICON. Μη κατευθυνόμενο, χωρίς βάρη. Δίκτυο που αναπαριστά το Ηλεκτρικό Δίκτυο Δυτικών Πολιτειών των ΗΠΑ. Σημείωση: Χρησιμοποιήθηκε από τους Watts και Strogatz για την δημοσίευση στην οποία εισήγαγαν την έννοια των Small-World δικτύων [21].

- **R-Dependancies:** Τεχνολογικό δίκτυο, με 2471 κόμβους και 5451 ακμές, από το ICON. Κατευθυνόμενο, χωρίς βάρη. Δίκτυο εξαρτήσεων μεταξύ πακέτων της γλώσσας προγραμματισμού R.

Τα παραπάνω δίκτυα υπέστησαν επεξεργασία πριν την είσοδό τους στον ταξινομητή. Αρχικά, όσοι γράφοι ήταν κατευθυνόμενοι μετατράπηκαν σε μη κατευθυνόμενους, αγνοώντας την κατεύθυνση των ακμών τους. Επιπλέον, απομονώθηκε η μεγαλύτερη συνεκτική συνιστώσα όλων τους, πάνω στην οποία πραγματοποιήθηκε αργότερα η ταξινόμηση και η τμηματοποίηση. Οι τροποποιήσεις αυτές έγιναν έτσι ώστε να είναι δυνατό να υπολογιστούν οι μετρικές που επιλέχθηκαν για τους συγκεκριμένους γράφους, αφού ορισμένες από αυτές δεν ορίζονται για μη συνδεδεμένους, ή κατευθυνόμενους γράφους.

4.7 Επιλογή Μεθόδου Τμηματοποίησης Δικτύων

Με σκοπό την εύρεση της βέλτιστης μεθόδου τμηματοποίησης δικτύων σε υποδίκτυα με διαφορετικά εξελικτικά μοντέλα, κατασκευάστηκαν αρχικά 5 σύνθετα τεχνητά δίκτυα, συνενώνοντας άλλα τεχνητά δίκτυα που χαρακτηρίζονταν από ένα μοναδικό εξελικτικό μοντέλο το κάθε ένα. Συγκεκριμένα, τα μοντέλα δικτύων τα οποία συνδυάστηκαν ήταν τα εξής: Barabasi-Albert με Random Geometric, Barabasi-Albert με Watts-Strogatz, Watts-Strogatz με Erdos-Renyi, Waxman με Random Geometric, και Gilbert με Waxman. Οι ενώσεις αυτές πραγματοποιήθηκαν προσθέτοντας ακμές μεταξύ των επιμέρους γράφων, επιλέγοντας τους κόμβους στους οποίους αυτές προσέπιπταν με τυχαίο τρόπο. Ύστερα εφαρμόστηκαν διάφορες μέθοδοι τμηματοποίησης πάνω στους 5 αυτούς γράφους, απορρίφθηκαν όσοι υπογράφοι από αυτούς που προέκυψαν ήταν πολύ μικροί για να ταξινομηθούν (δηλαδή όσοι είχαν λιγότερους από 20 κόμβους), και οι υπόλοιποι ταξινομήθηκαν σε εξελικτικά μοντέλα, με την βοήθεια του επιλεγμένου ταξινομητή.

Για κάθε μέθοδο υπολογίστηκε επιπλέον το πλήθος των υπογράφων που προέκυψαν, το μέγεθος των υπογράφων αυτών, το πλήθος των ορθά ταξινομημένων υπογράφων καθώς και το πλήθος των ορθά ταξινομημένων κόμβων συνολικά για τον γράφο, το modularity της διαμέρισης, και ο χρόνος εκτέλεσης. (Σημείωση: Ένας κόμβος θεωρήθηκε ορθά ταξινομημένος αν ο υπογράφος στον οποίο ανήκε ταξινομήθηκε στο ίδιο μοντέλο με τον γράφο από τον οποίο αυτός προερχόταν, ενώ ένας υπογράφος θεωρήθηκε ορθά ταξινομημένος αν τουλάχιστον 70% των κόμβων του ήταν ορθά ταξινομημένοι.) Το πιο σημαντικό κριτήριο ήταν οι υπογράφοι να ανήκουν στα εξελικτικά μοντέλα των συνιστωσών του γράφου από τον οποίον προήλθαν, και μόνο σε αυτά. Ταυτόχρονα, αξιολογήθηκαν θετικά τα μεγάλα πλήθη ορθά κατανεμημένων υπογράφων και κόμβων, και τα μεγάλα modularity, ενώ οι μεγάλοι χρόνοι εκτέλεσης αξιολογήθηκαν αρνητικά.

4.7.1 Εντοπισμός Κοινοτήτων

Όλες οι μέθοδοι που χρησιμοποιήθηκαν ανήκαν στην κατηγορία των αλγορίθμων εντοπισμού κοινοτήτων. Η επιλογή των αλγορίθμων που δοκιμάστηκαν έγινε με κριτήρια την διαφορετικότητά τους μεταξύ τους ως προς τον τρόπο που λειτουργούν (έτσι ώστε να ελεγχθούν αρκετές διαφορετικές προσεγγίσεις), και την δυνατότητα εύρεσης κάποιας υλοποίησης τους. Αυτοί που χρησιμοποιήθηκαν τελικά ήταν οι Girvan-Newman, Walktrap, Infomap, Leiden, και Spectral Clustering, ενώ οι υλοποιήσεις τους βρέθηκαν στις βιβλιοθήκες cdlib, networkx, και sklearn της Python. Μετά την αξιολόγησή τους, επιλέχθηκε ο Leiden.

Σημείωση: Λόγω του υπερβολικά μεγάλου χρόνου εκτέλεσης του αλγορίθμου Girvan-Newman πάνω στα δίκτυα που κατασκευάστηκαν (πάνω από 1 ώρα σε ορισμένες περιπτώσεις), χρειάστηκε να κατασκευαστούν άλλα, μικρότερα δίκτυα με τα ίδια μοντέλα, έτσι ώστε να ελεγχθεί η επίδοσή του. Το γεγονός αυτό αυτομάτως απέκλεισε ως ενδεχόμενο την χρήση του Girvan-Newman για την τμηματοποίηση των πραγματικών δικτύων, καθώς τα περισσότερα από αυτά έχουν παρόμοιο πλήθος με τα πρώτα τεχνητά δίκτυα που κατασκευάστηκαν. Παρ' όλα αυτά, αποφασίστηκε να ελεγχθεί και η δική του επίδοση, καθώς θα μπορούσε να αποδειχθεί χρήσιμος σε μία εφαρμογή με μικρότερα δίκτυα, περισσότερο διαθέσιμο χρόνο, ή καλύτερους υπολογιστικούς πόρους.

4.8 Τμηματοποίηση Πραγματικών Δικτύων

Αφότου επιλέχθηκε η βέλτιστη μέθοδος τμηματοποίησης δικτύων, αυτή εφαρμόστηκε και στα πραγματικά δίκτυα που αναφέρθηκαν στην παράγραφο 4.6. Έτσι, εντοπίστηκαν τα τμήματά τους που χαρακτηρίζονται από διαφορετικά εξελικτικά μοντέλα (μαζί με το μέγεθός τους), και το ποια μοντέλα είναι αυτά.

Κεφάλαιο 5

Αποτελέσματα

Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα των μεθόδων και των πειραμάτων που περιγράφηκαν στο προηγούμενο κεφάλαιο, μαζί με κάποιες παρατηρήσεις πάνω σε αυτά.

5.1 Επιλογή Μετρικών Δικτύων

Παρακάτω παραθέτονται τα αποτελέσματα για κάθε μετρική ξεχωριστά (μετά από στρογγυλοποίηση σε δύο δεκαδικά ψηφία εκτός αν αναφέρεται κάτι διαφορετικό), καθώς και μια συνολική σύγκριση μεταξύ αυτών. Οι κατανομές δεν περιλαμβάνονται για οικονομία χώρου, αλλά μπορούν να βρεθούν στο αντίστοιχο αρχείο.

5.1.1 Κατανομή Βαθμών Κόμβων

Πίνακας 5.1: *Αποτελέσματα για την Κατανομή Βαθμών Κόμβων*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	10	9.64	0.51
Gilbert	499.47	246.78	0.28
Random Geometric	480.22	18438.10	0.33
Small-World	4	1.47	0.57
Scale-Free	9.95	98.23	0.31
Waxman	40.23	42.55	0.38

5.1.2 Μέσο Μήκος Μονοπατιού

Πίνακας 5.2: *Αποτελέσματα για το Μέσο Μήκος Μονοπατιού*

	Τιμή	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	3.25	2.95
Gilbert	1.50	2.21
Random Geometric	1.55	13.80
Small-World	5.58	0.55
Scale-Free	2.99	0.55
Waxman	2.15	0.55

5.1.3 Διάμετρος

Πίνακας 5.3: *Αποτελέσματα για την Διάμετρο*

	Τιμή	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	5	0.6
Gilbert	2	0.43
Random Geometric	3	6.43
Small-World	10	0.92
Scale-Free	5	0.96
Waxman	3	0.93

5.1.4 Ακτίνα

Πίνακας 5.4: *Αποτελέσματα για την Ακτίνα*

	Τιμή	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	4	0.58
Gilbert	2	0.42
Random Geometric	2	6.47
Small-World	7	0.53
Scale-Free	3	0.55
Waxman	3	0.52

5.1.5 Συντελεστής Συσταδοποίησης

Πίνακας 5.5: *Αποτελέσματα για τον Συντελεστή Συσταδοποίησης*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.01	0	0.22
Gilbert	0.50	0	64.08
Random Geometric	0.77	0.01	59.61
Small-World	0.07	0.02	0.18
Scale-Free	0.04	0	0.28
Waxman	0.04	0	0.66

5.1.6 Ενδιαμεσική Κεντρικότητα

Πίνακας 5.6: *Αποτελέσματα για την Ενδιαμεσική Κεντρικότητα*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.0023	0	4.24
Gilbert	0.0005	0	93.76
Random Geometric	0.0005	0	69.68
Small-World	0.0046	0	3.60
Scale-Free	0.0020	0	5.03
Waxman	0.0012	0	9.21

Η στρογγυλοποίηση της μέσης τιμής έγινε στα 4 δεκαδικά ψηφία, λόγω της πολύ μικρής τιμής της. Η τιμή της διασποράς κυμαινόταν από την τάξη του 10^{-9} έως την τάξη του 10^{-5} για όλους τους γράφους, επομένως στρογγυλοποιήθηκε στο 0 παντού.

5.1.7 Κεντρικότητα της Εγγύτητας

Πίνακας 5.7: *Αποτελέσματα για την Κεντρικότητα της Εγγύτητας*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.31	0.0002	1.26
Gilbert	0.67	0	0.51
Random Geometric	0.65	0.0052	6.60
Small-World	0.18	0.0001	0.71
Scale-Free	0.36	0.0007	0.71
Waxman	0.47	0.0002	0.67

Η στρογγυλοποίηση της διασποράς έγινε στα 4 δεκαδικά ψηφία, λόγω της πολύ μικρής τιμής της.

5.1.8 Κεντρικότητα της Πληροφορίας

Πίνακας 5.8: *Αποτελέσματα για την Κεντρικότητα της Πληροφορίας*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.0041	0	2.71
Gilbert	0.2498	0	4.75
Random Geometric	0.2249	0.0010	4.46
Small-World	0.0012	0	1.08
Scale-Free	0.0034	0	2.80
Waxman	0.0193	0	2.64

Η στρογγυλοποίηση της μέσης τιμής και της διασποράς έγινε στα 4 δεκαδικά ψηφία, λόγω των πολύ μικρών τιμών τους.

5.1.9 Κεντρικότητα της Δρομολόγησης

Πίνακας 5.9: *Αποτελέσματα για την Κεντρικότητα της Δρομολόγησης*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.0053	0	27.10
Gilbert	0.0018	0	1108.39
Random Geometric	0.0022	0	880.34
Small-World	0.0095	0	6.06
Scale-Free	0.0049	0	22.61
Waxman	0.0034	0	90.05

Η στρογγυλοποίηση της μέσης τιμής έγινε στα 4 δεκαδικά ψηφία, λόγω της πολύ μικρής τιμής της. Η τιμή της διασποράς κυμαινόταν από την τάξη του 10^{-9} έως την τάξη του 10^{-5} για όλους τους γράφους, επομένως στρογγυλοποιήθηκε στο 0 παντού.

5.1.10 Κεντρικότητα της Γεφύρωσης

Πίνακας 5.10: *Αποτελέσματα για την Κεντρικότητα της Γεφύρωσης*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.000207	0	4.1
Gilbert	0.000001	0	92.83
Random Geometric	0.000001	0	70.04
Small-World	0.001095	0	3.25
Scale-Free	0.000123	0	4.07
Waxman	0.000028	0	9.19

Η στρογγυλοποίηση της μέσης τιμής έγινε στα 6 δεκαδικά ψηφία, λόγω της πολύ μικρής τιμής της. Η τιμή της διασποράς κυμαινόταν από την τάξη του 10^{-17} έως την τάξη του 10^{-7} για όλους τους γράφους, επομένως στρογγυλοποιήθηκε στο 0 παντού.

5.1.11 Φασματική Κεντρικότητα

Ο υπολογισμός της συγκεκριμένης μετρικής δεν ήταν δυνατός, καθώς παρουσιάστηκαν προβλήματα κατά την εκτέλεσή του, και δεν βρέθηκε διαφορετική υλοποίηση. Ως αποτέλεσμα, η φασματική κεντρικότητα απορρίφθηκε αυτομάτως.

5.1.12 Κεντρικότητα των Ιδιοδιανυσμάτων

Πίνακας 5.11: *Αποτελέσματα για την Κεντρικότητα των Ιδιοδιανυσμάτων*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.0297	0.0001	0.22
Gilbert	0.0316	0	0.49
Random Geometric	0.0301	0.0001	0.94
Small-World	0.0279	0.0002	0.27
Scale-Free	0.0219	0.0005	0.20
Waxman	0.0312	0	0.23

Η στρογγυλοποίηση της μέσης τιμής και της διασποράς έγινε στα 4 δεκαδικά ψηφία, λόγω των πολύ μικρών τιμών τους.

5.1.13 Κεντρικότητα Katz

Πίνακας 5.12: *Αποτελέσματα για την Κεντρικότητα Katz*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.0297	0.0001	0.17
Gilbert	0.0316	0	0.48
Random Geometric	0.0301	0.0001	0.90
Small-World	0.0279	0.0002	0.39
Scale-Free	0.0219	0.0005	0.27
Waxman	0.0312	0	0.29

Η στρογγυλοποίηση της μέσης τιμής και της διασποράς έγινε στα 4 δεκαδικά ψηφία, λόγω των πολύ μικρών τιμών τους.

5.1.14 Λαπλασιανή Κεντρικότητα

Ο υπολογισμός της συγκεκριμένης μετρικής διήρκησε πάνω από 60 λεπτά, κι έτσι αποφασίστηκε να σταματήσει η εκτέλεση, και να απορριφθεί αυτομάτως.

5.1.15 Κεντρικότητα των Αρμονικών

Πίνακας 5.13: *Αποτελέσματα για την Κεντρικότητα των Αρμονικών*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	324.49	276.91	1.62
Gilbert	749.24	61.70	0.83
Random Geometric	734.74	5401.68	7.32
Small-World	192.49	141.64	0.97
Scale-Free	351.10	928.90	0.91
Waxman	487.90	141.17	1.35

5.1.16 Κεντρικότητα της Διήθησης

Πίνακας 5.14: *Αποτελέσματα για την Κεντρικότητα της Διήθησης*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.0023	0	4.97
Gilbert	0.0005	0	92.28
Random Geometric	0.0005	0	70.73
Small-World	0.0046	0	3.47
Scale-Free	0.0020	0	4.88
Waxman	0.0012	0	10.23

Η στρογγυλοποίηση της μέσης τιμής έγινε στα 4 δεκαδικά ψηφία, λόγω της πολύ μικρής τιμής της. Η τιμή της διασποράς κυμαινόταν από την τάξη του 10^{-9} έως την τάξη του 10^{-5} για όλους τους γράφους, επομένως στρογγυλοποιήθηκε στο 0 παντού.

5.1.17 Κεντρικότητα του Φορτίου

Πίνακας 5.15: *Αποτελέσματα για την Κεντρικότητα του Φορτίου*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.0023	0	4.79
Gilbert	0.0005	0	72.20
Random Geometric	0.0005	0	60.41
Small-World	0.0046	0	2.73
Scale-Free	0.0020	0	4.39
Waxman	0.0012	0	7.97

Η στρογγυλοποίηση της μέσης τιμής έγινε στα 4 δεκαδικά ψηφία, λόγω της πολύ μικρής τιμής της. Η τιμή της διασποράς κυμαινόταν από την τάξη του 10^{-9} έως την τάξη του 10^{-5} για όλους τους γράφους, επομένως στρογγυλοποιήθηκε στο 0 παντού.

5.1.18 Συντελεστής Πλουσίου Συλλόγου

Πίνακας 5.16: *Αποτελέσματα για τον Συντελεστή Πλουσίου Συλλόγου*

	Μέση Τιμή	**Διασπορά**	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.0023	0	5.41
Gilbert	0.0005	0	92.64
Random Geometric	0.0005	0	69.79
Small-World	0.0046	0	4.80
Scale-Free	0.0020	0	4.20
Waxman	0.0012	0	9.68

Η στρογγυλοποίηση της μέσης τιμής έγινε στα 4 δεκαδικά ψηφία, λόγω της πολύ μικρής τιμής της. Η τιμή της διασποράς κυμαινόταν από την τάξη του 10^{-9} έως την τάξη του 10^{-5} για όλους τους γράφους, επομένως στρογγυλοποιήθηκε στο 0 παντού.

5.1.19 Ζωτικότητα της Εγγύτητας

Ο υπολογισμός της συγκεκριμένης μετρικής διήρκησε πάνω από 60 λεπτά, κι έτσι αποφασίστηκε να σταματήσει η εκτέλεση, και να απορριφθεί αυτομάτως.

5.1.20 Συνάφεια

Πίνακας 5.17: *Αποτελέσματα για την Συνάφεια*

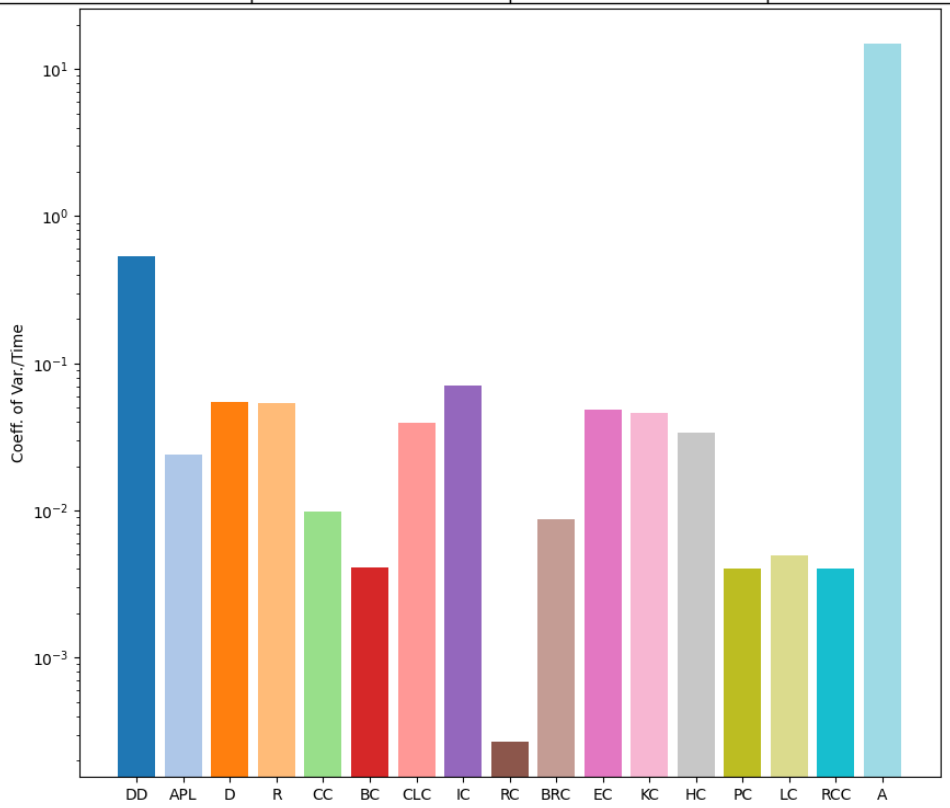
	Τιμή	**Χρόνος Εκτέλεσης (s)**
Erdos-Renyi	0.0115	0.53
Gilbert	-0.0017	1.50
Random Geometric	0.1094	1.56
Small-World	-0.0906	0.02
Scale-Free	-0.0496	0.03
Waxman	0.0141	0.08

Η στρογγυλοποίηση της τιμής έγινε στα 4 δεκαδικά ψηφία, λόγω του πολύ μικρού μεγέθους της.

5.1.21 Συνολική Σύγκριση

Στον πίνακα της εικόνας 5.1 παρουσιάζονται για κάθε μετρική που δεν απορρίφθηκε ήδη οι εξής τιμές, οι οποίες συνδυάζουν τα αποτελέσματα για όλα τα δίκτυα: Η διασπορά της τιμής της ή της μέσης τιμής της (Variance), ο συνολικός χρόνος εκτέλεσής της (Total Time), ο συντελεστής διασποράς της τιμής της ή της μέσης τιμής της (Coeff. of Var.), και ο λόγος του συντελεστή διασποράς ως προς τον συνολικό χρόνο εκτέλεσής της (Coeff. of Var./Time).

	Variance	Total Time	Coeff. of Var.	Coeff. of Var./Time
Degree Distribution	50050.4995	2.3932	1.2859	0.5373
Average Path Length	1.938	20.5993	0.4909	0.0238
Diameter	6.8889	10.2608	0.5624	0.0548
Radius	2.9167	9.0784	0.488	0.0537
Clustering Coefficient	0.0853	125.0222	1.2314	0.0098
Betweenness Centrality	0.0	185.5171	0.7578	0.0041
Closeness Centrality	0.0323	10.4559	0.4131	0.0395
Information Centrality	0.0119	18.4427	1.3009	0.0705
Routing Centrality	0.0	2134.5533	0.5695	0.0003
Bridging Centrality	0.0	183.4815	1.6022	0.0087
Eigenvector Centrality	0.0	2.334	0.114	0.0488
Katz Centrality	0.0	2.4962	0.114	0.0457
Harmonic Centrality	43439.5006	12.9946	0.4403	0.0339
Percolation Centrality	0.0	186.5606	0.7578	0.0041
Load Centrality	0.0	152.4865	0.7578	0.005
Rich Club Coefficient	0.0	186.5178	0.7578	0.0041
Assortativity	0.0038	3.7265	55.0923	14.7838



Σχήμα 5.1: ***Πάνω:*** *Πίνακας που περιλαμβάνει την διασπορά της τιμής της ή της μέσης τιμής, του συνολικό χρόνο εκτέλεσης, του συντελεστή διασποράς της τιμής της ή της μέσης τιμής, και τον λόγο του συντελεστή διασποράς ως προς τον συνολικό χρόνο εκτέλεσης κάθε μετρική.* ***Κάτω:*** *Γραφική παράσταση του λόγου του συντελεστή διασποράς ως προς τον συνολικό χρόνο εκτέλεσης κάθε μετρικής, σε λογαριθμική κλίμακα.*

Βάσει των αποτελεσμάτων αυτών, απορρίφθηκε η Ενδιαμεσική Κεντρικότητα, η Κεντρικότητα της Δρομολόγησης, η Κεντρικότητα της Δρομολόγησης, η Κεντρικότητα της Γεφύρωσης, η Κεντρικότητα της Διήθησης, η Κεντρικότητα του Φορτίου, και ο Συντελεστής Πλουσίου Συλλόγου, εξαιτίας των μεγάλων συνολικών χρόνων εκτέλεσής τους. Ενώ ο Συντελεστής Συσταδοποίησης έχει και αυτός σχετικά μεγάλο χρόνο εκτέλεσης, αποφασίστηκε αρχικά να χρησιμοποιηθεί, λόγω του μεγάλου συντελεστή διασποράς του και της εκτεταμένης παρου-

σίας του στην βιβλιογραφία. Όμως η συμπερίληψή του είχε ως αποτέλεσμα η δημιουργία του συνόλου δεδομένων να μην ολοκληρωθεί εντός αποδεκτού χρονικού πλαισίου (καθώς υπερέβαινε τις 6 ώρες), επομένως και ο Συντελεστής Συσταδοποίησης τελικά απορρίφθηκε.

Επιπλέον απορρίφθηκαν η Κεντρικότητα των Ιδιοδιανυσμάτων, και η Κεντρικότητα Katz, εξαιτίας των πολύ μικρών τιμών της διασποράς της μέσης τιμής τους και του συντελεστή διασποράς αυτής. Τέλος, παρόλο που ο συνολικός χρόνος εκτέλεσης του υπολογισμού της είναι μικρός και ο συντελεστής διασποράς των τιμών της δεν είναι πολύ μικρός, απορρίφθηκε και η ακτίνα, καθώς λαμβάνει ακριβώς την ίδια τιμή για 2 διαφορετικά ζεύγη μοντέλων, γεγονός που θεωρήθηκε ένδειξη ότι δύναται να επηρεάσει αρνητικά την επίδοση του ταξινομητή.

Έτσι, οι μετρικές που επιλέχθηκαν τελικά ήταν η **Κατανομή Βαθμών Κόμβων**, το **Μέσο Μήκος Μονοπατιού**, η **Διάμετρος**, η **Κεντρικότητα της Εγγύτητας**, η **Κεντρικότητα της Πληροφορίας**, η **Κεντρικότητα των Αρμονικών**, και η **Συνάφεια**.

5.2 Επιλογή Ταξινομητή

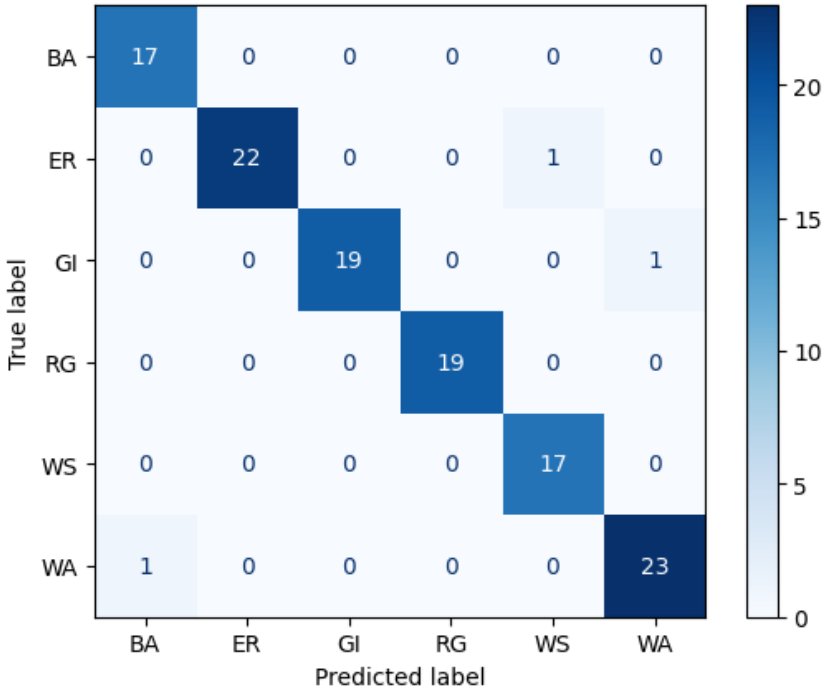
Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των μετρικών precision, recall, accuracy, και F1 score για κάθε ταξινομητή, στρογγυλοποιημένα σε 4 δεκαδικά ψηφία.

Πίνακας 5.18: *Πίνακας επιδόσεων των ταξινομητών*

	Precision	**Recall**	**Accuracy**	**F1 Score**
Linear SVC	0.7577	0.7608	0.7583	0.7579
k-Nearest Neighbors	0.5935	0.5805	0.5833	0.5700
SVC	0.4728	0.4423	0.4333	0.3943
Ada Boost	0.7779	0.6999	0.7333	0.6984
Gradient Boosting	0.9512	0.9510	0.9500	0.9498
Random Forest	0.9745	0.9775	0.9750	0.9756
Bagging	0.5494	0.5535	0.5667	0.5357

Όπως φαίνεται στον πίνακα 5.18, ο ταξινομητής **Random Forest** είχε την καλύτερη επίδοση σε όλες τις μετρικές, επομένως ήταν αυτός που επιλέχθηκε για την ταξινόμηση των πραγματικών δικτύων. Ο πίνακας σύγχυσης του ταξινομητή αυτού και τα αποτελέσματα της μελέτη αφαίρεσης που πραγματοποιήθηκε με την βοήθειά του παρατίθενται στις εικόνες 5.2 και 5.3.

Παρατηρούμε πως ο ταξινομητής έχει πολύ καλή ικανότητα διάκρισης όλων των μοντέλων. Επιπλέον, καμία μετρική δεν φαίνεται να επιβαρύνει τον ταξινομητή, καθώς η αφαίρεση κάθε μίας προκαλεί είτε μείωση στην επίδοσή του, είτε πολύ μικρή αύξηση (της τάξης του 1% ή και μικρότερη) που πιθανότατα είναι τυχαία και δεν οφείλεται στην φύση της μετρικής. Η Συνάφεια, μάλιστα, αποδεικνύεται ιδιαίτερα σημαντική για την καλή επίδοση του ταξινομητή, γεγονός αναμενόμενο δεδομένου του πολύ υψηλού συντελεστή διασποράς της μετρικής αυτής. Ταυτόχρονα, το γεγονός αυτό αποτελεί ένδειξη πως η χρήση του συντελεστή διασποράς ως το κύριο κριτήριο επιλογής μετρικών δικτύων πέρα από τον χρόνο εκτέλεσης, ήταν ορθή.



Σχήμα 5.2: *Πίνακας σύγχυσης του ταξινομητή Random Forest. Οι ετικέτες αντιστοιχούν στα αρχικά των μοντέλων.*

	removed_feature	accuracy	precision	recall	f1
0	Degree Distribution	0.975000	0.976754	0.975000	0.975282
1	Average Path Length	0.983333	0.984259	0.983333	0.983390
2	Diameter	0.966667	0.968259	0.966667	0.966604
3	Information Centrality	0.975000	0.976754	0.975000	0.975282
4	Closeness Centrality	0.983333	0.984259	0.983333	0.983390
5	Harmonic Centrality	0.983333	0.984259	0.983333	0.983390
6	Assortativity	0.791667	0.788154	0.791667	0.788812

Σχήμα 5.3: *Μελέτη αφαίρεσης των μετρικών Κατανομή Βαθμών Κόμβων, το Μέσο Μήκος Μονοπατιού, Διάμετρος, Κεντρικότητα της Εγγύτητας, Κεντρικότητα της Πληροφορίας, Κεντρικότητα των Αρμονικών, και Συνάφεια, με την βοήθεια του ταξινομητή Random Forest*

5.3 Ταξινόμηση Πραγματικών Δικτύων

Ακολουθούν τα αποτελέσματα της ταξινόμησης των πραγματικών δικτύων που αναφέρθηκαν στο Κεφάλαιο 4.

- **Football:** Waxman
- **GOT:** Barabasi-Albert

- **Email-EU:** Waxman
- **AdvogatoTrust:** Barabasi-Albert
- **MorenoCrime:** Watts-Strogatz
- **NetScience:** Watts-Strogatz
- **Euroroad:** Watts-Strogatz
- **ChicagoRegional:** Watts-Strogatz
- **Flights-FAA:** Watts-Strogatz
- **Bristol:** Watts-Strogatz
- **Coach:** Watts-Strogatz
- **YuYeast:** Barabasi-Albert
- **BacillusMegaterium:** Gilbert
- **Malaria:** Random Geometric
- **Fullerene:** Watts-Strogatz
- **PowerNet:** Watts-Strogatz
- **R-Dependancies:** Gilbert

Παρατηρούμε πως η συχνότητα εμφάνισης του μοντέλου Watts-Strogatz είναι με διαφορά μεγαλύτερη από όλων των υπολοίπων (9 εμφανίσεις), ειδικά όταν πρόκειται για δίκτυα μεταφορών. Ακολουθεί σε συχνότητα το μοντέλο Barabasi-Albert (2 εμφανίσεις σε κοινωνικά δίκτυα και 1 σε βιολογικό), και ύστερα το Waxman (2 εμφανίσεις σε κοινωνικά δίκτυα) και το Gilbert (1 εμφάνιση σε βιολογικό δίκτυο και 1 σε τεχνολογικό). Το μοντέλο Random Geometric εμφανίζεται 1 φορά σε βιολογικό δίκτυο, ενώ το Erdos-Renyi απουσιάζει.

Επιπλέον, αξίζει να σημειωθεί πως το δίκτυο PowerNet ταξινομήθηκε στο μοντέλο Watts-Strogatz, σε συμφωνία με τα ευρήματα των Watts και Strogatz [21].

5.4 Επιλογή Μεθόδου Τμηματοποίησης Δικτύων

Παρατίθενται τα πλήθη κόμβων των απλών δικτύων:

- **Για τον Girvan Newman:** Erdos-Renyi: 401,Gilbert: 200, Random Geometric: 392, Watts-Strogatz: 299, Barabasi-Albert: 519, Waxman: 333
- **Για τους υπόλοιπους αλγορίθμους:** Erdos-Renyi: 1401,Gilbert: 1200, Random Geometric: 1392, Watts-Strogatz: 1299, Barabasi-Albert: 1519, Waxman: 1333

Ακολουθούν τα αποτελέσματα της διαμέρισης βάσει εντοπισμού κοινοτήτων στα τεχνητά σύνθετα δίκτυα που κατασκευάστηκαν, όπως και οι παρατηρήσεις μας πάνω σε αυτά. Δίπλα σε κάθε μοντέλο βρίσκεται σε παρένθεση το συνολικό πλήθος των κόμβων που ταξινομήθηκαν σε αυτό. Όλες οι τιμές είναι στρογγυλοποιημένες στα 2 δεκαδικά ψηφία.

5.4.1 Barabasi-Albert και Random Geometric

Πίνακας 5.19: *Αποτελέσματα του Girvan Newman για τον συνδυασμό Barabasi-Albert και Random Geometric*

	Girvan Newman
Πλήθος κοινοτήτων	2
Μοντέλα κοινοτήτων	Barabasi-Albert (519), Random Geometric (392)
Μέγεθος κοινοτήτων	519, 392
Πλήθος διαγεγραμμένων κόμβων	0
Πλήθος σωστά ταξινομημένων κόμβων	911 (από τους 911)
Πλήθος σωστά ταξινομημένων κοινοτήτων	2
Modularity	0.27
Χρόνος εκτέλεσης (s)	656.18

Πίνακας 5.20: *Αποτελέσματα των Walktrap και Infomap για τον συνδυασμό Barabasi-Albert και Random Geometric*

	Walktrap	**Infomap**
Πλήθος κοινοτήτων	4	5
Μοντέλα κοινοτήτων	Barabasi-Albert (1517), Random Geometric (1392)	Barabasi-Albert (1519), Random Geometric (1392)
Μέγεθος κοινοτήτων	1517, 599, 485, 308	1519, 418, 374, 324, 276
Πλήθος διαγεγραμμένων κόμβων	0	2
Πλήθος σωστά ταξινομημένων κόμβων	2909 (από τους 2909)	2911 (από τους 2911)
Πλήθος σωστά ταξινομημένων κοινοτήτων	4	5
Modularity	0.45	0.50
Χρόνος εκτέλεσης (s)	300.65	139.42

Πίνακας 5.21: *Αποτελέσματα των Leiden και Spectral Clustering για τον συνδυασμό Barabasi-Albert και Random Geometric*

	Leiden	**Spectral Clustering**
Πλήθος κοινοτήτων	5	2
Μοντέλα κοινοτήτων	Barabasi-Albert (1519), Random Geometric (1392)	Barabasi-Albert (1519), Random Geometric (1392)
Μέγεθος κοινοτήτων	1519, 364, 361, 336, 331	1519, 1392
Πλήθος διαγεγραμμένων κόμβων	0	0
Πλήθος σωστά ταξινομημένων κόμβων	2911 (από τους 2911)	2911 (από τους 2911)
Πλήθος σωστά ταξινομημένων κοινοτήτων	5	2
Modularity	0.50	0.08
Χρόνος εκτέλεσης (s)	127.97	1049.18

Παρατηρήσεις

Όλοι οι αλγόριθμοι διαμερίζουν σωστά τον γράφο στις δύο συνιστώσες του, με μοναδική εξαίρεση τον Walktrap που οδηγεί στην παράλειψη 2 μόλις κόμβων. Ο Spectral Clustering είναι αισθητά αργότερος από τους υπόλοιπους, ενώ ο συντομότερος είναι ο Leiden.

5.4.2 Barabasi-Albert και Watts-Strogatz

Πίνακας 5.22: *Αποτελέσματα του Girvan Newman για τον συνδυασμό Barabasi-Albert και Watts-Strogatz*

	Girvan Newman
Πλήθος κοινοτήτων	2
Μοντέλα κοινοτήτων	Watts-Strogatz (818)
Μέγεθος κοινοτήτων	752, 66
Πλήθος διαγεγραμμένων κόμβων	0
Πλήθος σωστά ταξινομημένων κόμβων	299 (από τους 818)
Πλήθος σωστά ταξινομημένων κοινοτήτων	1
Modularity	0.04
Χρόνος εκτέλεσης (s)	41.28

Πίνακας 5.23: *Αποτελέσματα των Walktrap και Infomap για τον συνδυασμό Barabasi-Albert και Watts-Strogatz*

	Walktrap	**Infomap**
Πλήθος κοινοτήτων	1	1
Μοντέλα κοινοτήτων	Barabasi-Albert (1578)	Barabasi-Albert (1495)
Μέγεθος κοινοτήτων	1578	1495
Πλήθος διαγεγραμμένων κόμβων	1240	1323
Πλήθος σωστά ταξινομημένων κόμβων	1519 (από τους 1578)	1495 (από τους 1495)
Πλήθος σωστά ταξινομημένων κοινοτήτων	1	1
Modularity	0.18	0.21
Χρόνος εκτέλεσης (s)	92.15	61.96

Παρατηρήσεις

Οι περισσότεροι αλγόριθμοι εντοπίζουν μόνο ένα από τα μοντέλα Barabasi-Albert και Watts-Strogatz. Ο μόνος που αναγνωρίζει και τα δύο είναι ο Leiden, ο οποίος εντοπίζει λανθασμένα και το μοντέλο Erdos-Renyi, όμως ταξινομεί σχετικά λίγους κόμβους σε αυτό (212). Ταυτόχρονα, ο Leiden και ο Spectral Clustering είναι οι μόνοι υποψήφιοι αλγόριθμοι που δεν δημιουργούν ένα τεράστιο πλήθος μικρών κοινοτήτων (≤ 20 κόμβοι) που δεν μπορούν να ταξινομηθούν. Ο Leiden έχει, επιπλέον, το μεγαλύτερο πλήθος ορθά ταξινομημένων κόμβων συνολικά, και τον μικρότερο χρόνο εκτέλεσης.

Πίνακας 5.24: *Αποτελέσματα των Leiden και Spectral Clustering για τον συνδυασμό Barabasi-Albert και Watts-Strogatz*

	Leiden	**Spectral Clustering**
Πλήθος κοινοτήτων	20	1
Μοντέλα κοινοτήτων	Barabasi-Albert (1461), Watts-Strogatz (1139), Erdos-Renyi (212)	Watts-Strogatz (2808)
Μέγεθος κοινοτήτων	441, 409, 299, 212, 151, 126, 125, 118, 95, 93, 92, 87, 86, 77, 76, 74, 68, 67, 62, 54	2808
Πλήθος διαγεγραμμένων κόμβων	6	10
Πλήθος σωστά ταξινομημένων κόμβων	2089 (από τους 2812)	1289 (από τους 2808)
Πλήθος σωστά ταξινομημένων κοινοτήτων	13	0
Modularity	0.36	0.00
Χρόνος εκτέλεσης (s)	16.76	225.68

5.4.3 Watts-Strogatz και Erdos-Renyi

Πίνακας 5.25: *Αποτελέσματα του Girvan Newman για τον συνδυασμό Watts-Strogatz και Erdos-Renyi*

	Girvan Newman
Πλήθος κοινοτήτων	1
Μοντέλα κοινοτήτων	Erdos-Renyi (699)
Μέγεθος κοινοτήτων	699
Πλήθος διαγεγραμμένων κόμβων	1
Πλήθος σωστά ταξινομημένων κόμβων	400 (από τους 699)
Πλήθος σωστά ταξινομημένων κοινοτήτων	0
Modularity	0.00
Χρόνος εκτέλεσης (s)	14.45

Παρατηρήσεις

Και σε αυτή την περίπτωση, οι περισσότεροι αλγόριθμοι εντοπίζουν μόνο ένα (ή κανένα στην περίπτωση του Infomap) από τα μοντέλα Watts-Strogatz και Erdos-Renyi. Την εξαίρεση αποτελεί ο Walktrap, οποίος εντοπίζει λανθασμένα και το Barabasi-Albert. Όμως ο Walktrap οδηγεί στην παράβλεψη 1861 κόμβων, και στην σωστή ταξινόμηση μόνο 281 από τους 839 που απομένουν. Έτσι ο βέλτιστος αλγόριθμος σε αυτήν την περίπτωση είναι Leiden, ο οποίος επιδεικνύει την καλύτερη επίδοση μαζί με τον Spectral Clustering, και ολοκληρώνεται σε πολύ λιγότερο χρόνο από αυτόν.

Πίνακας 5.26: *Αποτελέσματα των Walktrap και Infomap για τον συνδυασμό Watts-Strogatz και Erdos-Renyi*

	Walktrap	**Infomap**
Πλήθος κοινοτήτων	12	10
Μοντέλα κοινοτήτων	Watts-Strogatz (688), Barabasi-Albert (122), Erdos-Renyi (29)	Barabasi-Albert (236)
Μέγεθος κοινοτήτων	367, 149, 54, 48, 46, 29, 29, 25, 24, 24, 23, 21	29, 26, 25, 25, 24, 22, 22, 21, 21, 21
Πλήθος διαγεγραμμένων κόμβων	1861	2464
Πλήθος σωστά ταξινομημένων κόμβων	281 (από τους 839)	0 (από τους 236)
Πλήθος σωστά ταξινομημένων κοινοτήτων	0	0
Modularity	0.27	0.33
Χρόνος εκτέλεσης (s)	4.73	12.21

Πίνακας 5.27: *Αποτελέσματα των Leiden και Spectral Clustering για τον συνδυασμό Watts-Strogatz και Erdos-Renyi*

	Leiden	**Spectral Clustering**
Πλήθος κοινοτήτων	21	2
Μοντέλα κοινοτήτων	Watts-Strogatz (2700)	Watts-Strogatz (2700)
Μέγεθος κοινοτήτων	179, 168, 150, 135, 132, 127, 125, 125, 124, 124, 123, 122, 122, 122, 120, 120, 119, 118, 118, 114, 113	1745, 955
Πλήθος διαγεγραμμένων κόμβων	0	0
Πλήθος σωστά ταξινομημένων κόμβων	1299 (από τους 2700)	1299 (από τους 2700)
Πλήθος σωστά ταξινομημένων κοινοτήτων	0	0
Modularity	0.38	0.22
Χρόνος εκτέλεσης (s)	8.67	102.95

5.4.4 Waxman και Random Geometric

Πίνακας 5.28: *Αποτελέσματα του Girvan Newman για τον συνδυασμό Waxman και Random Geometric*

	Girvan Newman
Πλήθος κοινοτήτων	2
Μοντέλα κοινοτήτων	Random Geometric (392), Waxman (333)
Μέγεθος κοινοτήτων	392, 333
Πλήθος διαγεγραμμένων κόμβων	0
Πλήθος σωστά ταξινομημένων κόμβων	725 (από τους 725)
Πλήθος σωστά ταξινομημένων κοινοτήτων	2
Modularity	0.23
Χρόνος εκτέλεσης (s)	502.36

Πίνακας 5.29: *Αποτελέσματα των Walktrap και Infomap για τον συνδυασμό Waxman και Random Geometric*

	Walktrap	**Infomap**
Πλήθος κοινοτήτων	8	8
Μοντέλα κοινοτήτων	Random Geometric (1392), Waxman (1333)	Random Geometric (1392), Waxman (1333)
Μέγεθος κοινοτήτων	538, 512, 323, 300, 291, 273, 257, 231	422, 389, 382, 336, 323, 321, 285, 267
Πλήθος διαγεγραμμένων κόμβων	0	0
Πλήθος σωστά ταξινομημένων κόμβων	2725 (από τους 2725)	2725 (από τους 2725)
Πλήθος σωστά ταξινομημένων κοινοτήτων	8	8
Modularity	0.51	0.54
Χρόνος εκτέλεσης (s)	151.42	100.20

Πίνακας 5.30: *Αποτελέσματα των Leiden και Spectral Clustering για τον συνδυασμό Waxman και Random Geometric*

	Leiden	**Spectral Clustering**
Πλήθος κοινοτήτων	5	2
Μοντέλα κοινοτήτων	Waxman (1333), Random Geometric (1011), Gilbert (381)	Random Geometric (1392), Waxman (1333)
Μέγεθος κοινοτήτων	1333, 381, 354, 340, 317	1392, 1333
Πλήθος διαγεγραμμένων κόμβων	0	0
Πλήθος σωστά ταξινομημένων κόμβων	2344 (από τους 2725)	2725 (από τους 2725)
Πλήθος σωστά ταξινομημένων κοινοτήτων	4	2
Modularity	0.57	0.28
Χρόνος εκτέλεσης (s)	186.16	1091.43

Παρατηρήσεις

Σε αυτήν την περίπτωση όλοι οι αλγόριθμοι εκτός από τον Leiden έχουν την καλύτερη δυνατή επίδοση, όσον αφορά τα αποτελέσματά τους. Παρ' όλα αυτά, το σφάλμα του Leiden είναι σχετικά μικρό (381 κόμβοι από τους 2725). Ο ελάχιστος χρόνος εκτέλεσης ανήκει στον Infomap, με μικρή διαφορά από τους Walktrap και Leiden.

5.4.5 Gilbert και Waxman

Πίνακας 5.31: *Αποτελέσματα του Girvan Newman για τον συνδυασμό Gilbert και Waxman*

	Girvan Newman
Πλήθος κοινοτήτων	2
Μοντέλα κοινοτήτων	Gilbert (200), Waxman (333)
Μέγεθος κοινοτήτων	333, 200
Πλήθος διαγεγραμμένων κόμβων	0
Πλήθος σωστά ταξινομημένων κόμβων	533 (από τους 533)
Πλήθος σωστά ταξινομημένων κοινοτήτων	2
Modularity	0.32
Χρόνος εκτέλεσης (s)	208.19

Πίνακας 5.32: *Αποτελέσματα των Walktrap και Infomap για τον συνδυασμό Gilbert και Waxman*

	Walktrap	**Infomap**
Πλήθος κοινοτήτων	2	5
Μοντέλα κοινοτήτων	Waxman (1333), Gilbert (1200)	Waxman (1333), Gilbert (1200)
Μέγεθος κοινοτήτων	1333, 1200	1200, 384, 332, 328, 289
Πλήθος διαγεγραμμένων κόμβων	0	0
Πλήθος σωστά ταξινομημένων κόμβων	2533 (από τους 2533)	2533 (από τους 2533)
Πλήθος σωστά ταξινομημένων κοινοτήτων	2	5
Modularity	0.19	0.16
Χρόνος εκτέλεσης (s)	347.12	241.99

Πίνακας 5.33: *Αποτελέσματα των Leiden και Spectral Clustering για τον συνδυασμό Gilbert και Waxman*

	Leiden	**Spectral Clustering**
Πλήθος κοινοτήτων	2	2
Μοντέλα κοινοτήτων	Waxman (1333), Gilbert (1200)	Waxman (1333), Gilbert (1200)
Μέγεθος κοινοτήτων	1333, 1200	1333, 1200
Πλήθος διαγεγραμμένων κόμβων	0	0
Πλήθος σωστά ταξινομημένων κόμβων	2533 (από τους 2533)	2533 (από τους 2533)
Πλήθος σωστά ταξινομημένων κοινοτήτων	2	2
Modularity	0.19	0.19
Χρόνος εκτέλεσης (s)	332.01	327.99

Παρατηρήσεις

Και για αυτόν τον συνδυασμό δικτύων, όλοι οι αλγόριθμοι οδηγούν στην σωστή αναγνώριση των μοντέλων των συνιστωσών του σύνθετου γράφου, με όλους τους κόμβους να ταξινομούνται ορθά. Τον μικρότερο χρόνο εκτέλεσης επιδεικνύει ο Infomap, με σχετικά μικρή διαφορά με τους υπόλοιπους (ο Girvan Newman εξαιρείται από την σύγκριση, καθώς τρέχει σε μικρότερα δίκτυα από τους υπόλοιπους).

5.4.6 Τελική απόφαση

Στις περιπτώσεις των συνδυασμών των μοντέλων Barabasi-Albert με Random Geometric, και Gilbert με Waxman όλοι οι αλγόριθμοι είχαν την βέλτιστη επίδοση χωρίς να ξεχωρίζει κάποιος. Στις περιπτώσεις των συνδυασμών των Barabasi-Albert με Watts-Strogatz, και Watts-Strogatz με Erdos-Renyi ο Leiden παρουσιάζει με διαφορά καλύτερη επίδοση από τους υπόλοιπους, για τους λόγους που προαναφέρθηκαν. Τέλος, στην περίπτωση του συνδυασμού Waxman με Random Geometric ο Leiden είναι ο μοναδικός που παρουσιάζει κάποιο σφάλμα, όμως αυτό είναι μικρό. Έτσι, επιλέχθηκε ο Leiden για την ταξινόμηση των πραγματικών δικτύων, καθώς θεωρήθηκε πως έχει τα συνολικά καλύτερα αποτελέσματα.

Θα πρέπει να σημειωθεί, επίσης, ότι και ο αλγόριθμος Spectral Clustering είχε επίδοση συγκρίσιμη με του Leiden όσον αφορά την ορθότητα των αποτελεσμάτων του. Δεν επιλέχθηκε, όμως, αφού οι χρόνοι εκτέλεσής του ήταν υπερβολικά μεγάλοι, και το γεγονός ότι ένα από τα ορίσματά του είναι το επιθυμητό πλήθος κοινοτήτων είναι αρκετά περιοριστικό όταν πρόκειται για πραγματικά δίκτυα (καθώς δεν γνωρίζουμε εκ των προτέρων το πλήθος των μοντέλων από τα οποία αποτελείται ένα δίκτυο, και μία λανθασμένη εκτίμηση μπορεί να οδηγήσει σε απώλεια πληροφορίας).

Για την ακρίβεια, υπάρχει πιθανότητα η καλή επίδοση του αλγορίθμου Spectral Clustering να οφείλεται κυρίως, ή και εξ ολοκλήρου, στο γεγονός ότι «γνωρίζει» εκ των προτέρων το πλήθος το πλήθος των κοινοτήτων στις οποίες πρέπει να διαμεριστεί ο γράφος. Έτσι, ενθαρρύνεται η διαμέριση στις σωστές συνιστώσες, και αποφεύγεται η διάσπαση σε πολλές μικρές κοινότητες που πιθανώς δεν μπορούν να ταξινομηθούν σωστά. Ίσως εάν ζητούσαμε από τον Spectral Clustering να διαχωρίσει τα δίκτυα σε 6 συνιστώσες (το μέγιστο δυνατό πλήθος διαφορετικών μοντέλων που μπορούν να εμφανιστούν σε έναν γράφο), τα αποτελέσματά του να ήταν πολύ διαφορετικά. Από την άλλη, είναι πιθανό η λειτουργία του βάσει της ομοιότητας μεταξύ των κόμβων να είναι ο κύριος λόγος της καλής επίδοσής του.

Όσον αφορά τους αλγορίθμους Walktrap και Infomap, παρατηρούμε πως μοιράζονται τις εξής δύο ιδιότητες: βασίζονται στους τυχαίους περιπάτους, και δεν έχουν καλή επίδοση σε σύγκριση με τους υπόλοιπους που εξετάσαμε. Μπορούμε να συμπεράνουμε, λοιπόν, πως η αναζήτηση κοινοτήτων σε έναν γράφο μέσω τυχαίων περιπάτων σε αυτόν δεν είναι η βέλτιστη προσέγγιση όταν στοχεύουμε στην διαμέριση αυτού στις συνιστώσες που αντιστοιχούν στα διαφορετικά μοντέλα τα οποία περιέχει. Τέλος, μπορούμε να υποθέσουμε πως ο λόγος που ο Leiden επιτυγχάνει τόσο καλά αποτελέσματα είναι ότι βασίζεται στην μεγιστοποίηση του modularity, η οποία (εκ του ορισμού του modularity) οδηγεί στην αφαίρεση των γεφυρών ενός γράφο έτσι ώστε να σχηματιστούν οι κοινότητές του, ενώ ο τρόπος που έχουμε ενώσει τους

επί μέρους γράφους ώστε να σχηματίσουν τους σύνθετους είναι μέσω γεφυρών. Εξάλλου, και ο Girvan Newman, που βασίζεται στην αφαίρεση των γεφυρών, παρουσιάζει πολύ καλή επίδοση (ανεξάρτητα από τον απαγορευτικό χρόνο εκτέλεσής του). Βέβαια, αυτό σημαίνει πως η σύνδεση των διαφορετικών συνιστωσών ενός γράφου μέσω γεφυρών μπορεί να είναι προϋπόθεση για την καλή επίδοση του Leiden, το οποίο δεν γνωρίζουμε αν ισχύει εν γένει για τα πραγματικά δίκτυα. Όμως, βάσει των αποτελεσμάτων μας, αυτός παραμένει η καλύτερή μας επιλογή για την τμηματοποίησή τους.

5.5 Τμηματοποίηση Πραγματικών Δικτύων

Ακολουθούν τα μοντέλα στα οποία ταξινομήθηκαν οι κοινότητες στις οποίες διαμερίστηκαν τα πραγματικά δίκτυα με την χρήση του αλγορίθμου Leiden, μαζί με το πλήθος κόμβων που ταξινομήθηκαν σε αυτά.

- **Football:** – (Ο συγκεκριμένος γράφος είναι υπερβολικά μικρός για να διαιρεθεί σε κοινότητες με παραπάνω από 20 κόμβους)
- **GOT:** Barabasi-Albert (58)
- **Email-EU:** Barabasi-Albert (444), Gilbert (348), Waxman (194)
- **AdvogatoTrust:** Barabasi-Albert (4302), Waxman (278), Watts-Strogatz (218), Erdos-Renyi (200)
- **MorenoCrime:** Barabasi-Albert (461), Watts-Strogatz (315), Erdos-Renyi (36)
- **NetScience:** Barabasi-Albert (260), Waxman (24)
- **Euroroad:** Watts-Strogatz (947), Barabasi-Albert (32)
- **ChicagoRegional:** Watts-Strogatz (12979)
- **Flights-FAA:** Watts-Strogatz (579), Barabasi-Albert (363), Erdos-Renyi (284)
- **Bristol:** Watts-Strogatz (2529)
- **Coach:** Watts-Strogatz (1961), Erdos-Renyi (287), Barabasi-Albert (48)
- **YuYeast:** Barabasi-Albert (810), Watts-Strogatz (788)
- **BacillusMegaterium:** Barabasi-Albert (4319)
- **Malaria:** Random Geometric (31)
- **Fullerene:** Watts-Strogatz (3840)
- **PowerNet:** Watts-Strogatz (4796), Erdos-Renyi (95), Barabasi-Albert (31)
- **R-Dependancies:** Gilbert (29)

Παρατηρούμε πως τα περισσότερα δίκτυα είτε περιέχουν ένα μοναδικό εξελικτικό μοντέλο (αυτό στο οποίο είχαν ταξινομηθεί εξ αρχής), είτε αποτελούνται από ένα συνδυασμό μοντέλων ο οποίος εμπεριέχει το αρχικό μοντέλο στο οποίο ταξινομήθηκαν. Εξαίρεση αποτελούν μόνο τα δίκτυα NetScience και BacillusMegaterium στα οποία δεν εντοπίζεται κάποια συνιστώσα που αντιστοιχεί στο μοντέλο το οποίο είχαν ταξινομηθεί αρχικά. Επιπλέον, σε όλα τα κοινωνικά δίκτυα, και τα περισσότερα βιολογικά, η μεγαλύτερη συνιστώσα αντιστοιχεί στο μοντέλο Barabasi-Albert (Scale-Free). Μάλιστα, η Barabasi-Albert είναι η πιο κοινή συνιστώσα, με εμφανίσεις σε 11 δίκτυα, ενώ ακολουθεί με μικρή διαφορά (10 εμφανίσεις) η συνιστώσα Watts-Strogatz. Η τρίτη στη σειρά σε πλήθος εμφανίσεων είναι η συνιστώσα Erdos-Renyi, με μεγάλη διαφορά από τις προηγούμενες (μόνο 5 εμφανίσεις). Τα υπόλοιπα μοντέλα εμφανίζονται ακόμη πιο σπάνια, με το Waxman να εμφανίζεται 3 φορές (όλες σε κοινωνικά δίκτυα), το Gilbert 2, και το Random Geometric 1.

Εδώ θα πρέπει να σημειώσουμε πως υπάρχει πιθανότητα η σχετικά μεγάλη συχνότητα εμφάνισης της συνιστώσας Erdos-Renyi να οφείλεται σε σφάλμα του Leiden, και να μην αντανακλά την πραγματικότητα. Παρατηρούμε πως κατά την αξιολόγηση των αλγορίθμων εντοπισμού κοινοτήτων, ο Leiden εντόπισε λανθασμένα δύο φορές κάποια συνιστώσα που αντιστοιχούσε σε Τυχαίο Γράφο (μία φορά Erdos-Renyi, και μία φορά Gilbert). Οι λανθασμένες αυτές συνιστώσες ήταν οι μικρότερες του αντίστοιχου γράφου και στις δύο περιπτώσεις, όπως είναι και η Erdos-Renyi τις περισσότερες φορές που εμφανίζεται στην παρούσα διαδικασία τμηματοποίησης και ταξινόμησης. Λόγω αυτού του μοτίβου, δημιουργείται αμφιβολία σχετικά με την αξιοπιστία των αποτελεσμάτων που αφορούν το μοντέλο Erdos-Renyi.

Μέρος III

Επίλογος

Κεφάλαιο 6

Κατακλείδα

Το κεφάλαιο αυτό αποτελεί την κατακλείδα της παρούσας εργασίας. Περιέχει μια σύνοψη των μεθόδων και των αποτελεσμάτων μας, τα συμπεράσματα στα οποία καταλήξαμε βάση των προαναφερθέντων, όπως και ιδέες για μελλοντικές επεκτάσεις.

6.1 Σύνοψη και Συμπεράσματα

Στα πλαίσια της παρούσης διπλωματικής εργασίας και στηριζόμενοι σε παρεμφερή άρθρα, διεξήγαμε αρχικά μία έρευνα στην βιβλιογραφία, όπου εντοπίσαμε 6 εξελικτικά μοντέλα, και 20 μετρικές δικτύων οι οποίες θα μπορούσαν δυνητικά να χρησιμοποιηθούν για την ταξινόμηση δικτύων στα μοντέλα αυτά. Ύστερα εκτελέσαμε μία σειρά ελέγχων πάνω σε αυτές τις μετρικές (προσδίδοντας ιδιαίτερη βαρύτητα στην διασπορά που παρουσιάζει η κάθε μετρική συνολικά για τα μοντέλα, και στον χρόνο εκτέλεσής της), και επιλέξαμε 7 για να χρησιμοποιηθούν ως είσοδοι σε έναν ταξινομητή. Ο ταξινομητής επίσης επιλέχθηκε μετά από έλεγχο της επίδοσής του, με την χρήση των μετρικών precision, recall, accuracy, και F1 score. Αφού εκπαιδεύτηκε σε ένα σύνολο 600 συνθετικών δικτύων που κατασκευάσαμε οι ίδιοι, χρησιμοποιήθηκε για την ταξινόμηση 17 πραγματικών δικτύων, με είσοδο τις προαναφερθείσες μετρικές. Επιπλέον, αξιολογήσαμε 5 αλγορίθμους τμηματοποίησης δικτύων βάσει του χρόνου εκτέλεσής τους, και της ακρίβειας των αποτελεσμάτων τους σε 5 συνθετικά δίκτυα που αποτελούσαν το καθένα σύνθεση δύο άλλων δικτύων τα οποία είχαν κατασκευαστεί με την χρήση διαφορετικών εξελικτικών μοντέλων. Αυτός που παρήγαγε τα ορθότερα αποτελέσματα, σε ένα εύλογο χρονικό διάστημα, επιλέχθηκε ώστε να χρησιμοποιηθεί για τα προαναφερθέντα πραγματικά δίκτυα. Τα υποδίκτυα που προέκυψαν ταξινομήθηκαν ύστερα με την βοήθεια του επιλεγμένου ταξινομητή.

Συγκεκριμένα, τα μοντέλα που χρησιμοποιήθηκαν ήταν τα Erdos-Renyi, Gilbert, Random Geometric, Watts-Strogatz, Barabasi-Albert, και Waxman. Οι μετρικές που επιλέχθηκαν ήταν η Κατανομή Βαθμών Κόμβων, το Μέσο Μήκος Μονοπατιού, η Διάμετρος, η Κεντρικότητα της Εγγύτητας, η Κεντρικότητα της Πληροφορίας, η Κεντρικότητα των Αρμονικών, και η Συνάφεια, ενώ ο ταξινομητής ήταν ο Random Forest, και ο αλγόριθμος τμηματοποίησης ο Leiden. Όσον αφορά τα αποτελέσματα των ταξινομήσεων, αυτά παρουσιάζονται αναλυτικά στο Κεφάλαιο 5, όμως μπορούμε να παρατηρήσουμε ότι το μοντέλο που εμφανίζεται πιο συχνά όταν τα δίκτυα εξετάζονται αυτούσια είναι το Watts-Strogatz (Small-World). Όταν τα δίκτυα τμηματοποιούνται, παρατηρούμε πως σχεδόν όλα περιέχουν μια συνιστώσα που

αντιστοιχεί στο μοντέλο στο οποίο είχαν ταξινομηθεί αρχικά. Επιπλέον, οι συνιστώσες που αντιστοιχούν στο μοντέλο Barabasi-Albert είναι ιδιαίτερα κοινές, καθώς εντοπίζονται στα περισσότερα δίκτυα, και έχουν την μεγαλύτερη συχνότητα εμφάνισης συγκριτικά με αυτές που αντιστοιχούν σε άλλα μοντέλα (ακολουθούμενες από αυτές που αντιστοιχούν στο μοντέλο Watts-Strogatz). Μάλιστα, σε όλα τα κοινωνικά δίκτυα, και τα περισσότερα βιολογικά, η συνιστώσα Barabasi-Albert είναι η μεγαλύτερη, ανεξάρτητα του μοντέλου στο οποίο είχε ταξινομηθεί αρχικά το κάθε δίκτυο. Αντιστοίχως, σε όλα τα δίκτυα μεταφορών η μεγαλύτερη συνιστώσα είναι αυτή που αντιστοιχεί στο μοντέλο Watts-Strogatz, γεγονός αναμενόμενο αφού όλα είχαν ταξινομηθεί σε αυτό το μοντέλο εξ αρχής.

Βάσει αυτών των αποτελεσμάτων, συμπεραίνουμε ότι τα μοντέλα που εμφανίζονται πιο συχνά στα πραγματικά δίκτυα είναι το Small-World και το Scale-Free. Αναλύοντας τα δίκτυα σε κατηγορίες, θεωρούμε ασφαλές να διατυπώσουμε τον ισχυρισμό ότι τα δίκτυα μεταφορών χαρακτηρίζονται ως επί το πλείστον από το μοντέλο Small-World, ο οποίος συνάδει και με τα συμπεράσματα των Jiang και Claramunt [68] και των Sen et al. [69]. Επιπλέον, η θεώρηση πως τα πραγματικά δίκτυα μπορούν να αναλυθούν σε συνιστώσες που αντιστοιχούν σε διαφορετικά εξελικτικά μοντέλα φαίνεται πως είναι σωστή, καθώς σχεδόν όλα τα δίκτυα περιέχουν μία συνιστώσα που αντιστοιχεί στο μοντέλο όπου είχαν ταξινομηθεί όταν μελετήθηκαν αυτούσια, και η παρουσία του μοντέλου Scale-Free ως η μεγαλύτερη συνιστώσα σε κάθε κοινωνικό δίκτυο, και μόνο σε αυτά (με την εξαίρεση του δικτύου YuYeast), δεν μπορεί να οφείλεται σε σφάλμα ή σε σύμπτωση. Ακόμη, λόγω της παρουσίας αυτής σε όλα τα κοινωνικά δίκτυα που εξετάσαμε, συμπεραίνουμε ότι τα κοινωνικά δίκτυα περιέχουν εν γένει μία μεγάλη Scale-Free συνιστώσα, και ότι ίσως μπορούν να χαρακτηριστούν ολόκληρα ως Scale-Free εφόσον αναζητούμε ένα ενιαίο μοντέλο, αφού η συνιστώσα αυτή είναι με διαφορά η μεγαλύτερη. Το γεγονός ότι τα περισσότερα από αυτά ταξινομήθηκαν ως Small-World όταν τα δίκτυα εξετάστηκαν αυτούσια, πιθανώς να οφείλεται σε σφάλμα του ταξινομητή λόγω της παρουσίας πολλών μοντέλων, και της ομοιότητας του μοντέλου Scale-Free με το Small-World (σε σημείο που ορισμένοι επιστήμονες θεωρούν το Scale-Free ως υποκατηγορία του Small-World [81]). Πρέπει, επίσης, να σημειώσουμε πως το μοντέλο Waxman εμφανίζεται ως συνιστώσα στα περισσότερα κοινωνικά δίκτυα, ενώ ορισμένα ταξινομήθηκαν ολόκληρα σε αυτό. Το γεγονός αυτό υποδηλώνει πως και το Waxman απαντάται συχνά στα κοινωνικά δίκτυα. Στις υπόλοιπες κατηγορίες δικτύων δεν παρουσιάζεται κάποιο ισχυρό μοτίβο ώστε να μπορέσουμε να εξάγουμε ανάλογα συμπεράσματα, όμως αξίζει να αναφερθεί πως το ηλεκτρικό δίκτυο που εξετάσαμε (PowerNet) ταξινομήθηκε στην κατηγορία των Small-World δικτύων, σε συμφωνία με τα ευρήματα σχετικών ερευνών που αναφέρθηκαν σε προηγούμενα κεφάλαια [21] [65] [3].

Όσον αφορά τον προβληματισμό σχετικά με την συχνότητα εμφάνισης του μοντέλου Scale-Free, καταλήγουμε στο συμπέρασμα πως ενώ δεν είναι πανταχού παρόν, δεν είναι ούτε σπάνιο. Η κατηγορία δικτύων στην οποία εμφανίζεται πιο συχνά είναι τα κοινωνικά δίκτυα, όμως εμφανίζεται και σε άλλες κατηγορίες ως δευτερεύουσα συνιστώσα (δίκτυα μεταφορών, τεχνολογικά), ή ακόμη και ως πρωτεύουσα (βιολογικά δίκτυα).

6.2 Μελλοντικές Επεκτάσεις

Στα πλαίσια της παρούσας διπλωματικής, ενώ χρησιμοποιήσαμε πληθώρα μεθόδων και τεχνολογιών για την απάντηση των ερωτημάτων που θέσαμε, δεν τις εξαντλήσαμε. Παρατίθενται ορισμένοι τρόποι με τους οποίους θα μπορούσε να εμπλουτιστεί το περιεχόμενο της εργασίας:

- Χρήση περισσότερων μοντέλων. Ενώ χρησιμοποιήσαμε τα περισσότερα μοντέλα που εντοπίσαμε στην βιβλιογραφία, υπήρξαν και κάποια που αγνοήσαμε εσκεμμένα, όπως αναφέρεται και στο Κεφάλαιο 4. Θα ήταν ενδιαφέρον να εξεταστεί και η περίπτωση που όλα τα μοντέλα συμπεριλαμβάνονται στην δημιουργία του συνόλου εκπαίδευσης, και αργότερα στην ταξινόμηση.

- Χρήση περισσότερων μετρικών δικτύων. Κατά την διαδικασία επιλογής μετρικών, απορρίψαμε ορισμένες λόγω του υπερβολικά μεγάλου χρόνου εκτέλεσης του υπολογισμού τους, ενώ εμφάνιζαν μεγάλη διασπορά. Εφόσον υπήρχε περισσότερος διαθέσιμος χρόνος ή κάποιος τρόπος να επισπευσθεί ο υπολογισμός τους, θα μπορούσαν να συμπεριληφθούν με σκοπό την καλύτερη επίδοση του ταξινομητή. Θα μπορούσαν να χρησιμοποιηθούν, επιπλέον, κάποιες μετρικές πέρα από τις 20 που εμείς εξετάσαμε.

- Χρήση διαφορετικού ταξινομητή. Η επιλογή των ταξινομητών που εξετάστηκαν έγινε ακολουθώντας διεξοδικά το διάγραμμα 4.1. Όμως η κατηγορία των Ταξινομητών Συνόλου είναι παρουσιάζει μεγάλο εύρος, με αποτέλεσμα να μην εξεταστούν όλοι οι ταξινομητές που ανήκουν σε αυτή, παρά μόνο οι 4 σημαντικότεροι. Επιπλέον, το διάγραμμα αναφέρεται μόνο σε κλασικούς ταξινομητές, ενώ και τα GNN χρησιμοποιούνται για την ταξινόμηση δικτύων με πολύ καλά αποτελέσματα. Αν και ο ταξινομητής που επιλέξαμε παρουσιάζει πολύ καλή επίδοση, είναι πιθανό κάποιος άλλος να είχε ακόμη καλύτερη.

- Χρήση περισσότερων δικτύων για την εκπαίδευση του ταξινομητή. Λόγω χρονικών περιορισμών, χρησιμοποιήσαμε μόνο 100 δίκτυα ανά μοντέλο για την δημιουργία του συνόλου δεδομένων εκπαίδευσης του ταξινομητή. Εάν υπήρχε η χρονική ευχέρεια να χρησιμοποιήσουμε περισσότερα δίκτυα ανά μοντέλο, ίσως να παρατηρούσαμε μία καλύτερη επίδοση του ταξινομητή.

- Εμπλουτισμός των δεδομένων εκπαίδευσης με κατευθυνόμενα δίκτυα και δίκτυα με βάρη. Όλα τα συνθετικά δίκτυα που κατασκευάσαμε με σκοπό την ένταξή τους στα σύνολα εκπαίδευσης και ελέγχου του ταξινομητή ήταν μη κατευθυνόμενα και χωρίς βάρη. Είναι πιθανό η συμπερίληψη κατευθυνόμενων δικτύων και δικτύων με βάρη στο σύνολο εκπαίδευσης του ταξινομητή, να οδηγούσε σε καλύτερη επίδοσή του σε περιπτώσεις πραγματικών δικτύων με αυτά τα χαρακτηριστικά.

- Διαφορετική διαχείριση των κατευθυνόμενων γράφων. Ο τρόπος που διαχειριστήκαμε τα πραγματικά δίκτυα που ήταν κατευθυνόμενα ήταν να αγνοήσουμε την ιδιότητά τους αυτή. Όμως θα μπορούσαμε να τα έχουμε αντιμετωπίσει με διαφορετικό τρόπο, παραδείγματος χάρη όπως οι Broido και Clauset στην αντίστοιχη έρευνά τους [70].

Ενδεχομένως σε αυτήν την περίπτωση και τα αποτελέσματα της ταξινόμησης και της τμηματοποίησής τους να ήταν διαφορετικά.

- Χρήση διαφορετικού αλγορίθμου τμηματοποίησης δικτύων. Για την εύρεση του βέλτιστου αλγορίθμου τμηματοποίησης των πραγματικών δικτύων σε συνιστώσες που χαρακτηρίζονται από διαφορετικά μοντέλα, εξετάσαμε 5 αλγορίθμους εντοπισμού κοινοτήτων. Όμως υπάρχει ένα πολύ μεγάλο πλήθος αλγορίθμων εντοπισμού κοινοτήτων, συνεπώς είναι πιθανό ο βέλτιστος αλγόριθμος για το πρόβλημα αυτό να είναι κάποιος με τον οποίο δεν ασχοληθήκαμε.

- Επαλήθευση των αποτελεσμάτων με την χρήση του συντελεστή μικρού κόσμου. Ένα μεγάλο μέρος των πραγματικών δικτύων που εξετάστηκαν ταξινομήθηκε στο μοντέλο Small-World. Θα ήταν ενδιαφέρον να υπολογίσουμε τον συντελεστή μικρού κόσμου σ για τα δίκτυα αυτά, ώστε να διαπιστώσουμε εάν τα αποτελέσματά μας επαληθεύονται και με αυτή την μέθοδο εντοπισμού Small-World δομής.

6.3 Διαθεσιμότητα προγραμματιστικού περιεχομένου

Το σύνολο των αρχείων που περιέχουν τον κώδικα και τα αποτελέσματά μας είναι διαθέσιμο στον παρακάτω σύνδεσμο:

https://github.com/FaidraA/Evolution_model_identification

Βιβλιογραφία

[1] M. E. J. Newman. *The Structure and Function of Complex Networks*. *SIAM Review*, 45(2):167–256, 2003.

[2] Pedro J. Zufiria και Iker Barriales-Valbuena. *Evolution models for dynamic networks*. *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*, σελίδες 252–256, 2015.

[3] Rafael Espejo, Sara Lumbreras και Andres Ramos. *Analysis of transmission-power-grid topology and scalability, the European case study*. *Physica A: Statistical Mechanics and its Applications*, 509:383–395, 2018.

[4] Hans Hinterberger. *Graph*, σελίδες 1260–1261. Springer US, Boston, MA, 2009.

[5] Mrs Kothimbire, D. Shelke, Mrs Gaikwad, A. Yelpale και R. Shinde. *A Comprehensive Review of Graph Theory Applications in Network Analysis*. *International Journal of Mathematics And Computer Research*, 13, 2025.

[6] Geoffrey Canright. *Complex NetworksComplex networksand Graph Theory*, σελίδες 1244–1253. Springer New York, New York, NY, 2009.

[7] NY Comdori, 2020.

[8] M. E. J. Newman. *Analysis of weighted networks*. *Phys. Rev. E*, 70:056131, 2004.

[9] Wikipedia contributors. *Path (graph theory) — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Path_(graph_theory)&oldid=1275103758, 2025.

[10] Wikipedia contributors. *Connectivity (graph theory) — Wikipedia, The Free Encyclopedia*, 2025.

[11] Geeksforgeeks authors. *Graph and its representations*, 2024.

[12] Arnab Chakraborty, 2020.

[13] Ernesto Estrada. *The structure of complex networks: theory and applications*. American Chemical Society, 2012.

[14] Mikhail Drobyshevskiy και Denis Turdakov. *Random Graph Modeling: A Survey of the Concepts*. *ACM Comput. Surv.*, 52(6), 2019.

[15] Frédéric Amblard, Audren Bouadjio-Boulic, Carlos Sureda Gutiérrez και Benoit Gaudou. *Which models are used in social simulation to generate social networks? a review of 17 years of publications in JASSS. 2015 Winter Simulation Conference (WSC)*, σελίδες 4021–4032, 2015.

[16] Quentin Duchemin και Yohann De Castro. *Random Geometric Graph: Some Recent Developments and Perspectives. High Dimensional Probability IX*Radosław Adamczak, Nathael Gozlan, Karim Lounici και Mokshay Madiman, επιμελητές, σελίδες 347–392, Cham, 2023. Springer International Publishing.

[17] Matthew Roughan, Jonathan Tuke και Eric Parsonage. *Estimating the Parameters of the Waxman Random Graph. Algorithms and Models for the Web Graph*Konstantin Avrachenkov, Paweł Prałat και Nan Ye, επιμελητές, σελίδες 71–86, Cham, 2019. Springer International Publishing.

[18] Albert László Barabási, Erzsébet Ravasz και Tamás Vicsek. *Deterministic scale-free networks. Physica A: Statistical Mechanics and its Applications*, 299(3):559–564, 2001.

[19] Albert Laszlo Barabasi και Reka Albert. *Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509-512. Science (New York, N.Y.)*, 286:509–12, 1999.

[20] Cristina Guzman, Daphna Keidar, Tristan Meynier, Andreas Opedal και Niklas Stoehr. *Recovering Barabási-Albert Parameters of Graphs through Disentanglement*, 2021.

[21] Duncan J. Watts και Steven H. Strogatz. *Collective dynamics of 'small-world' networks. Nature*, 393(6684):440–442, 1998.

[22] Javier Martın Hernández και Piet Van Mieghem. *Classification of graph metrics. Delft University of Technology: Mekelweg, The Netherlands*, 1, 2011.

[23] Charles Perez και Rony Germon. *Chapter 7 - Graph Creation and Analysis for Linking Actors: Application to Social Data. Automating Open Source Intelligence*Robert Layton και Paul A. Watters, επιμελητές, σελίδες 103–129. Syngress, Boston, 2016.

[24] Simon Mukwembi. *A note on diameter and the degree sequence of a graph. Applied Mathematics Letters*, 25(2):175–178, 2012.

[25] Chintan Amrit και Joanneter Maat. *Understanding Information Centrality Metric: A Simulation Approach*, 2018.

[26] Karen Stephenson και Marvin Zelen. *Rethinking centrality: Methods and examples. Social Networks*, 11(1):1–37, 1989.

[27] Paolo Boldi και Sebastiano Vigna. *Axioms for Centrality*, 2013.

[28] M. Barthélemy. *Betweenness centrality in large complex networks. The European Physical Journal B*, 38(2):163–168, 2004.

[29] Alfredo Cuzzocrea, Alexis Papadimitriou, Dimitrios Katsaros και Yannis Manolopoulos. *Edge betweenness centrality: A novel algorithm for QoS-based topology control over wireless sensor networks. Journal of Network and Computer Applications*, 35(4):1210–1217, 2012. Intelligent Algorithms for Data-Centric Sensor Networks.

[30] M. Girvan και M. E. J. Newman. *Community structure in social and biological networks. Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[31] Gnana Thedchanamoorthy, Mahendra Piraveenan, Dharshana Kasthuriratna και Upul Senanayake. *Node Assortativity in Complex Networks: An Alternative Approach. Procedia Computer Science*, 29:2449–2461, 2014. 2014 International Conference on Computational Science.

[32] Punam Bedi και Chhavi Sharma. *Community detection in social networks. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6:ν/α–ν/α, 2016.

[33] Natalie R Smith, Paul N Zivich, Leah M Frerichs, James Moody και Allison E Aiello. *A guide for choosing community detection algorithms in social network studies: The Question Alignment approach. Am. J. Prev. Med.*, 59(4):597–605, 2020.

[34] Vincent D Blondel, Jean Loup Guillaume, Renaud Lambiotte και Etienne Lefebvre. *Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[35] Pascal Pons και Matthieu Latapy. *Computing Communities in Large Networks Using Random Walks. Journal of Graph Algorithms and Applications*, 10(2):191–218, 2006.

[36] Martin Rosvall και Carl T Bergstrom. *Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. U. S. A.*, 105(4):1118–1123, 2008.

[37] V. A. Traag, L. Waltman και N. J.van Eck. *From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports*, 9(1), 2019.

[38] Zhixin Zhou και Arash A. Amini. *Analysis of spectral clustering algorithms for community detection: the general bipartite setting*, 2018.

[39] Ulrikevon Luxburg. *A tutorial on spectral clustering. Statistics and Computing*, 17(4):395–416, 2007.

[40] Harry Sevi, Matthieu Jonckheere και Argyris Kalogeratos. *Generalized Spectral Clustering for Directed and Undirected Graphs*, 2022.

[41] Gopinath Rebala, Ajay Ravi και Sanjay Churiwala. *Machine Learning Definition and Basics*, σελίδες 1–17. Springer International Publishing, Cham, 2019.

[42] Weiche Hsieh, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang, Keyu Chen, Pohsun Feng, Yizhu Wen, Xinyuan Song, Tianyang Wang, Ming Liu, Junjie Yang, Ming Li, Bowen Jing, Jintao Ren, Junhao Song, Hong Ming Tseng, Yichao Zhang, Lawrence K. Q. Yan, Qian Niu, Silin Chen,

Yunze Wang και Chia Xin Liang. *A Comprehensive Guide to Explainable AI: From Classical Models to LLMs*, 2024.

[43] Julianna Delua. *Supervised versus unsupervised learning: What's the difference?*, 2021.

[44] Gopinath Rebala, Ajay Ravi και Sanjay Churiwala. *Classification*, σελίδες 57–66. Springer International Publishing, Cham, 2019.

[45] Gireen Naidu, Tranos Zuva και Elias Mmbongeni Sibanda. *A review of evaluation metrics in machine learning algorithms. Computer science on-line conference*, σελίδες 15–25. Springer, 2023.

[46] Mohammad Hossin και Md Nasir Sulaiman. *A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process*, 5(2):1, 2015.

[47] Margherita Grandini, Enrico Bagli και Giorgio Visani. *Metrics for Multi-Class Classification: an Overview*, 2020.

[48] Himani Bhavsar και Mahesh H Panchal. *A review on support vector machine for data classification. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 1(10):185–189, 2012.

[49] Gongde Guo, Hui Wang, David Bell, Yaxin Bi και Kieran Greer. *KNN Model-Based Approach in Classification. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*Robert Meersman, Zahir Tari και Douglas C. Schmidt, επιμελητές, σελίδες 986–996, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[50] Ruihu Wang. *AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review. Physics Procedia*, 25:800–807, 2012. International Conference on Solid State Devices and Materials Science, April 1-2, 2012, Macao.

[51] Alexey Natekin και Alois Knoll. *Gradient boosting machines, a tutorial. Frontiers in neurorobotics*, 7:21, 2013.

[52] Wei Fan και Kun Zhang. *Bagging*, σελίδες 206–210. Springer US, Boston, MA, 2009.

[53] Lior Rokach και Oded Maimon. *Decision Trees*, σελίδες 165–192. Springer US, Boston, MA, 2005.

[54] Aakash Parmar, Rakesh Katariya και Vatsal Patel. *A Review on Random Forest: An Ensemble Classifier. International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*Jude Hemanth, Xavier Fernando, Pavel Lafata και Zubair Baig, επιμελητές, σελίδες 758–763, Cham, 2019. Springer International Publishing.

[55] Travis Meyer, Cesar Ramirez, Matthew Tamasi και Adam Gormley. *A User's Guide to Machine Learning for Polymeric Biomaterials. ACS Polymers Au*, 3, 2022.

[56] Max Bramer. *Ensemble Classification*, σελίδες 209–220. Springer London, London, 2020.

[57] Linhong Zhu, Wee Keong Ng και Shuguo Han. *Classifying Graphs Using Theoretical Metrics: A Study of Feasibility. Database Systems for Adanced Applications*Jianliang Xu, Ge Yu, Shuigeng Zhou και Rainer Unland, επιμελητές, σελίδες 53–64, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[58] Réka Albert και Albert László Barabási. *Statistical mechanics of complex networks. Reviews of Modern Physics*, 74(1):47–97, 2002.

[59] Sergey Dorogovtsev και José Fernando Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*, τόμος 57. Oxford University Press, 2003.

[60] Gábor Szárnyas, Zsolt Kővári, Ágnes Salánki και Dániel Varró. *Towards the characterization of realistic models: evaluation of multidisciplinary graph metrics. Proceedings of the ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems*, MODELS '16, σελίδα 87–94, New York, NY, USA, 2016. Association for Computing Machinery.

[61] John Boaz Lee, Ryan Rossi και Xiangnan Kong. *Graph Classification using Structural Attention. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, σελίδα 1666–1674, New York, NY, USA, 2018. Association for Computing Machinery.

[62] Federico Errica, Marco Podda, Davide Bacciu και Alessio Micheli. *A Fair Comparison of Graph Neural Networks for Graph Classification*, 2022.

[63] Keyulu Xu, Weihua Hu, Jure Leskovec και Stefanie Jegelka. *How Powerful are Graph Neural Networks?*, 2019.

[64] Qawi K. Telesford, Karen E. Joyce, Satoru Hayasaka, Jonathan H. Burdette και Paul J. Laurienti. *The ubiquity of small-world networks*, 2011.

[65] Bálint Hartmann και Viktória Sugár. *Searching for small-world and scale-free behaviour in long-term historical data of a real-world power grid. Scientific Reports*, 11(1):6575, 2021.

[66] Michael Stumpf, Piers Ingram, I. Nouvel και C. Wiuf. *Statistical Model Selection Methods Applied to Biological Networks. Transactions on Computational Systems Biology III*, 3, 2005.

[67] Raya Khanin και Ernst Wit. *How Scale-Free Are Biological Networks. Journal of Computational Biology*, 13(3):810–818, 2006. PMID: 16706727.

[68] Bin Jiang και Christophe Claramunt. *Topological Analysis of Urban Street Networks. Environment and Planning B: Planning and Design*, 31(1):151–162, 2004.

[69] Parongama Sen, Subinay Dasgupta, Arnab Chatterjee, P. A. Sreeram, G. Mukherjee και S. S. Manna. *Small-world properties of the Indian railway network. Phys. Rev. E*, 67:036106, 2003.

[70] Anna D. Broido και Aaron Clauset. *Scale-free networks are rare. Nature Communications*, 10(1):1017, 2019.

[71] Albert László Barabási. *Love is all you need. Clauset's fruitless search for scale-free networks. unpublished*, 2018.

[72] Ivan Voitalov, Pimvan der Hoorn, Remcovan der Hofstad και Dmitri Krioukov. *Scale-free networks well done. Phys. Rev. Res.*, 1:033034, 2019.

[73] Linqing Liu, Mengyun Shen και Chang Tan. *Scale free is not rare in international trade networks. Scientific Reports*, 11(1):13359, 2021.

[74] I. Artico, I. Smolyarenko, V. Vinciotti και E. C. Wit. *How rare are power-law networks really? Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2241):20190742, 2020.

[75] Gipsi Lima-Mendez και Jacquesvan Helden. *The powerful law of the power law and other myths in network biology. Molecular bioSystems*, 5:1482–93, 2009.

[76] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel και Bobby Bhattacharjee. *Measurement and Analysis of Online Social Networks*. σελίδες 29–42, 2007.

[77] Petter Holme. *Rare and everywhere: Perspectives on scale-free networks. Nature Communications*, 10(1):1016, 2019.

[78] Markus Meringer και Eric W. Weisstein. *Regular Graph*. From MathWorld–A Wolfram Resource.

[79] M.E.J. Newman και D.J. Watts. *Renormalization group analysis of the small-world network model. Physics Letters A*, 263(4):341–346, 1999.

[80] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot και E. Duchesnay. *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, 12:2825–2830, 2011.

[81] L. A. N. Amaral, A. Scala, M. Barthélémy και H. E. Stanley. *Classes of small-world networks. Proceedings of the National Academy of Sciences*, 97(21):11149–11152, 2000.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

π.χ.	παραδείγματος χάριν
ΗΒ	Ηνωμένο Βασίλειο
ΗΠΑ	Ηνωμένες Πολιτείες της Αμερικής
AI	Artificial Intelligence
GNN	Graph Neural Network
k-NN	k-Nearest Neighbors
Rk-NN	Reverse k-Nearest Neighbors
SVC	Support Vector Classifier
SVM	Support Vector Machine

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
Ακρίβεια	Accuracy
Αληθώς αρνητικό	True negative
Αληθώς θετικό	True positive
Ανάκληση	Recall
Αρθρότητα	Modularity
Βαθμοί διαχωρισμού	Degrees of separation
Δέντρο απόφασης	Decision tree
Δίκτυα μικρού-κόσμου	Small-world networks
Διανύσμα στήριξης	Support vector
Εκθετικός τυχαίος γράφος	Exponential random graph
Ενδιαμεσική κεντρικότητα	Betweenness centrality
Εντοπισμός κοινοτήτων	Community detection
Επιβλεπόμενη μάθηση	Supervised learning
Επισημασμένα δεδομένα	Labeled data
Εξηγήσιμη τεχνητή νοημοσύνη	Explainalble artificial intelligence
Ετικέτα	Label
Ευστοχία	Precision
Ζωτικότητα της εγγύτητας	Closeness vitality
Ιεραρχικός γράφος	Hierarchical graph
Κανονικός γράφος	Lattice
Κεντρικότητα της γεφύρωσης	Bridging centrality
Κεντρικότητα της δρομολόγησης	Routing centrality
Κεντρικότητα της διήθησης	Percolation centrality
Κεντρικότητα της εγγύτητας	Closeness centrality
Κεντρικότητα της εγγύτητας της τρέχουσας ροής	Current flow closeness centrality
Κεντρικότητα της πληροφορίας	Information centrality
Κεντρικότητα των αρμονικών	Harmonic centrality
Κεντρικότητα των ιδιοδιανυσμάτων	Eigenvector centrality
Κεντρικότητα του φορτίου	Load centrality
Λαπλασιανή κεντρικότητα	Laplacian centrality
Μελέτη αφαίρεσης	Ablation study
Μείωση της διαστατικότητας	Dimensionality reduction
Μη επιβλεπόμενη μάθηση	Unsupervised learning
Μη επισημασμένα δεδομένα	Unlabeled data

Μη-κλιμακούμενο δίκτυο	Scale-free network
Νόμος δύναμης	Power law
Παλινδρόμηση	Regression
Πίνακας σύγχυσης	Confusion matrix
Πρόβλημα του μαύρου κουτιού	Black box problem
Συνάφεια	Assortativity
Συνάρτηση χάρτη	Map function
Συνάρτησης απώλειας	Loss function
Συντελεστής πλουσίου συλλόγου	Rich club coefficient
Συντελεστής Συσταδοποίησης	Clustering Coefficient
Συσταδοποίηση	Clustering
Συσχέτιση	Association
Ταξινομητής	Classifier
Ταξινομητής Συνόλου	Ensemble classifier
Ταξινόμηση	Classification
Τυχαίο δάσος	Random forest
Τυχαίος γράφος	Random graph
Τυχαίος γεωμετρικός γράφος	Random geometric graph
Φαινόμενο του μικρού κόσμου	Small-world phenomenon
Φασματική κεντρικότητα	Spectral centrality
Χωρικό δίκτυο	Spatial network
Ψευδώς αρνητικό	False negative
Ψευδώς θετικό	False positive
Ψηφοφορία	Voting

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Communication, Electronic and Information Engineering

Evolution model identification in synthetic and real-world networks

Diploma Thesis

of

FAIDRA ANASTASIA ANTHOPOULOU

Supervisor: Symeon Papavassileiou
NTUA Professor

Athens, October 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Communication, Electronic and Information Engineering

Evolution model identification in synthetic and real-world networks

Diploma Thesis

of

FAIDRA ANASTASIA ANTHOPOULOU

Supervisor: Symeon Papavassileiou
NTUA Professor

Approved by the examination committee on 2 October 2025.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Symeon Papavassileiou	Eleni Stai	Vasileios Karyotis
NTUA Professor	NTUA Assistant Professor	IU Professor

Athens, October 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Communication, Electronic and Information Engineering

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....

Faidra Anastasia Anthopoulou

Electrical and Computer Engineering Graduate of NTUA

2 October 2025

Abstract

Over the past two decades, the scientific community has engaged in vigorous discourse regarding the evolution models into which various types of real-world networks are classified, accompanied by a wealth of sometimes conflicting research on the subject. Moreover, the view that a significant proportion of real-world networks consist of components characterized by different evolution models has recently emerged. These developments motivated the focus of the present thesis, which aims to investigate the prevalence of each evolution model in real-world networks, as well as the validity of the aforementioned claim. To achieve this objective, 20 network metrics were first evaluated with respect to their suitability for use as inputs to a classifier. Subsequently, a set of 600 synthetic networks was generated, and the values of the selected metrics were computed for them. These values were then used as training and test data for a series of classifiers, which were evaluated to select the optimal one. The optimal classifier was employed to classify 17 real-world networks, which were then segmented using a selected partitioning algorithm, and the resulting components were classified as well. Analysis of the results showed that Small-World and Scale-Free networks occur most frequently. Moreover, the Small-World model is particularly prevalent in transportation networks, whereas the Scale-Free model is predominantly observed as the largest component in social networks. The Waxman model also frequently appears as a component in social networks, while no clear pattern was observed for the remaining models. Finally, our findings support the notion that networks can indeed consist of components characterized by different evolution models.

Keywords

Complex Networks, Evolution Models, Machine Learning, Network Partitioning, Network Classification

Acknowledgements

I would like to express my gratitude to Professor S. Papavassiliou for the trust he placed in me by assigning this thesis to me, as well as for his guidance and invaluable advice. I would also like to thank Dr. K. Tsitseklis for his suggestions and his assistance throughout its preparation.

Athens, October 2025

Faidra Anastasia Anthopoulou

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter serves as an introduction to the present diploma thesis. It includes a concise presentation of its subject matter, a description of the approach we followed in order to address the problem under consideration, and information regarding its overall structure.

1.1 Subject Matter

The presence of networks is profoundly pervasive across the vast majority of aspects of human life. From our social interactions, transactions, and transportation, to the very cells that make up our bodies, networks are ubiquitous. For this reason, their study has long been a central focus of the scientific community, with the aim of better understanding their structure and behavior.

An important tool in Network Science for accomplishing this task is network models, which (as their name suggests) aim to model networks and their behavior [1]. Specifically, evolution models, which are the focus of the present work, concern the evolution of a network over time, that is, the way in which new connections are formed and consequently the resulting structure of the network based on its past. Thus, classifying a network into the correct evolution model based on its current state can provide a wealth of information about its future behavior and structure [2]. For instance, knowing the evolution model (and by extension the topology) of a power grid allows us to simulate it through a synthetic network constructed according to that model, and facilitates robustness analysis [3].

There are several different approaches for performing this classification. In the literature, one can find methods based on theoretical metrics, as well as approaches that utilize Artificial Intelligence tools, such as Machine Learning models and Neural Networks. Furthermore, it has been suggested that many networks are better described by two or more models corresponding to different components of the network, rather than by a single model representing the network as a whole.

1.2 Our Approach

Motivated by the above, we decided to investigate the identification of evolution models in both real-world and synthetic networks, within the framework of the present thesis.

Our approach involves the selection of a set of network metrics, following an evaluation of their suitability for our specific problem. These metrics are subsequently calculated for the network to be classified, and their values are used as inputs to a classifier that has been shown to be the most effective. The classifier output provides an answer to the question of which evolution model best describes the network. Additionally, we examine the notion that networks may sometimes consist of components characterized by different evolution models, by segmenting them into subnetworks using an appropriate partitioning method (which we have also selected based on performance evaluation), and then classifying these subnetworks following the same procedure described for entire networks.

1.3 Structure

The present thesis is organized into the following chapters:

- **Chapter 2 (Theoretical Framework):** Provides the reader with definitions and certain properties of the concepts used, with the aim of facilitating a smooth introduction to the subject of the thesis.
- **Chapter 3 (Literature Overview):** Presents a concise overview of a range of studies relevant to the topic of this thesis, thereby outlining the scientific context in which the research was conducted.
- **Chapter 4 (Methodology):** Details the methods and technologies employed to achieve the objectives of the thesis, as well as the criteria that guided their selection.
- **Chapter 5 (Results):** Contains the results of the experimental part of the thesis and certain observations on them, along with the decisions made based on them.
- **Chapter 6 (Conclusion):** Summarizes the methodology and results, presents the conclusions drawn from our findings, and proposes directions for potential future research.

Part I

Theoretical Foundations

Chapter 2

Theoretical Framework

In this chapter, we provide information regarding the key concepts used in this work, as well as the technologies required for the implementation of its experimental part. To be exact, we will analyze concepts and algorithms related to Graphs, Complex networks, and Machine Learning.

2.1 Graph Theory Fundamentals

2.1.1 Graph Definition

A graph is the mathematical representation of a set of objects, and the relationships between them. These objects constitute the **vertices** V of the graph, while the relationships between them are represented by the graph's **edges** E. Therefore, a graph can be mathematically represented by the pair of sets $\{V, E\}$. [4] [5]

Graphs are frequently referred to as **networks**, as they are a highly important tool for the study and representation of both real-world and synthetic networks, to the extent that these two concepts are at times regarded as equivalent. For example, a group of classmates constitutes a social network that can be represented by a graph, wherein the nodes correspond to the students and the edges denote the friendships among them. [6] [5]

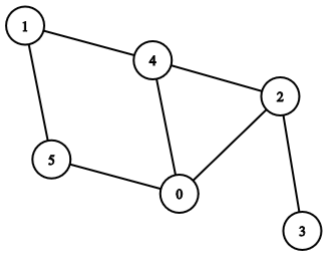


Figure 2.1. *Example of a graph with 6 nodes and 7 edges*

2.1.2 Graph Properties

Directedness

A graph is called **undirected** if each edge (v_i, v_j) is equivalent to the edge (v_j, v_i), where v_i and v_j are the nodes incident to the edge. In such a graph, each edge denotes a bidirectional relationship. An example of a network represented by an undirected graph is a social networking platform where symmetric friendships are formed between members (e.g. Facebook). [5]

Conversely, a graph is called **directed** if the edges v_i and v_j are distinct from each other. In this case, each edge is considered to have a specific direction, indicated by the order in which the incident nodes are listed in the pair. Thus, the relationship represented by each edge is one-way. Similarly, an example of a network represented by a directed graph is a social networking platform where asymmetric "follower–following" relationships are formed between members (e.g. X). [5]

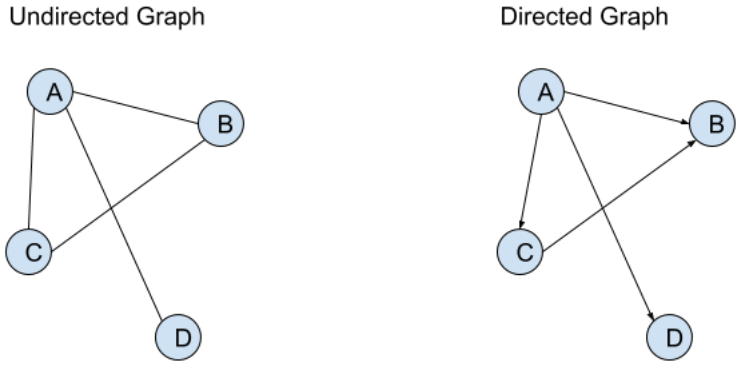


Figure 2.2. *Two graphs: On the left an undirected one, On the right a directed one* [7]

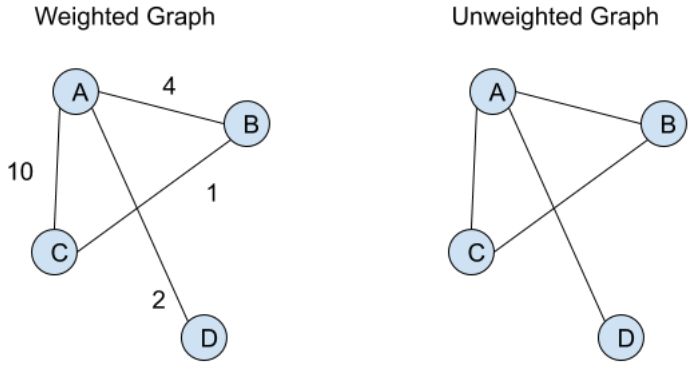


Figure 2.3. *Two graphs: On the left a weighted one, On the right an unweighted one* [7]

Weightedness

A graph is termed **weighted** when each edge is assigned a numerical value, which constitutes the **weight** of the edge, and **unweighted** in the opposite case. The weight of an

edge represents the value of some characteristic of the connection to which it corresponds, thereby differentiating the connections and enabling the modeling of phenomena such as cost, distance, capacity, and others. In unweighted graphs, all edges (and therefore all connections) are considered equivalent to one another. [5]

Adjacency and Walks

Two nodes v_i, v_j are called **neighbors** if there exists an edge (v_i, v_j) or (v_j, v_i) connecting them. The number of neighbors of a node constitutes its **degree** in the case of an unweighted graph. In a weighted graph, the degree of a node is defined as the sum of the weights of all the edges that connect it to its neighbors [8]. In directed graphs, the in-degree and out-degree of a node are defined as the number of edges directed toward or away from the node, respectively.

If there exists a sequence of edges (and corresponding vertices) that can be followed to travel from node v_i to node v_j, then this sequence is called a **walk** between v_i and v_j. If no edge is repeated along the walk, it is called a **trail**, and if, in addition, no vertex is repeated, it is called a **path**. [9]

Connectivity and Components

A graph is called **connected** if there exists at least one walk between any two of its nodes (ignoring the possible directions of the edges), and **disconnected** otherwise. If the graph is directed and connected, then it is called **strongly connected** if it remains connected when the edge directions are taken into account, otherwise it is called **weakly connected**. [10]

A **connected component** of a graph consists of a subset of the graph's vertices and edges (i.e. a **subgraph**) that is connected. Every graph contains at least one connected component. By removing edges from a graph, it can be split into multiple connected components, a technique commonly used in many graph algorithms (for example in those aimed at community detection). [10]

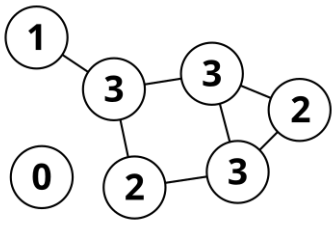


Figure 2.4. *A disconnected graph with two connected components* [10]

2.1.3 Graph Representation

Graphs are typically represented either by means of an **adjacency matrix** or by means of an **adjacency list**.

Adjacency Matrix

The adjacency matrix of an unweighted graph is a matrix of size $N \times N$, where N is the number of vertices in the graph. If the graph is undirected, then the entries in the matrix with coordinates (i,j) and (j,i) contain the number 1 if there exists an edge connecting nodes v_i and v_j, and 0 otherwise. If the graph is directed, then the entry with coordinates (i,j) contains 1 only if there exists an edge (v_i, v_j) (that is, from node i to node j), and 0 otherwise. [11]

For a weighted graph, the same applies, except that the presence of an edge in the adjacency matrix is not denoted by the number 1, but by its weight. [12]

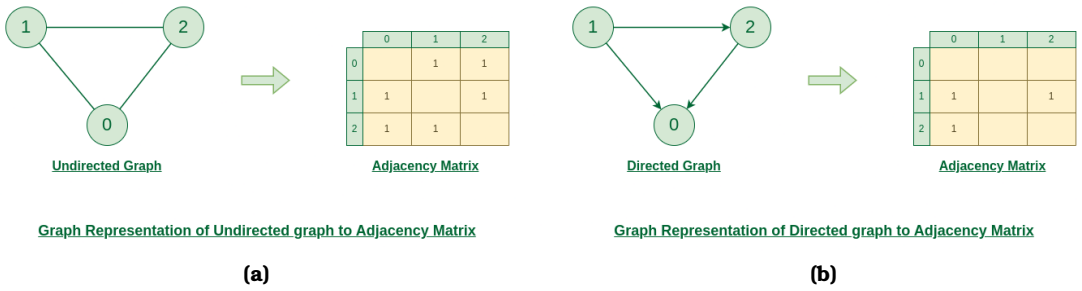


Figure 2.5. *Examples of unweighted graphs and their corresponding adjacency matrices:* ***(a)*** *Undirected graph,* ***(b)*** *Directed graph*

[11]

Adjacency List

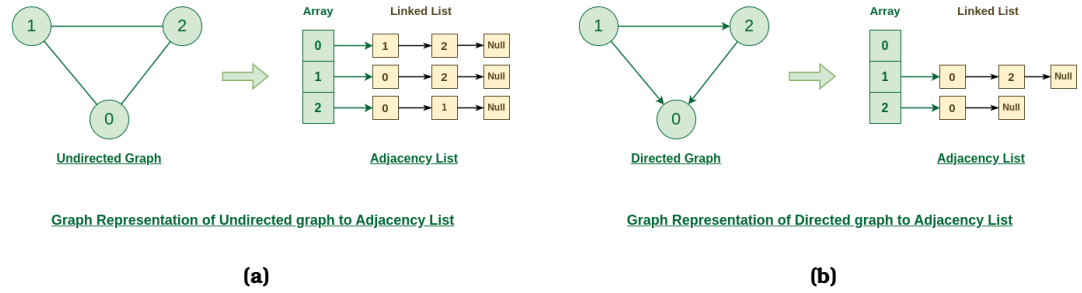


Figure 2.6. *Examples of graphs and their corresponding adjacency lists:* ***(a)*** *Undirected graph,* ***(b)*** *Directed graph*

[11]

The adjacency list of an unweighted graph consists of an array of size $N \times 1$, where at each position i there is a linked list containing all the neighbors of v_i. If the graph is directed, then the neighbor list of a node v_i includes only those neighbors v_j that are connected to it by an edge directed from v_i to v_j (that is, edge (v_i, v_j)). If the graph is undirected, then all neighbors of a node are included in the corresponding list, since the relationships are bidirectional. [11]

For a weighted graph, all the above applies, along with the fact that, together with each neighbor v_j of a node v_i recorded in the adjacency list of v_i, the weight of the edge (v_i, v_j) is also recorded. Thus, each element of the neighbor list of a node contains two values: the identifier of a neighboring node, and the weight of the edge connecting them. [12]

2.2 Complex Networks

2.2.1 Complex Network Definition

A network is a representation of a system, through its nodes which represent the entities of the system, and its edges which represent the relationships or interactions between these entities. A **complex network** is a network whose structure is intricate, in terms of the volume of information it represents and contains. Examples of complex networks include road networks, protein–protein interaction networks, the World Wide Web, trade networks, collaboration networks, citation networks, and others. [13]

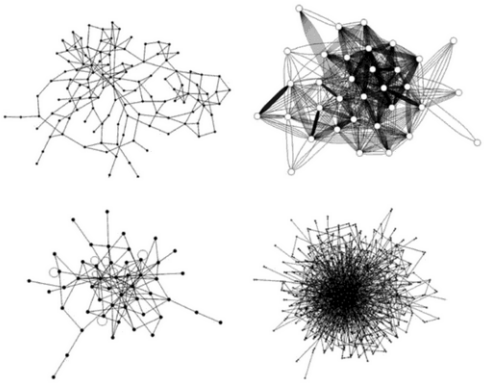


Figure 2.7. *Different types of networks:* ***(Top left)*** *Representation of an electronic circuit,* ***(Top right)*** *Representation of the interactions of a group of researchers at a West Virginia university,* ***(Bottom left)*** *Representation of the protein-protein interactions of a strain of the Herpes virus,* ***(Bottom right)*** *Representation of a connected component of the metabolic network of Helicobacter pylori as a directed network*
[13]

2.2.2 Evolution Models of Complex Networks

In this subsection, we will present the evolution models that this work aims to identify in complex networks.

Random Graphs

A Random Graph is a graph in which the probability of an edge existing between two vertices is independent of the properties of those vertices and depends only on some probability measure. In the vast majority of cases, a Random Graph is represented using one of the following two models: the Erdos-Renyi model ($G(n, m)$) or the Gilbert model ($G(n, p)$). [14]

- **Erdos-Renyi (G(n, m)):** In this model, one graph is chosen uniformly at random from all possible graphs with n vertices and m edges. [14]

- **Gilbert (G(n, p)):** In this model, the graph consists of n vertices, and each possible edge is independently formed with probability p.[14]

Some examples of network categories where these models find application are social networks, biological networks, and the World Wide Web. [14]

Spatial Networks

A Spatial Network is a network (or graph) in which the probability of the existence of an edge is determined by some measure of distance between the nodes it connects. [15] In the context of this work, among all models for representing Spatial Networks, the ones that were selected were the Random Geometric Graph (G(n, r)) and the Waxman model (G(n, q, s)).

- **Random Geometric Graph (G(n, r)):** In this model, n points are chosen at random positions in a plane, which will form the nodes of the graph, and any nodes whose distance is less than or equal to r are connected by an edge. [16]

- **Waxman (G(n, q, s)):** In this model, n points are chosen at random positions in a plane (usually within the unit square), which will form the nodes of the graph, and each possible edge between two nodes i, j is formed with probability $p(d_{ij}) = qe^{-sd_{ij}}$, where d_{ij} is the distance between them. [17]

These models find application in network types where geographical distance is important, such as mobile phone networks, road networks, and networks of interactions between certain animals. [16] [17]

Finally, it should be noted that randomness is introduced in both of these models, and therefore they can be considered a subcategory of Random Graphs. In other words, they are both Random and Spatial graphs at the same time. [16] [17]

Scale-Free Networks

A Scale-Free network is a network, or graph, in which the probability that a randomly selected node has exactly k neighbors decreases exponentially as a function of k, that is, it follows the Power Law distribution $P(k) \sim k^{-\gamma}$, where γ is a parameter typically in the range $2 < \gamma < 3$. [18] [19] The name "Scale-Free Network" comes from the fact that this exponential degree distribution is independent of the network's size (scale).

The model used for these networks is named **Barabasi-Albert G(n, m)**, where n is the number of nodes and m is the number of edges added each time a new node enters, connecting it to the existing nodes. [5] [20]

Such networks, wherein new nodes are more likely to connect to existing nodes with a high degree (preferential attachment phenomenon) [5], are very common in nature (especially in the case of social networks). This model, for example, can be used to

represent various types of networks such as the World Wide Web, human sexual contact networks, and cellular chemical networks. [18]

Small-World Networks

Small-World networks constitute an intermediate model between Regular Graphs (graphs in which all nodes have the same degree) and Random Graphs. This is achieved via the **Watts-Strogatz** $G(n, k, p)$ model as follows: starting from a Regular Ring Graph with n nodes where each is connected to its k nearest neighbors, each edge (v_i, v_j) is rewired with probability p to become (v_i, v_w), where v_w is a different randomly selected node of the graph, provided that the edge (v_i, v_w) does not already exist. [21]

These networks get their name from the "small-world phenomenon", also known as "6 degrees of separation", since the average distance between two nodes in such a graph is approximately 6 edges long. Some examples of networks represented by this model are the neural network of the Caenorhabditis elegans worm, the US electrical grid, and actor collaboration networks. [21]

2.2.3 Complex Network Metrics

In this subsection, we present the complex network metrics employed for the purposes of the present work.

Degree Distribution

The **degree distribution** of a network shows the number of nodes exhibiting a specific degree, for each degree observed in the network. This metric is of considerable analytical importance, as several of the models under examination display distinctive degree distribution patterns (for instance, the degree distribution of a Scale-Free network is characterized by the Power Law). The **average degree** is the mean of the degrees of all nodes in the network and likewise constitutes valuable information about it. [22]

Average Path Length

The **shortest path** between two nodes is the path with the shortest length (i.e. the fewest number of edges) that connects them. The term **average path length** refers to the mean length of the shortest paths between any two nodes in the graph. [23] It can be mathematically defined as $l = \frac{1}{N(N-1)} \sum_{i \neq j} d(v_i, v_j)$, where $d(v_i, v_j)$ is the distance (shortest path) between nodes v_i and v_j, and N is the number of nodes in the graph. [5]

It constitutes a valuable metric that can provide indicative insight into a graph. An illustrative example is the six-degrees-of-separation property, which pertains to the average path length and is characteristic of Small-World graphs. [23] [5]

Diameter

The **diameter** of a graph is the maximum distance between any two nodes in the graph. It can be mathematically defined as $diam = \max_{v_i, v_j \in V}(d(v_i, v_j))$, where $d(v_i, v_j)$

denotes the distance between the nodes v_i and v_j, and V is the set of nodes of the graph.

The diameter provides important information about a network, as the upper bound of distances within it. For example, in the case of a communication network, the diameter is indicative of the transmission time and the signal attenuation involved in sending a message in the worst-case scenario. [24]

Centrality Measures

Centrality measures are metrics that aim to identify the most important (or most influential) nodes, according to various criteria. [5] The centrality measures employed in the present study are as follows:

- **Closeness Centrality:** Refers to the proximity of a node to all other nodes in the graph to which it belongs, and is therefore inversely proportional to the sum of its distances to all other nodes. It is mathematically defined as $C_C(v_i) = \frac{1}{\sum_j d(v_i,v_j)}$. Starting from a node with high closeness centrality allows one to reach other parts of the network more quickly. It constitutes one of the most widely used metrics for network analysis. [5]

- **Information Centrality:** Concerns the flow of information within a network, as it considers all possible paths between two nodes and evaluates the amount of information passing through them (which is inversely proportional to the path length). The information centrality of a node is defined as the harmonic mean of the information contained in all paths originating from that node, and is mathematically defined as $C_I(i) = \frac{n}{\sum_{j=1}^{n} \frac{1}{I_{ij}}}$, where n is the number of nodes in the graph, and I_{ij} is the amount of information in the path between v_i and v_j (represented as a matrix when multiple paths exist). Information centrality is related to closeness centrality, as it coincides with the current flow closeness centrality. [25] [26]

- **Harmonic Centrality:** Aims to address the issue introduced by the possible absence of a path from one node to another in the computation of closeness centrality. This is achieved by replacing the mean distance with the harmonic mean of the distances, therefore harmonic centrality is mathematically defined as $C_h(i) = \sum_{i \neq j} \frac{1}{d(v_i,v_j)}$. Although the difference from closeness centrality may appear minor, in practice this modification yields substantially different results and facilitates the analysis of weakly connected networks. [27]

- **Betweenness Centrality:** The betweenness centrality of a node v equals the number of shortest paths between other nodes that pass through it, and serves as a measure of a node's influence on the flow of information between other nodes. It is mathematically defined as $C_{NB}(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$, where σ_{st} denotes the total number of shortest paths from node s to node t, and $\sigma_{st}(v)$ denotes the number of shortest paths from s to t that pass through v. [28] Respectively, the betweenness centrality of an edge e is defined as $C_{EB}(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}$. [29] [30]

Assortativity

Assortativity constitutes a measure of the tendency of a graph's nodes to connect to other similar nodes. Various criteria for determining node similarity exist, with the most common one being their degree. In this case, degree assortativity constitutes a *Pearson correlation*, and can be mathematically defined as $r = \frac{1}{\sigma_q^2}[(\sum_{jk} jke_{j,k}) - \mu_q^2]$, where $e_{j,k}$ denotes the joint probability distribution of two neighboring nodes having degrees j and k respectively, while σ_q^2 and μ_q^2 represent the standard deviation and the mean value of the degree probability distribution of the network. [31]

Assortativity assumes values in the interval $[-1, 1]$, with $r = 1$ meaning that all nodes have neighbors exclusively of the same degree as their own, $r = -1$ meaning that no node has a neighbor of the same degree as its own, and $r = 0$ indicating that any node may be connected randomly to any other. In general, positive assortativity values indicate that nodes tend to connect with similar ones, whereas negative values suggest the opposite. This tendency in a network depends on its nature: for example, a social network of friendships between individuals is likely to exhibit positive assortativity, whereas a network representing a food chain is likely to exhibit negative assortativity. Thus, assortativity can serve as a strong indicator of the structure, and therefore the evolution model that a network follows. [31]

2.2.4 Network Partitioning Methods

In this subsection, we will discuss several methods by which a network can be partitioned into subnetworks (or subgraphs).

Community Detection

A **community** can be defined as a group of nodes that interact more frequently and are more strongly connected to each other compared to the other nodes in the network. These nodes may also be closer to each other, either in terms of distance or in terms of similarity. [32] [33]

Community detection is the process that aims to identify the communities within a network and can be implemented using many different algorithms. [33] The metric most frequently used to evaluate the quality of a network's partition into communities is called **modularity**, which measures the density of edges within communities compared to the density of edges whose endpoints belong to different communities. Modularity takes values in the range $[-1, 1]$ and is mathematically expressed as $Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j)$, where m is the number of edges in the graph, A_{ij} is the value at position (i, j) of the graph's adjacency matrix, k_i and k_j are the degrees of nodes v_i and v_j respectively, c_i and c_j are the communities to which v_i and v_j belong respectively, and δ is the function $\delta(u, v)$, which equals 1 if $u = v$ and 0 otherwise. [34] [8]

The following section provides a description of the algorithms used in the context of the present work:

- **Girvan-Newmann:** This algorithm is based on betweenness centrality. Specifically, it removes edges with the highest betweenness, under the assumption that these edges act as bridges between different communities, thereby partitioning the graph into separate communities. Edge removal stops either when the graph has been divided into a specific number of communities defined by the programmer, or when all edges have been removed from the graph. The quality of each resulting partition can be evaluated using a metric (e.g., modularity) to identify the optimal one. The Girvan-Newmann algorithm has a time complexity of $O(m^2n)$, where n is the number of nodes and m the number of edges. Although originally designed for undirected, unweighted graphs, it has been generalized to work with directed and weighted graphs as well. [30] [33] [8]

- **Walktrap:** This algorithm is based on the observation that random walks on a graph tend to become "trapped" within densely connected parts of the graph, which correspond to communities. Random walks are walks in a graph formed by starting at a node and, at each step, randomly selecting the next node from the neighbors of the current node. If the current node is v_i, the probability of transitioning to v_j is $P_{ij} = \frac{A_{ij}}{d(i)}$, where A_{ij} is the entry at position (i,j) of the graph's adjacency matrix, and $d(i)$ is the degree of v_i. The sequence of nodes that constitutes the walk is a *Markov chain*, with states corresponding to the graph's nodes and transition probability equal to P_{ij} at each step. The algorithm applies only to undirected graphs, whether weighted or unweighted. Its worst-case time and space complexities are $O(mn^2)$ and $O(n^2)$ respectively, while its average-case time and space complexities are $O(n^2 \log(n))$ and $O(n^2)$ respectively. [35]

- **Infomap:** This algorithm uses the probability flow of random walks on a graph (i.e., the entry and exit probabilities for each node) to partition it according to the flow of information within it. Groups of nodes through which information moves quickly and easily are considered strongly connected and form communities. More precisely, the algorithm uses a map function that takes a graph partition as input and calculates the average length of a random walk within it, aiming to identify the partition that minimizes the average walk length. The subgraphs in this optimal partition correspond to communities, since shorter random walks indicate denser connections and faster information flow. Infomap can be applied to any type of graph (directed or undirected, weighted or unweighted). [36]

- **Leiden:** This algorithm constitutes an improved version of a different community detection algorithm, Louvain. Louvain is based on modularity maximization. Initially, each node is assigned to a separate community, and iterative attempts to merge communities are made, so that the resulting community has higher modularity than the initial ones. When modularity can no longer be increased, each community is replaced by a single node and the process is repeated until modularity is maximized as much as possible. This process can sometimes result in final communities that are poorly connected or even disconnected. Leiden does not suf-

fer from the problem of disconnected communities and additionally produces better results while running faster. [34] [37] This algorithm applies only to undirected graphs, with or without weights. [33]

- **Spectral Clustering:** This algorithm groups nodes into k communities based on their similarity, which is computed from the first k eigenvectors of the graph Laplacian ($L = D - W$, where D is the diagonal matrix containing the degrees of the nodes and W is the adjacency matrix of the graph). An interesting observation, here, is that the eigenvalues of the Laplacian matrix are qualitatively related to the graph, in various ways. For example, in a connected graph, the eigenvalue 0 always has multiplicity 1, while the first nonzero eigenvalue indicates how well-connected the graph is, with larger values corresponding to more connections. It is also noteworthy that the eigenvalues are in ascending order of value, with the first one always equal to 0. Spectral Clustering is not used exclusively for community detection in graphs, but for a wide range of problems. However, when executed on a graph's adjacency matrix, it can provide a very good partition of the graph. [38] [39] It is mainly applied to undirected graphs, but it can be adapted for directed graphs as well. [40] Additionally, it can be applied to both weighted and unweighted graphs. [39]

2.3 Machine Learning

2.3.1 Machine Learning Definition

Machine Learning is a computer science field that focuses on algorithms and methods for automating the solution of complex problems, which are otherwise difficult to address through conventional programming approaches. [41]

More precisely, in contrast to the conventional approach in which the programmer defines a set of instructions to solve a problem, machine learning algorithms address the problem autonomously, by generating a set of rules called a **model** and which is trained on data to make predictions for new inputs. This facilitates the resolution of complex tasks such as data classification (e.g., the categorization of email as spam or not spam), text recognition from images, speech recognition, and autonomous driving, among others. [41]

Machine learning algorithms tend to produce more accurate results, as the rules they generate are based solely on the data and are therefore more objective than those defined by a potentially biased programmer. On the other hand, the fact that we do not know exactly how they arrived at a particular decision poses a challenge, as it reduces our trust in it, especially in critical contexts such as medical diagnoses. [41] This issue (known as the Black Box Problem) is being actively studied and addressed within the research field of Explainable Artificial Intelligence. [42]

2.3.2 Supervised and Unsupervised Learning

The data on which a machine learning algorithm is trained can be either **labeled** or **unlabeled**, depending on whether they have an associated label or not. A **label** is a feature or category to which a data sample belongs, and which is intended to be predicted for a new input in a machine learning problem based on labeled data. Accordingly, machine learning is classified as **Supervised** when it involves algorithms designed for problems with labeled data, and as **Unsupervised** when the data are unlabeled. [43]

Supervised Learning deals with **classification** problems, which are discussed in the following subsection, and **regression** problems, where the task is to predict a numerical output (the label) from the input features. Examples of such problems include the separation of emails into spam and not spam (classification), and the prediction of a company's future revenues based on its past revenues (regression). [43]

Unsupervised Learning primarily deals with **clustering** problems, where data need to be grouped according to their similarity, **association** problems, where the goal is to discover relationships and correlations among data based on the values of their features, and **dimensionality reduction** problems, where the number of features must be reduced without compromising the quality of the information they provide. Examples of such problems include market segmentation based on similarities among consumers (clustering), the implementation of a recommendation system for an e-Commerce website (association), and denoising image data to improve clarity (dimensionality reduction). [43]

2.3.3 Classifiers

Classifier Definition

The problem of **classification** involves selecting the category (**class**) to which a new input belongs, from a set of possible categories. A **classifier** is a model trained on a set of labeled data, in order to solve the classification problem. The set of possible classes into which a classifier can place a new input is the set of the different labels present in the training data. Many different machine learning techniques can be used to build a classifier, such as Neural Networks, Random Forests, Markov models, and others. [44]

Classifier Performance Metrics

Classifier performance is typically evaluated using **precision**, **recall**, **accuracy**, and the **F1 score** [45]. Before proceeding, it is necessary to define the following concepts for a binary classifier (i.e., one that chooses between only two classes), where one class is designated as positive and the other as negative:

- **True Positive:** The input is positive, and is determined to be positive
- **False Positive:** The input is negative, and is determined to be positive
- **True Negative:** The input is negative, and is determined to be negative
- **False Negative:** The input is positive, and is determined to be negative

[46] [47]

	Actual Positive Class	Actual Negative Class
Predicted Positive Class	True positive (*tp*)	False negative (*fn*)
Predicted Negative Class	False positive (*fp*)	True negative (*tn*)

Figure 2.8. *Table illustrating the definitions of the above terms* [46]

We can now define the aforementioned metrics as follows:

- **Precision** = $\frac{TP}{TP+FP}$ (the proportion of true positives among all inputs classified as positive)
- **Recall** = $\frac{TP}{TP+TN}$ (the proportion of true positives among all actual positive inputs)
- **Accuracy** = $\frac{TP+TN}{TP+FP+TN+FN}$ (the proportion of correctly classified inputs among all inputs)
- **F1 Score** = $2\frac{precision \cdot recall}{precision+recall}$ (the harmonic mean of precision and recall)

[46] [47]

For classifiers that categorize inputs into more than two classes (**multi-class classification**), precision and recall can be computed using either the macroscopic approach (macro precision and macro recall) or the microscopic approach (micro precision and micro recall). The macroscopic approach is generally preferred, since with the microscopic approach, precision equals recall, and accuracy equals the F1 score. [47]

With the **macroscopic approach**, precision and recall are calculated for each class, and then their average is calculated to obtain the overall metric for the classifier. The concepts of true positives, false positives, and false negatives are generalized as follows: **true positives** for a class k are the instances that actually belong to that class and are correctly classified as such, **false positives** are those classified as k but in fact belong to a different class, and **false negatives** are those classified as a different class but actually belong to k. (Note: the table containing the counts of true positives, false positives, and false negatives for all classes is called a **confusion matrix**.) The formulas for calculating overall **precision** and **recall** remain the same, using these generalized definitions. Similarly, the **F1 score** is computed as the harmonic mean of overall precision and recall, while **accuracy** continues to be defined as the proportion of correctly classified inputs among all inputs. [47]

Classifier Examples

The classifiers employed in this study are summarized below:

- **Support Vector Classifier (SVC) and Linear Support Vector Classifier (Linear SVC):** This classifier is a type of **Support Vector Machine (SVM)**, hence its name.

An SVM is a model whose training algorithm's objective is to find the **optimal hyperplane** that separates the regions formed by the members of each class in a geometric space, where the axes correspond to the features of the data (e.g., if the data have 7 features, we obtain a 7-dimensional space, with each axis representing the values of one feature). The members of each class that lie closest to the hyperplane constitute the **support vectors**, from which the classifier derives its name. The optimal hyperplane is defined as the one that maximizes the distance between itself and the support vectors. If the classes are linearly separable, meaning the boundary between them can be a straight line, then we can use a (**Linear SVC**) which specifically seeks a straight line as the boundary (and not a non-linear boundary as in the general case of the **SVC**). In both cases, the space is divided into regions corresponding to each class, and a new input is classified based on the region in which it is placed according to its feature values. [48]

- ***k*-Nearest Neighbors Classifier (*k*-NN):** This classifier also relies on the position of a new input relative to the training data in a geometric space defined exactly as described for SVMs. In this case, the ***k* nearest neighbors** of a new input are identified (i.e., the training samples with the smallest distance from the new input), and the class of the new input is determined according to the majority class among these k nearest neighbors. In some cases, the exact distances of the neighbors from the input are also taken into account as weights in the decision process, with smaller distances corresponding to greater weights. The selection of the appropriate value of k is of crucial importance for the performance of this classifier and must be done carefully. Several methods exist for this purpose, the simplest being to repeat the classification with different values of k and select the one that yields the best results according to the chosen classifier performance metrics. [49]

- **Ensemble Classifiers:** Ensemble classifiers are a category of classifiers that consist of and utilize a combination of other, simpler classifiers in order to achieve better overall performance. [48] Some of them are the following:

  a) AdaBoost: This classifier relies on the assignment of weights to all training samples, which are initially equal. After the training of each (weak) classifier that it consists of, the error is calculated for each of the training samples. If an error occurs for a sample, its weight is increased, whereas if no error occurs, it is decreased. This forces the classifiers to focus on their weak points, resulting in their improvement. Thus, a stronger overall classifier eventually emerges. [50]

  b) Gradient Boosting: This classifier consists of many individual simple classifiers, which are trained sequentially. All the individual classifiers are correlated with the negative gradient of the ensemble classifier's loss function, so that after training each one, the loss value decreases. Thus, the overall classifier continuously improves. The loss function can be arbitrary, and its choice depends on the problem at hand. [51]

  c) Bagging (Bootstrap Aggregating): This classifier consists of many (usually un-

stable) classifiers, each trained on a different bootstrap sample of the training data. A bootstrap sample is a subset of the training data selected uniformly at random, while a sample selected for a bootstrap is not removed from the set and can be reselected for another bootstrap. Thus, each bootstrap contains a mixture of unique and repeated training samples. The final decision of the ensemble classifier is obtained through voting, that is, the input is assigned to the class chosen by the majority of the individual classifiers. In this way, the ensemble classifier is more accurate than the individual ones, as errors due to overfitting (i.e., excessive adaptation of the classifier to the training data) and statistical variance are reduced. [52]

d) Random Forest: This classifier consists of many **Decision Trees**, which are simple classifiers that operate by recursively splitting the data into subsets. Specifically, a tree is constructed starting from the root, and depending on the value of one of the data features (chosen based on a gain criterion), each node branches out to two children until all features are exhausted and we reach the leaves of the tree. A new input is classified by following the path determined by its feature values, from the root of the tree to a leaf. [53] A Random Forest trains the Trees it consists of on different subsets of both the training data and the features, which are selected from the pool of training data and features exactly as described for the Bagging classifier. In the end, the classification result is again determined by voting, meaning the decision of the majority of the Trees is selected. This procedure yields much better results than using a single Decision Tree. [54]

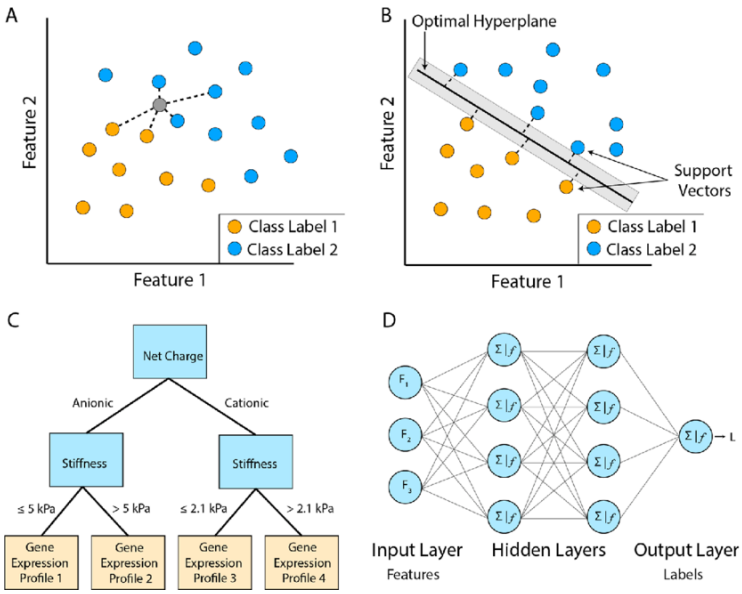


Figure 2.9. *A: A k-nearest neighbors classifier with k = 4 in a classification problem where the data have two features, B: A linear SVC in a classification problem where the data have two features, C: A Decision Tree in a classification problem where the data have two features, D: A Neural Network with one input layer, two hidden layers, and one output layer (not examined in the present section)*

[55]

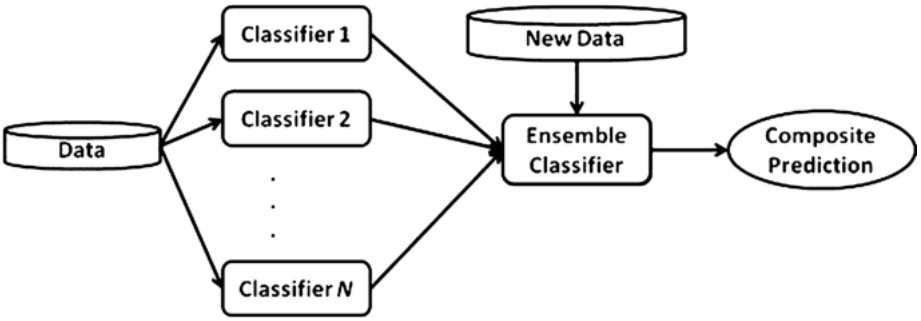


Figure 2.10. *A generalized description of the structure and operation of Ensemble Classifiers: An Ensemble Classifier consists of multiple individual classifiers, which are trained on the problem's dataset. When a new input is to be classified, the operation or the results of the individual classifiers are appropriately combined, so that the Ensemble Classifier reaches an overall decision.*

[56]

Chapter 3

Literature Overview

This chapter provides a concise overview of much of the existing research and literature on the classification of networks into evolution models, which is quite limited. Furthermore, it presents the article on the prevalence of the Scale-Free model in real-world networks, which inspired the research topic of the present thesis, along with several other articles related to this subject.

3.1 General references on the classification of networks into models

3.1.1 Graph classification using theoretical metrics

Linhong Zhu, Wee Keong Ng, and Shuguo Han investigated the classification of graphs based on some of their features and attempted to determine the optimal features for this task. To achieve their goal, they used the following 12 metrics as the features representing each graph: Clustering Coefficient, Average Path Length, Global Efficiency, Node Degree Distribution, Coreness, Density, Girth, Circumference, Connectivity, Vertex Compression Ratio, Edge Compression Ratio, and Acyclicity. Furthermore, they developed an algorithm for selecting the most representative and discriminative features, treating this selection as an optimization problem. [57]

The algorithm returned 8 of the above features, which were then used to represent a total of 300 graphs (both real-world and synthetic), assigning a feature vector to each graph. Each of these graphs was also assigned a label from three possible options (INTERVAL, HOPI, and DUAL), depending on which of these was evaluated as the optimal index-solution to the reachability index selection problem (that is, the problem of choosing the appropriate data structure for storing information about the feasibility of reaching any node from any other). This evaluation was based on computing the total access time of one million random queries on the graph for each possible choice. The majority of these 300 graphs were used as training data for two classifiers, namely a k-nearest neighbors classifier (k-NN) and a reverse k-nearest neighbors classifier (Rk-NN), while the remainder were used as test data for the aforementioned classifiers. [57]

Finally, the performance of the two classifiers was compared based on their accuracy and micro precision, as well as their execution time and memory consumption. It was

found that k-NN has shorter execution time and consumes less memory, but Rk-NN performs better in terms of accuracy and micro precision (though its results are less stable). The results demonstrated that graph classification based on their appropriate features is indeed feasible. [57]

The previous publication did not directly address graph classification into evolution models, but it demonstrated that graph classification through metrics is feasible, a technique also used in the present work. Regarding the selection of suitable metrics for our problem, we can refer to sources such as the article by Reka Albert and Albert-Laszlo Barabasi [58] or the book by Sergey Dorogovtsev and José Fernando Mendes [59], which describe in detail several of the evolution models we used and identify their most representative features. Additionally, the publication by Gábor Szárnyas, Zsolt Kővári, Ágnes Salánki, and Dániel Varró [60] examines a wide range of metrics regarding how characteristic they are for a graph, based on criteria of homogeneity (i.e., the extent to which networks belonging to the same category have similar values of the metric) and distinctiveness (i.e., the extent to which networks in different categories have different metric values).

3.1.2 Graph classification using Neural Networks

Another important tool for graph classification, which has been the subject of extensive research, is Neural Networks. These offer various advantages in cases where the method of selecting theoretical metrics to represent the graphs is not effective. For example, John Boaz Lee, Ryan Rossi, and Xiangnan Kong [61] attempt to address the issues that may arise with the aforementioned method due to the presence of noise in real-world networks by constructing a Recurrent Neural Network. This model draws information only from the neighborhoods of certain characteristic nodes in the network, which are selected via walks guided by an *attention mechanism*. This approach allows noise to be ignored and reduces computational costs. Furthermore, in the examined datasets, it was found to have greater accuracy than other methods that consider the entire graph, which in fact increases even further when external memory is incorporated.

Another type of Neural Network that is extremely useful for network classification is Graph Neural Networks (or GNNs), which are specifically designed for processing data in graph format. GNNs automatically and efficiently extract the representative features of a graph, a fact which has increased their popularity and consequently the number of existing models. Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli [62] have conducted a detailed evaluation and comparison of various popular GNN models for graph classification (with the GIN model by Xu et al. [63] appearing to have the highest overall accuracy among those studied), which can prove very useful for selecting the appropriate model for a classification problem. Finally, it should be noted that although these models are designed for graph classification across arbitrary categories, they can also be employed for classifying graphs into evolution models, making them directly relevant to the research topic of the present work.

3.1.3 Evolution model identification in real-world networks

Technological Networks

Bálint Hartmann and Viktória Sugár, motivated by the controversial view that power grids can be represented either by the Small-World model or the Scale-Free model, decided to study the Hungarian power grid using data spanning 70 years in order to determine whether it can indeed be represented by one of these two models and whether this changes over time. To achieve their goal, they constructed a graph for each year (a total of 70 graphs), with nodes corresponding to generators, transformers, and substations, and edges representing transmission lines. To assess whether the graphs could be classified as Small-World, they employed the small-world coefficient σ and the ω metric (as proposed by Telesford et al. [64]), while for Scale-Free classification they used the degree distribution. In addition, they calculated several other metrics to study the evolution of the power grid over time. The results revealed significant changes during the first 20 years and a stabilization of metric values over the subsequent 50 years, which they linked to historical events such as the construction of a new power plant and the addition of new transmission lines. They concluded that these events also influenced the network's evolution model, as they found that the grid could be categorized as Small-World during the last 50 years of their study, but not in the first 20. This shift was associated with the addition of different voltage levels to the network. Furthermore, they concluded that the Scale-Free model is not representative of this network at any point in time. [65]

The above study is not the only one that aims to detect an evolution model in power grids, as this knowledge is useful for solving important problems such as the construction of synthetic power grids to simulate real-world ones and robustness analysis. Among the numerous studies on European, American, and Asian power grids, we choose to also present, as an additional example, the study of Rafael Espejo, Sara Lumbreras, and Andres Ramos on 15 European electrical distribution networks. This research is thorough and investigates in depth the structure of these 15 networks, examining the voltage levels of each distribution network (220 kV and 400 kV) both as separate networks and combined. They employ six metrics to analyze the network topology, of which the one most directly related to the subject of the present thesis is the small-world coefficient σ, as it is used to evaluate whether a network can be considered Small-World. The results show that when the two voltage levels are examined together (that is, combined into a single graph), the distribution networks of all the countries can be considered Small-World. In contrast, when the voltage levels are examined separately (that is, represented by two different graphs), a minority of them cannot be considered Small-World, which is attributed to differences in the structure of each country's network. These findings appear to be in agreement with the conclusions of Hartmann and Sugár. [3]

Biological Networks

Another field in which efforts are being made to detect evolution models is that of biological networks, with the aim of better understanding them. Given that various types

of biological networks such as metabolic networks, protein–protein interaction networks, and gene interaction and expression networks have been observed to display Scale-Free behavior, Raya Khanin and Ernst Wit set out to investigate whether biological networks can indeed be considered Scale-Free. To this end, they focused on networks that had already been classified as Scale-Free by other researchers, who had relied on fitting a straight line to the degree distribution on a log–log plot. First, they estimated the exponent γ of the power law distribution for each of these networks using the statistical method of maximum likelihood. Next, they performed a goodness-of-fit test to determine whether the networks truly followed the power law distribution corresponding to the estimated γ. The results of this analysis showed that the networks under consideration cannot ultimately be characterized as Scale-Free, since their degree distributions do not follow a power law distribution. Khanin and Wit further note that their findings are consistent with those of Stumpf et al. [66], emphasizing the need for further research on the topology of biological networks. Indeed, the question of which evolution model biological networks follow remains open. [67]

Transportation Networks

Efforts to identify evolution models are also being carried out for transportation networks. Bin Jiang and Christophe Claramunt, motivated by the utility of network analysis in the field of Geographic Information Systems, examined a set of characteristics of certain transportation networks, including whether they exhibit Small-World or Scale-Free behavior. The networks studied were the street networks of Gävle (Sweden), Munich (Germany), and San Francisco (USA), while the criterion for the presence of Scale-Free behavior was the shape of the cumulative probability distribution of street connectivity on a log-log plot. The results showed that none of the street networks of these cities are characterized by the Scale-Free model. Regarding the assessment of Small-World behavior, this was carried out using the values of the networks' average path length and clustering coefficient, which indicated that these networks can indeed be classified as Small-World. Consequently, Jiang and Claramunt concluded that urban street networks are not Scale-Free but Small-World [68].

Another study addressing the identification of an evolution model in a transportation network is that of Sen et al., which focuses on the structural analysis of the Indian Railway Network. In this context, they examined whether the network can be classified as Small-World or Scale-Free, using criteria similar to those applied by Jiang and Claramunt. Specifically, the average path length and clustering coefficient were again used to detect Small-World behavior, while the shape of the degree distribution and Clustering Coefficient were used for Scale-Free behavior. Based on the results, Sen et al. concluded that the Indian Railway Network belongs to the category of Small-World networks, but not to that of Scale-Free networks, and they argued that these findings can be generalized to the railway network of any country [69].

3.2 The prevalence of the Scale-Free model in real-world networks

3.2.1 Scale-free networks are rare

Scale-Free networks are a popular subject of study in network science, as they are considered to represent a large portion of real-world networks. Although there is a plethora of research supporting this claim, there are also many studies that refute it, making this issue a controversial one. Furthermore, the criteria used to classify a network as Scale-Free are not consistent across studies, but instead present several variations. These facts led Anna D. Broido and Aaron Clauset to conduct their own research on the prevalence of the Scale-Free model in real-world networks, titled *Scale-free networks are rare*. [70]

Broido and Clauset worked on 928 networks, the data for which they obtained from the Index of Complex Networks (ICON). Of these, 53% were biological networks, 22% technological networks, 16% social networks, 7% transportation networks, and 2% information networks, thus encompassing enough categories to allow for a more general conclusion. These networks also varied in size, with the number of nodes ranging from a few dozen to over a million, and differed in a range of properties such as directionality, weights, temporal variability, and whether they were multiplex, bipartite, or multigraphs. Since the presence of many of these properties leads to multiple degree distributions for the same graph (for example, for a directed graph there is both the in-degree distribution and the out-degree distribution), they decided to decompose each non-simple graph into a set of simple graphs that would represent it. For instance, a multiplex graph with n layers is broken down into n simpler graphs corresponding to its layers, and then each of these is further divided into even simpler graphs according to its other properties, and so on, until all resulting graphs are simple (that is, lacking any of the aforementioned properties).

The main criterion used to decide which networks can be considered Scale-Free was the extent to which their degree distribution resembles the power law distribution $P(k) = Ck^{-a}$, $k \geq k_{min} \geq 1$, where k corresponds to the degree, C is the normalization constant, a the distribution parameter, and k_{min} the minimum degree for which the degrees are required to follow the power law distribution. The fact that only the upper tail of the distribution is examined, from k_{min} onwards, makes the choice of k_{min} highly important. This choice was made using the standard Kolmogorov-Smirnov minimization method, while the equally important choice of the parameter a was made via discrete maximum likelihood estimation. The evaluation of how well the power law distribution resulting from the estimated values of $\hat{k_{min}}$ and $\hat{a}$ fits the actual degree distribution of the network was carried out through a standard goodness-of-fit test, which returned a p-value as a measure of similarity between the two distributions. In addition, the power law distribution corresponding to the parameter pair $(\hat{k_{min}}, \hat{a})$ was compared against four other distributions (exponential, log-normal, power law with exponential cutoff, and Weibull) in the region $k \geq k_{min}$ using Vuong's normalized likelihood ratio test, in order to determine which one best approximates the actual degree distribution. It should be noted here that

the application of these methods, and therefore the classification, was performed on the simplified graphs rather than on the original ones.

Another key part of the work of Broido and Clauset was the construction of five categories of Scale-Free networks, thus summarizing the different criteria found in the literature while at the same time creating a scale for the strength of the Scale-Free property in a network. These categories are as follows:

- **Super-Weak:** For at least 50% of the simple graphs, no other distribution is preferred over the power law
- **Weakest):** For at least 50% of the simple graphs, the power law distribution cannot be rejected ($p \geq 0.1$)
- **Weak):** The requirements of Weakest are satisfied, and the the power law region contains at least 50 nodes
- **Strong):** The requirements of Weakest and Weak are satisfied, and at the same time $2 < a < 3$ for at least 50% of the simple graphs
- **Strongest):** The requirements of Strong are satisfied for at least 90% of the simple graphs, and the requirements of Super-Weak are satisfied for at least 95% of the simple graphs

Classifying a network in the first category (Super Weak) provides an indirect indication of the presence of a Scale-Free structure, whereas classification in any of the subsequent categories provides a direct indication. It should also be noted that the last four categories are nested, meaning that if a network belongs to one of them, it will also belong to all the ones above it (except, of course, Super Weak). Finally, to represent the not Scale-Free networks, an additional category is defined:

- **Not Scale Free):** Networks that are neither Super-Weak nor Weakest.

The results of classifying the networks in the dataset into the above categories, following the method described, are summarized in figure 3.1.

As shown in figure 3.1, 49% of the networks cannot be considered Scale-Free, 46% belong to the Super-Weak category, 29% to Weakest, 19% to Weak, 10% to Strong, and 4% to Strongest. Broido and Clauset also analyzed these results by network category. For **biological networks**, they found that 63% of them are Not Scale Free, 33% are Super-Weak, 19% are Weakest, and 6% are Strongest. They observed that most of the Not Scale Free biological networks were fungal networks. In fact, all the fungal networks they examined were Not Scale Free, which may have been expected given their spatially dependent structure. For this reason, they decided to investigate what the percentages for each category would be if fungal networks were excluded from the dataset. They found that in this case, while the percentages of the Super-Weak, Weakest, and Weak categories would increase significantly, the increases in the Strong and Strongest categories would be negligible, meaning that the inclusion of fungal networks ultimately does not affect

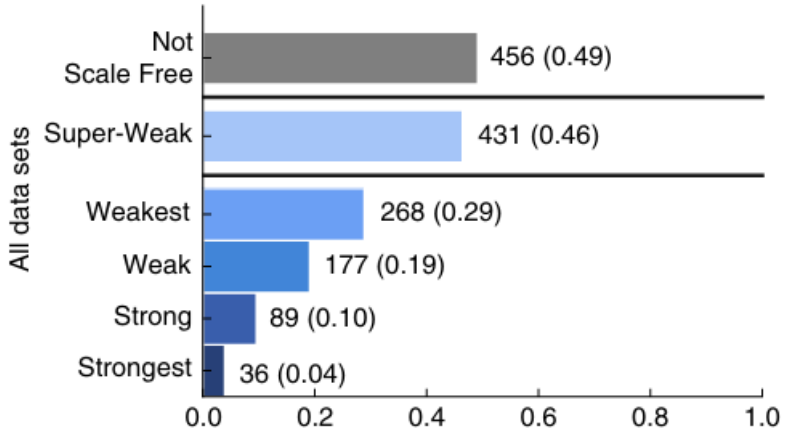


Figure 3.1. *Proportion of networks by Scale-Free evidence category. Bars separate the Super-Weak category from the nested definitions, and from the Not Scale Free category, defined as networks that are neither Weakest or Super-Weak*
[70]

the qualitative conclusions of the study. As for **social networks**, the results differed considerably from the overall ones, as 50% were classified as Not Scale Free, 41% as Super-Weak, 48% as Weakest, and 31% as Weak, but none as Strong or Strongest. The results also differed from the overall ones in the case of **technological networks**, as only 8% of them were classified as Not Scale Free, while 90% were classified as Super-Weak, 42% as Weakest, 37% as Weak, 28% as Strong, and only 1% as Strongest. Finally, **transportation networks** were not numerous enough for a similar percentage analysis, but it was observed that all of them belonged to the Not Scale Free, Super-Weak, or Weakest categories, while none belonged to Strong or Strongest.

Moreover, to ensure that these results were not influenced by the evaluation method itself, four robustness checks were carried out. Specifically, it was examined whether the outcomes were affected under the following conditions: (a) all networks considered are by nature simple (i.e., unweighted, undirected, single-layer, and without multiedges), (b) the power law distribution with exponential cutoff is removed from the alternative distributions, (c) the percentage thresholds of the categories are lowered so that classification into them is allowed if even a single simple graph meets the criteria, and (d) the scaling behavior of the first and second moments of the degree distribution is assessed. The results of these checks showed that the methods employed were indeed not biased.

Ultimately, Broido and Clauset concluded that Scale-Free networks are not ubiquitous but rather much rarer than suggested by the existing literature, and that their use as the initial stage of modeling and structural analysis of real-world networks is unfounded. They also emphasized that the likelihood of Scale-Free structure varies depending on the category of networks. For instance, it is more likely to occur in certain types of biological and technological networks, and much less likely in social networks. Transportation and information networks were too few in number to draw firm conclusions about them, but there was no strong evidence that their structure differs significantly from that of the others.

3.2.2 Are Scale-Free networks truly rare? A controversial issue

The article by Broido and Clauset [70] sparked quite a controversy, and even elicited a response from Barabasi himself [71], the creator of the Scale-Free model, who objected to the methodology and conclusions of the aforementioned work. This article, combined with the extensive discussion within the scientific community regarding the prevalence of Scale-Free networks, inspired Voitalov et al. [72] to conduct their own research on this topic, also providing a rigorous definition of Scale-Free networks in response to the ambiguity present in the literature, and performing a statistical analysis on a set of real-world networks. The results of this study confirmed that many of the networks previously found to be Scale-Free are indeed Scale-Free (such as the World Wide Web, human protein–protein interaction networks, social group membership networks, and others), thereby painting a different picture of the prevalence of Scale-Free networks than that suggested by Broido and Clauset. The same developments also motivated Liu et al. [73] to investigate the prevalence of the Scale-Free model, this time in trade networks. Their results showed that the structure of trade networks displays great diversity due to the different nature of the products involved in each case, but the appearance of Scale-Free structure is not rare. They further noted that mishandling edge weights, when these represent capacities, may lead to overlooking the potentially Scale-Free structure of a network.

Another noteworthy study on the prevalence of the Scale-Free model in real-world networks is that of Artico et al. [74], who argued that Broido and Clauset [70] and Khanin and Wit [67] imposed overly strict criteria on the degree distribution of a network in order for it to be considered Scale-Free. Inspired by the approach of Voitalov et al. [72], as well as by other researchers, they proposed their own, less restrictive criteria. In addition, Artico et al. modified the standard Kolmogorov–Smirnov and goodness-of-fit tests employed by Broido and Clauset, and subsequently applied these methods to the same dataset in order to re-evaluate their conclusions. The results of Artico et al. indicated that 64% of the examined networks were Scale-Free, therefore they concluded that Scale-Free networks are, in fact, not rare, thereby challenging the claim of Broido and Clauset.

The debate regarding the prevalence of Scale-Free structure in real-world networks did not, of course, emerge only with the publication of Broido and Clauset's article [70], but had already existed almost since the very introduction of the Scale-Free model in 1999. An example of earlier research on this subject is the article of Gipsi Lima-Mendez and Jacques van Helden [75], which examined the validity of popular beliefs of that time (2000–2010) concerning the prevalence of the power law distribution and the Scale-Free model in biological networks. Following statistical analysis, Lima-Mendez and van Helden concluded that the following five beliefs are myths: (a) the degree distribution of biological networks follows a power law, (b) biological networks are Scale-Free, (c) metabolic networks are Small-World, (d) Small-World networks are vulnerable to targeted attacks but robust to random deletions, (e) biological networks grow through preferential attachment. Another similar study from that period is that of Khanin and Wit [67], already

discussed in section 3.1.3, which likewise concluded that biological networks cannot in general be characterized as Scale-Free.

Around the same time, the study of Mislove et al. [76] was published, in which four online social networks (Flickr, LiveJournal, Orkut, and YouTube) were examined with respect to various properties, including Scale-Free behavior. Based on log–log plots of the degree distributions of these networks, Mislove et al. concluded that three out of the four (specifically Flickr, LiveJournal, and Orkut) exhibit strong Scale-Free behavior. They also investigated the extent to which the degree distributions of these networks resemble a power law distribution by performing a statistical analysis following the same methodology as in previously mentioned studies, and found that all four networks display power law degree distributions.

The previously mentioned studies are, of course, not the only ones conducted within the framework of the long-standing and lively discussion of the scientific community regarding the prevalence of the Scale-Free model. Beyond the contents of the present thesis, a summary of this discussion can be found in the article by Petter Holme [77], along with his own perspective on the matter.

Part II

Empirical Study

Chapter 4

Methodology

4.1 Overview

In the present chapter, we introduce the set of methods employed in this thesis for the identification of the evolution models characterizing a collection of real-world and synthetic networks, as well as for the partitioning of networks into subnetworks characterized by different evolution models.

The core approach to classifying networks into evolution models involved training a classifier, with graph feature values as its input and the corresponding evolution model as its output. The first step in this process was to select the evolution models to be considered as possible classifier output. This was followed by the selection of the metrics to be used as inputs, which was primarily based on the extent to which their values varied across the various models. A training dataset was then constructed by generating multiple synthetic networks for each model and computing the selected metrics for all of them. The next step was to test the performance of various classifiers on this dataset and select the one that achieved the highest performance. This classifier was ultimately employed for the classification of a number of real-world networks.

Regarding the identification of network segments characterized by different evolution models, a number of partitioning methods were tested in order to determine the most accurate one. More precisely, each method was applied to synthetic networks created by merging two or more networks of different types, after which the selected classifier was used to identify the evolution model characterizing each subnetwork. The method yielding the most accurate results (i.e., those most closely approximating the actual structure of the corresponding networks) was then employed to identify subnetworks within the real-world networks that were classified earlier, and the resulting subnetworks were subsequently classified into evolution models using the selected classifier.

4.2 Evolution Model Selection

The selection of the evolution models that we were aiming to identify in real-world and synthetic networks constitutes a fundamental part of this work. An initial survey of the literature identified the following models: Regular Graph, Erdos–Renyi, Gilbert, Random Geometric Graph, Watts–Strogatz, Newman–Watts, Barabasi–Albert, Waxman,

and Hierarchic Graph. Among these, the Erdos–Renyi and Gilbert models belong to the category of Random Graphs, the Watts–Strogatz and Newman–Watts models fall into the Small-World category, the Waxman model belongs to the category of Spatial Networks, and finally, the Random Geometric Graph is considered both a Random Graph and a Spatial Network.

Subsequently, some of these models were excluded based on two criteria: their frequency of occurrence in real-world networks, and their similarity to other models. Specifically, Regular Graphs exhibit a very particular structure (all their nodes have the same degree), which is found in very few types of real-world networks, for example those representing crystal structures [78]. Likewise, Hierarchic Graphs take the form of Trees, which also occur only in specific types of networks that describe hierarchical relationships (for example, kinship networks) [15]. With regard to model similarity, we decided to allow the inclusion of two or more models belonging to the same category, with the exception of Watts–Strogatz and Newman–Watts, since the latter is an (improved) variation of the former [79]. Between the two, we chose to retain the Watts–Strogatz model, as it appears much more frequently in the literature and in other studies concerning the classification of networks into evolution models.

Thus, the models ultimately employed were **Erdos–Renyi**, **Gilbert**, **Random Geometric Graph**, **Watts–Strogatz**, **Barabasi–Albert**, and **Waxman**, while Regular Graphs, Hierarchical Graphs, and the Newman–Watts model were excluded.

4.3 Network Metrics Selection

The selection of network metrics to be used as input for the classifier also began with a thorough review of the literature. This process identified 20 metrics that can provide significant insights into the structure and behavior of a network, and thus aid in its classification into a specific evolution model. These metrics were as follows: Degree Distribution, Average Path Length, Diameter, Radius, Clustering Coefficient, Betweenness Centrality, Closeness Centrality, Information Centrality, Routing Centrality, Bridging Centrality, Spectral Centrality, Eigenvector Centrality, Katz Centrality, Laplacian Centrality, Harmonic Centrality, Percolation Centrality, Load Centrality, Rich Club Coefficient, Closeness Vitality, and Assortativity.

Subsequently, 6 synthetic networks were generated, each constructed so as to be characterized by one of the 6 evolution models selected earlier. It was further decided that all these networks should be built with the same number of nodes, in order to ensure that the metric values would not be affected by network size when computed. Moreover, an effort was made to assign identical values to the remaining parameters, although this was not always feasible since not all models share the same parameters (for example, some use the probability of edge formation as a parameter, while others use the number of new connections that originate from each node).

The 20 aforementioned metrics were computed for all 6 synthetic graphs, in order to assess both the extent to which each metric exhibits variation across different models, and whether it can be feasibly computed for a large number of networks within a reasonable

time frame. More specifically, for each metric we computed either its single value per graph (as, for example, in the case of Assortativity), or its distribution, mean, and variance when it is defined over all nodes of a graph (for example, centrality measures). In addition, for each metric we calculated the variance of its mean (or single) value, as well as the coefficient of variation of this value, as a more objective measure of variability across different network models (since the variance depends on the magnitude of the mean value, whereas the coefficient of variation does not). Finally, for each metric we recorded its execution time per network, its total execution time across all 6 synthetic graphs, as well as the ratio of its coefficient of variation to its total execution time, as a performance measure.

Following this, an overall review of the results was conducted, taking into account both the computed values and the qualitative differences in the shapes of the distributions, wherever these existed. More precisely, for each metric, differences in the distributions and in the mean or single values across networks were evaluated positively, as well as comparatively high values of the variance of the mean or single value, the coefficient of variation, and the ratio of the coefficient of variation to execution time, while high execution times were evaluated negatively. Thus, after discarding certain metrics due to excessively high execution times or other computational issues, the following 7 were selected, which were deemed to exhibit sufficiently large differences across models to achieve good classifier performance: **Degree Distribution**, **Average Path Length**, **Diameter**, **Closeness Centrality**, **Information Centrality**, **Harmonic Centrality**, and **Assortativity**.

4.4 Dataset Construction

The dataset for training and testing the classifier was constructed by generating 100 graphs for each evolution model using the corresponding functions of the NetworkX library in the Python programming language, which was the language used for the implementation of the experimental part of this work. Additionally, for each of these 600 graphs the values of the metrics selected as classifier inputs were computed, so that each graph was represented in the dataset by its model (serving as its label) and the values of the aforementioned metrics computed on it. 80% of the dataset graphs (i.e., 480 graphs) was selected at random to be used as the training set for the classifier, while the remaining 20% (120 graphs) was allocated to the test set.

4.5 Classifier Selection

The selection of the appropriate classifier was guided by the flowchart shown in Figure 4.1, which is taken from the documentation of the scikit-learn library in Python [80]. Specifically, starting from the beginning (START) of the diagram, the recommended path was followed by answering "yes" the questions of whether there are more than 50 samples, whether a class is to be predicted, whether the data are labeled, and whether the number of samples is fewer than 100,000. This path points to the Linear SVC classifier, which

therefore became one of the candidate classifiers for the present classification problem. To determine the optimal classifier for this problem, the path corresponding to the poor performance of Linear SVC was followed, then the absence of text data, and finally the poor performance of the *k*-Nearest Neighbors classifier (which was, of course, also added to the list of candidates). This path ultimately leads to the SVC classifier and the Ensemble Classifiers category, from which AdaBoost, Gradient Boosting, Random Forest, and Bagging were selected as candidates.

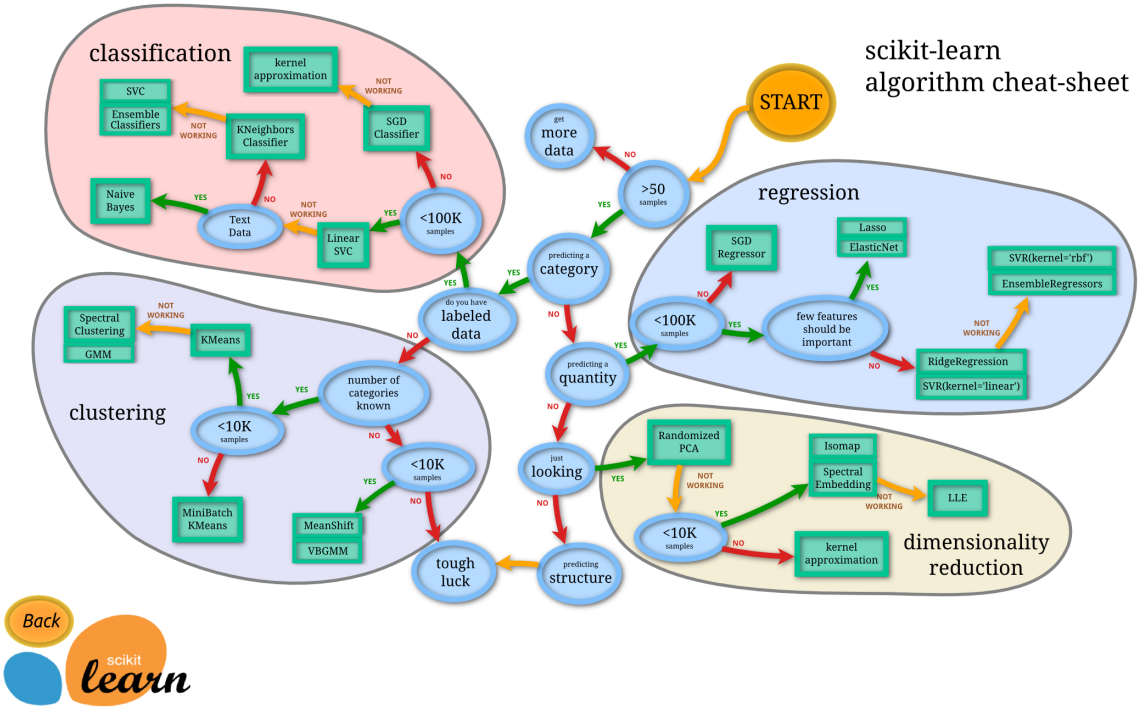


Figure 4.1. *Flowchart on the selection of the appropriate machine learning algorithm depending on the type of problem and its parameters*
[80]

Thus, the set of candidate classifiers consisted of Linear SVC, *k*-Nearest Neighbors, SVC, AdaBoost, Gradient Boosting, Random Forest, and Bagging. These classifiers were then evaluated on the dataset of the problem using the metrics precision, recall, accuracy, and F1 score. The classifier with the highest values across these metrics (in our case, **Random Forest**) was selected, and its Confusion Matrix was computed in order to gain further insight into its weaknesses. In addition, an Ablation Study was conducted, that is, the classifier's performance was reevaluated after removing each input metric individually, in order to examine whether any of the employed metrics actually reduced its performance.

4.6 Real-world Network Classification

The real-world networks classified by their evolution model in the present work were primarily selected from the Index of Complex Networks (ICON, https://icon.colorado.edu/networks), as well as from the Stanford Large Network Dataset Collection (SNAP, https://snap.stanford.edu/data/), the KONECT project archives of the University of Koblenz–Landau

(http://konect.cc/), and the Kaggle platform (https://www.kaggle.com/). These networks exhibit considerable variety in their structure, with the number of nodes and edges ranging from hundreds to tens of thousands, and with their properties and categories displaying significant diversity.

The selected networks were as follows:

- **Football:** Social network with 115 nodes and 613 edges, sourced from KONECT. Undirected and unweighted. A network of American football games between colleges during regular season Fall 2000.
- **GOT:** Social network with 114 nodes and 396 edges, sourced from Kaggle. Undirected and unweighted. A network of character interactions from the 5th season of the television series Game of Thrones.
- **Email-EU:** Social network with 1005 nodes and 25571 edges, sourced from SNAP. Directed and unweighted. An email network.
- **AdvogatoTrust:** Social network with 6541 nodes and 51127 edges, sourced from ICON. Directed and weighted. A network of trust relationships among users on Advogato.
- **MorenoCrime:** Social network with 1380 nodes and 1476 edges, sourced from KONECT. Undirected, unweighted, and bipartite. A network that records the involvement of certain individuals (left part of the graph) with crimes (right part of the graph).
- **NetScience:** Social network with 1589 nodes and 2742 edges, sourced from ICON. Undirected and weighted. A coauthorship network of scientists working in network science.
- **Euroroad:** Transportation network with 1174 nodes and 1417 edges, sourced from ICON. Undirected and unweighted. A network of international "E-roads", mostly in Europe.
- **ChicagoRegional:** Transportation network with 12982 nodes and 39018 edges, sourced from ICON. Directed and unweighted. A network representing traffic flow between intersections in the Chicago, Illinois, area.
- **Flights-FAA:** Transportation network with 1226 nodes and 2615 edges, sourced from ICON. Directed and unweighted. A network of air traffic routes, from the preferred routes database of the Federal Aviation Administration (FAA) National Flight Data Center (NFDC) of the USA.
- **Bristol:** Transportation network with 3731 nodes and 3101 edges, sourced from ICON. Undirected and unweighted. A network representing bus lines in Bristol, England.

- **Coach:** Transportation network with 3915 nodes and 3228 edges, sourced from ICON. Undirected and unweighted. A network representing coach lines in the UK.
- **YuYeast:** Biological network with 2018 nodes and 2930 edges, sourced from ICON. Undirected and unweighted. A network of protein-protein interactions in the fungus S. cerevisiae.
- **BacillusMegaterium:** Biological network with 4179 nodes and 5644 edges, sourced from ICON. Undirected, unweighted, and bipartite. A metabolic network of the bacterium Bacillus megaterium.
- **Malaria:** Biological network with 307 nodes and 1446 edges, sourced from ICON. Undirected, unweighted, and it consists one of the layers of a bigger, multiplex graph. A networks of recombinant antigen genes from the human malaria parasite P. falciparum.
- **Fullerene:** Chemical network with 3840 nodes and 5760 edges, sourced from ICON. Undirected and unweighted. A network of carbon atoms and the atomic bonds that connect them within fullerene molecules.
- **PowerNet:** Technological network with 4941 nodes and 6594 edges, sourced from ICON. Undirected and unweighted. A network that represents the Western States Power Grid of the US. Note: Used by Watts and Strogatz in the paper where they introduced the concept of Small-World networks [21].
- **R-Dependancies:** Technological network with 2471 nodes and 5451 edges, sourced from ICON. Directed and unweighted. A network of package dependencies in the R programming language.

The above networks were preprocessed before being input into the classifier. Initially, any directed graphs were converted to undirected ones by ignoring the direction of their edges. In addition, the largest connected component of each network was extracted, on which the classification and partitioning were subsequently performed. These modifications were made to ensure that the selected metrics could be computed for the graphs, as some of them are not defined for disconnected or directed graphs.

4.7 Network Partitioning Method Selection

In order to identify the optimal method for partitioning networks into subnetworks characterized by different evolution models, five composite synthetic networks were initially constructed by merging other synthetic networks, each of which was characterized by a single evolution model. Specifically, the network models that were combined were as follows: Barabasi-Albert with Random Geometric Graph, Barabasi-Albert with Watts-Strogatz, Watts-Strogatz with Erdos-Renyi, Waxman with Random Geometric Graph, and Gilbert with Waxman. These merges were performed by adding edges between the component graphs, with the nodes to which these edges were attached being chosen at random.

Afterward, various partitioning methods were applied to these five graphs, any resulting subgraphs that were too small to be classified (i.e., those with fewer than 20 nodes) were discarded, and the remaining ones were classified into evolution models using the selected classifier.

For each method, we additionally computed the number of resulting subgraphs, their sizes, the number of correctly classified subgraphs, as well as the total number of correctly classified nodes in the graph, the modularity of the partition, and the execution time. (Note: A node was considered correctly classified if the subgraph to which it belonged was classified into the same model as the graph from which it originated, while a subgraph was considered correctly classified if at least 70% of its nodes were correctly classified.) The most important criterion was that the subgraphs belonged to the evolution models of the components of the graph from which they originated, and only to these. At the same time, large numbers of correctly classified subgraphs and nodes, as well as high modularity values, were evaluated positively, whereas long execution times were evaluated negatively.

4.7.1 Community Detection

All the methods used belonged to the category of community detection algorithms. The selection of the algorithms tested was based on their diversity in terms of how they operate (so that several different approaches could be examined), as well as the availability of an implementation. The algorithms that were ultimately used were Girvan–Newman, Walktrap, Infomap, Leiden, and Spectral Clustering, with their implementations found in the cdlib, networkx, and sklearn Python libraries. After their evaluation, Leiden was selected.

Note: Due to the excessively long execution time of the Girvan–Newman algorithm on the constructed networks (exceeding 1 hour in some cases), it was necessary to generate additional smaller networks with the same models in order to evaluate its performance. This fact automatically ruled out the use of Girvan–Newman for the partitioning of the real-world networks, since most of them have a size comparable to that of the initial synthetic networks. Nevertheless, its performance was still tested, as it could prove useful in applications involving smaller networks, more available time, or greater computational resources.

4.8 Real-world Network Partitioning

After the optimal network partitioning method was selected, it was applied to the real-world networks described in section 4.6. In this way, the segments characterized by different evolution models (along with their sizes) were identified, as well as which models these were.

Chapter 5

Results

This chapter presents the results of the methods and experiments described in the previous chapter, along with some observations on them.

5.1 Network Metrics Selection

The results for each metric are presented below separately (rounded to two decimal places unless otherwise specified), along with an overall comparison between them. The distributions are not included here for the sake of brevity, but can be found in the corresponding file.

5.1.1 Degree Distribution

Table 5.1. *Results for the Degree Distribution*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	10	9.64	0.51
Gilbert	499.47	246.78	0.28
Random Geometric	480.22	18438.10	0.33
Small-World	4	1.47	0.57
Scale-Free	9.95	98.23	0.31
Waxman	40.23	42.55	0.38

5.1.2 Average Path Length

Table 5.2. *Results for the Average Path Length*

	Value	Execution Time (s)
Erdos-Renyi	3.25	2.95
Gilbert	1.50	2.21
Random Geometric	1.55	13.80
Small-World	5.58	0.55
Scale-Free	2.99	0.55
Waxman	2.15	0.55

5.1.3 Diameter

Table 5.3. *Results for the Diameter*

	Value	Execution Time (s)
Erdos-Renyi	5	0.6
Gilbert	2	0.43
Random Geometric	3	6.43
Small-World	10	0.92
Scale-Free	5	0.96
Waxman	3	0.93

5.1.4 Radius

Table 5.4. *Results for the Radius*

	Value	Execution Time (s)
Erdos-Renyi	4	0.58
Gilbert	2	0.42
Random Geometric	2	6.47
Small-World	7	0.53
Scale-Free	3	0.55
Waxman	3	0.52

5.1.5 Clustering Coefficient

Table 5.5. *Results for the Clustering Coefficient*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.01	0	0.22
Gilbert	0.50	0	64.08
Random Geometric	0.77	0.01	59.61
Small-World	0.07	0.02	0.18
Scale-Free	0.04	0	0.28
Waxman	0.04	0	0.66

5.1.6 Betweenness Centrality

Table 5.6. *Results for the Betweenness Centrality*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.0023	0	4.24
Gilbert	0.0005	0	93.76
Random Geometric	0.0005	0	69.68
Small-World	0.0046	0	3.60
Scale-Free	0.0020	0	5.03
Waxman	0.0012	0	9.21

The mean value was rounded to 4 decimal places, due to its very small magnitude. The variance ranged from the order of 10^{-9} to the order of 10^{-5} for all graphs, and was therefore rounded to 0 in all cases.

5.1.7 Closeness Centrality

Table 5.7. *Results for the Closeness Centrality*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.31	0.0002	1.26
Gilbert	0.67	0	0.51
Random Geometric	0.65	0.0052	6.60
Small-World	0.18	0.0001	0.71
Scale-Free	0.36	0.0007	0.71
Waxman	0.47	0.0002	0.67

The variance was rounded to 4 decimal places, due to its very small magnitude.

5.1.8 Information Centrality

Table 5.8. *Results for the Information Centrality*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.0041	0	2.71
Gilbert	0.2498	0	4.75
Random Geometric	0.2249	0.0010	4.46
Small-World	0.0012	0	1.08
Scale-Free	0.0034	0	2.80
Waxman	0.0193	0	2.64

The mean value and the variance were rounded to 4 decimal places, due to their very small magnitude.

5.1.9 Routing Centrality

Table 5.9. *Results for the Routing Centrality*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.0053	0	27.10
Gilbert	0.0018	0	1108.39
Random Geometric	0.0022	0	880.34
Small-World	0.0095	0	6.06
Scale-Free	0.0049	0	22.61
Waxman	0.0034	0	90.05

The mean value was rounded to 4 decimal places, due to its very small magnitude. The variance ranged from the order of 10^{-9} to the order of 10^{-5} for all graphs, and was therefore rounded to 0 in all cases.

5.1.10 Bridging Centrality

Table 5.10. *Results for the Bridging Centrality*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.000207	0	4.1
Gilbert	0.000001	0	92.83
Random Geometric	0.000001	0	70.04
Small-World	0.001095	0	3.25
Scale-Free	0.000123	0	4.07
Waxman	0.000028	0	9.19

The mean value was rounded to 6 decimal places, due to its very small magnitude. The variance ranged from the order of 10^{-17} to the order of 10^{-7} for all graphs, and was therefore rounded to 0 in all cases.

5.1.11 Spectral Centrality

The calculation of this metric was not possible, as execution issues arose and no alternative implementation was found. As a result, spectral centrality was automatically excluded.

5.1.12 Eigenvector Centrality

Table 5.11. *Results for the Eigenvector Centrality*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.0297	0.0001	0.22
Gilbert	0.0316	0	0.49
Random Geometric	0.0301	0.0001	0.94
Small-World	0.0279	0.0002	0.27
Scale-Free	0.0219	0.0005	0.20
Waxman	0.0312	0	0.23

The mean value and the variance were rounded to 4 decimal places, due to their very small magnitude.

5.1.13 Katz Centrality

Table 5.12. *Results for the Katz Centrality*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.0297	0.0001	0.17
Gilbert	0.0316	0	0.48
Random Geometric	0.0301	0.0001	0.90
Small-World	0.0279	0.0002	0.39
Scale-Free	0.0219	0.0005	0.27
Waxman	0.0312	0	0.29

The mean value and the variance were rounded to 4 decimal places, due to their very small magnitude.

5.1.14 Laplacian Centrality

The calculation of this metric took over 60 minutes, so execution was halted and it was automatically excluded.

5.1.15 Harmonic Centrality

Table 5.13. *Results for the Harmonic Centrality*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	324.49	276.91	1.62
Gilbert	749.24	61.70	0.83
Random Geometric	734.74	5401.68	7.32
Small-World	192.49	141.64	0.97
Scale-Free	351.10	928.90	0.91
Waxman	487.90	141.17	1.35

5.1.16 Percolation Centrality

Table 5.14. *Results for the Percolation Centrality*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.0023	0	4.97
Gilbert	0.0005	0	92.28
Random Geometric	0.0005	0	70.73
Small-World	0.0046	0	3.47
Scale-Free	0.0020	0	4.88
Waxman	0.0012	0	10.23

The mean value was rounded to 4 decimal places, due to its very small magnitude. The variance ranged from the order of 10^{-9} to the order of 10^{-5} for all graphs, and was therefore rounded to 0 in all cases.

5.1.17 Load Centrality

Table 5.15. *Results for the Load Centrality*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.0023	0	4.79
Gilbert	0.0005	0	72.20
Random Geometric	0.0005	0	60.41
Small-World	0.0046	0	2.73
Scale-Free	0.0020	0	4.39
Waxman	0.0012	0	7.97

The mean value was rounded to 4 decimal places, due to its very small magnitude. The variance ranged from the order of 10^{-9} to the order of 10^{-5} for all graphs, and was therefore rounded to 0 in all cases.

5.1.18 Rich Club Coefficient

Table 5.16. *Results for the Rich Club Coefficient*

	Mean	Variance	Execution Time (s)
Erdos-Renyi	0.0023	0	5.41
Gilbert	0.0005	0	92.64
Random Geometric	0.0005	0	69.79
Small-World	0.0046	0	4.80
Scale-Free	0.0020	0	4.20
Waxman	0.0012	0	9.68

The mean value was rounded to 4 decimal places, due to its very small magnitude. The variance ranged from the order of 10^{-9} to the order of 10^{-5} for all graphs, and was therefore rounded to 0 in all cases.

5.1.19 Closeness Vitality

The calculation of this metric took over 60 minutes, so execution was halted and it was automatically excluded.

5.1.20 Assortativity

Table 5.17. *Results for the Assortativity*

	Value	Execution Time (s)
Erdos-Renyi	0.0115	0.53
Gilbert	-0.0017	1.50
Random Geometric	0.1094	1.56
Small-World	-0.0906	0.02
Scale-Free	-0.0496	0.03
Waxman	0.0141	0.08

The value was rounded to 4 decimal places, due to its very small magnitude.

5.1.21 Overall Comparison

In the table of Figure 5.1, the following values are presented for each metric that was not already excluded, combining the results across all networks: the variance of its value or mean (Variance), its total execution time (Total Time), the coefficient of variation of its value or mean (Coeff. of Var.), and the ratio of the coefficient of variation to the total execution time (Coeff. of Var./Time).

	Variance	Total Time	Coeff. of Var.	Coeff. of Var./Time
Degree Distribution	50050.4995	2.3932	1.2859	0.5373
Average Path Length	1.938	20.5993	0.4909	0.0238
Diameter	6.8889	10.2608	0.5624	0.0548
Radius	2.9167	9.0784	0.488	0.0537
Clustering Coefficient	0.0853	125.0222	1.2314	0.0098
Betweenness Centrality	0.0	185.5171	0.7578	0.0041
Closeness Centrality	0.0323	10.4559	0.4131	0.0395
Information Centrality	0.0119	18.4427	1.3009	0.0705
Routing Centrality	0.0	2134.5533	0.5695	0.0003
Bridging Centrality	0.0	183.4815	1.6022	0.0087
Eigenvector Centrality	0.0	2.334	0.114	0.0488
Katz Centrality	0.0	2.4962	0.114	0.0457
Harmonic Centrality	43439.5006	12.9946	0.4403	0.0339
Percolation Centrality	0.0	186.5606	0.7578	0.0041
Load Centrality	0.0	152.4865	0.7578	0.005
Rich Club Coefficient	0.0	186.5178	0.7578	0.0041
Assortativity	0.0038	3.7265	55.0923	14.7838

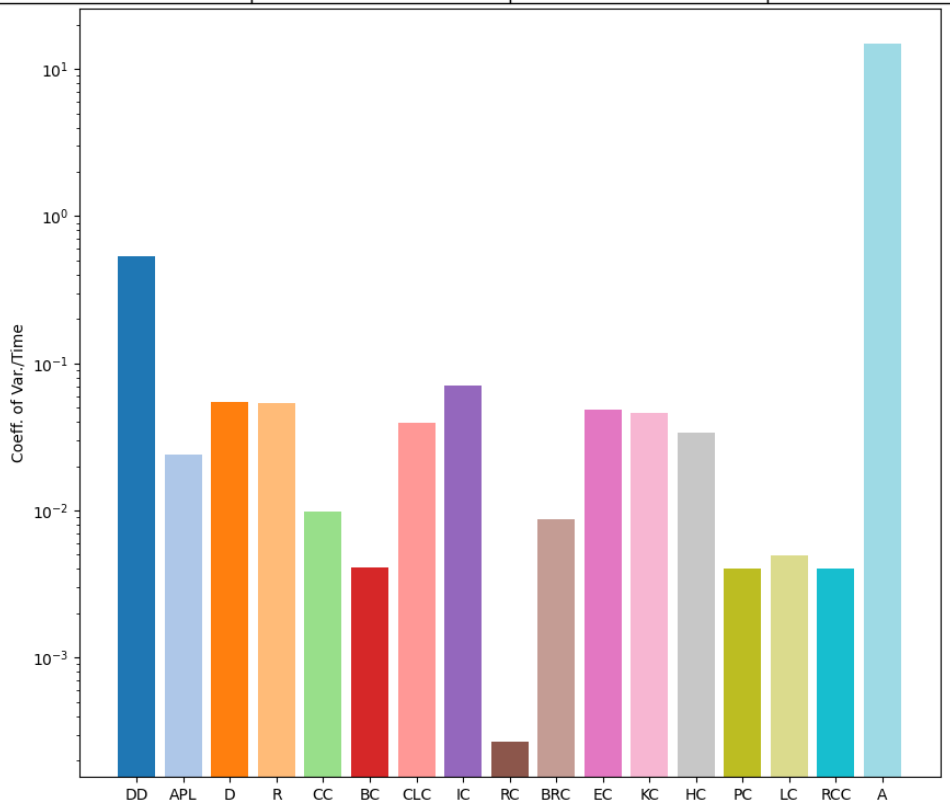


Figure 5.1. ***Top:*** *Table showing, for each metric, the variance of its value or mean, the total execution time, the coefficient of variation of its value or mean, and the ratio of the coefficient of variation to the total execution time.* ***Bottom:*** *Graphical representation of the ratio of the coefficient of variation to the total execution time for each metric, plotted on a logarithmic scale.*

Based on these results, the Betweenness Centrality, the Routing Centrality, the Bridging Centrality, the Percolation Centrality, the Load Centrality, and the Rich Club Coefficient were discarded due to their high total execution times. Although the Clustering Coefficient also has a relatively long execution time, it was initially retained because of its high coefficient of variation and its widespread use in the literature. However, including it caused the dataset generation to exceed an acceptable time frame (taking over 6 hours), and thus the Clustering Coefficient was ultimately discarded as well.

Additionally, the Eigenvector Centrality and the Katz Centrality were discarded due to the very low variance and coefficient of variation of their mean values. Finally, although the total execution time for computing the Radius is small and its coefficient of variation is not too low, it was also discarded because it takes exactly the same value for two different pairs of models, which was considered an indication that it could negatively affect the classifier's performance.

Thus, the metrics that were ultimately selected were the **Degree Distribution**, the **Average Path Length**, the **Diameter**, the **Closeness Centrality**, the **Information Centrality**, the **Harmonic Centrality**, and the **Assortativity**.

5.2 Classifier Selection

In the table below, the results for the precision, recall, accuracy, and F1 score are presented for each classifier, rounded to 4 decimal places.

Table 5.18. *Classifier performance table*

	Precision	**Recall**	**Accuracy**	**F1 Score**
Linear SVC	0.7577	0.7608	0.7583	0.7579
k-Nearest Neighbors	0.5935	0.5805	0.5833	0.5700
SVC	0.4728	0.4423	0.4333	0.3943
Ada Boost	0.7779	0.6999	0.7333	0.6984
Gradient Boosting	0.9512	0.9510	0.9500	0.9498
Random Forest	0.9745	0.9775	0.9750	0.9756
Bagging	0.5494	0.5535	0.5667	0.5357

As shown in Table 5.18, the **Random Forest** classifier achieved the best performance across all metrics and was therefore selected for the classification of the real-world networks. The confusion matrix of this classifier and the results of the ablation study conducted with its assistance are presented in Figures 5.2 and 5.3.

It can be observed that the classifier exhibits a very strong ability to distinguish between all models. Moreover, no metric appears to hinder the classifier, as the removal of any one results either in a decrease in its performance or in a very slight increase (on the order of 1% or less), which is likely due to chance rather than the nature of the metric. In fact, Assortativity proves to be particularly important for the strong performance of the classifier, which is expected given the very high coefficient of variation of this metric. At the same time, this finding indicates that using the coefficient of variation as the primary criterion for network metric selection apart from the execution time, was indeed appropriate.

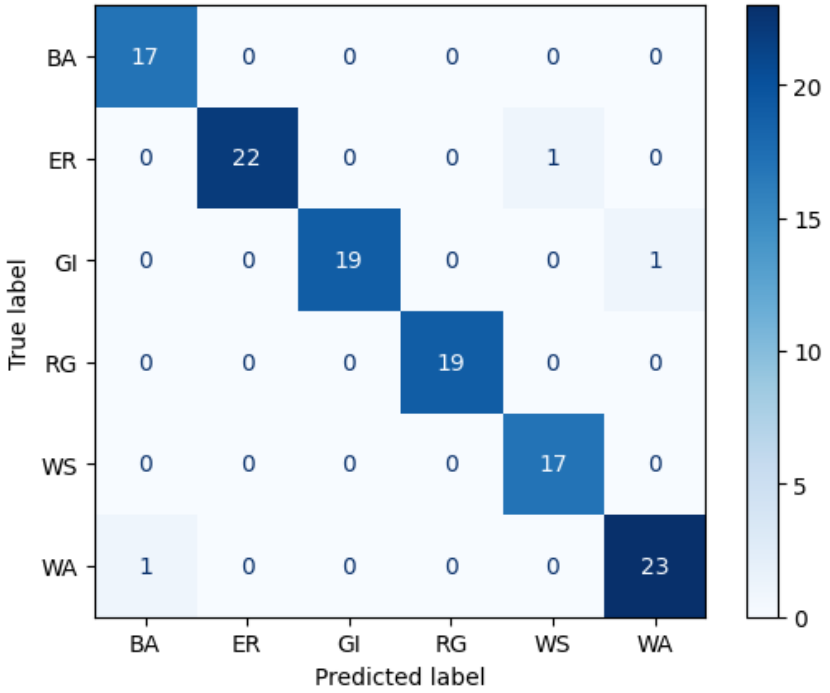


Figure 5.2. *Confusion matrix for the Random Forest classifier. The labels correspond to the acronyms of the respective models.*

	removed_feature	accuracy	precision	recall	f1
0	Degree Distribution	0.975000	0.976754	0.975000	0.975282
1	Average Path Length	0.983333	0.984259	0.983333	0.983390
2	Diameter	0.966667	0.968259	0.966667	0.966604
3	Information Centrality	0.975000	0.976754	0.975000	0.975282
4	Closeness Centrality	0.983333	0.984259	0.983333	0.983390
5	Harmonic Centrality	0.983333	0.984259	0.983333	0.983390
6	Assortativity	0.791667	0.788154	0.791667	0.788812

Figure 5.3. *Ablation study of the metrics Degree Distribution, Average Path Length, Diameter, Closeness Centrality, Information Centrality, Harmonic Centrality, and Assortativity, using the Random Forest classifier*

5.3 Real-world Network Classification

Presented below are the results of the classification of the real-world networks discussed in Chapter 4.

- **Football:** Waxman
- **GOT:** Barabasi-Albert

- **Email-EU:** Waxman
- **AdvogatoTrust:** Barabasi-Albert
- **MorenoCrime:** Watts-Strogatz
- **NetScience:** Watts-Strogatz
- **Euroroad:** Watts-Strogatz
- **ChicagoRegional:** Watts-Strogatz
- **Flights-FAA:** Watts-Strogatz
- **Bristol:** Watts-Strogatz
- **Coach:** Watts-Strogatz
- **YuYeast:** Barabasi-Albert
- **BacillusMegaterium:** Gilbert
- **Malaria:** Random Geometric Graph
- **Fullerene:** Watts-Strogatz
- **PowerNet:** Watts-Strogatz
- **R-Dependancies:** Gilbert

It can be observed that the Watts-Strogatz model occurs far more frequently than all the others (9 occurrences), particularly in transportation networks. The Barabasi–Albert model follows with 2 occurrences in social networks and 1 in a biological network, then the Waxman model (2 occurrences in social networks) and the Gilbert model (1 occurrence in a biological network and 1 in a technological network). The Random Geometric model appears once in a biological network, while the Erdos–Renyi model is absent.

Furthermore, it is worth noting that the PowerNet network was classified into the Watts–Strogatz model, consistent with the findings of Watts and Strogatz [21].

5.4 Network Partitioning Method Selection

The node counts of the component networks are as follows:

- **For Girvan Newman:** Erdos-Renyi: 401,Gilbert: 200, Random Geometric: 392, Watts-Strogatz: 299, Barabasi-Albert: 519, Waxman: 333
- **For the other algorithms:** Erdos-Renyi: 1401,Gilbert: 1200, Random Geometric: 1392, Watts-Strogatz: 1299, Barabasi-Albert: 1519, Waxman: 1333

The results of the community detection-based partitioning on the composite synthetic networks are presented below, along with our observations. Next to each model, the total number of nodes assigned to it is indicated in parentheses. All values are rounded to 2 decimal places.

5.4.1 Barabasi-Albert and Random Geometric

Table 5.19. *Results of Girvan Newman for the combination of Barabasi-Albert and Random Geometric*

	Girvan Newman
Number of communities	2
Community models	Barabasi-Albert (519), Random Geometric (392)
Community sizes	519, 392
Number of discarded nodes	0
Number of correctly classified nodes	911 (out of 911)
Number of correctly classified communities	2
Modularity	0.27
Execution Time (s)	656.18

Table 5.20. *Results of Walktrap and Infomap for the combination of Barabasi-Albert and Random Geometric*

	Walktrap	**Infomap**
Number of communities	4	5
Community models	Barabasi-Albert (1517), Random Geometric (1392)	Barabasi-Albert (1519), Random Geometric (1392)
Community sizes	1517, 599, 485, 308	1519, 418, 374, 324, 276
Number of discarded nodes	0	2
Number of correctly classified nodes	2909 (out of 2909)	2911 (out of 2911)
Number of correctly classified communities	4	5
Modularity	0.45	0.50
Execution Time (s)	300.65	139.42

Table 5.21. *Results of Leiden and Spectral Clustering for the combination of Barabasi-Albert and Random Geometric*

	Leiden	**Spectral Clustering**
Number of communities	5	2
Community models	Barabasi-Albert (1519), Random Geometric (1392)	Barabasi-Albert (1519), Random Geometric (1392)
Community sizes	1519, 364, 361, 336, 331	1519, 1392
Number of discarded nodes	0	0
Number of correctly classified nodes	2911 (out of 2911)	2911 (out of 2911)
Number of correctly classified communities	5	2
Modularity	0.50	0.08
Execution Time (s)	127.97	1049.18

Observations

All algorithms correctly partition the graph into its two components, with the sole exception of Walktrap, which leads to the omission of only 2 nodes. Spectral Clustering is noticeably slower than the others, while Leiden is the fastest.

5.4.2 Barabasi-Albert and Watts-Strogatz

Table 5.22. *Results of Girvan Newman for the combination of Barabasi-Albert and Watts-Strogatz*

	Girvan Newman
Number of communities	2
Community models	Watts-Strogatz (818)
Community sizes	752, 66
Number of discarded nodes	0
Number of correctly classified nodes	299 (out of 818)
Number of correctly classified communities	1
Modularity	0.04
Execution Time (s)	41.28

Table 5.23. *Results of Walktrap and Infomap for the combination of Barabasi-Albert and Watts-Strogatz*

	Walktrap	**Infomap**
Number of communities	1	1
Community models	Barabasi-Albert (1578)	Barabasi-Albert (1495)
Community sizes	1578	1495
Number of discarded nodes	1240	1323
Number of correctly classified nodes	1519 (out of 1578)	1495 (out of 1495)
Number of correctly classified communities	1	1
Modularity	0.18	0.21
Execution Time (s)	92.15	61.96

Observations

Most algorithms detect only one of the Barabasi-Albert and Watts-Strogatz models. The only one that recognizes both is Leiden, which also incorrectly identifies the Erdos–Renyi model, though it assigns relatively few nodes to it (212). At the same time, Leiden and Spectral Clustering are the only candidate algorithms that do not generate a vast number of small communities ($\leq$ 20 nodes) that cannot be classified. Moreover, Leiden has the highest number of correctly classified nodes overall and the shortest execution time.

Table 5.24. *Results of Leiden and Spectral Clustering for the combination of Barabasi-Albert and Watts-Strogatz*

	Leiden	**Spectral Clustering**
Number of communities	20	1
Community models	Barabasi-Albert (1461), Watts-Strogatz (1139), Erdos-Renyi (212)	Watts-Strogatz (2808)
Community sizes	441, 409, 299, 212, 151, 126, 125, 118, 95, 93, 92, 87, 86, 77, 76, 74, 68, 67, 62, 54	2808
Number of discarded nodes	6	10
Number of correctly classified nodes	2089 (out of 2812)	1289 (out of 2808)
Number of correctly classified communities	13	0
Modularity	0.36	0.00
Execution Time (s)	16.76	225.68

5.4.3 Watts-Strogatz and Erdos-Renyi

Table 5.25. *Results of Girvan Newman for the combination of Watts-Strogatz and Erdos-Renyi*

	Girvan Newman
Number of communities	1
Community models	Erdos-Renyi (699)
Community sizes	699
Number of discarded nodes	1
Number of correctly classified nodes	400 (out of 699)
Number of correctly classified communities	0
Modularity	0.00
Execution Time (s)	14.45

Observations

In this case as well, most algorithms detect only one (or none, in the case of Infomap) of the Watts-Strogatz and Erdos–Renyi models. The exception is Walktrap, which also incorrectly identifies the Barabasi-Albert model. However, Walktrap leads to the omission of 1861 nodes and correctly classifies only 281 of the remaining 839. Therefore, the optimal algorithm in this case is Leiden, which demonstrates the best performance along with Spectral Clustering, while completing the task in significantly less time than the latter.

Table 5.26. *Results of Walktrap and Infomap for the combination of Watts-Strogatz and Erdos-Renyi*

	Walktrap	**Infomap**
Number of communities	12	10
Community models	Watts-Strogatz (688), Barabasi-Albert (122), Erdos-Renyi (29)	Barabasi-Albert (236)
Community sizes	367, 149, 54, 48, 46, 29, 29, 25, 24, 24, 23, 21	29, 26, 25, 25, 24, 22, 22, 21, 21, 21
Number of discarded nodes	1861	2464
Number of correctly classified nodes	281 (out of 839)	0 (out of 236)
Number of correctly classified communities	0	0
Modularity	0.27	0.33
Execution Time (s)	4.73	12.21

Table 5.27. *Results of Leiden and Spectral Clustering for the combination of Watts-Strogatz and Erdos-Renyi*

	Leiden	**Spectral Clustering**
Number of communities	21	2
Community models	Watts-Strogatz (2700)	Watts-Strogatz (2700)
Community sizes	179, 168, 150, 135, 132, 127, 125, 125, 124, 124, 123, 122, 122, 122, 120, 120, 119, 118, 118, 114, 113	1745, 955
Number of discarded nodes	0	0
Number of correctly classified nodes	1299 (out of 2700)	1299 (out of 2700)
Number of correctly classified communities	0	0
Modularity	0.38	0.22
Execution Time (s)	8.67	102.95

5.4.4 Waxman and Random Geometric

Table 5.28. *Results of Girvan Newman for the combination of Waxman and Random Geometric*

	Girvan Newman
Number of communities	2
Community models	Random Geometric (392), Waxman (333)
Community sizes	392, 333
Number of discarded nodes	0
Number of correctly classified nodes	725 (out of 725)
Number of correctly classified communities	2
Modularity	0.23
Execution Time (s)	502.36

Table 5.29. *Results of Walktrap and Infomap for the combination of Waxman and Random Geometric*

	Walktrap	**Infomap**
Number of communities	8	8
Community models	Random Geometric (1392), Waxman (1333)	Random Geometric (1392), Waxman (1333)
Community sizes	538, 512, 323, 300, 291, 273, 257, 231	422, 389, 382, 336, 323, 321, 285, 267
Number of discarded nodes	0	0
Number of correctly classified nodes	2725 (out of 2725)	2725 (out of 2725)
Number of correctly classified communities	8	8
Modularity	0.51	0.54
Execution Time (s)	151.42	100.20

Table 5.30. *Results of Leiden and Spectral Clustering for the combination of Waxman and Random Geometric*

	Leiden	**Spectral Clustering**
Number of communities	5	2
Community models	Waxman (1333), Random Geometric (1011), Gilbert (381)	Random Geometric (1392), Waxman (1333)
Community sizes	1333, 381, 354, 340, 317	1392, 1333
Number of discarded nodes	0	0
Number of correctly classified nodes	2344 (out of 2725)	2725 (out of 2725)
Number of correctly classified communities	4	2
Modularity	0.57	0.28
Execution Time (s)	186.16	1091.43

Observations

In this case, all algorithms except Leiden achieve optimal performance in terms of their results. Nevertheless, Leiden's error is relatively small (381 nodes out of 2,725). The shortest execution time belongs to Infomap, followed closely by Walktrap and Leiden.

5.4.5 Gilbert and Waxman

Table 5.31. *Results of Girvan Newman for the combination of Gilbert and Waxman*

	Girvan Newman
Number of communities	2
Community models	Gilbert (200), Waxman (333)
Community sizes	333, 200
Number of discarded nodes	0
Number of correctly classified nodes	533 (out of 533)
Number of correctly classified communities	2
Modularity	0.32
Execution Time (s)	208.19

Table 5.32. *Results of Walktrap and Infomap for the combination of Gilbert and Waxman*

	Walktrap	**Infomap**
Number of communities	2	5
Community models	Waxman (1333), Gilbert (1200)	Waxman (1333), Gilbert (1200)
Community sizes	1333, 1200	1200, 384, 332, 328, 289
Number of discarded nodes	0	0
Number of correctly classified nodes	2533 (out of 2533)	2533 (out of 2533)
Number of correctly classified communities	2	5
Modularity	0.19	0.16
Execution Time (s)	347.12	241.99

Table 5.33. *Results of Leiden and Spectral Clustering for the combination of Gilbert and Waxman*

	Leiden	**Spectral Clustering**
Number of communities	2	2
Community models	Waxman (1333), Gilbert (1200)	Waxman (1333), Gilbert (1200)
Community sizes	1333, 1200	1333, 1200
Number of discarded nodes	0	0
Number of correctly classified nodes	2533 (out of 2533)	2533 (out of 2533)
Number of correctly classified communities	2	2
Modularity	0.19	0.19
Execution Time (s)	332.01	327.99

Observations

For this combination of networks, all algorithms correctly identify the models of the composite graph's components, with all nodes classified correctly as well. Infomap exhibits the shortest execution time, followed closely by the other algorithms (Girvan Newman is excluded from this comparison, as it was run on smaller networks than the rest).

5.4.6 Final Decision

In the cases of the model combinations of Barabasi-Albert with Random Geometric Graph and Gilbert with Waxman, all algorithms achieved optimal performance, with no single one standing out. In the cases of the Barabasi-Albert with Watts-Strogatz and Watts-Strogatz with Erdos-Renyi combinations, Leiden clearly outperforms the others for the reasons discussed earlier. Finally, in the case of the Waxman with Random Geometric Graph combination, Leiden is the only algorithm to exhibit any error, however it is minor. Thus, Leiden was selected for the classification of the real-world networks, as it was deemed to have the best overall performance.

It should also be noted that the Spectral Clustering algorithm achieved performance comparable to that of Leiden in terms of accuracy. However, it was not selected, as its execution times were excessively long, and the requirement of specifying the desired number of communities as an input parameter is quite restrictive in the case of real-world networks (since the number of models composing a network is not known in advance, and an incorrect estimate could lead to a loss of information).

In fact, it is possible that the good performance of the Spectral Clustering algorithm is primarily, or even entirely, due to the fact that it "knows" in advance the number of communities into which the graph should be partitioned. This encourages partitioning into the correct components and prevents fragmentation into many small communities that may not be classifiable. If, for instance, we had tasked Spectral Clustering to divide the networks into six components (the maximum possible number of different models that can appear in a graph), its results might have been very different. However, it is also possible that its node similarity-based approach is the main reason for its strong performance.

With regard to the Walktrap and Infomap algorithms, it can be observed that they share the following two characteristics: they are both based on random walks, and they perform poorly compared to the other algorithms we examined. We may therefore conclude that the use of random walks as a basis for community detection in graphs is a suboptimal approach when the goal is to partition it into the components corresponding to the different models it contains. Finally, we can hypothesize that the reason Leiden achieves such strong results is that it is based on modularity maximization, which (by definition of modularity) leads to community formation by the removal of bridges, while our composite graphs were constructed by connecting their component graphs via bridges. In fact, the Girvan-Newman algorithm, which is also based on bridge removal, also exhibits very strong performance (regardless of its prohibitive execution time). However, this sug-

gests that the connection of the different components of a graph via bridges may be a prerequisite for the strong performance of Leiden, something we do not know whether it holds in general for real-world networks. Nevertheless, based on our results, Leiden remains our best choice for their partitioning.

5.5 Real-world Network Partitioning

The models to which the communities, resulting from the partitioning of the real-world networks using the Leiden algorithm, were assigned are presented bellow, along with the number of nodes in each model component.

- **Football:** – (This graph is too small to be partitioned into communities containing over 20 nodes)
- **GOT:** Barabasi-Albert (58)
- **Email-EU:** Barabasi-Albert (444), Gilbert (348), Waxman (194)
- **AdvogatoTrust:** Barabasi-Albert (4302), Waxman (278), Watts-Strogatz (218), Erdos-Renyi (200)
- **MorenoCrime:** Barabasi-Albert (461), Watts-Strogatz (315), Erdos-Renyi (36)
- **NetScience:** Barabasi-Albert (260), Waxman (24)
- **Euroroad:** Watts-Strogatz (947), Barabasi-Albert (32)
- **ChicagoRegional:** Watts-Strogatz (12979)
- **Flights-FAA:** Watts-Strogatz (579), Barabasi-Albert (363), Erdos-Renyi (284)
- **Bristol:** Watts-Strogatz (2529)
- **Coach:** Watts-Strogatz (1961), Erdos-Renyi (287), Barabasi-Albert (48)
- **YuYeast:** Barabasi-Albert (810), Watts-Strogatz (788)
- **BacillusMegaterium:** Barabasi-Albert (4319)
- **Malaria:** Random Geometric Graph (31)
- **Fullerene:** Watts-Strogatz (3840)
- **PowerNet:** Watts-Strogatz (4796), Erdos-Renyi (95), Barabasi-Albert (31)
- **R-Dependancies:** Gilbert (29)

We observe that most networks either contain a single evolution model (the one which they were initially classified as) or consist of a combination of models that includes the model they were originally classified as. The only exceptions are the NetScience and BacillusMegaterium networks, in which no component corresponding to the initial model

is identified. Moreover, in all social networks and in most biological ones, the largest component corresponds to the Barabasi-Albert (Scale-Free) model. In fact, Barabasi-Albert is the most prevalent component, appearing in 11 networks, followed closely by Watts-Strogatz with 10 occurrences. The third most prevalent component is Erdos-Renyi, with only 5 occurrences, noticeably fewer than the previous two. The remaining models appear even more rarely, with Waxman appearing 3 times (all in social networks), Gilbert twice, and Random Geometric Graph once.

It should be noted that the relatively high prevalence of the Erdos-Renyi component may be due to an error of the Leiden algorithm and might not reflect the true structure of the networks. It can be observed that during the evaluation of the community detection algorithms, Leiden incorrectly identified a component corresponding to a Random Graph twice (once to Erdos-Renyi and an other time to Gilbert). In both cases, the misclassified components were the smallest in their respective networks, as is generally the case for the Erdos-Renyi components observed in this partitioning and classification procedure. This pattern casts doubt about the reliability of the results concerning the Erdos-Renyi model.

Part III

Conclusion

Chapter 6

Conclusion

This chapter serves as the conclusion of the present work. It provides a summary of our methods and results, the conclusions we reached based on the latter, as well as ideas for future extensions.

6.1 Summary and Conclusions

In the context of this thesis, and building on related studies, we first conducted a literature review, through which we identified 6 evolution models and 20 network metrics that could potentially be used for classifying networks into these models. We then carried out a series of tests on these metrics (placing particular weight on the variance exhibited by each metric across all 6 models and on their execution time), and selected 7 to be used as input to a classifier. The classifier itself was also chosen after an assessment of its performance, using the metrics precision, recall, accuracy, and F1 score. After being trained on a set of 600 synthetic networks that we constructed, it was used to classify 17 real-world networks, with the aforementioned metrics as input. In addition, we evaluated 5 community detection algorithms based on their execution time and accuracy on 5 synthetic networks, each of which was a composite of two other networks generated using different evolution models. The algorithm that produced the most accurate results within a reasonable time frame was then chosen to be applied on the aforementioned real-world networks. The resulting subnetworks were subsequently classified with the assistance of the selected classifier.

Specifically, the models used were Erdos-Renyi, Gilbert, Random Geometric, Watts-Strogatz, Barabasi-Albert, and Waxman. The selected metrics were Degree Distribution, Average Path Length, Diameter, Closeness Centrality, Information Centrality, Harmonic Centrality, and Assortativity, while the chosen classifier was Random Forest, and the community detection algorithm was Leiden. Regarding the classification results, which are presented in detail in Chapter 5, it can be observed that the model most frequently identified when networks are examined intact is the Watts-Strogatz (Small-World) model. When the networks are partitioned, it can be observed that nearly all contain a component corresponding to the model into which they were originally classified. Furthermore, components corresponding to the Barabasi-Albert model are particularly common, as they are present in most networks under examination and exhibit the highest prevalence

compared to those corresponding to other models (followed by the Watts-Strogatz model). Notably, in all social networks, as well as in most biological ones, the Barabasi-Albert component is the largest, regardless of the model into which the network was originally classified. Similarly, in all transportation networks, the largest component corresponds to the Watts-Strogatz model, an expected outcome given that all of them had initially been classified into this model.

Based on these results, we conclude that the models that appear most frequently in real-world networks are the Small-World and Scale-Free ones. When analyzing networks by category, we deem it safe to assert that transportation networks are predominantly characterized by the Small-World model, which is also consistent with the findings of Jiang and Claramunt [68] and Sen et al. [69]. Furthermore, the assumption that real-world networks can be decomposed into components corresponding to different evolution models appears valid, as almost all networks contain at least one component corresponding to the model assigned to them when they were studied intact, and the presence of the Scale-Free model as the largest component in every social network (with the sole exception of the YuYeast network) is unlikely to be due to error or coincidence. Moreover, given its presence across all examined social networks, we infer that social networks generally contain a large Scale-Free component, and that they could potentially be characterized entirely as Scale-Free if a single model were to be assigned, since this component is clearly the largest. The fact that most of these networks were classified as Small-World when considered as a whole may be attributed to classifier error, arising from the coexistence of multiple models and the similarity between the Scale-Free and Small-World models (so much so that some researchers consider the Scale-Free model a subcategory of Small-World [81]). It should also be noted that the Waxman model appears as a component in most social networks, with some networks classified as such in their entirety, indicating that Waxman structures are also frequently found in social networks. No strong patterns were observed in the remaining network categories that would allow similar conclusions; however, it is worth noting that the power grid we examined (PowerNet) was classified as a Small-World network, in agreement with the findings of related studies mentioned in previous chapters [65] [3] [21].

Regarding the prevalence of the Scale-Free model, we conclude that while it is not ubiquitous, it is not rare either. The category of networks in which it is most prevalent is social networks; however, it also occurs in other categories as a secondary component (transportation and technological networks), and even as a primary one (biological networks).

6.2 Future Research

Within the framework of this thesis, although we employed a plethora of methods and technologies to address the questions we posed, we did not exhaust them. Below, we present several ways in which the content of this work could be further enriched.

- Incorporation of additional models. While we employed most of the models identified in the literature, some were deliberately omitted, as mentioned in Chapter 4. It would be interesting to also examine the case where all models are included in the creation of the training set, and subsequently in the classification process.

- Incorporation of additional network metrics. During the metric selection process, certain metrics were excluded due to their excessively long computation times, despite exhibiting high variance. Should more time be available, or their computation be accelerated, these metrics could be included to potentially enhance the classifier's performance. Furthermore, metrics beyond the 20 we examined could also be employed.

- Use of a different classifier. The selection of the classifiers examined was carried out by meticulously following the diagram in Figure 4.1. However, the category of Ensemble Classifiers is very broad, resulting in not all classifiers within it being considered, but only the 4 most important ones. Moreover, the diagram pertains exclusively to traditional classifiers, whereas Graph Neural Networks (GNNs) are also used for network classification with very promising results. Although the classifier we selected demonstrated very strong performance, it is possible that a different one could have performed even better.

- Use of more networks for classifier training. Due to time constraints, we only used 100 networks per model to create the classifier's training dataset. Had those constraints not existed, and more networks per model been used, perhaps we would observe an improved classifier performance.

- Incorporation of directed and weighted networks in the training data. All synthetic networks we generated for inclusion in the classifier's training and test sets were undirected and unweighted. Including directed and weighted networks in the training dataset could potentially improve the classifier's performance on real-world networks exhibiting these characteristics.

- Alternative handling of directed networks. The approach we used for real-world directed networks was to ignore their directionality. However, they could have been handled differently, for instance as in the study by Broido and Clauset [70]. Perhaps, in that case, the results of both the classification and partitioning would have been different.

- Use of a different network partitioning algorithm. To identify the optimal algorithm for partitioning real-world networks into components corresponding to different models, we evaluated 5 community detection algorithms. However, given the very large number of community detection algorithms available, it is possible that the optimal algorithm for this task is one of those we did not consider.

- Validation of results using the small-world coefficient. A large portion of the real-world networks we examined were classified as Small-World. It would be interesting

to compute the small-world coefficient σ for these networks, in order to determine whether our results are corroborated by this alternative method of detecting Small-World structure.

6.3 Code availability

All files containing our code and results can be accessed through the following link:
https://github.com/FaidraA/Evolution_model_identification

Bibliography

[1] M. E. J. Newman. *The Structure and Function of Complex Networks*. *SIAM Review*, 45(2):167–256, January 2003.

[2] Pedro J. Zufiria and Iker Barriales-Valbuena. *Evolution models for dynamic networks*. in *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*, pages 252–256, 2015.

[3] Rafael Espejo, Sara Lumbreras and Andres Ramos. *Analysis of transmission-power-grid topology and scalability, the European case study*. *Physica A: Statistical Mechanics and its Applications*, 509:383–395, 2018.

[4] Hans Hinterberger. *Graph*, pages 1260–1261. Springer US, Boston, MA, 2009.

[5] Mrs Kothimbire, D. Shelke, Mrs Gaikwad, A. Yelpale and R. Shinde. *A Comprehensive Review of Graph Theory Applications in Network Analysis*. *International Journal of Mathematics And Computer Research*, 13, 03 2025.

[6] Geoffrey Canright. *Complex NetworksComplex networksand Graph Theory*, pages 1244–1253. Springer New York, New York, NY, 2009.

[7] NY Comdori, 2020.

[8] M. E. J. Newman. *Analysis of weighted networks*. *Phys. Rev. E*, 70:056131, Nov 2004.

[9] Wikipedia contributors. *Path (graph theory) — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Path_(graph_theory)&oldid=1275103758, 2025.

[10] Wikipedia contributors. *Connectivity (graph theory) — Wikipedia, The Free Encyclopedia*, 2025.

[11] Geeksforgeeks authors. *Graph and its representations*, 2024.

[12] Arnab Chakraborty, 2020.

[13] Ernesto Estrada. *The structure of complex networks: theory and applications*. American Chemical Society, 2012.

[14] Mikhail Drobyshevskiy and Denis Turdakov. *Random Graph Modeling: A Survey of the Concepts*. *ACM Comput. Surv.*, 52(6), dec 2019.

[15] Frédéric Amblard, Audren Bouadjio-Boulic, Carlos Sureda Gutiérrez and Benoit Gaudou. *Which models are used in social simulation to generate social networks? a review of 17 years of publications in JASSS.* in *2015 Winter Simulation Conference (WSC)*, pages 4021–4032, 2015.

[16] Quentin Duchemin and Yohann De Castro. *Random Geometric Graph: Some Recent Developments and Perspectives.* in *High Dimensional Probability IX*Radosław Adamczak, Nathael Gozlan, Karim Lounici and Mokshay Madiman, editors, pages 347–392, Cham, 2023. Springer International Publishing.

[17] Matthew Roughan, Jonathan Tuke and Eric Parsonage. *Estimating the Parameters of the Waxman Random Graph.* in *Algorithms and Models for the Web Graph*Konstantin Avrachenkov, Paweł Prałat and Nan Ye, editors, pages 71–86, Cham, 2019. Springer International Publishing.

[18] Albert László Barabási, Erzsébet Ravasz and Tamás Vicsek. *Deterministic scale-free networks.* *Physica A: Statistical Mechanics and its Applications*, 299(3):559–564, 2001.

[19] Albert Laszlo Barabasi and Reka Albert. *Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509-512. Science (New York, N.Y.)*, 286:509–12, 11 1999.

[20] Cristina Guzman, Daphna Keidar, Tristan Meynier, Andreas Opedal and Niklas Stoehr. *Recovering Barabási-Albert Parameters of Graphs through Disentanglement*, 2021.

[21] Duncan J. Watts and Steven H. Strogatz. *Collective dynamics of 'small-world' networks.* *Nature*, 393(6684):440–442, Jun 1998.

[22] Javier Martın Hernández and Piet Van Mieghem. *Classification of graph metrics. Delft University of Technology: Mekelweg, The Netherlands*, 1, 2011.

[23] Charles Perez and Rony Germon. *Chapter 7 - Graph Creation and Analysis for Linking Actors: Application to Social Data.* in *Automating Open Source Intelligence*Robert Layton and Paul A. Watters, editors, pages 103–129. Syngress, Boston, 2016.

[24] Simon Mukwembi. *A note on diameter and the degree sequence of a graph.* *Applied Mathematics Letters*, 25(2):175–178, 2012.

[25] Chintan Amrit and Joanneter Maat. *Understanding Information Centrality Metric: A Simulation Approach*, 2018.

[26] Karen Stephenson and Marvin Zelen. *Rethinking centrality: Methods and examples.* *Social Networks*, 11(1):1–37, 1989.

[27] Paolo Boldi and Sebastiano Vigna. *Axioms for Centrality*, 2013.

[28] M. Barthélemy. *Betweenness centrality in large complex networks.* *The European Physical Journal B*, 38(2):163–168, Mar 2004.

[29] Alfredo Cuzzocrea, Alexis Papadimitriou, Dimitrios Katsaros and Yannis Manolopoulos. *Edge betweenness centrality: A novel algorithm for QoS-based topology control over wireless sensor networks. Journal of Network and Computer Applications*, 35(4):1210–1217, 2012. Intelligent Algorithms for Data-Centric Sensor Networks.

[30] M. Girvan and M. E. J. Newman. *Community structure in social and biological networks. Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[31] Gnana Thedchanamoorthy, Mahendra Piraveenan, Dharshana Kasthuriratna and Upul Senanayake. *Node Assortativity in Complex Networks: An Alternative Approach. Procedia Computer Science*, 29:2449–2461, 2014. 2014 International Conference on Computational Science.

[32] Punam Bedi and Chhavi Sharma. *Community detection in social networks. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6:n/a–n/a, 02 2016.

[33] Natalie R Smith, Paul N Zivich, Leah M Frerichs, James Moody and Allison E Aiello. *A guide for choosing community detection algorithms in social network studies: The Question Alignment approach. Am. J. Prev. Med.*, 59(4):597–605, October 2020.

[34] Vincent D Blondel, Jean Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre. *Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[35] Pascal Pons and Matthieu Latapy. *Computing Communities in Large Networks Using Random Walks. Journal of Graph Algorithms and Applications*, 10(2):191–218, Jan. 2006.

[36] Martin Rosvall and Carl T Bergstrom. *Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. U. S. A.*, 105(4):1118–1123, January 2008.

[37] V. A. Traag, L. Waltman and N. J.van Eck. *From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports*, 9(1), March 2019.

[38] Zhixin Zhou and Arash A. Amini. *Analysis of spectral clustering algorithms for community detection: the general bipartite setting*, 2018.

[39] Ulrikevon Luxburg. *A tutorial on spectral clustering. Statistics and Computing*, 17(4):395–416, Dec 2007.

[40] Harry Sevi, Matthieu Jonckheere and Argyris Kalogeratos. *Generalized Spectral Clustering for Directed and Undirected Graphs*, 2022.

[41] Gopinath Rebala, Ajay Ravi and Sanjay Churiwala. *Machine Learning Definition and Basics*, pages 1–17. Springer International Publishing, Cham, 2019.

[42] Weiche Hsieh, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang, Keyu Chen, Pohsun Feng, Yizhu Wen, Xinyuan Song, Tianyang Wang, Ming Liu, Junjie Yang, Ming Li, Bowen Jing, Jintao Ren, Junhao Song, Hong Ming Tseng, Yichao Zhang, Lawrence K. Q. Yan, Qian Niu, Silin Chen, Yunze Wang and Chia Xin Liang. *A Comprehensive Guide to Explainable AI: From Classical Models to LLMs*, 2024.

[43] Julianna Delua. *Supervised versus unsupervised learning: What's the difference?*, 2021.

[44] Gopinath Rebala, Ajay Ravi and Sanjay Churiwala. *Classification*, pages 57–66. Springer International Publishing, Cham, 2019.

[45] Gireen Naidu, Tranos Zuva and Elias Mmbongeni Sibanda. *A review of evaluation metrics in machine learning algorithms*. in *Computer science on-line conference*, pages 15–25. Springer, 2023.

[46] Mohammad Hossin and Md Nasir Sulaiman. *A review on evaluation metrics for data classification evaluations*. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[47] Margherita Grandini, Enrico Bagli and Giorgio Visani. *Metrics for Multi-Class Classification: an Overview*, 2020.

[48] Himani Bhavsar and Mahesh H Panchal. *A review on support vector machine for data classification*. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 1(10):185–189, 2012.

[49] Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer. *KNN Model-Based Approach in Classification*. in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*Robert Meersman, Zahir Tari and Douglas C. Schmidt, editors, pages 986–996, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[50] Ruihu Wang. *AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review*. *Physics Procedia*, 25:800–807, 2012. International Conference on Solid State Devices and Materials Science, April 1-2, 2012, Macao.

[51] Alexey Natekin and Alois Knoll. *Gradient boosting machines, a tutorial*. *Frontiers in neurorobotics*, 7:21, 2013.

[52] Wei Fan and Kun Zhang. *Bagging*, pages 206–210. Springer US, Boston, MA, 2009.

[53] Lior Rokach and Oded Maimon. *Decision Trees*, pages 165–192. Springer US, Boston, MA, 2005.

[54] Aakash Parmar, Rakesh Katariya and Vatsal Patel. *A Review on Random Forest: An Ensemble Classifier*. in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*Jude Hemanth, Xavier Fernando,

Pavel Lafata and Zubair Baig, editors, pages 758–763, Cham, 2019. Springer International Publishing.

[55] Travis Meyer, Cesar Ramirez, Matthew Tamasi and Adam Gormley. *A User's Guide to Machine Learning for Polymeric Biomaterials*. *ACS Polymers Au*, 3, 11 2022.

[56] Max Bramer. *Ensemble Classification*, pages 209–220. Springer London, London, 2020.

[57] Linhong Zhu, Wee Keong Ng and Shuguo Han. *Classifying Graphs Using Theoretical Metrics: A Study of Feasibility*. in *Database Systems for Adanced Applications*Jianliang Xu, Ge Yu, Shuigeng Zhou and Rainer Unland, editors, pages 53–64, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[58] Réka Albert and Albert László Barabási. *Statistical mechanics of complex networks*. *Reviews of Modern Physics*, 74(1):47–97, January 2002.

[59] Sergey Dorogovtsev and José Fernando Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*, volume 57. Oxford University Press, 03 2003.

[60] Gábor Szárnyas, Zsolt Kővári, Ágnes Salánki and Dániel Varró. *Towards the characterization of realistic models: evaluation of multidisciplinary graph metrics*. in *Proceedings of the ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems*, MODELS '16, page 87–94, New York, NY, USA, 2016. Association for Computing Machinery.

[61] John Boaz Lee, Ryan Rossi and Xiangnan Kong. *Graph Classification using Structural Attention*. in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1666–1674, New York, NY, USA, 2018. Association for Computing Machinery.

[62] Federico Errica, Marco Podda, Davide Bacciu and Alessio Micheli. *A Fair Comparison of Graph Neural Networks for Graph Classification*, 2022.

[63] Keyulu Xu, Weihua Hu, Jure Leskovec and Stefanie Jegelka. *How Powerful are Graph Neural Networks?*, 2019.

[64] Qawi K. Telesford, Karen E. Joyce, Satoru Hayasaka, Jonathan H. Burdette and Paul J. Laurienti. *The ubiquity of small-world networks*, 2011.

[65] Bálint Hartmann and Viktória Sugár. *Searching for small-world and scale-free behaviour in long-term historical data of a real-world power grid*. *Scientific Reports*, 11(1):6575, Mar 2021.

[66] Michael Stumpf, Piers Ingram, I. Nouvel and C. Wiuf. *Statistical Model Selection Methods Applied to Biological Networks*. *Transactions on Computational Systems Biology III*, 3, 07 2005.

[67] Raya Khanin and Ernst Wit. *How Scale-Free Are Biological Networks. Journal of Computational Biology*, 13(3):810–818, 2006. PMID: 16706727.

[68] Bin Jiang and Christophe Claramunt. *Topological Analysis of Urban Street Networks. Environment and Planning B: Planning and Design*, 31(1):151–162, 2004.

[69] Parongama Sen, Subinay Dasgupta, Arnab Chatterjee, P. A. Sreeram, G. Mukherjee and S. S. Manna. *Small-world properties of the Indian railway network. Phys. Rev. E*, 67:036106, Mar 2003.

[70] Anna D. Broido and Aaron Clauset. *Scale-free networks are rare. Nature Communications*, 10(1):1017, Mar 2019.

[71] Albert László Barabási. *Love is all you need. Clauset's fruitless search for scale-free networks. unpublished*, 2018.

[72] Ivan Voitalov, Pimvan der Hoorn, Remcovan der Hofstad and Dmitri Krioukov. *Scale-free networks well done. Phys. Rev. Res.*, 1:033034, Oct 2019.

[73] Linqing Liu, Mengyun Shen and Chang Tan. *Scale free is not rare in international trade networks. Scientific Reports*, 11(1):13359, Jun 2021.

[74] I. Artico, I. Smolyarenko, V. Vinciotti and E. C. Wit. *How rare are power-law networks really? Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2241):20190742, 2020.

[75] Gipsi Lima-Mendez and Jacquesvan Helden. *The powerful law of the power law and other myths in network biology. Molecular bioSystems*, 5:1482–93, 12 2009.

[76] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel and Bobby Bhattacharjee. *Measurement and Analysis of Online Social Networks.* pages 29–42, 10 2007.

[77] Petter Holme. *Rare and everywhere: Perspectives on scale-free networks. Nature Communications*, 10(1):1016, Mar 2019.

[78] Markus Meringer and Eric W. Weisstein. *Regular Graph.* From MathWorld–A Wolfram Resource.

[79] M.E.J. Newman and D.J. Watts. *Renormalization group analysis of the small-world network model. Physics Letters A*, 263(4):341–346, 1999.

[80] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, 12:2825–2830, 2011.

[81] L. A. N. Amaral, A. Scala, M. Barthélémy and H. E. Stanley. *Classes of small-world networks. Proceedings of the National Academy of Sciences*, 97(21):11149–11152, 2000.

List of Abbreviations

e.g.	for example
i.e.	that is
GNN	Graph Neural Network
k-NN	k-Nearest Neighbors
Rk-NN	Reverse k-Nearest Neighbors
SVC	Support Vector Classifier
SVM	Support Vector Machine
UK	United Kingdom
USA	United States of America