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

Σύγκριση Σύγχρονων Μεθόδων Εξεύρεσης Κοινοτήτων και Πρακτικές Εφαρμογές

Διπλωματικη Εργασια

της

ΑΝΤΩΝΙΑΣ ΡΑΦΑΕΛΑΣ Σ. ΚΑΤΑΡΑ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2022

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

Σύγκριση Σύγχρονων Μεθόδων Εξεύρεσης Κοινοτήτων και Πρακτικές Εφαρμογές

Διπλωματικη Εργασια

της

ΑΝΤΩΝΙΑΣ ΡΑΦΑΕΛΑΣ Σ. ΚΑΤΑΡΑ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την κάτωθι τριμελή επιτροπή την 21η Μαρτίου 2022.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Συμεών Παπαβασιλείου	Ιωάννα Ρουσσάκη	Γεώργιος Ματσόπουλος
Καθηγητής Ε.Μ.Π.	Επίκουρη Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2022

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

(Υπογραφή)

.....
Αντωνία Ραφαέλα Κατάρα

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

Περίληψη

Η Ανίχνευση Κοινοτήτων είναι μία από τις πιο δημοφιλείς μεθόδους για την ανάλυση Μεγάλων Δεδομένων που προκύπτουν από τοπολογίες Σύνθετων Δικτύων. Τα περισσότερα πραγματικά δίκτυα εμφανίζουν κοινοτική δομή, δηλαδή οι κόμβοι είναι οργανωμένοι σε ομάδες και μοιράζονται τις περισσότερες φορές μια κοινή ιδιότητα. Όσο το μέγεθος και η πολυπλοκότητα των δικτύων αυξάνεται, το πρόβλημα της Ανίχνευσης Κοινοτήτων σε γράφους καθίσταται ένα δύσκολο υπολογιστικά πρόβλημα, δημιουργώντας την ανάγκη για ανάπτυξη μεθόδων που θα δίνουν ικανοποιητικά αποτελέσματα σε μικρό χρόνο.

Στην παρούσα διπλωματική εργασία, αρχικά, ερευνήθηκαν συγκεκριμένες μέθοδοι Ανίχνευσης Κοινοτήτων σε μεγάλους γράφους. Στη συνέχεια, και με βάση την υπάρχουσα βιβλιογραφία, επιλέχθηκαν δύο αλγόριθμοι, από δύο διαφορετικές κατηγορίες μεθόδων. Σε πρώτο στάδιο μελετήθηκε ο Hyperbolic Girvan-Newman, ο οποίος αποτελεί μια παραλλαγή του κλασικού Girvan-Newman αλγορίθμου. Η διαφορά τους έγκειται, στο ότι ο HGN κάνει πρώτα χρήση δικτύων, ενσωματωμένων στον Υπερβολικό χώρο. Σε δεύτερο στάδιο υλοποιήθηκε ο αλγόριθμος Walktrap, ο οποίος βασίζεται σε Τυχαίους Περιπάτους.

Έπειτα μελετήθηκε τόσο η απόδοση του HGN όσο και του Walktrap σε δεδομένα από Κοινωνικά Δίκτυα μικρής-κλίμακας και επιχειρήθηκε μία σύγκριση των αποτελεσμάτων μεταξύ αυτών των δύο μεθόδων. Τα αποτελέσματα αυτά παρουσιάζονται αναλυτικά τόσο με πίνακες όσο και με διαγράμματα, στο αντίστοιχο κεφάλαιο της Διπλωματικής Εργασίας.

Λέξεις Κλειδιά

Ανίχνευση Κοινοτήτων, Υπερβολικός Χώρος, Υπερβολική Γεωμετρία, Κοινωνικά Δίκτυα, Ενσωμάτωση Δικτύου, Τυχαίοι Περίπατοι.

Abstract

Community Detection is one of the most popular methods for analysis of Big Data sets that stem from complex network topologies. Most real networks have a community structure, i.e., their nodes are organised in groups and most of the time they share a common property. As the size and complexity of networks increases, detecting communities in graphs is becoming a difficult computational problem, creating the need to develop methods that will yield satisfactory results in a short time.

In the framework of this diploma thesis, the basic techniques of Community Detection in big graphs were initially studied. Thereafter, based on the existing literature, two algorithms were selected, representing two different types of approaches. The first is Hyperbolic Girvan-Newman, which is a variant of the classical Girvan-Newman algorithm. Their difference is that HGN employs network graph embedding in the Hyperbolic Space. At the second stage, Walktrap algorithm was implemented, which is based on Random Walks.

Then the efficiency of both HGN and Walktrap was studied on data from small-scale Social Networks and a comparison of the results between these two methods was attempted. These results are presented thoroughly using tables and diagrams in the respective chapter of this Diploma Thesis.

Keywords

Community Detection, Clustering, Hyperbolic Space, Hyperbolic Geometry, Online Social Networks, Network Embedding, Random Walks.

Ευχαριστίες

Κατ' αρχάς, θα ήθελα να ευχαριστήσω θερμά τον καθηγητή ΕΜΠ Συμεών Παπαβασιλείου για την επίβλεψη της διπλωματικής μου εργασίας, καθώς και για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα.

Ευχαριστώ επίσης ιδιαίτερα τον Αναπληρωτή Καθηγητή του Τμήματος Πληροφορικής του Ιονίου Παν/μίου Βασίλη Καρυώτη για την πολύτιμη βοήθεια του και τον χρόνο που διέθεσε για την ολοκλήρωση της εργασίας μου.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, την αδελφή μου και αγαπημένα μου πρόσωπα για τη διαρκή στήριξη τους, καθ' όλη την διάρκεια των σπουδών μου.

Αθήνα, 21η Μαρτίου 2022

Αντωνία Ραφαέλα Κατάρα

Περιεχόμενα

Περίληψη **1**

Abstract **3**

Ευχαριστίες **5**

1 Εισαγωγή **15**
1.1 Περιγραφή του προβλήματος 15
1.1.1 Προσεγγίσεις 16
1.2 Αντικείμενο της διπλωματικής 17
1.3 Δομή της Διπλωματικής Εργασίας 17

2 Βασικά Στοιχεία Θεωρίας Γραφημάτων **19**
2.1 Βασική ορολογία, ορισμοί και προκαταρκτικές γνώσεις. 19
2.1.1 Περίπατοι-Μονοπάτια 21
2.2 Ανίχνευση Κοινοτήτων σε Γράφους 22
2.2.1 Ορισμός Κοινότητας 22
2.2.2 Τύποι Κοινοτήτων 24

3 Τεχνητά Σύνθετα Δίκτυα **25**
3.1 Μοντέλα Κατασκευής 25
3.1.1 Τυχαίοι Γράφοι 26
3.1.2 Τυχαίοι Γεωμετρικοί Γράφοι 26
3.1.3 Δίκτυα Μικρού Κόσμου 27
3.1.4 Δίκτυα Ελεύθερης Κλίμακας 28
3.2 Μετρικές Σύνθετων Δικτύων 29
3.2.1 Συντελεστής Ομαδοποίησης 29
3.2.2 Κεντρικότητα Ενδιαμεσικότητας Ακμής 30
3.2.3 Κανονικοποιημένη Αμοιβαία Πληροφορία 30

4 Μεγιστοποίηση της Αρθρωτότητας **33**

5 Ανακάλυψη Κοινοτήτων με την Μέθοδο Hyperbolic Girvan-Newman **35**
5.1 Ο αλγόριθμος Girvan-Newman 35
5.2 Ενσωμάτωση Rigel 36
5.3 Υπολογισμός Υπερβολικής Κεντρικότητας Ενδιαμεσικότητας Ακμής 37

5.4 Ο Αλγόριθμος Hyperbolic Girvan-Newman 39

6 Ανακάλυψη Κοινοτήτων με τον Αλγόριθμο Walktrap 41
6.1 Βασικές Αρχές Τυχαίων Περιπάτων 41
6.1.1 Ορισμοί 41
6.1.2 Ιδιότητες 42
6.1.3 Μέτρο Δομικής Ομοιότητας 43
6.2 Περιγραφή του Αλγόριθμου Walktrap 44

7 Πειραματική Αξιολόγηση 47
7.1 Τοπολογίες Δικτύων Πειραμάτων 47
7.1.1 Πραγματικά Δίκτυα 47
7.1.2 Τεχνητά Σύνθετα Δίκτυα 48
7.2 Σύγκριση των Μεθόδων Hyperbolic Girvan-Newman και Walktrap 50

8 Επίλογος 59
8.1 Σύνοψη και Συμπεράσματα 59
8.2 Μελλοντικές Επεκτάσεις 60

Παραρτήματα 61

Α΄ Πηγαίος Κώδικας 63
Α΄.0.1 Κώδικας παραγωγής LFR δικτύων και εξαγωγής των κοινοτήτων τους . 63
Α΄.0.2 Η μέθοδος Modularity Maximization 64
Α΄.0.3 Βοηθητικός κώδικας για τον αλγόριθμο HGN 65
Α΄.0.4 Ο αλγόριθμος Walktrap 66
Α΄.0.5 Υπολογισμός Μετρικής NMI 67

Β΄ Κατηγορίες Βιβλιογραφικών Αναφορών 69

Βιβλιογραφία 74

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια 75

Απόδοση ξενόγλωσσων όρων 77

Ευρετήριο ελληνικών όρων 79

Ευρετήριο ξενόγλωσσων όρων 81

Κατάλογος Εικόνων

2.1 Παράδειγμα γράφου, πρόκειται για το γράφο του δικτύου the polbooks network [1]. . . . 20
2.2 Παράδειγμα αναπαράστασης γειτονιάς. Η γειτονιά του κόμβου 1 είναι το σύνολο κόμβων που αποτελείται από τους 2,3,5,6 και 7.[1] . . . 21
2.3 Στον παραπάνω γράφο η ακμή (4,5) αποτελεί γέφυρα, η αφαίρεση της οποίας δημιουργεί δύο συνεκτικές συνιστώσες. . . . 22
2.4 Ο υπογράφος με τους κόμβους 1,2,3,4,5,16 είναι μία κλίκα [2]. . . . 23

3.1 Τυχαίος Γράφος βάσει του μοντέλου Gilbert, για τρεις διαφορετικές τιμές του p [3]. . . . 26
3.2 Παράδειγμα ενός Τυχαίου Γεωμετρικού Γράφου [4]. . . . 27
3.3 Εξέλιξη ενός γράφου Watts-Strogatz, όσο αυξάνεται η πιθανότητα p. . . . 28
3.4 Εξέλιξη ενός δικτύου επιστημονικών συνεργασιών σε χρονικό διάστημα τεσσάρων μηνών, καθώς εισέρχονται νέα μέλη-κόμβοι [5]. . . . 28
3.5 Ο αλγόριθμος του Brandes [6]. . . . 31

4.1 Σχηματική Αναπαράσταση της μεθόδου **Blondel**. . . . 34

5.1 Διάγραμμα Ροής του αλγορίθμου Girvan-Newman[2]. . . . 36
5.2 Αποτέλεσμα διαμέρισης του Girvan-Newman, στο δίκτυο Zachary's Karate Club [7]. . . . 37
5.3 Διάγραμμα Ροής του αλγορίθμου HGN. . . . 40

6.1 Σχηματική Αναπαράσταση Τυχαίου Περιπάτου σε Γράφο. . . . 42
6.2 Σχηματική Αναπαράσταση του **Walktrap**. . . . 45

7.1 Απεικόνιση του δικτύου Social circles: Facebook. . . . 48
7.2 Σύγκριση Χρόνου Εκτέλεσης των δύο αλγορίθμων. . . . 51
7.3 Αρθρωτότητα που προκύπτει από την εφαρμογή των HGN και Walktrap. . . . 51
7.4 Σύγκριση Χρόνου Εκτέλεσης των δύο αλγορίθμων. . . . 53
7.5 Αρθρωτότητα που προκύπτει από την εφαρμογή των HGN και Walktrap. . . . 53
7.6 Σύγκριση Χρόνου Εκτέλεσης για 4 Δίκτυα Ελεύθερης Κλίμακας. . . . 54
7.7 Παραγόμενη Αρθρωτότητα για τα 4 Δίκτυα Ελεύθερης Κλίμακας του Πίνακα 7.6. . . . 54
7.8 Σύγκριση Χρόνου Εκτέλεσης για 4 Δίκτυα Μικρού Κόσμου. . . . 55
7.9 Παραγόμενη Αρθρωτότητα για τα 4 Δίκτυα Μικρού Κόσμου του Πίνακα 7.6. . . . 55

7.10 Σύγκριση Χρόνου Εκτέλεσης για 4 Τυχαίους Γεωμετρικούς Γράφους. 56
7.11 Παραγόμενη Αρθρωτότητα για τα 4 Δίκτυα Τυχαίου Γεωμετρικού Γράφου του Πίνακα 7.6. 56
7.12 Σύγκριση τιμών Αρθρωτότητας και μετρικής NMI του αλγορίθμου Walktrap, για τα 5 δίκτυα του Πίνακα 7.7 57
7.13 Σύγκριση τιμών Αρθρωτότητας και μετρικής NMI του αλγορίθμου HGN, για τα 5 δίκτυα του Πίνακα 7.7 58

Α΄.1 Αλγόριθμος παραγωγής LFR δικτύων. 63
Α΄.2 Η μέθοδος Modularity Maximization εφαρμοσμένη σε τεχνητά σύνθετα δίκτυα [8]. 64
Α΄.3 Παραγωγή Συνδεδεμένου Γράφου 65
Α΄.4 Κώδικας Αλγορίθμου Walktrap 66
Α΄.5 Κώδικας Υπολογισμού Μετρικής NMI 67

Κατάλογος Πινάκων

7.1 Χαρακτηριστικά Τυχαίων Γεωμετρικών Γράφων 49
7.2 Χαρακτηριστικά Δικτύων Ελεύθερης Κλίμακας 49
7.3 Χαρακτηριστικά Δικτύων Μικρού Κόσμου 49
7.4 Χαρακτηριστικά Δικτύων Με Γνωστές Κοινότητες 49
7.5 Σύγκριση Χρόνου Εκτέλεσης Αλγορίθμων HGN και Walktrap για Δίκτυα με Άγνωστες Κοινότητες. 50
7.6 Σύγκριση Χρόνου Εκτέλεσης Αλγορίθμων HGN και Walktrap για Δίκτυα με Άγνωστες Κοινότητες. 52
7.7 Απόδοση Αλγορίθμων HGN και Walktrap ως προς τη μετρική NMI 57

Κατάλογος Αλγορίθμων

5.1 Αλγόριθμος Υπολογισμού HEBC [9] 38

Κεφάλαιο 1

Εισαγωγή

1.1 Περιγραφή του προβλήματος

Τα τελευταία χρόνια, ο ολοένα αυξανόμενος όγκος πληροφορίας πληροφορίας που διακινείται ηλεκτρονικά, η ραγδαία ανάπτυξη των δικτύων επικοινωνιών, του Διαδικτύου, καθώς και των Κοινωνικών Δικτύων, είναι μερικοί από τους λόγους που η μελέτη και η ανάλυση των δικτύων καθίσταται αναγκαία, για ένα πλήθος εφαρμογών, σε ένα ευρύ φάσμα τομέων. Οι σχέσεις που αναπτύσσονται μεταξύ οντοτήτων, σε δίκτυα που προκύπτουν ως αποτέλεσμα των παραπάνω και η μελέτη τους, εμπίπτουν στην επιστήμη της "Ανάλυσης Κοινωνικών Δικτύων", για την οποία υπάρχει σημαντικός όγκος βιβλιογραφίας. Ενδεικτικά αναφέρονται οι παραπομπές [10], [11], [12], [13]. Τα δίκτυα αυτά ορίζονται και ως Σύνθετα Δίκτυα (Complex Networks), εξαιτίας της πολυπλοκότητας των δεδομένων τους και των σχέσεων μεταξύ τους. Τέτοια δίκτυα που αναπαριστούν πολύπλοκα συστήματα του πραγματικού κόσμου, εμφανίζουν συνήθως δομή, η οποία δεν είναι τυχαία, αλλά χαρακτηρίζεται από ομάδες κόμβων, οι οποίοι είναι πυκνά συνδεδεμένοι μεταξύ τους, ενώ έχουν συγκριτικά λιγότερες συνδέσεις με τους υπόλοιπους κόμβους του δικτύου. Ταυτόχρονα, οι κόμβοι μιας ομάδας εμφανίζουν σημαντικές θεματικές ή/και λειτουργικές ομοιότητες. Οι ομάδες αυτές, συχνά αναφέρονται ως Κοινότητες.

Η ανίχνευση κοινοτήτων αποτελεί ένα σημαντικό ζήτημα της σύγχρονης επιστήμης δικτύων, αφού τα περισσότερα πραγματικά συστήματα παρουσιάζουν κοινοτική δομή, δηλαδή οι κορυφές των γράφων που τα αναπαριστούν μπορούν να ομαδοποιηθούν (clusters), ώστε το πλήθος των ακμών εντός της ομάδας να είναι μεγαλύτερο από το πλήθος των ακμών που ενώνουν κορυφές διαφορετικών ομάδων. Η ανίχνευση κοινοτήτων είναι ιδιαίτερα σημαντική για διαφορετικούς τομείς της επιστήμης, όπως η βιολογία, η επιστήμη των υπολογιστών και οι κοινωνικές επιστήμες. Τον τελευταίο καιρό υπάρχει αυξημένο ενδιαφέρον για τον εντοπισμό κοινοτήτων σε κοινωνικά δίκτυα, όχι μόνο από ερευνητικής σκοπιάς, αλλά και ως μέσο για την αξιοποίηση των αποτελεσμάτων σε ένα ευρύ φάσμα ευφυών υπηρεσιών και εφαρμογών. Χαρακτηριστικά παραδείγματα τέτοιων εφαρμογών αποτελούν, τα συστήματα συστάσεων (recommender systems), η εύρεση χρηστών που έχουν παρόμοια ενδιαφέροντα και η στοχευμένη προώθηση αγαθών (marketing).

1.1.1 Προσεγγίσεις

Η επεξεργασία και η διαχείριση πολύ μεγάλων γράφων (Big Data) είναι μια δύσκολη και πολύπλοκη υπολογιστικά, διαδικασία, η οποία απαιτεί τεχνικές ανάλυσης ικανές να διαχειριστούν το μεγάλο όγκο δεδομένων αλλά και να παράγουν αποτελέσματα σε αποδεκτό χρονικό διάστημα. Στην προσπάθεια για γρήγορη ανάλυση, μελετήθηκαν αλγόριθμοι όπως ο HGN [14], [9], όπου πραγματοποιείται απονομή συντεταγμένων σε κάθε παρατήρηση ή κόμβο σε έναν Υπερβολικό Χώρο συντεταγμένων.

Αρκετοί αλγόριθμοι έχουν προταθεί τα τελευταία χρόνια, για την προσέγγιση του προβλήματος της ανίχνευσης κοινοτήτων. Πολλές από τις κλασικές μεθόδους ανίχνευσης - ή αλλιώς μέθοδοι συσταδοποίησης (clustering methods) απαιτούν τον ορισμό συγκεκριμένων ιδιοτήτων των κοινοτήτων, πριν την εκτέλεση τους, όπως για παράδειγμα το πλήθος των κοινοτήτων [15], την πυκνότητα [16] κ.ά.. Σε άλλες μεθόδους υπάρχει η δυνατότητα εξαγωγής κοινοτήτων από ένα γράφο χωρίς προηγούμενη γνώση ή πληροφορία, καθώς αντλούν τις πληροφορίες αυτές από την ίδια την τοπολογία [17], [18]. Επίσης, αρκετές προσεγγίσεις αλγορίθμων λαμβάνουν υπόψη τους συγκεκριμένα χαρακτηριστικά των κοινοτήτων που συναντώνται στον πραγματικό κόσμο, όπως είναι οι επικαλυπτόμενες (overlapping) [19] και οι ιεραρχικές (hierarchical) κοινότητες [20], [21].

Πολλοί αλγόριθμοι δίνουν ικανοποιητικά αποτελέσματα, όταν δοκιμάζονται υπό ορισμένες συνθήκες και παραμέτρους. Ωστόσο, εξετάζοντας πιο λεπτομερώς, οι ίδιοι αλγόριθμοι μπορεί να παρουσιάζουν σημαντικές αδυναμίες, αποτυγχάνοντας σε κάποιο όριο. Για παράδειγμα, μια δημοφιλής μέθοδος ανίχνευσης κοινοτήτων, η βελτιστοποίηση τμηματικότητας (modularity optimization) [22], είναι πιθανόν να παρουσιάσει προβλήματα στην ανάλυση μεγάλων γράφων, αδυνατώντας να ανιχνεύσει τις μικρές κοινότητες, λόγω του ορίου ανάλυσης (resolution limit). Παράλληλα όμως, μπορεί κανείς να αντλήσει χρήσιμες ενδείξεις από αυτές τις αδυναμίες: για παράδειγμα, θα μπορούσε κανείς να εξαγάγει τους πυρήνες των πραγματικών κοινοτήτων, ανεξαρτήτως του γεγονότος ότι η μέθοδος δεν έχει καταφέρει να ομαδοποιήσει όλους τους κόμβους σε σωστές κοινότητες [23]. Επίσης είναι σημαντικό, οι αλγόριθμοι να μπορούν να αποδίδουν τις πραγματικά συνεκτικές δομές του γράφου ως κοινότητες, ανιχνεύοντας παράλληλα κόμβους που δεν ανήκουν σε κάποια κοινότητα (outliers), λόγω ασθενών συνδέσεων, καθώς και κόμβους που λειτουργούν ως ¨γέφυρες¨ μεταξύ διαφορετικών κοινοτήτων (hubs).

Το πρόβλημα της ανίχνευσης κοινοτήτων, καθώς και η γενικότερη ερευνητική περιοχή της ανάλυσης δεδομένων γράφων, καλείται να απαντήσει σε ανοικτές προκλήσεις, όπως οι παρακάτω (ενδεικτικά):

- Είναι σε θέση οι υπάρχουσες μέθοδοι ανίχνευσης κοινοτήτων, να χειριστούν πολύ μεγάλους γράφους εκατομμυρίων και δισεκατομμυρίων κόμβων και ακμών;
- Είναι οι μέθοδοι παραλληλοποιήσιμοι[1];

[1] Ο όρος παραλληλοποίηση περιγράφει τη δυνατότητα ταυτόχρονης εκτέλεσης δύο -ή και περισσότερων διεργασιών μέσω πολυνηματικής ή κατανεμημένης επεξεργασίας, με σκοπό τη μείωση κόστους της υπολογιστικής πολυπλοκότητας.

1.2 Αντικείμενο της διπλωματικής

Στα πλαίσια της παρούσας εργασίας, μελετήθηκαν δύο βασικοί αλγόριθμοι ανίχνευσης κοινοτήτων, οι οποίοι ανήκουν σε διαφορετικές κατηγορίες ο καθένας, με βάση τις μεθοδολογικές τους αρχές. Η πρώτη κατηγορία χρησιμοποιεί μια από πάνω-προς-τα-κάτω προσέγγιση, για την ομαδοποίηση των κόμβων (Hierarchical Divisive Clustering). Σε αυτήν εμπίπτει και ο αλγόριθμος Hyperbolic Girvan-Newman ([14], [9], [13]), μια παραλλαγή του κλασικού Girvan-Newman [7], ώστε να χρησιμοποιεί την Υπερβολική Κεντρικότητα Ενδιαμεσικότητας Ακμής. Η επιλογή του Υπερβολικού Χώρου στον συγκεκριμένο αλγόριθμο, γίνεται γιατί από μελέτες άλλων επιστημόνων προκύπτει ότι η υπερβολική γεωμετρία είναι εκείνη που ταιριάζει καλύτερα σε ιεραρχικές δομές, που εμφανίζονται και σε σχεσιακά μοντέλα Συνθέτων Δικτύων, όπως τα Δίκτυα Ελεύθερης Κλίμακας (Scale-Free Networks) [24]. Η μεθοδολογία αυτή δίνει αρκετά ικανοποιητικά αποτελέσματα και σε πολλές περιπτώσεις είναι ταχύτερη από τον αλγόριθμο Girvan-Newman.

Η δεύτερη κατηγορία κάνει χρήση παραδοσιακών μεθόδων συσταδοποίησης, όπως τα μέτρα ομοιότητας (similarity measures). Σε αυτήν εμπίπτει ο αλγόριθμος Walktrap των *Pons* και *Latapy* [25], όπου και επιλέγεται να αναλυθεί στα πλαίσια της εργασίας. Ο εν λόγω αλγόριθμος χρησιμοποιεί τυχαίους περιπάτους για την εξερεύνηση του γράφου, λαμβάνοντας υπόψη τις ομοιότητες μεταξύ των κόμβων, κατά την εξερεύνηση.

Τέλος, στα πλαίσια της διπλωματικής εργασίας επιχειρήθηκε μια σύγκριση και πειραματική αξιολόγηση μεταξύ των προαναφερθέντων αλγορίθμων. Για το σκοπό αυτό, χρησιμοποιήθηκαν πραγματικά Κοινωνικά Δίκτυα μικρής κλίμακας, τεχνητά Σύνθετα Δίκτυα μεγαλύτερου μεγέθους, για τα οποία χρησιμοποιήθηκαν μερικά από τα γνωστότερα μοντέλα κατασκευής τους, αλλά και Δίκτυα Ελεύθερης Κλίμακας (scale-free networks).

1.3 Δομή της Διπλωματικής Εργασίας

Η Διπλωματική Εργασία αποτελείται από οχτώ Κεφάλαια. Στο Κεφάλαιο 2 έως 4 γίνεται η απαιτούμενη εισαγωγή σε έννοιες που χρησιμοποιούνται και παρουσιάζεται το θεωρητικό υπόβαθρο της εργασίας. Στα Κεφάλαια 5, 6 παρουσιάζονται οι δύο προτεινόμενες μέθοδοι, ενώ στο 7ο Κεφάλαιο παρουσιάζονται τα πειραματικά αποτέλεσμα για διάφορους τύπους δικτύων.

Πιο συγκεκριμένα, η διάρθρωση της εργασίας είναι η ακόλουθη:

- Στο Κεφάλαιο 2 γίνεται μια θεωρητική επισκόπηση στις βασικές έννοιες της Θεωρίας Γραφημάτων, οι οποίες είναι απαραίτητες για την καλύτερη κατανόηση της εργασίας. Επίσης παρουσιάζεται μερική ορολογία και συμβολισμοί που θα ακολουθούνται στη συνέχεια της εργασίας. Ακόμα, εστιάζουμε στην έννοια της ανίχνευσης κοινοτήτων, ορίζοντας την κοινότητα σε ένα γράφο και τα βασικά της στοιχεία.
- Στο Κεφάλαιο 3 παρουσιάζονται τα μοντέλα κατασκευής Σύνθετων Δικτύων που χρησιμοποιούνται στην εργασία και αναλύονται βασικές έννοιες και μετρικές τους, όπως η Κεντρικότητα Ενδιαμεσικότητας Ακμής που θα μας απασχολήσει στο Κεφάλαιο 5.

- Στο Κεφάλαιο 4 παρουσιάζεται η έννοια της Αρθρωτότητας μιας Ομαδοποίησης, καθώς και ο αλγόριθμος Ομαδοποίησης που εξασφαλίζει τη μέγιστη Αρθρωτότητα.
- Το Κεφάλαιο 5 περιλαμβάνει αρχικά την παρουσίαση του σημαντικού αλγόριθμου ομαδοποίησης Girvan-Newman και έπειτα την ανάλυση της μεθόδου Ενσωμάτωσης ενός γράφου στον Υπερβολικό Χώρο, που χρησιμοποιείται στην παρούσα εργασία. Τέλος παρουσιάζεται ο τροποποιημένος αλγόριθμος Girvan-Newman.
- Το Κεφάλαιο 6 περιλαμβάνει την παρουσίαση του αλγορίθμου Walktrap. Γίνεται αναφορά επίσης, σε βασικές αρχές και ιδιότητες των τυχαίων περιπάτων σε γράφους.
- Στο Κεφάλαιο 7 παρουσιάζεται η πειραματική αξιολόγηση των εξεταζόμενων αλγορίθμων με μετρήσεις της ακρίβειας και της αποδοτικότητας, ως προς τον χρόνο εκτέλεσης και το μέγεθος των δεδομένων.
- Τέλος, στο Κεφάλαιο 8 συνοψίζονται τα σημαντικότερα αποτελέσματα της εργασίας και προτείνονται ενδιαφέροντα θέματα για περαιτέρω μελλοντική διερεύνηση.

Κεφάλαιο 2

Βασικά Στοιχεία Θεωρίας Γραφημάτων

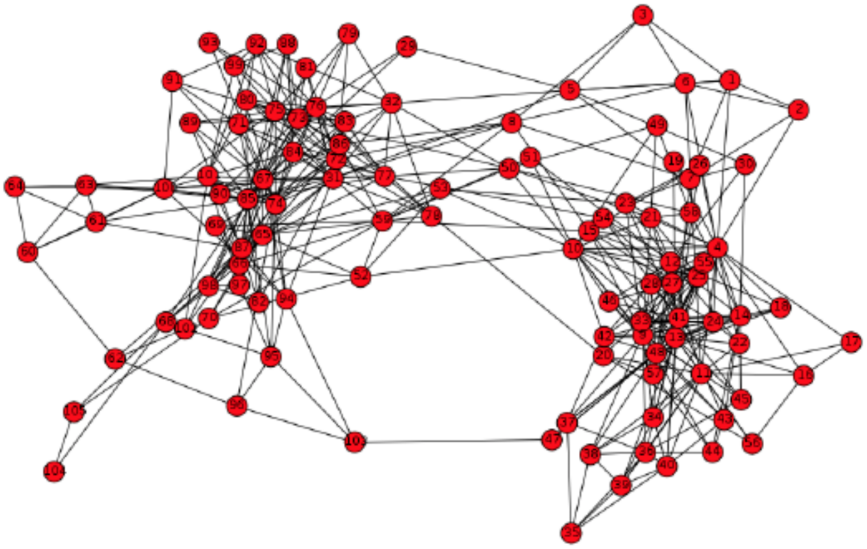
Η Θεωρία Γραφημάτων είναι ένα πολύτιμο εργαλείο τη σημερινή εποχή, που η ανάπτυξη σύνθετων συστημάτων όπως κοινωνικών, βιολογικών, ηλεκτρικών δικτύων, δικτύων υπολογιστών είναι ραγδαία. Συνεπώς, καθίσταται αναγκαία η αναπαράσταση τους για τη μελέτη και την ανάλυση των σχέσεων που αναπτύσσονται μεταξύ οντοτήτων σε δίκτυα. Η Θεωρία Γραφημάτων αποτελεί το κατάλληλο εργαλείο για το σκοπό αυτό. Στο κεφάλαιο αυτό θα δοθούν διευκρινήσεις και ορισμοί της θεωρίας γράφων, οι οποίοι χρησιμοποιούνται εκτενώς στο πλαίσιο της παρούσας εργασίας και θα αποτελέσουν τη βάση για την κατασκευή της έννοιας της Κοινότητας.

2.1 Βασική ορολογία, ορισμοί και προκαταρκτικές γνώσεις.

Συχνά οι έννοιες γράφος ή γράφημα (graph) και δίκτυο συγχέονται και αναφέρονται στην βιβλιογραφία, χωρίς να διακρίνεται κάποια σημαντική διαφορά. Συνήθως, στα δίκτυα η σχέση μεταξύ των αντικειμένων (κόμβων) είναι πιο ισχυρά καθορισμένη, ενώ αντίθετα στους γράφους εντοπίζεται μία λιγότερο αυστηρή δομή. Παρακάτω παρουσιάζονται κάποιες από τις βασικές έννοιες της Θεωρίας Γράφων, που αντλήθηκαν από τα [26], [27], [28].

Γράφος ή γράφημα (graph) είναι μία δομή που αποτελείται από ένα σύνολο κορυφών (vertices) ή κόμβων (nodes) ή σημείων (points) που συνδέονται μεταξύ τους με ένα σύνολο ακμών (edges) ή γραμμών (lines). Ένας γράφος ορίζεται ως ένα ζεύγος G = (V, E), όπου V αντιπροσωπεύει το σύνολο των κόμβων και E το σύνολο των ακμών του γράφου G. Το πλήθος των κόμβων ενός γράφου συμβολίζεται με n = |V| και ονομάζεται τάξη (order) του γράφου. Το πλήθος των ακμών συμβολίζεται με m = |E| και ονομάζεται μέγεθος (size) του γράφου. Κάθε ακμή προσδιορίζεται από δύο κόμβους που ονομάζονται τερματικά σημεία (end points). Αν η ακμή e έχει τα u, v ως τερματικά σημεία τότε η e ονομάζεται προσπίπτουσα (incident) στα σημεία u, v, ή λέγεται ότι η e συνδέει (connects) τα u, v. Η ακμή e συμβολίζεται με (u, v) ή (v, u). Αν δύο κόμβοι δεν συνδέονται, ονομάζονται ως μη γειτονικοί, ή αλλιώς ανεξάρτητοι (independent).

Η πυκνότητα (density) ενός γράφου, ορίζεται από το σύνολο των ακμών του και των κόμβων του. Εάν δύο γράφοι έχουν την ίδια τάξη, αλλά διαφορετικό μέγεθος, τότε έχουν διαφορετική



Εικόνα 2.1: *Παράδειγμα γράφου, πρόκειται για το γράφο του δικτύου the polbooks network [1].*

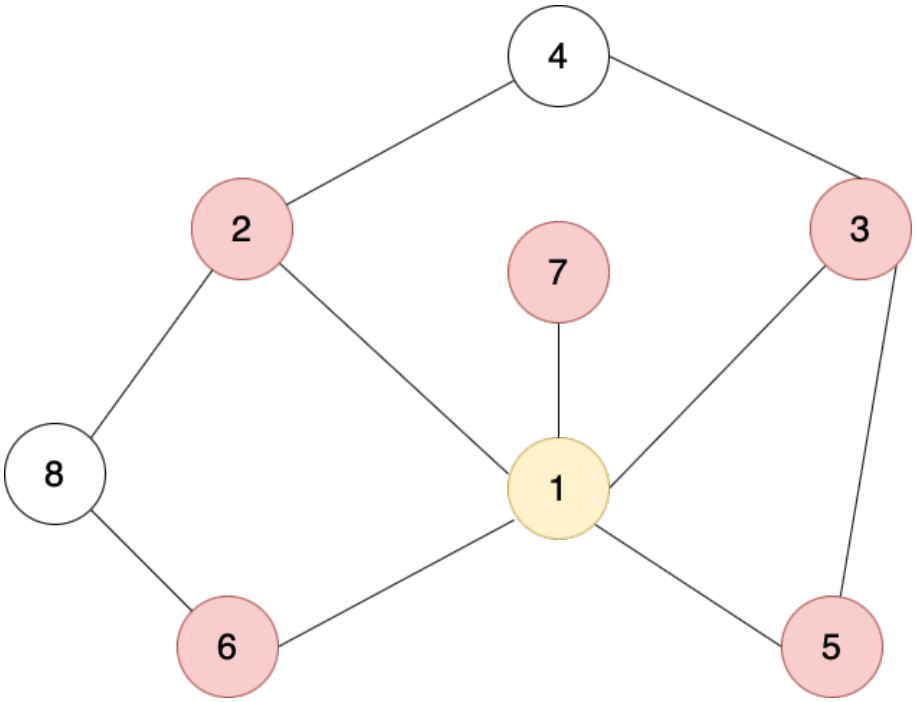
πυκνότητα (ο γράφος με το μεγαλύτερο μέγεθος λέμε ότι είναι πιο πυκνός από αυτόν με το μικρότερο μέγεθος).

Ένας γράφος ονομάζεται κατευθυνόμενος (directed), εάν περιλαμβάνει αποκλειστικά κατευθυνόμενες ακμές. Οι ακμές ενός γράφου ονομάζονται κατευθυνόμενες, όταν τα ζεύγη των ακμών (u, v) και (v, u) λαμβάνονται ως διατεταγμένα, δηλαδή $(u, v) \neq (v, u)$. Στην περίπτωση που οι ακμές δεν θεωρούνται διατεταγμένες, τότε ονομάζονται μη διατεταγμένες ή απλές. Αντίστοιχα, ονομάζεται μη κατευθυνόμενος (undirected) εάν περιλαμβάνει μόνο απλές ακμές.

Δεδομένης μιας ακμής (u, v), ο κόμβος u ονομάζεται γείτονας (ή γειτονικός) του κόμβου v, και αντίστροφα. Η γειτονιά (neighborhood) ενός κόμβου u (Εικόνα 2.2), συμβολίζεται με N(u) και είναι το σύνολο των κόμβων που ορίζεται από την σχέση: $N(u) = [u \in V(G) | (u, v) \in E(G)]$. Το πλήθος των γειτόνων-ακμών που προσπίπτουν στο κόμβο u, ονομάζεται βαθμός (degree) του κόμβου u και συμβολίζεται με d(u). Αν για κάποιον κόμβο ισχύει $d(u) = 0$, ο κόμβος ονομάζεται απομονωμένος (isolated) και αντίστοιχα, αν ισχύει $d(u) = 1$, ο κόμβος ονομάζεται εκκρεμής (pendant). Ένας γράφος, για τον οποίο κάθε κόμβος του έχει βαθμό d ονομάζεται d-κανονικός.

Δύο κόμβοι u και v λέγονται συνδεδεμένοι στον γράφο G, εάν υπάρχει ένα μονοπάτι μεταξύ αυτών των δύο κόμβων. Ένας γράφος G λέγεται συνδεδεμένος ή συνεκτικός, εάν όλα τα ζεύγη κόμβων του είναι συνδεδεμένα. Σε διαφορετική περίπτωση, ο γράφος δεν είναι συνδεδεμένος και χωρίζεται σε συνεκτικές συνιστώσες (connected components), οι οποίες αποτελούνται από συνδεδεμένους υπογράφους του G. Επίσης ένας γράφος $H' = (V', E')$, όπου $V' \subseteq V$ και $E' \subseteq E$, ονομάζεται υπογράφος (subgraph) του G. Ένας υπογράφος του οποίου όλοι οι κόμβοι συνδέονται μεταξύ τους, ονομάζεται κλίκα (clique).

Ένας γράφος G, ονομάζεται σταθμισμένος (weighted), όταν ένας πραγματικός αριθμός w, που ονομάζεται βάρος, συνδέεται με κάθε ένα από τα άκρα του. Επίσης ένας γράφος $G = (V1, V2, E)$, καλείται διμερής (bipartite) εάν η ομάδα κόμβων V, χωρίζεται σε δύο ξεχωριστές



Εικόνα 2.2: *Παράδειγμα αναπαράστασης γειτονιάς. Η γειτονιά του κόμβου 1 είναι το σύνολο κόμβων που αποτελείται από τους 2,3,5,6 και 7.*[1]

(disjoint) υποομάδες V1, V2 και κάθε άκρο συνδέει μία κορυφή του V1 με μια κορυφή του V2, δηλαδή δεν υπάρχουν ακμές μεταξύ κόμβων της ίδιας υποομάδας.

Η πληροφορία αναφορικά με την τοπολογία ενός γράφου $G = (V, E)$ εμπεριέχεται στον Πίνακα Γειτνίασης (Adjacency Matrix) A, ο οποίος είναι ένας n × n (|V| × |V|) πίνακας που ορίζεται ως εξής:

$$A = (a_{ij})_{nxn} = \begin{cases} 1 & , \text{εάν } (i,j) \in E, \forall i,j \in 1,....,n \\ 0 & , \text{διαφορετικά} \end{cases} \tag{2.1}$$

Τα διαγώνια στοιχεία του πίνακα αυτού είναι μηδενικά. Για έναν μη-κατευθυνόμενο γράφο (undirected graph), ο A είναι επίσης συμμετρικός ($A = A^T$). Εάν οι ακμές είναι σταθμισμένες (weighted), ορίζεται αντίστοιχα ο Πίνακας Βαρών (Weight Matrix) του οποίου το στοιχείο w_{ij} εκφράζει το βάρος της ακμής μεταξύ των κορυφών i και j.

2.1.1 Περίπατοι-Μονοπάτια

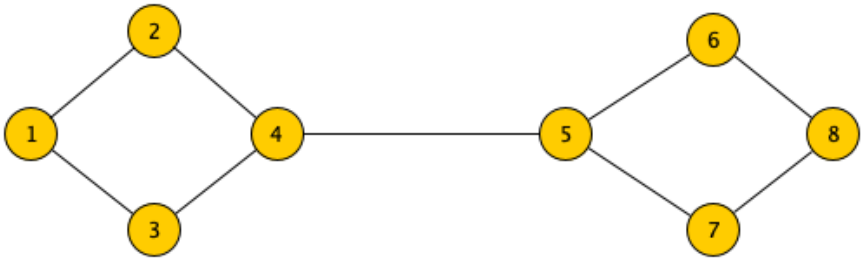
Ως περίπατος σε ένα γράφημα G, ορίζεται μία ακολουθία εναλλασσόμενων ακμών και κορυφών $[u1, u2....u_k]$, τέτοια ώστε $(u_i, u_{i+1}) \in E$ για κάθε $i = 1, ..., k-1$.

Ένα μονοπάτι είναι στην ουσία ένας περίπατος, χωρίς επαναλαμβανόμενες κορυφές. Αν η αρχική και η τελική κορυφή ταυτίζονται, τότε λέμε ότι έχουμε ένα κύκλο στο G. Αν το μονοπάτι P, που συνδέει δύο κορυφές u και v σε ένα γράφο G, έχει το μικρότερο μήκος

[1] Για το σχεδιασμό γραφημάτων τόσο σε αυτό το Κεφάλαιο όσο και στα υπόλοιπα χρησιμοποιήθηκε το πακέτο λογισμικού yEd graph editor [29].

από οποιοδήποτε άλλο μονοπάτι που μπορεί να τις συνδέει, τότε ορίζεται ως ελάχιστο ή συντομότερο μονοπάτι (shortest path) μεταξύ των δύο κορυφών.

Γέφυρα (bridge) ονομάζεται η ακμή, που η αφαίρεση της από ένα συνδεδεμένο γράφο, προκαλεί την διαμέριση του σε περισσότερες από μία συνεκτικές συνιστώσες, Εικόνα 2.3.



Εικόνα 2.3: *Στον παραπάνω γράφο η ακμή (4,5) αποτελεί γέφυρα, η αφαίρεση της οποίας δημιουργεί δύο συνεκτικές συνιστώσες.*

2.2 Ανίχνευση Κοινοτήτων σε Γράφους

Συχνά, οι κόμβοι σε ένα γράφο οργανώνονται σε ομάδες, οι οποίες φαίνεται να υπάρχουν μερικώς ανεξάρτητα από το υπόλοιπο κομμάτι του γράφου, με το οποίο μοιράζονται μόνο λίγες ακμές. Η σχέση μεταξύ των μελών της ομάδας είναι ισχυρότερη, όπως φαίνεται από το μεγάλο αριθμό των αμοιβαίων συνδέσεων. Τέτοιες ομάδες κόμβων, μπορούν να θεωρηθούν ως ανεξάρτητα συστατικά του γράφου.

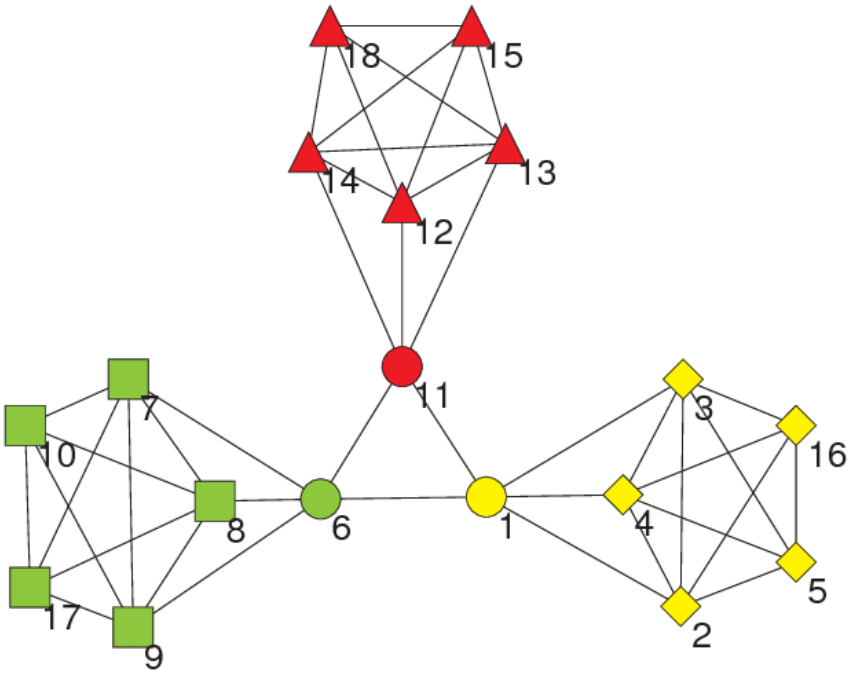
2.2.1 Ορισμός Κοινότητας

Μία κοινότητα σε έναν γράφο είναι μία ομάδα κόμβων που παρουσιάζουν μεγάλη ομοιότητα μεταξύ τους, σύμφωνα με πολύ καλά ορισμένα και μετρήσιμα κριτήρια. Ανά τη βιβλιογραφία, έχουν προταθεί πολλοί ορισμοί. Μια κοινή προσέγγιση, είναι ο ορισμός της κοινότητας ως ομάδας κόμβων, με υψηλότερη πυκνότητα ακμών μεταξύ κόμβων εντός της ομάδας, από τη μέση πυκνότητα ακμών στο γράφο. Γενικότερα, οι ορισμοί της κοινότητας μπορούν να κατηγοριοποιηθούν στις εξής κατηγορίες [30]: τους τοπικούς ορισμούς (local), τους καθολικούς ορισμούς (global) και τους ορισμούς που βασίζονται στην ομοιότητα κόμβων.

Τοπικοί Ορισμοί: Οι τοπικοί ορισμοί επικεντρώνονται σε ένα υπογράφο συμπεριλαμβανομένης, πιθανώς, και της άμεσης γειτονιάς του, τον οποίον και μελετούν, αγνοώντας το υπόλοιπο δίκτυο. Η λογική πίσω από αυτό είναι ότι οι κοινότητες έχουν λίγους δεσμούς με το υπόλοιπο δίκτυο, άρα, σε κάποιο βαθμό, μπορούν να θεωρηθούν ως ξεχωριστές οντότητες με δική τους αυτονομία. Οι ορισμοί αυτοί αφορούν κατηγορίες υπογράφων, όπως κλίκες (cliques), n-κλίκες (n-cliques), k-πλέγματα (k-plexes), κλπ.. Η έννοια της κλίκας είναι πολύ σημαντική. Μια κλίκα είναι ένας μέγιστος υπογράφος όπου κάθε κόμβος είναι γειτονικός με όλους τους άλλους κόμβους του υπογράφου. Ένα παράδειγμα κλίκας παρουσιάζεται στην Εικόνα 2.4. Ένα k-πλέγμα είναι ένας μέγιστος υπογράφος, έτσι ώστε κάθε κόμβος να είναι γειτονικός με όλους τους άλλους, εκτός από το πολύ k από αυτούς. Σε αντίθεση, ένας

k-πυρήνας (k-core) είναι ένας μέγιστος υπογράφος όπου κάθε κόμβος είναι γειτονικός με τουλάχιστον k κορυφές εντός του υπογράφου.

Καθολικοί Ορισμοί: Οι κοινότητες αποτελούνται από ένα σύνολο κόμβων και ακμών επομένως έχει νόημα να συγκρίνουμε την τοπολογία της κοινότητας με αυτήν του συνολικού γράφου. Σε αυτή τη κατηγορία ορισμού της κοινότητας συχνά χρησιμοποιείται το μηδενικό μοντέλο (null model) [31] το οποίο είναι ένα αντίγραφο του γράφου με το ίδιο πλήθος κόμβων και ακμών αλλά διαφορετική κατανομή ακμών. Στο μηδενικό μοντέλο δεν υπάρχουν κοινοτικές δομές και ο απλούστερος τρόπος δημιουργίας του είναι να κατανεμηθούν οι ακμές τυχαία ανάμεσα στους κόμβους. Το μηδενικό μοντέλο χρησιμοποιείται για την σύγκριση της δομής των κοινοτήτων του αρχικού γράφου με το αντίστοιχο σύνολο κόμβων και ακμών του μηδενικού μοντέλου. Έτσι μπορεί να εκτιμηθεί πόσο διαφορετική είναι η δομή της κάθε κοινότητας σε σχέση με μια δομή γράφου όπου δεν υπάρχουν κοινότητες, όπως για παράδειγμα να παρατηρηθεί το πλήθος ακμών που υπάρχουν εσωτερικά της κοινότητας και το πλήθος ακμών μεταξύ των κοινοτήτων. Επίσης, υπάρχει μια κλάση ορισμών που βασίζονται στην ιδέα ότι ένας γράφος έχει κοινοτική δομή, εάν διαφέρει σημαντικά από έναν τυχαίο γράφο (κατά Erdős–Rényi- οποιεσδήποτε δύο κορυφές του γράφου έχουν την ίδια πιθανότητα να είναι συνδεδεμένες).



Εικόνα 2.4: *Ο υπογράφος με τους κόμβους 1,2,3,4,5,16 είναι μία κλίκα [2].*

Ορισμοί βασιζόμενοι στην ομοιότητα κόμβων: Οι εν λόγω ορισμοί βασίζονται στην υπόθεση ότι οι κοινότητες είναι ομάδες κόμβων παρόμοιων μεταξύ τους. Η ομοιότητα μεταξύ κάθε ζεύγους κόμβων μπορεί να υπολογιστεί με βάση κάποια ιδιότητα, τοπική ή καθολική, ανεξάρτητα από το αν υπάρχει ακμή που συνδέει τους κόμβους ή όχι. Κάθε κόμβος καταλήγει στην ομάδα με τους πιο όμοιους με αυτόν κόμβους. Η χρήση μέτρων ομοιότητας για

τη συσταδοποίηση αποτελεί τη βάση των παραδοσιακών μεθόδων, όπως είναι η ιεραρχική, η μερική και η φασματική συσταδοποίηση.

2.2.2 Τύποι Κοινοτήτων

Ο εντοπισμός κοινοτήτων σε δίκτυα είναι μια διαδικασία που περιέχει μια εγγενή ασάφεια, η οποία κάθε φορά αποκρυσταλλώνεται από διαφορετικούς παράγοντες. Τέτοιοι παράγοντες μπορεί να είναι διαδικαστικοί (εξαρτώνται από την αλγοριθμική προσέγγιση που ο ενασχολούμενος προτίθεται να ακολουθήσει, ανάλογα με τα δεδομένα του) ή/και δομικοί (εξαρτώνται από τη μορφολογία του δικτύου). Ακολουθεί μία βασική κατηγοριοποίηση τύπων κοινοτήτων, όπως αυτή προκύπτει από δίκτυα πραγματικού χρόνου:

- **Αλληλοεπικαλυπτόμενες Κοινότητες (Overlapping Communities):** Σε πολλά δίκτυα πραγματικού χρόνου, οι κοινότητες μπορεί να έχουν κοινό έναν η περισσότερους κόμβους. Στα κοινωνικά δίκτυα για παράδειγμα, ένα -ή και περισσότερα- άτομα μπορεί να ανήκει σε διαφορετικές κοινότητες όπως της δουλειάς του, των φίλων του και της οικογένειάς του, οι οποίες τον έχουν ως κοινό μέλος.
- **Κατευθυνόμενες Κοινότητες (Directed Communities):** Σε αντιστοιχία με τους τύπους κατευθυνόμενων γράφων, πολλά φαινόμενα πραγματικού χρόνου αναπαρίστανται από μεταξύ τους συνδέσμους, οι οποίοι δεν είναι «αμοιβαίοι», όπως για παράδειγμα ένα hyperlink στον ιστό που οδηγεί το χρήστη από μια ιστοσελίδα σε άλλη, χωρίς να υπάρχει αντίστοιχο για να τον επιστρέφει στην προηγούμενη.
- **Σταθμισμένες Κοινότητες (Weighted Communities):** Σε αντιστοιχία με την κατηγοριοποίηση γράφων, μια ομάδα συνδεδεμένων κορυφών μπορεί να θεωρηθεί κοινότητα μόνο εάν τα βάρη των συνδέσμων των κορυφών είναι αρκετά ισχυρά, δεδομένου ενός κατωφλιού.
- **Δυναμικές Κοινότητες (Dynamic Communities):** Η συγκεκριμένη κατηγορία κοινοτήτων αφορά ομάδες συνδέσμων που εμφανίζονται και εξαφανίζονται, επομένως και οι αντίστοιχες κοινότητες εξελίσσονται και αναδιαμορφώνονται με την πάροδο του χρόνου.

Κεφάλαιο 3

Τεχνητά Σύνθετα Δίκτυα

Τα τελευταία χρόνια, ο ολοένα αυξανόμενος όγκος πληροφορίας που διακινείται ηλεκτρονικά, σε συνδυασμό με την αυξημένη δημοτικότητα των κοινωνικών δικτύων, καθώς και με την ταχεία ανάπτυξη των δικτύων επικοινωνιών και υπολογιστών γενικότερα, οδήγησε στην ανάγκη για περαιτέρω μελέτη των δικτύων αυτών. Τα δίκτυα αυτά συνοψίζονται με τον όρο Σύνθετα Δίκτυα (Complex Networks). Τα Σύνθετα Δίκτυα παρουσιάζουν συγκεκριμένες ιδιότητες, όπως ο σχηματισμός κοινοτήτων, ο υψηλός συντελεστής ομαδοποίησης και άλλες.

Διάφορα μοντέλα έχουν προταθεί για την μελέτη και την προσομοίωση των παραπάνω δικτύων. Τα μοντέλα αυτά μπορούν να χωριστούν στις εξής δύο κατηγορίες: χωρικά (spatial) και σχεσιακά μοντέλα (relational). Χωρικά λέγονται τα δίκτυα, των οποίων οι κόμβοι συνδέονται μεταξύ τους ανάλογα με τη θέση τους σε κάποιο γεωμετρικό χώρο. Παράδειγμα τέτοιων δικτύων αποτελούν οι Τυχαίοι Γεωμετρικοί Γράφοι (Random Geometric Graph - RGG). Σχεσιακά λέγονται τα δίκτυα, των οποίων οι κόμβοι του δικτύου συνδέονται ανάλογα με τις τοπολογικές ιδιότητες που μπορεί να παρουσιάζουν στο δίκτυο, όπως βάσει του βαθμού κόμβου. Παράδειγμα τέτοιων μοντέλων αποτελούν τα Δίκτυα Ελεύθερης Κλίμακας (Scale-Free networks), καθώς και τα Δίκτυα Μικρού Κόσμου (Small-World networks).

Σε αυτή την ενότητα, θα γίνει μία παρουσίαση των προαναφερθέντων μοντέλων, τα οποία θα χρησιμοποιηθούν αργότερα και στο κομμάτι της πειραματικής διαδικασίας (Κεφάλαιο 7), ως μέτρο σύγκρισης για τις μεθοδολογίες που αναλύονται στην παρούσα εργασία. Ακόμα, θα παρουσιαστούν ορισμένες μετρικές που είναι απαραίτητες για τη μελέτη των Σύνθετων Δικτύων, όπως η Κεντρικότητα Ενδιαμεσικότητας Ακμής και ο Συντελεστής Ομαδοποίησης.

3.1 Μοντέλα Κατασκευής

Σε αυτή την ενότητα, θα εξεταστούν μοντέλα κατασκευής τεχνητών Σύνθετων Δικτύων. Συγκεκριμένα, θα παρουσιαστούν το μοντέλο των Τυχαίων Γράφων των Gilbert [3] και Erdos-Renyi [32], οι Τυχαίοι Γεωμετρικοί Γράφοι [33], τα Δίκτυα Ελεύθερης Κλίμακας που παράγονται σύμφωνα με το μοντέλο των Barabasi και Albert [24] και τέλος τα Δίκτυα Μικρού Κόσμου με την μεθοδολογία κατασκευής των Watts-Strogatz [34].

3.1.1 Τυχαίοι Γράφοι

Όταν οι κόμβοι ενός γράφου συνδέονται τυχαία μεταξύ τους, τότε παράγεται ένας τυχαίος γράφος. Από τα πιο γνωστά μοντέλα κατασκευής τους είναι τα μοντέλα του Gilbert και των Erdos-Renyi. Ακολουθούν συνοπτικές περιγραφές τους:

- Το μοντέλο του Gilbert
  Στο μοντέλο που πρότεινε ο Gilbert [3] για την κατασκευή ενός τυχαίου γράφου, ξεκινώντας από ένα σύνολο n απομονωμένων κόμβων, σταδιακά προστίθενται τυχαία, ακμές μεταξύ τους. Κάθε ακμή έχει πιθανότητα να εμφανιστεί, ίση με p, ανεξάρτητα από τις άλλες. Όσο η μεταβλητή p τείνει στη τιμή 1, τόσο ο γράφος τείνει στο να γίνει ένας πλήρης γράφος n κόμβων.

- Το μοντέλο των Erdos-Renyi
  Σε αυτό το μοντέλο [32], ένας τυχαίος γράφος G(n,M) επιλέγεται τυχαία, από όλους τους πιθανούς γράφους με n κορυφές και M ακμές, με ομοιόμορφη κατανομή πιθανότητας. Στην περίπτωση που $pn^2 \to \infty$ το μοντέλο του Gilbert γίνεται ισοδύναμο με αυτό των Erdos-Renyi, για $M = \binom{n}{2}p$.

Figure 2: For n=10, p=0

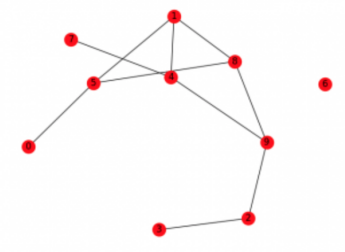


Figure 3: For n=10, p=0.25

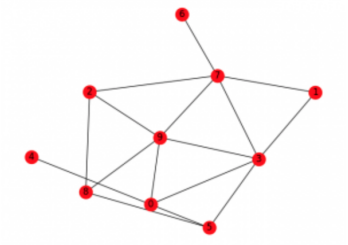


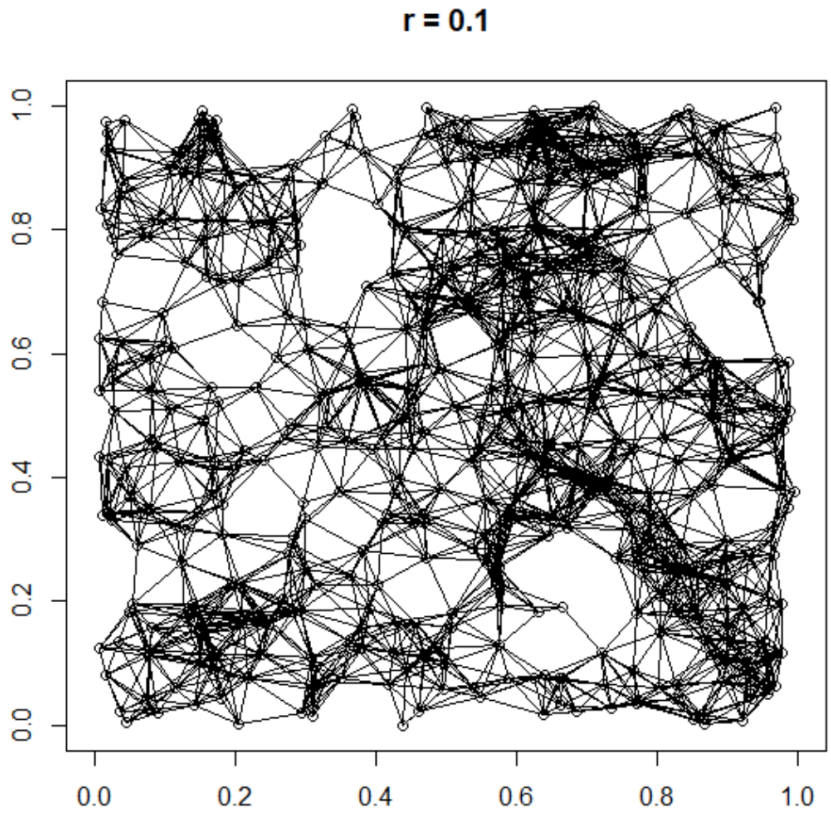
Figure 4: For n=10, p=0.5

Εικόνα 3.1: *Τυχαίος Γράφος βάσει του μοντέλου Gilbert, για τρεις διαφορετικές τιμές του p [3].*

3.1.2 Τυχαίοι Γεωμετρικοί Γράφοι

Ένας Τυχαίος Γεωμετρικός Γράφος [35], [36] (RGG) G(N,r) είναι ένας γράφος N κόμβων, των οποίων οι συντεταγμένες βρίσκονται σε ένα Γεωμετρικό Χώρο και οι κόμβοι συνδέονται μεταξύ τους, αν η απόσταση τους δεν υπερβαίνει τη τιμή της μεταβλητής r. Το μοντέλο του Τυχαίου Γεωμετρικού Γράφου αποτελεί ένα χωρικό μοντέλο κατασκευής γράφων και

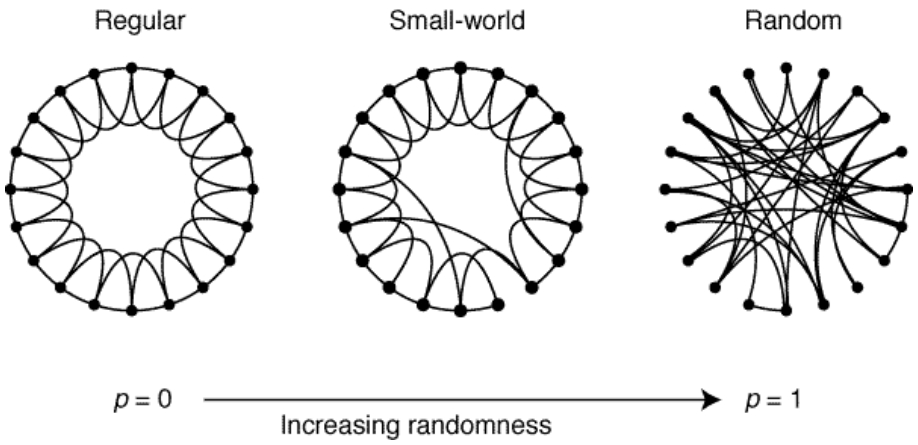
οι συνδέσεις μεταξύ των κόμβων εξαρτώνται από τη θέση τους στο χώρο. Παράδειγμα ένος τέτοιου γράφου δίνεται στην Εικόνα 3.2.



Εικόνα 3.2: *Παράδειγμα ενός Τυχαίου Γεωμετρικού Γράφου [4].*

3.1.3 Δίκτυα Μικρού Κόσμου

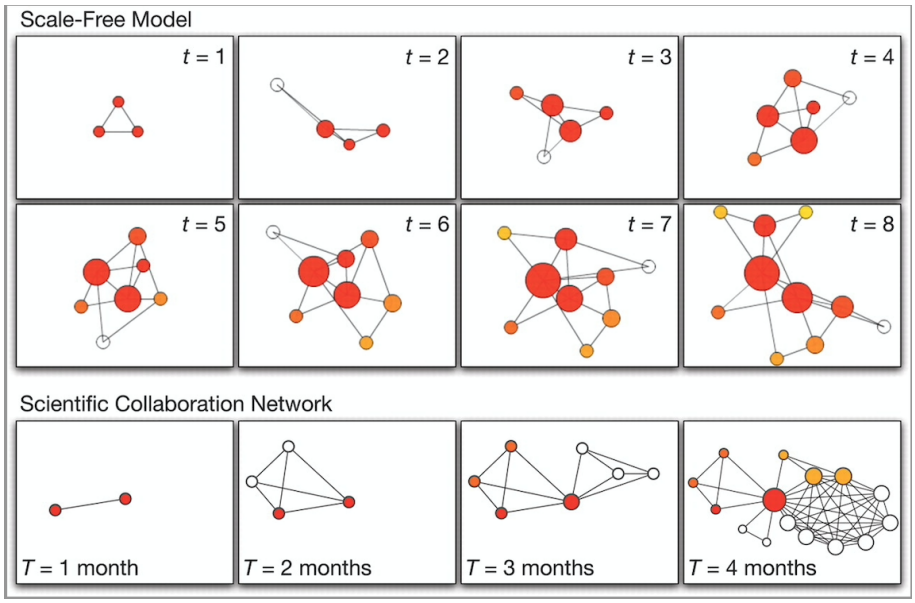
Στα Δίκτυα Μικρού Κόσμου (small-world networks), αν και οι περισσότεροι κόμβοι δεν είναι γείτονες μεταξύ τους, καθένας από αυτούς συνδέεται με τους υπόλοιπους (μη γειτονικούς) μέσω ενός μικρού μήκους μονοπατιού. Στις ιδιότητες τους λοιπόν, πέρα από το μικρό μέσο μήκος μονοπατιού, συγκαταλέγονται ο σχετικά υψηλός συντελεστής ομαδοποίησης, η σχετικά ομοιόμορφη τοπολογία (οι περισσότεροι κόμβοι παρουσιάζουν τον ίδιο αριθμό ακμών), καθώς και ο υψηλός αριθμός τριγώνων που σχηματίζονται εντός του δικτύου.
Στη παρούσα εργασία, το μοντέλο κατασκευής τεχνητών Δικτύων Μικρού Κόσμου που χρησιμοποιείται, είναι το μοντέλο των Watts-Strogatz [34]. Στο συγκεκριμένο, ξεκινώντας από ένα διατεταγμένο πλέγμα N κόμβων με πιθανότητα p, κόμβοι που δεν ήταν γειτονικοί επανασυνδέονται, ανεξάρτητα από την πρότερη μεταξύ τους απόσταση. Όσο η πιθανότητα p αυξάνει, τόσο ο παραγόμενος γράφος τείνει στο να γίνει ένας Τυχαίος Γράφος (Εικόνα 3.3).



Εικόνα 3.3: *Εξέλιξη ενός γράφου Watts-Strogatz, όσο αυξάνεται η πιθανότητα p.*

3.1.4 Δίκτυα Ελεύθερης Κλίμακας

Τα δίκτυα, στα οποία η κατανομή των βαθμών όλων των κόμβων ακολουθεί ένα νόμο δύναμης (power-law), όπου $P(\kappa) \sim \kappa^{-\gamma}$, χαρακτηρίζονται ως Δίκτυα Ελεύθερης Κλίμακας (scale-free networks). Η ποσότητα $P(\kappa)$ εκφράζει την πιθανότητα ένας κόμβος να έχει βαθμό ίσο με κ. Η παράμετρος γ έχει προσδιοριστεί ότι συνήθως κυμαίνεται μεταξύ $2 \leq \gamma \leq 3$, αν και έχουν παρατηρηθεί και Δίκτυα Ελεύθερης Κλίμακας με υψηλότερη τιμή. Η κατανομή νόμου δύναμης που παρουσιάζουν δίκτυα τέτοιου τύπου, εξηγείται με μηχανισμούς όπως η προτίμηση στη δύναμη (preferential attachment). Η συγκεκριμένη έννοια εκφράζει, ότι ένας κόμβος που εισέρχεται σε ένα δίκτυο, είναι πιθανότερο να συνδεθεί με κόμβους υψηλού βαθμού. Συνεπώς, οι παλιοί κόμβοι καθίστανται πιο ισχυροί (με περισσότερες συνδέσεις), ενώ οι νεότεροι λιγότερο. Μια περαιτέρω εξήγηση, μπορεί να δοθεί με το παράδειγμα της Εικόνας 3.4 , όπου οι αρχικοί κόμβοι αποκτούν όλο και περισσότερες συνδέσεις, καθώς νέοι κόμβοι εισέρχονται στο δίκτυο.



Εικόνα 3.4: *Εξέλιξη ενός δικτύου επιστημονικών συνεργασιών σε χρονικό διάστημα τεσσάρων μηνών, καθώς εισέρχονται νέα μέλη-κόμβοι [5].*

Το μοντέλο κατασκευής τεχνητών Δικτύων Ελεύθερης Κλίμακας που θα χρησιμοποιηθεί στα πλαίσια της εργασίας, είναι το μοντέλο των Barabasi και Albert [24]. Σύμφωνα με αυτό, αρχικά το δίκτυο αποτελείται από m_0 κόμβους. Σε κάθε χρονική στιγμή t ένας νέος κόμβος εισέρχεται στο δίκτυο. Ο κόμβος αυτός θα συνδεθεί με $m < m_0$ κόμβους. Η πιθανότητα Π ένας κόμβος i να συνδέεται με τον νέο κόμβο, εξαρτάται από το βαθμό του k_i και είναι ίση με $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$. Πραγματικά δίκτυα, όπως μερικά Κοινωνικά Δίκτυα, ο Παγκόσμιος Ιστός καθώς και Δίκτυα Μεταφορών, μπορούν να θεωρηθούν Δίκτυα Ελεύθερης Κλίμακας και συνεπώς το μοντέλο των Barabasi και Albert αποτελεί μία καλή προσέγγιση για τη μελέτη τους.

3.2 Μετρικές Σύνθετων Δικτύων

3.2.1 Συντελεστής Ομαδοποίησης

Ο Συντελεστής Ομαδοποίησης (clustering coefficient) είναι ένας δείκτης που η τιμή του υποδηλώνει το βαθμό στον οποίο οι κόμβοι ενός δικτύου τείνουν να σχηματίζουν τοπικές κοινότητες μεταξύ τους. Εμπειρικά έχει παρατηρηθεί ότι στα περισσότερα πραγματικά δίκτυα, και συγκεκριμένα στα Κοινωνικά Δίκτυα, οι κόμβοι τείνουν να σχηματίζουν πυκνά συνδεδεμένες κοινότητες. Ο συντελεστής αυτός έχει διάφορους ορισμούς, ανάλογα με το επίπεδο του δικτύου που εξετάζουμε κάθε φορά. Δύο τέτοιες διαφοροποιήσεις είναι και οι παρακάτω: Στη μία περίπτωση, εξετάζεται το δίκτυο συνολικά και στην άλλη περίπτωση, εξετάζεται τοπικά ένας συγκεκριμένος κόμβος του δικτύου.

- Ολικός Συντελεστής Ομαδοποίησης (Global Clustering Coefficient) [37]:

$$C = \frac{3\times\text{number of triangles}^{1}}{\text{number of all triplets}}$$

  Ο ολικός συντελεστής ομαδοποίησης εκφράζει την συνολική πιθανότητα σύνδεσης μεταξύ των γειτονικών κόμβων και βασίζεται για τον υπολογισμό του, σε πιθανές τριπλέτες[2]κόμβων που σχηματίζονται εντός του δικτύου.

- Τοπικός Συντελεστής Ομαδοποίησης (Local Clustering Coefficient) [34] ενός κόμβου u_i:

$$C_i = \frac{|\{e_{jk}: u_j, u_k \in N_i, e_{jk} \in E\}|}{k_i(k_i-1)}$$

  όπου e_{jk}: η ακμή που συνδέει το κόμβο j με το κόμβο k, N_i: η γειτονιά του κόμβου u_i και k_i: ο βαθμός του κόμβου u_i

  Ο τοπικός συντελεστής εκφράζει το πόσο απέχει η γειτονιά ενός κόμβου από το να γίνει κλίκα (2.1).

[1] τριγωνικός γράφος (triangle graph): Ένας τριγωνικός γράφος είναι ένας επίπεδος μη-κατευθυνόμενος γράφος που περιλαμβάνει τρεις κλειστές τριπλέτες.

[2] τριπλέτα (triplet): Ως τριπλέτα ορίζεται ένα υπογράφος με τρεις κόμβους, οι οποίοι μπορεί να συνδέονται με τους δύο ακόλουθους τρόπος: Είτε συνδέονται ανά δύο, οπότε σχηματίζουν μία ανοιχτή τριπλέτα (open triplet), είτε συνδέονται και οι τρεις μαζι, οπότε σχηματίζουν μία κλειστή τριπλέτα (closed triplet).

- Μέσος Συντελεστής Ομαδοποίησης Δικτύου (Network Average Clustering Coefficient) [34]:

$$\overline{C} = \frac{1}{n} \sum_{i=1}^{n} C_i$$

Ο μέσος συντελεστής ομαδοποίησης, μπορεί να χρησιμοποιηθεί αντί του ολικού συντελεστή ομαδοποίησης, ως ο μέσος όρος των τοπικών συντελεστών ομαδοποίησης για κάθε κόμβο του δικτύου.

3.2.2 Κεντρικότητα Ενδιαμεσικότητας Ακμής

Η Κεντρικότητα Ενδιαμεσικότητας Ακμής ορίστηκε από τους Girvan και Newman ως το πλήθος των ελάχιστων μονοπατιών που περνάνε από μία ακμή του δικτύου [7]. Μία ακμή με υψηλή τιμή ενδιαμεσικότητας είναι πολύ πιθανό να αντιπροσωπεύει μία γέφυρα στο δίκτυο. Τέτοιο παράδειγμα ακμής, αποτελεί η ακμή (4,5) της Εικόνας 2.3. Η συγκεκριμένη ακμή παρουσιάζει τη μεγαλύτερη τιμή Κεντρικότητας Ενδιαμεσικότητας και η αφαίρεση της θα δημιουργούσε δύο συνεκτικές συνιστώσες στο δίκτυο. Παρακάτω, παρατίθεται σε μορφή ψευδοκώδικα (Εικόνα 3.5), ο αλγόριθμος του Brandes [6], για τον υπολογισμό της εν λόγω μετρικής:

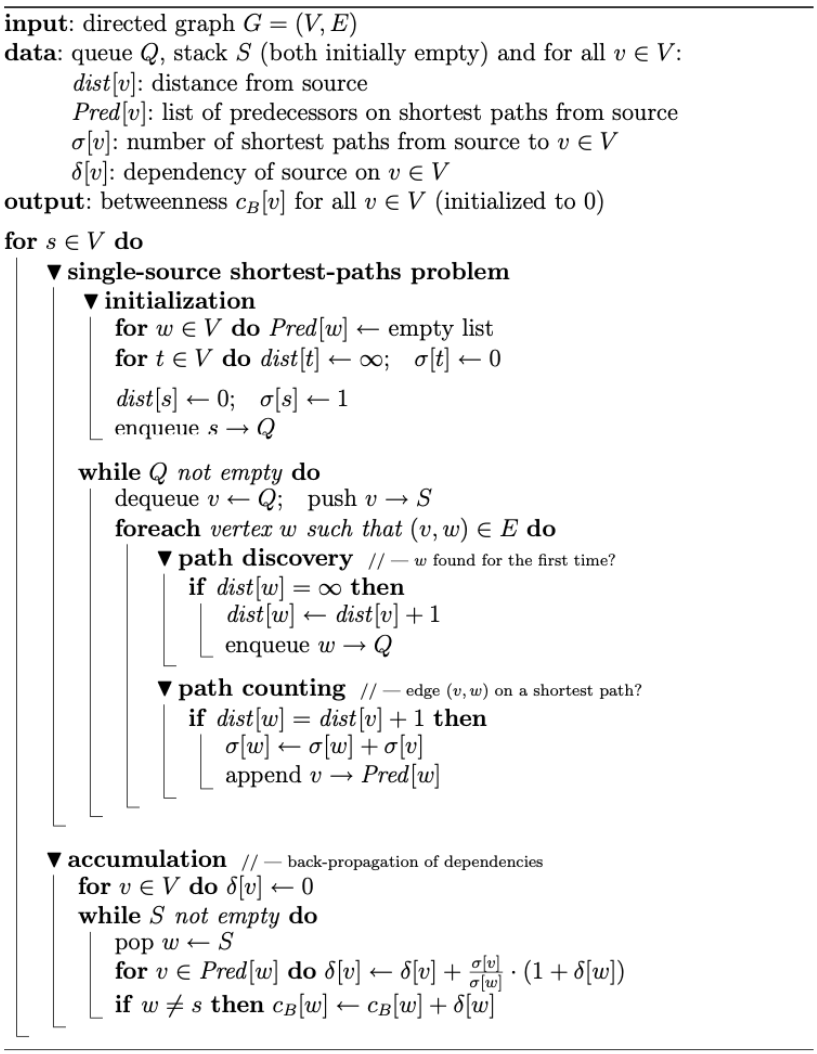
Ο συγκεκριμένος αλγόριθμος έχει πολυπλοκότητα O(m*n) για γράφους χωρίς βάρη, όπου m ο αριθμός των ακμών ενός γράφου και n ο αριθμός των κορυφών του. Ο υπολογισμός Κεντρικότητας Ενδιαμεσικότητας Ακμής, εκτελώντας κάθε φορά μία αναζήτηση κατά πλάτος BFS, για τον υπολογισμό του συντομότερου μονοπατιού μεταξύ κάθε ζεύγους κορυφών, οδηγεί σε πολυπλοκότητα της τάξης $O(m * n^2)$.

3.2.3 Κανονικοποιημένη Αμοιβαία Πληροφορία

Η αξιολόγηση ενός αλγορίθμου πάνω σε οποιοδήποτε γράφο, με ενσωματωμένη κοινοτική δομή, συνεπάγεται και το καθορισμό ενός ποσοτικού κριτηρίου, το οποίο εκτιμά πόσο καλή είναι η απάντηση που δίνει ο αλγόριθμος σε σύγκριση με την πραγματική κοινοτική δομή, που δίνεται από τις πληροφορίες του γράφου. Η σύγκριση αυτή μπορεί να γίνει με τη χρήση κατάλληλων μέτρων ομοιότητας. Ανά τη βιβλιογραφία, έχουν προταθεί διάφορες μετρικές [38]. Στη παρούσα εργασία θα γίνει χρήση ενός σύγχρονου και αξιόπιστου μέτρου, το οποίο ονομάζεται Κανονικοποιημένη Αμοιβαία Πληροφορία (Normalized Mutual Information-NMI) και εισήχθη από τους Danon et al.
Σύμφωνα με αυτή τη μετρική, δύο διαχωρισμοί X και Y λαμβάνονται υπόψιν ως τυχαίες διακριτές μεταβλητές με πεδία ορισμού [1 έως I] και [1 έως J] αντίστοιχα. Η από κοινού κατανομή πιθανότητας (joint probability distribution) τους υπολογίζεται, λαμβάνοντας υπόψη τις συχνότητες που μετρήθηκαν στα διαθέσιμα δεδομένα:

$$p_{ij} = \frac{n_{ij}}{n}$$

όπου η τιμή p_{ij} αντιπροσωπεύει την πιθανότητα, ένα τυχαίο στοιχείο να ανήκει ταυτόχρονα και στις δύο ομάδες x_i και x_j. Οι οριακές κατανομές (marginal distributions) λαμβάνονται αθροίζοντας τις από κοινού πιθανότητες:



Εικόνα 3.5: *Ο αλγόριθμος του Brandes [6].*

$$p_{i+} = \sum_j p_{ij}$$

$$p_{+j} = \sum_i p_{ij}$$

Η τιμή p_{i+} (αντίστοιχα, η p_{+j}) αντιπροσωπεύει την πιθανότητα ένα τυχαίο στοιχείο να ανήκει στην ομάδα x_i (αντίστοιχα, στην ομάδα y_j). Στη συνέχεια, μπορεί κανείς να υπολογίσει την αμοιβαία πληροφόρηση I(X,Y) αυτών των μεταβλητών, η οποία μετρά την πιθανοτική εξάρτηση τους [39]:

$$I(X, Y) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_i + p_{+j}}$$

Η αμοιβαία πληροφόρηση αντιπροσωπεύει το ποσό πληροφορίας που μοιράζεται μεταξύ των δύο μεταβλητών. Στην πραγματικότητα ισχύει ότι $I(X, Y) = H(X) - H(X|Y)$, όπου $H(X) = -\sum_x P(x) \log P(x)$ η Shannon εντροπία του X και $H(X|Y) = -\sum_{x,y} P(x, y) \log P(x|y)$ είναι η σχετική εντροπία του X δοθέντος του Y. Η αμοιβαία πληροφόρηση I(X,Y) ωστόσο δεν είναι ιδανική ως μέτρο ομοιότητας. Αυτό συμβαίνει διότι δοθείσας μίας διαμέρισης X, όλες οι διαμερίσεις που παράγονται από αυτή, μέσω περαιτέρω διαμερίσεων κάποιων από τις συστάδες της, θα είχαν όλες την ίδια κοινή πληροφορία με τη X, ενώ ενδέχεται να είναι πολύ

διαφορετικές μεταξύ τους. Στην περίπτωση αυτή η αμοιβαία πληροφόρηση απλά θα ήταν ίση με την εντροπία H(X), διότι η υπό συνθήκη εντροπία θα ήταν συστηματικά μηδέν. Για να αποφευχθεί αυτό, οι Danon et al υιοθέτησαν τη Κανονικοποιημένη Αμοιβαία Πληροφορία (Normalized Mutual Information-NMI), η οποία ορίζεται ως [40]:

$$NMI(X, Y) = \frac{2I(X,Y)}{H(X)+H(Y)}$$

Η παραπάνω ποσότητα ισούται με 1 εάν οι διαμερίσεις είναι πανομοιότυπες, ενώ έχει αναμενόμενη τιμή 0 εάν οι διαμερίσεις είναι ανεξάρτητες. Η Κανονικοποιημένη Αμοιβαία Πληροφορία χρησιμοποιείται πολύ συχνά σε ελέγχους και συγκρίσεις αλγορίθμων εντοπισμού κοινωνιών σε δίκτυα.
(Lancichinetti & Fortunato,2009 [41])

Κεφάλαιο 4

Μεγιστοποίηση της Αρθρωτότητας

Αντικείμενο της Διαμέρισης Γράφων (graph partitioning or graph clustering) είναι η ανακάλυψη ή ο εντοπισμός κοινοτήτων που απαρτίζουν το γράφο, χωρίς πρότερη γνώση τους. Για αυτό το σκοπό, υπάρχει ένα σύνολο από αλγορίθμους Εντοπισμού Κοινοτήτων, καθένας από τους οποίους χρησιμοποιεί διαφορετικά κριτήρια και είναι δυνατόν να παράγει διαφορετικά αποτελέσματα σε σχέση με τους υπόλοιπους.

Η μέθοδος Εντοπισμού Κοινοτήτων που θα παρουσιαστεί στην παρούσα ενότητα, διαμερίζει το αρχικό δίκτυο σε μη επικαλυπτόμενα σύνολα από κόμβους(network-centric), χρησιμοποιώντας την έννοια της **Αρθρωτότητας**, προκειμένου να αξιολογήσει την ποιότητα διαμέρισης ενός δικτύου.

Η **Αρθρωτότητα** (modularity) είναι μια μετρική της δομής ενός δικτύου. Συγκεκριμένα, η μετρική αυτή αξιολογεί κατά πόσο είναι "σωστή" μια διαμέριση, ενός δικτύου, σε κοινότητες. Αφού το δίκτυο έχει διαμεριστεί σε k κοινότητες, ορίζεται ένας συμμετρικός πίνακας e, διαστάσεων $k \times k$, όπου κάθε στοιχείο e_{ij} είναι το ποσοστό των ακμών του δικτύου που συνδέει τις κοινότητες i,j. Το ίχνος του πίνακα, $Tr(e) = \sum_I e_{ii}$, ορίζει το ποσοστό των ακμών που βρίσκονται εντός των κοινοτήτων. Ακόμα, ορίζεται το άθροισμα γραμμής ως $a_{ij} = \sum_j e_{ij}$, που αντιπροσωπεύει το ποσοστό των ακμών που συνδέονται στην κοινότητα i. Συνεπώς, η **Αρθρωτότητα** ορίζεται ως εξής:

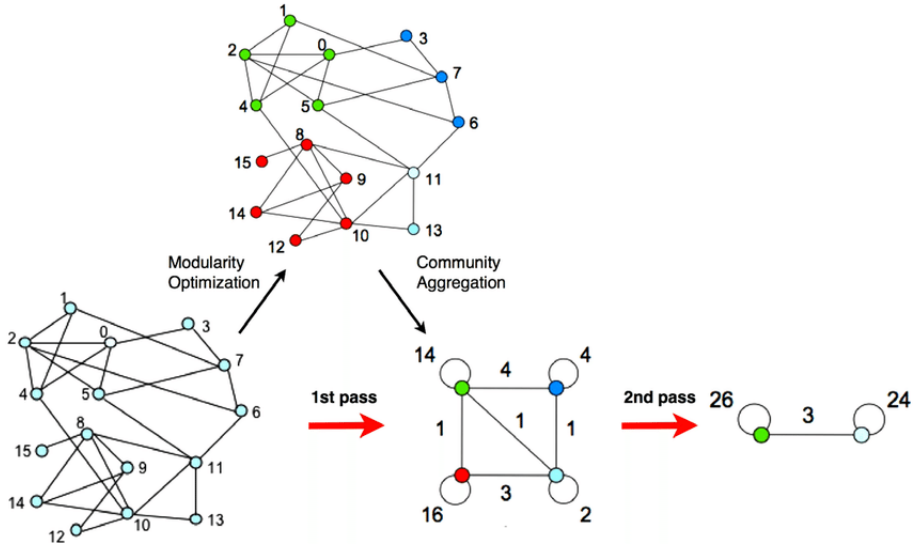
$$Q = \sum_i \left(e_{ii} - a_i^2\right) = Tr(e) - \left\|e^2\right\|$$

Το πεδίο τιμών του μεγέθους Q θα είναι από [-1,1]. Όσο πιο κοντά στο 1 είναι η τιμής της Q, τόσο πιο καλή είναι η διαμέριση του δικτύου σε κοινότητες.

Η **Μεγιστοποίηση Αρθρωτότητας** είναι η μέθοδος διαμέρισης που έχει ως στόχο την εύρεση του αριθμού κοινοτήτων, η οποία μεγιστοποιεί τη ποσότητα Q. Η μεθοδολογία που προτάθηκε στο [42] είναι η εξής: Αρχικά, όλοι οι κόμβοι του δικτύου θεωρούνται ξεχωριστές κοινότητες και υπάρχουν μηδενικές ακμές στο γράφο. Σε κάθε βήμα της μεθόδου, υπολογίζεται πρώτα ποια ακμή χρειάζεται να προστεθεί, προκειμένου να βελτιωθεί το modularity και στη συνέχεια γίνεται σύμπτυξη των κοινοτήτων, που βρίσκονται στα άκρα της. Η διαδικασία αυτή επαναλαμβάνεται τόσες φορές, όσοι είναι και οι κόμβοι του δικτύου.

Η πολυπλοκότητα της συγκεκριμένης μεθόδου είναι O((n+m)n) ή $O(n^2)$, όπου n ο αριθμός των κόμβων και m ο αριθμός των ακμών.

Στο πλαίσιο της παρούσας εργασίας, θα χρησιμοποιηθεί η μέθοδος του **Blondel** [43] για τη **Μεγιστοποίηση** της **Αρθρωτότητας**. Η συγκεκριμένη τεχνική βασίζεται σε μία τοπική βελτιστοποίηση της modularity, στη γειτονιά κάθε κόμβου. Θεωρώντας τους μεμονωμένους κόμβους ως κοινότητες, σε κάθε βήμα τοποθετείται ένας κόμβος i στην κοινότητα ενός γειτονικού του j, έτσι ώστε να αυξάνεται κατά το μέγιστο η τιμή Q. Στη συνέχεια, αφού παραχθεί, κατά αυτό το τρόπο, μία διαμέριση του δικτύου, οι κοινότητες αντικαθίστανται από υπερκόμβους (supernodes), διαδικασία που ουσιαστικά αντικαθιστά το δίκτυο με ένα μικρότερο και σταθμισμένο. Ταυτόχρονα, παράγεται μια ιεραρχική δομή του δικτύου, καθώς ανακαλύπτονται κοινότητες μεταξύ των υπερκόμβων. Η διαδικασία επαναλαμβάνεται μέχρι η modularity, η οποία υπολογίζεται πάντα σε σχέση με το αρχικό δίκτυο, να μη βελτιστοποιείται περαιτέρω.



Εικόνα 4.1: *Σχηματική Αναπαράσταση της μεθόδου* ***Blondel***.

Οπτικοποίηση των βημάτων του αλγορίθμου. Κάθε βήμα αποτελείται από δύο φάσεις: Στη πρώτη, επιχειρείται τοπική βελτιστοποίηση της modularity, κάνοντας τις απαραίτητες ενέργειες που αναφέρθηκαν παραπάνω, εντός κάθε κοινότητας. Σε δεύτερη φάση οι κοινότητες που βρέθηκαν συγχωνεύονται και αντικαθίστανται από υπερκόμβους, λαμβάνοντας ένα νέο δίκτυο κοινοτήτων. Η διαδικασία επαναλαμβάνεται μέχρι το σημείο, όπου η modularity δεν θα μπορεί να βελτιστοποιηθεί παραπάνω [43].

Κεφάλαιο 5

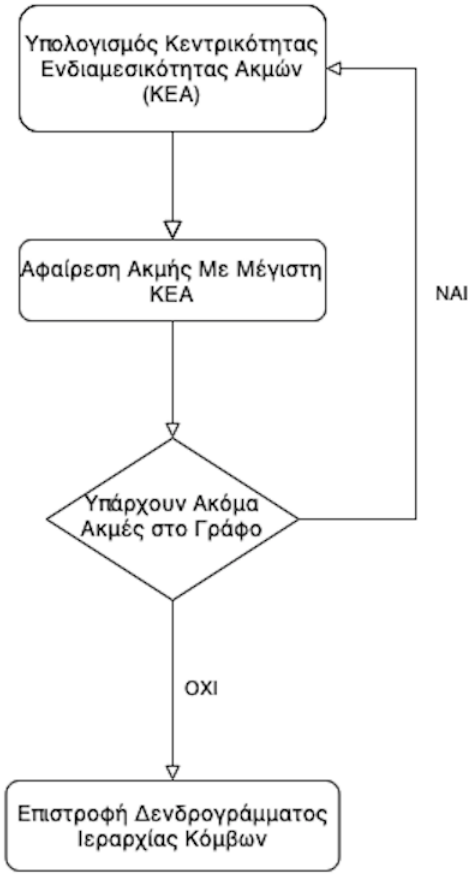
Ανακάλυψη Κοινοτήτων με την Μέθοδο Hyperbolic Girvan-Newman

Στο κεφάλαιο αυτό, παρουσιάζεται ο αλγόριθμος για τον εντοπισμό κοινοτήτων, **Hyperbolic Girvan-Newman** (**HGN**) [14],[9] ο οποίος αποτελεί μια τροποποίηση του κλασικού αλγορίθμου Girvan-Newman [7], έπειτα από χρήση της ενσωμάτωσης Rigel, που θα αναφερθεί παρακάτω.

5.1 Ο αλγόριθμος Girvan-Newman

Ο Girvan-Newman, πάνω στον οποίο βασίζεται ο HGN που αναλύεται παρακάτω, είναι ένας αλγόριθμος ιεραρχικής συσταδοποίησης, ο οποίος βασίζεται στην έννοια της Κεντρικότητας Ενδιαμεσικότητας Ακμής (Edge Betweeness Centrality) [7]. Κάθε φορά, ο αλγόριθμος υπολογίζει την ακμή με τη μεγαλύτερη τιμή κεντρικότητας και την αφαιρεί από το γράφο. Η επαναληπτική αυτή διαδικασία συνεχίζεται μέχρι το σημείο, όπου δεν υπάρχουν πια διαθέσιμες ακμές και στο τέλος επιστρέφει ένα δενδρόγραμμα, με την ιεραρχική δομή του δικτύου. Ο αλγόριθμος στηρίζεται στην υπόθεση, ότι οι ακμές που υπάρχουν μεταξύ διαφορετικών κοινοτήτων είναι λιγότερες από τις ακμές που συνδέουν κορυφές της ίδιας κοινότητας. Οι ακμές αυτές λειτουργούν ως "γέφυρες" μεταξύ των κοινοτήτων και παρουσιάζουν υψηλή τιμή Κεντρικότητας Ενδιαμεσικότητας Ακμής. Σταδιακά, η αφαίρεση τους οδηγεί στον εντοπισμό κοινοτήτων.

Όσον αφορά την απόδοση του, ο αλγόριθμος Girvan-Newman δίνει ικανοποιητικά αποτελέσματα σε γνωστά δίκτυα, όπως για παράδειγμα το Zachary's Karate Club (Εικόνα 5.2). Ωστόσο, σε μεγάλου μεγέθους δίκτυα που απαιτούνται περισσότεροι υπολογισμοί, ο αργός υπολογισμός της Κεντρικότητας Ενδιαμεσικότητας Ακμής για όλες τις ακμές του γράφου, έπειτα από κάθε αφαίρεση ακμής, αποτελεί ένα σημαντικό μειονέκτημα. Ο συνολικός χρόνος εκτέλεσης του αλγορίθμου είναι $O(m*n^2)$, όπου m ο αριθμός των ακμών και n ο αριθμός των κορυφών του δικτύου. Συνεπώς, λόγω της πολυπλοκότητας του, χρησιμοποιείται σε μεσαίου μεγέθους τοπολογίες. Στην Εικόνα 5.1 δίνεται το διάγραμμα ροής του αλγορίθμου.



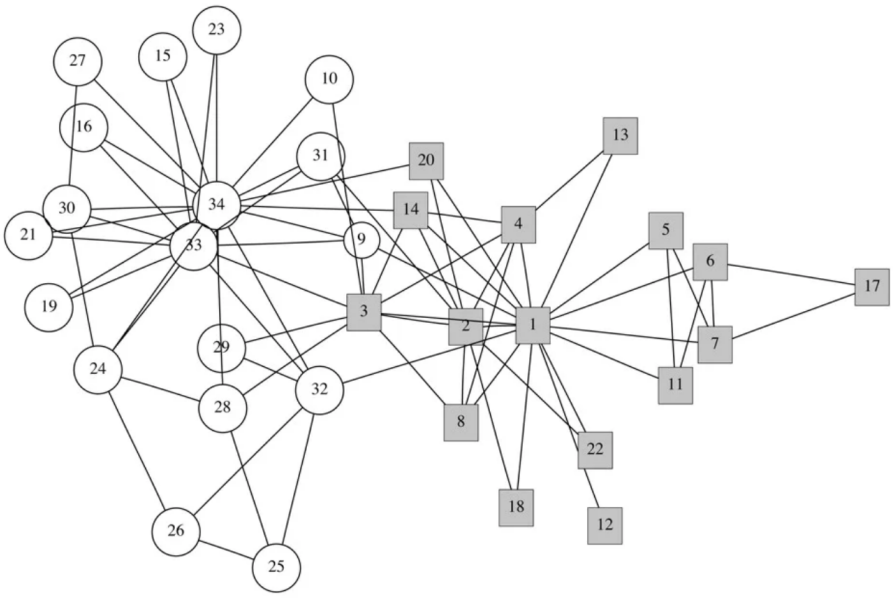
Εικόνα 5.1: *Διάγραμμα Ροής του αλγορίθμου Girvan-Newman*[1].

5.2 Ενσωμάτωση Rigel

Η Ενσωμάτωση (Embedding) γράφων στον Υπερβολικό Χώρο είναι μία διαδικασία κατά την οποία, αποδίδονται συντεταγμένες γεωμετρικού χώρου σε κάθε κορυφή ενός γράφου. Στόχος είναι η μετατροπή ενός γράφου σε μία αναπαράσταση χαμηλών διαστάσεων, όπου κάθε κόμβος αντιστοιχεί σε ένα διάνυσμα χαμηλών διαστάσεων και έχει την ιδιότητα να "κωδικοποιεί" πληροφορία, σχετικά με τη δομή ενός γράφου. Έχοντας ως αφετηρία, δεδομένα D διαστάσεων, σχηματίζουμε ένα γράφο ομοιότητας (Proximity Graph), όπου ενώνουμε με ακμές τα σημεία εκείνα που αντιπροσωπεύουν τις κορυφές του γράφου και είναι πλησιέστερα το ένα στο άλλο, σύμφωνα με κάποια μετρική ομοιότητας (π.χ. Ευκλείδεια απόσταση). Έπειτα, σε κάθε κορυφή αποδίδονται συντεταγμένες ενός χώρου διάστασης d, έτσι ώστε να διατηρούνται κατά το δυνατόν, οι αρχικές αποστάσεις μεταξύ των σημείων.
Σε πολύ μεγάλα Σύνθετα Δίκτυα, η ιδέα της Ενσωμάτωσης μπορεί να βελτιώσει αισθητά την ταχύτητα υπολογισμού σημαντικών μετρικών (π.χ. μήκος ελάχιστου μονοπατιού). Σε αυτή τη κατεύθυνση, μια προσπάθεια για την επίλυση του ζητήματος αποτελεί και η Ενσωμάτωση **Rigel** [44]. Ο συγκεκριμένος τύπος Ενσωμάτωσης διατηρεί τις αποστάσεις, δηλαδή η α-

[1]Για το σχεδιασμό των διαγραμμάτων χρησιμοποιήθηκε η σελίδα https://app.diagrams.net.

πόσταση των συντεταγμένων δύο κόμβων είναι συγκρίσιμη με την απόσταση τους στο δίκτυο (δηλαδή με το μήκος του μονοπατιού που τους συνδέει).
Η διαδικασία της Ενσωμάτωσης είναι η ακόλουθη: Αρχικά, για ένα δίκτυο N κόμβων, επιλέγονται $l \ll N$ κόμβοι, για να παίξουν το ρόλο των οροσήμων (landmarks). Οι κόμβοι που επιλέγονται ως ορόσημα είναι κόμβοι με μεγάλο βαθμό, έτσι ώστε να είναι, όσο το δυνατό, πιο κεντρικοί γίνεται (υψηλή κεντρικότητα βαθμού). Αυτοί αποτελούν τους πρώτους κόμβους, που θα υπολογίσουν αποστάσεις μεταξύ τους, με γνώμονα οι αποστάσεις αυτές να ανταποκρίνονται όσο το δυνατόν περισσότερο στα μήκη των μεταξύ τους συντομότερων μονοπατιών. Έπειτα, όλοι οι υπόλοιποι κόμβοι λαμβάνουν συντεταγμένες, τέτοιες ώστε οι αποστάσεις τους από τα ορόσημα να διατηρούνται κατά το δυνατό με μεγαλύτερη προσέγγιση.
Ο υπολογισμός του μήκους συντομότερου μονοπατιού, για κάθε ζεύγος κορυφών, είναι μία ακριβή υπολογιστικά διαδικασία που απαιτεί κάθε φορά την εκτέλεσης μίας Αναζήτησης-Κατά-Πλάτος (Breadth First Search, BFS). Πάνω σε αυτό το σημείο βασίζεται και η Ενσωμάτωση Rigel, όπου κάθε κόμβος υπολογίζει την απόσταση του με ένα υποσύνολο, επιλεγμένα τυχαίων, οροσήμων, εκτελώντας BFS.



Εικόνα 5.2: *Αποτέλεσμα διαμέρισης του Girvan-Newman, στο δίκτυο Zachary's Karate Club [7].*

5.3 Υπολογισμός Υπερβολικής Κεντρικότητας Ενδιαμεσικότητας Ακμής

Όπως αναφέρθηκε και παραπάνω, ο υπολογισμός της Κεντρικότητας Ενδιαμεσικότητας Ακμής για κάθε ακμή του γράφου, είναι μία ακριβή υπολογιστικά διαδικασία. Στο πλαίσιο αναζήτησης μιας ταχύτερης λύσης, για τον υπολογισμό της μετρικής αυτής, προτάθηκε αρχικά η ενσωμάτωση του δικτύου στον Υπερβολικό Χώρο (5.2) και έπειτα ο υπολογισμός της Υπερβολικής Κεντρικότητας Ενδιαμεσικότητας Ακμής (Hyperbolic Edge Betweeness Centrality, HEBC), όπως περιγράφεται στο [9]. Ο υπολογισμός της Υπερβολικής Κεντρικότητας Ενδιαμε-

σικότητας Ακμής προκύπτει έπειτα από τροποποίηση του αλγόριθμου που περιγράφεται στο [45], για τον υπολογισμό της Κεντρικότητας Ενδιαμεσικότητας Κορυφής (Hyperbolic Betweeness Centrality, HBC). Παρακάτω, ακολουθεί η περιγραφή του τροποποιημένου αλγορίθμου με χρήση ψευδοκώδικα:

Αλγοριθμος 5.1: *Αλγόριθμος Υπολογισμού HEBC [9]*

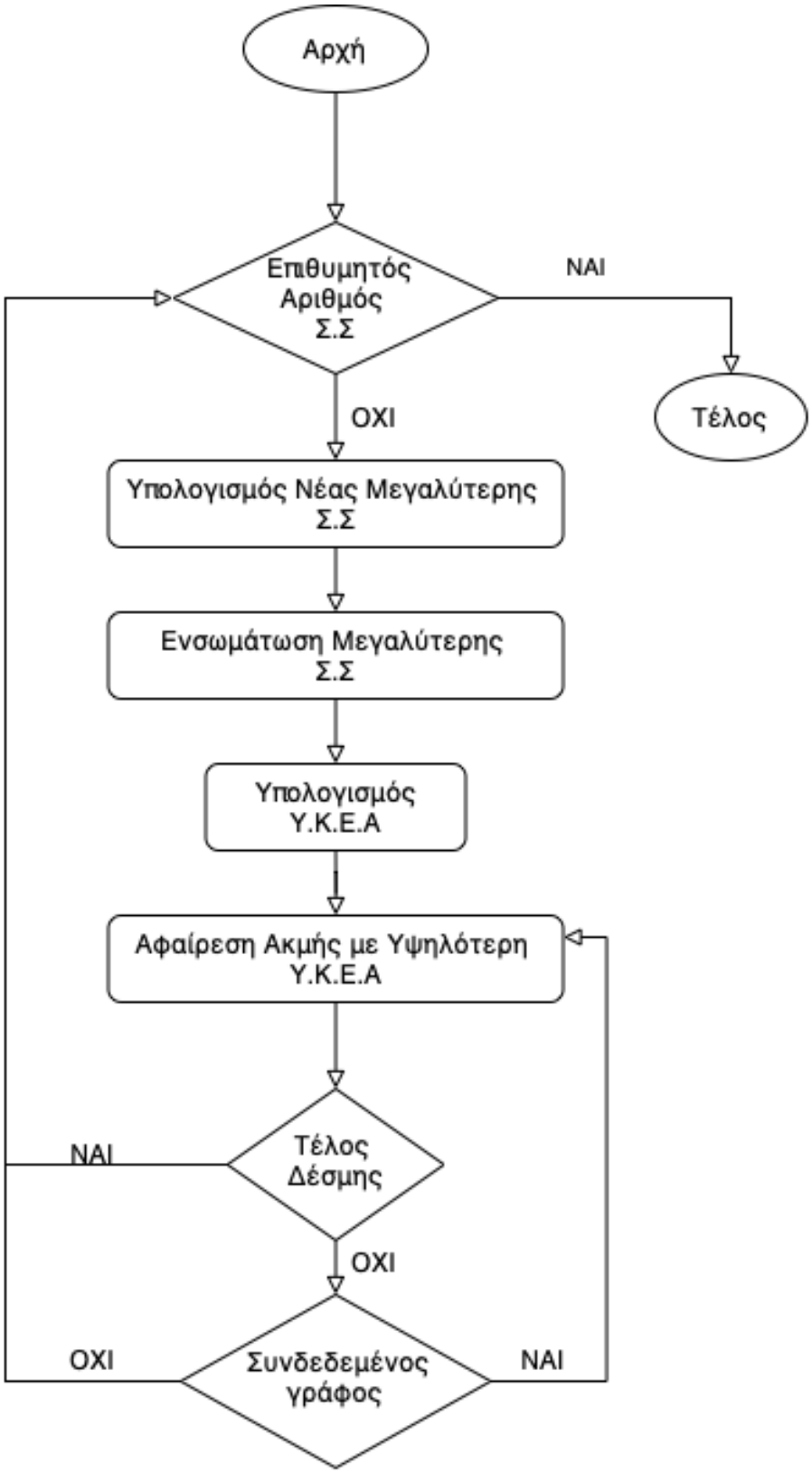
1: $HEBC(u, v) = 0, \forall u, v \in V$
2: για κάθε κόμβο $s \in V$:
3: %Μέρος I: Ταξινόμησε όλους τους κόμβους σε φθίνουσα σειρά σε σχέση με την απόσταση τους προς τον s, $u_N = s$
4: $S = \{u_1 \leqslant u_2 \leqslant ... \leqslant u_N = s\}, S_1 = S$
5: %Μέρος II:
6: $\sigma_s(u)$: ο αριθμός των άπληστων μονοπατιών με αφετηρία το κόμβο u και πέρας το κόμβο s
7: $\sigma_s(u) = 0, \forall u \in V, \sigma_s(s) = 1$
8: για $i = N : 1$ κάνε
9: για κάθε $u_j : u_i \in N_G(j, s)$:
10: $\sigma_s(u_j) = \sigma_s(u_j) + \sigma_s(u_i)$
11: αφαίρεσε το u_i από το S_1
12: %Μέρος III: Άθροισμα των εξαρτήσεων φορτίου (load dependencies)(δ) και των τιμών HEBC
13: $\delta(u) = 0, \forall u \in V$
14: για $i = 1 : N - 1$ κάνε
15: για κάθε $u_j \in N_G(i, s)$:
16: $c = \frac{\sigma_s(u_j)}{\sigma_s(u_i)} (\delta(u_i + 1))$;
17: $HEBC(u_i, u_j) = HEBC(u_i, u_j) + c$;
18: $HEBC(u_j, u_i) = HEBC(u_j, u_i) + c$;
19: $\delta(u_j) = \delta(u_i) + c$;
20: αφαίρεσε το u_i από το S

Στη γραμμή 2 του κώδικα, αρχίζει ένας εξωτερικός βρόχος, ο οποίος εκτελείται, θέτοντας κάθε κόμβο του δικτύου ως προορισμό. Έπειτα, εντός του βρόχου, στο Μέρος I του αλγορίθμου, οι κόμβοι ταξινομούνται σε μη-αύξουσα σειρά, σε σχέση με την υπερβολική απόσταση τους από το κόμβο s (προορισμό), έτσι ώστε στη συνέχεια να εξεταστούν με τη σωστή σειρά. Στο Μέρος II, υπολογίζεται ο αριθμός των άπληστων μονοπατιών μεταξύ πηγής-προορισμού. Τέλος στο Μέρος III, υπολογίζεται ο λόγος δ(u) για κάθε κόμβο u του δικτύου και στη συνέχεια, η Υπερβολική Κεντρικότητα Ενδιαμεσικότητα Ακμής (HEBC).
Στο σημείο αυτό, να διευκρινιστεί ότι η απόσταση μεταξύ δύο κόμβων εμπεριέχει σφάλμα σε σχέση με το μήκος του συντομότερου μονοπατιού. Για αυτό το λόγο, τα άπληστα μονοπάτια που υπολογίζονται δεν ταυτίζονται πάντα με τα συντομότερα μονοπάτια.

5.4 Ο Αλγόριθμος Hyperbolic Girvan-Newman

Συνοψίζοντας τα παραπάνω, ο αλγόριθμος Girvan-Newman, υπολογίζει κάθε φορά την τιμή της Κεντρικότητας Ενδιαμεσικότητας Ακμής (Κ.Ε.Α) για όλες τις ακμές και έπειτα αφαιρεί την ακμή με την υψηλότερη τιμή Κ.Ε.Α. Η διαδικασία επαναλαμβάνεται μέχρι το σημείο, όπου το σύνολο των κορυφών του αρχικού γράφου έχει διαμεριστεί στον επιθυμητό αριθμό κοινοτήτων.
Ο τροποποιημένος αλγόριθμος Girvan-Newman (HGN) ακολουθεί μια διαφορετική διαδικασία. Αρχικά, ο γράφος ενσωματώνεται στον Υπερβολικό Χώρο, με χρήση της Ενσωμάτωσης Rigel 5.2, επιλέγοντας ένα συγκεκριμένο αριθμό οροσήμων (landmarks). Έπειτα, μέσω του αλγορίθμου HEBC (5.1), υπολογίζονται οι τιμές της Υ.Κ.Ε.Α (Υπερβολικής Κεντρικότητας Ενδιαμεσικότητας Ακμής) για όλες τις ακμές του γράφου. Στη συνέχεια, αντί να αφαιρείται μια ακμή κάθε φορά, αφαιρείται ένα σύνολο-Δέσμη (batch) ακμών, ο αριθμός των οποίων καθορίζεται από το χρήστη και εξαρτάται από το πλήθος των ακμών του γράφου. Οι ακμές που επιλέγονται προς αφαίρεση, έχουν την υψηλότερη τιμή Υ.Κ.Ε.Α. Η διαδικασία επαναλαμβάνεται μέχρι είτε να υπάρξει αποσύνδεση του γράφου που εξετάζεται, είτε να έχουν αφαιρεθεί όλες οι ακμές που ανήκουν στη Δέσμη. Ο γράφος που θα προκύψει, είτε παρουσιάζει τον ίδιο αριθμό κορυφών με πριν αλλά με λιγότερες ακμές, είτε αφορά τη νέα μεγαλύτερη συνεκτική συνιστώσα, η οποία με τη σειρά της ενσωματώνεται στον Υπερβολικό Χώρο. Η διαδικασία ακολουθείται ώσπου να επιτευχθεί ο διαχωρισμός του γράφου σε ένα επιθυμητό αριθμό κοινοτήτων (συνεκτικών συνιστωσών). Παρακάτω δίνεται το διάγραμμα ροής του αλγορίθμου Hyperbolic Girvan-Newman (HGN) στην Εικόνα 5.3.

Τέλος, στα πλαίσια της παρούσας εργασίας, όσο αφορά την Ενσωμάτωση Rigel του γράφου, επιλέχθηκε Υπερβολικός Χώρος εννέα διαστάσεων και καμπυλότητας ίσης με μείον ένα. Επίσης, είναι σημαντικό να επιλεχθεί προσεχτικά το μέγεθος Δέσμης. Ένα μεγάλο μέγεθος μπορεί, ναι μεν, να οδηγήσει σε ταχύτερη εκτέλεση του προγράμματος, ωστόσο θα επιφέρει χειρότερα αποτελέσματα. Από την άλλη, η επιλογή ενός μικρότερου μεγέθους Δέσμης θα οδηγήσει μεν σε μια πιο χρονοβόρα διαδικασία εκτέλεσης του αλγορίθμου, από την άλλη όμως, ίσως δώσει καλύτερα αποτελέσματα.



Εικόνα 5.3: *Διάγραμμα Ροής του αλγορίθμου HGN.*

Κεφάλαιο 6

Ανακάλυψη Κοινοτήτων με τον Αλγόριθμο Walktrap

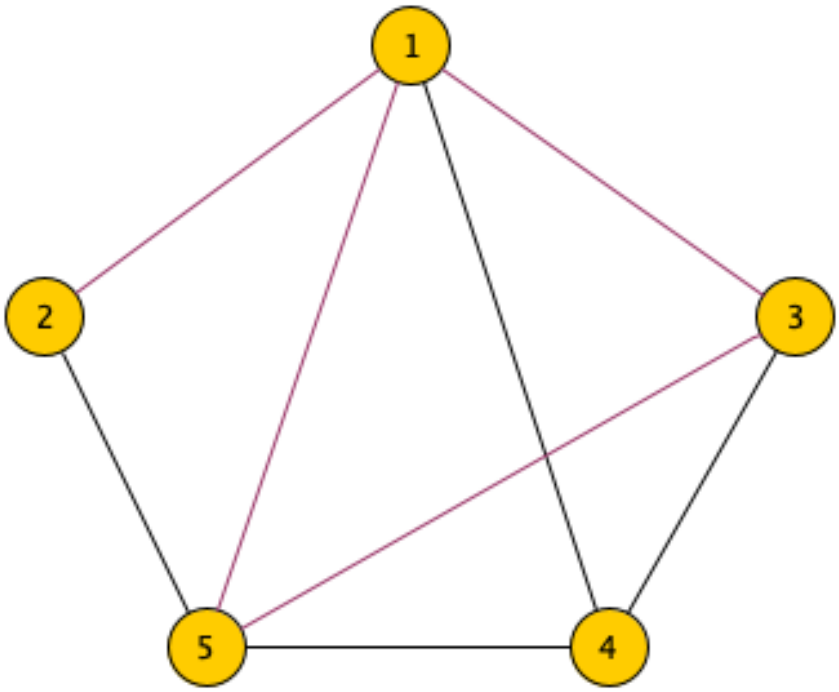
Στο κεφάλαιο αυτό, παρουσιάζεται ο αλγόριθμος για τον εντοπισμό κοινοτήτων, **Walktrap** των Pons and Latapy [25]. Αρχικά, γίνεται μια αναφορά σε βασικές αρχές και ιδιότητες των τυχαίων περιπάτων και έπειτα στην αξιοποίηση αυτών, για την κατασκευή του αλγορίθμου.

6.1 Βασικές Αρχές Τυχαίων Περιπάτων

Οι τυχαίοι περίπατοι αποτελούν ένα αποτελεσματικό εργαλείο για την εξερεύνηση των δικτύων. Είναι ιδιαίτερα αποτελεσματικοί, όταν ο γράφος είναι πολύ μεγάλος ή όταν υπάρχει πληροφορία μόνο για ορισμένα τμήματα του γράφου, καθώς οι τυχαίοι περίπατοι χρησιμοποιούν μόνο τοπικές πληροφορίες του δικτύου.
Ένα βασικό τους γνώρισμα, πάνω στο οποίο στηρίχθηκε και ο αλγόριθμος Walktrap είναι το εξής: Οι τυχαίοι περίπατοι σε γράφους τείνουν να "παγιδεύονται" σε "πυκνά" συνδεδεμένα μέρη που εμφανίζουν ισχυρή κοινοτική δομή, αναγκάζοντας έτσι τους τυχαίους περιπατητές να περάσουν περισσότερο χρόνο εντός των κοινοτήτων.

6.1.1 Ορισμοί

Η ανάλυση που θα ακολουθήσει, αφορά ένα συνεκτικό γράφημα $G = (V, E)$, με $n = |V|$ στο πλήθος κορυφές και $m = |E|$ στο πλήθος ακμές. Όπως αναφέρεται και στο Κεφάλαιο 2.1, η πληροφορία αναφορικά με την τοπολογία ενός γράφου G εμπεριέχεται στον Πίνακα Γειτνίασης του A. Συγκεκριμένα ισχύει το εξής: Αν $A_{ij} = 1$ τότε οι κορυφές i και j συνδέονται με κάποια ακμή, διαφορετικά $A_{ij} = 0$. Επίσης, ο βαθμός $d(i) = \sum_j A_{ij}$ της κορυφής i, ισούται με το σύνολο των κορυφών με τις οποίες αυτή συνδέεται (συμπεριλαμβανομένης και της ίδιας).
Έστω $(X_n)_n$, ένας τυχαίος περίπατος, δηλαδή μια στοχαστική διαδικασία διακριτού χρόνου, στις κορυφές του συνεκτικού, μη κατευθυνόμενου γράφου G. Σε κάθε βήμα, ο περιπατητής επιλέγει τυχαία μία από τις ακμές της κορυφής στην οποία βρίσκεται και μεταβαίνει σε γειτονική κορυφή, μέσω της ακμής που επιλέχθηκε τυχαία.



Εικόνα 6.1: *Σχηματική Αναπαράσταση Τυχαίου Περιπάτου σε Γράφο.*

$$1 \rightarrow 3 \rightarrow 5 \rightarrow 1 \rightarrow 2 \rightarrow ...$$

Η ακολουθία των κορυφών $v_0, v_1, v_2, ..., v_k, ...$, επιλεγμένη κατά αυτό το τρόπο, είναι ένας απλός τυχαίος περίπατος στον γράφο G. Σε κάθε βήμα k, η τυχαία μεταβλητή X_k παίρνει τιμές στο χώρο V. Η τυχαία ακολουθία $(X_n)_n = X_0, X_1, X_2, ..., X_k, ...$ λοιπόν, είναι μια στοχαστική διαδικασία διακριτού χρόνου που ορίζεται στο χώρο καταστάσεων V. Δεδομένου ότι ένας περιπατητής βρίσκεται σε μία κορυφή i με βαθμό d(i), όλες οι d(i) γειτονικές κορυφές της i είναι ισοπίθανες ως επόμενος προορισμός. Η ακολουθία λοιπόν των κορυφών που επισκέφθηκε ο περιπατητής, αποτελεί μία *Μαρκοβιανή Αλυσίδα*, οι καταστάσεις της οποίας είναι οι κορυφές του γράφου. Σε κάθε βήμα, η πιθανότητα μετάβασης από τη κορυφή i στη κορυφή j, είναι $P_{ij} = \frac{A_{ij}}{d(i)}$. Η πιθανότητα αυτή ορίζει το πίνακα μετάβασης P (transition matrix) των διαδικασιών τυχαίου περιπάτου. Επίσης ισχύει η σχέση: $P = D^{-1}A$, όπου D διαγώνιος πίνακας (των βαθμών $D_{ii} = d(i)$ και $D_{ij} = 0$ για $i \neq j, \forall i, j$).

6.1.2 Ιδιότητες

Η διαδικασία οδηγείται από τις δυνάμεις του πίνακα P:
Η πιθανότητα μετάβασης από μία κορυφή i σε μία κορυφή j, μέσω ενός τυχαίου περιπάτου μήκους t (δηλαδή αριθμού βημάτων), είναι $(P^t)_{ij}$. Η πιθανότητα αυτή, την οποία συμβολίζουμε με P^t_{ij} πληρεί δύο βασικές ιδιότητες της διαδικασίας τυχαίου περιπάτου, οι οποίες θα βοηθήσουν στην κατασκεύη της απόστασης r, ένα μέτρο το οποίο όρισαν οι συγγραφείς για τις ανάγκες του αλγορίθμου και χρησιμοποιείται για τη μέτρηση της ομοιότητας μεταξύ κορυφών. Οι ιδιότητες αυτές είναι οι ακόλουθες (οι αποδείξεις τους παρατίθονται στα

τμήματα 2 και 3.2 της δημοσίευσης [25]):

1. Όταν το μήκος t ενός τυχαίου περιπάτου, που ξεκινάει από μία κορυφή i τείνει στο άπειρο, η πιθανότητα να βρεθούμε σε μία κορυφή j, εξαρτάται μόνο από το βαθμό της κορυφής j (και όχι από τη κορυφή i):

$$\lim_{t\to\infty} P_{ij}^t = \frac{d(j)}{\sum_k d(k)}, \forall i$$

2. Ο λόγος των πιθανοτήτων να μεταβούμε από τη κορυφή i στη κορυφή j και από τη j στην i, μέσω ενός τυχαίου περιπάτου μήκους t, εξαρτάται μόνο από τους βαθμούς d(i) και d(j):

$$d(i)P_{ij}^t = d(j)P_{ji}^t, \forall i, \forall j$$

6.1.3 Μέτρο Δομικής Ομοιότητας

Χρησιμοποιώντας τις ιδιότητες των τυχαίων περιπάτων σε γράφους, οι συγγραφείς όρισαν ένα μέτρο δομικής ισοδυναμίας μεταξύ των κορυφών, δηλαδή ένα μέτρο απόστασης, το οποίο υπολογίζεται από τις πιθανότητες μετάβασης ενός τυχαίου περιπατητή από τη μία κορυφή στην άλλη, για ένα δεδομένο αριθμό βημάτων. Η απόσταση αυτή έχει το πλεονέκτημα ότι υπολογίζεται εύκολα και μπορεί να χρησιμοποιηθεί σε ένα συσσωρευτικό αλγόριθμο ιεραρχικής συσταδοποίησης (επαναληπτική συγχώνευση των κορυφών σε κοινότητες). Κατά το τρόπο αυτό, αποκτάται μία ιεραρχική κοινοτική δομή που αναπαρίσταται μέσω δενδρογράμματος (**dendrogram**).
Ανάλογα με το πλαίσιο στο οποίο χρησιμοποιείται ο αλγόριθμος, μπορούν να χρησιμοποιηθούν διαφορετικά κριτήρια για να επιλεγεί η καλύτερη ή οι καλύτερες διαμερίσεις. Ένα από τα κριτήρια αυτά, το οποίο χρησιμοποιήθηκε και από τους συγγραφείς στα πειράματα τους, είναι η Αρθρώτητα Q (**modularity** [31]), λόγω αξιοπιστίας των αποτελεσμάτων της, αλλά και λόγω της διευκόλυνσης που τους παρείχε στη σύγκριση του αλγόριθμου με άλλους. Η πολυπλοκότητα του αλγορίθμου, για τον υπολογισμό της κοινοτικής δομής, είναι O(mnH), όπου H το ύψος του αντίστοιχου δενδρογράμματος. Ο χρόνος εκτέλεσης χειρότερης περίπτωσης είναι $O(mn^2)$. Ωστόσο, τα περισσότερα πραγματικά δίκτυα είναι αραιά ($m = O(n)$) και το ύψος του δενδρογράμματος H είναι γενικά μικρό ($H = O(logn)$). Στην περίπτωση αυτή, η πολυπλοκότητα του αλγορίθμου είναι $O(n^2 logn)$.

Για να ομαδοποιηθούν οι κορυφές σε κοινότητες, εισάγεται η απόσταση r μεταξύ των κορυφών, η οποία αποτυπώνει τη κοινοτική δομή του γράφου. Εάν δύο κορυφές ανήκουν σε διαφορετικές κοινότητες, η απόσταση αυτή θα πρέπει να είναι μεγάλη, ενώ αν ανήκουν στην ίδια θα πρέπει να είναι μικρή. Ο υπολογισμός της προκύπτει από τη πληροφορία που παίρνουμε από τους τυχαίους περιπάτους στο γράφο. Η πληροφορία για τη κορυφή i, η οποία κωδικοποιείται μέσω του πίνακα μεταβάσεων P^t, βρίσκεται στις n πιθανότητες $(P_{ij}^t)_{1\le k\le n}$. Οι πιθανότητες αυτές είναι ουσιαστικά η i-οστή γραμμή του πίνακα μεταβάσεων P^t και συμβολίζεται με $P_{i\bullet}^t$. Δεδομένων των προαναφερθέντων ιδιοτήτων, εισάγονται παρακάτω δύο ορισμοί για την απόσταση μεταξύ κορυφών και κοινοτήτων στο γράφο, αντίστοιχα:

Ορισμός 1: Έστω, i και j δύο κορυφές στο γράφο. Τότε, η απόσταση τους r_{ij}, ορίζεται ως:

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{\left(P_{ik}^t - P_{jk}^t\right)^2}{d(k)}} = \left\| D^{-\frac{1}{2}} P_{i\bullet}^t - D^{-\frac{1}{2}} P_{j\bullet}^t \right\|$$

όπου $\|\cdot\|$ η Ευκλείδεια νόρμα του $\mathbb{R}^n$. Η απόσταση αυτή, εξαρτάται από το μήκος του τυχαίου περιπάτου t και μπορεί να γραφτεί και ως $(r_{ij})^t$.

Θεωρώντας τώρα, τυχαίους περιπάτους που ξεκινούν από μία κοινότητα, επιλέγοντας τον αρχικό κόμβο τυχαία μεταξύ των κορυφών της, ορίζεται η πιθανότητα μετάβασης από τη κοινότητα C στην κορυφή j σε t βήματα, ως:

$$P_{Cj}^t = \frac{1}{|C|} \sum_{i \in C} P_{ij}^t$$

Αυτή ορίζει ένα διάνυσμα πιθανοτήτων $P_{C\bullet}^t$, το οποίο επιτρέπει τη γενίκευση της απόστασης μεταξύ κοινοτήτων, μέσω του ακόλουθου ορισμού.

Ορισμός 2: Έστω, $C_1, C_2 \subset V$ δύο κοινότητες. Τότε, η απόσταση τους $r_{C_1 C_2}$ ορίζεται ως:

$$r_{C_1 C_2} = \left\| D^{-\frac{1}{2}} P_{C_1\bullet}^t - D^{-\frac{1}{2}} P_{C_2\bullet}^t \right\| = \sqrt{\sum_{k=1}^{n} \frac{\left(P_{C_1 k}^t - P_{C_2 k}^t\right)^2}{d(k)}}$$

6.2 Περιγραφή του Αλγόριθμου Walktrap

Έχοντας εισάγει την απόσταση μεταξύ κορυφών (αλλά και μεταξύ ομάδων κορυφών) για τον εντοπισμό δομικών ομοιοτήτων μεταξύ τους, το πρόβλημα εντοπισμού κοινοτήτων πλέον ανάγεται σε ένα πρόβλημα συσταδοποίησης. Οι συγγραφείς χρησιμοποιούν ένα συσσωρευτικό αλγόριθμο ιεραρχικής συσταδοποίησης, που βασίζεται στη μέθοδο του Ward [46]. Ο συγκεκριμένος αλγόριθμος δίνει πολύ καλά αποτελέσματα, ενώ μειώνει τον υπολογισμό των αποστάσεων.

Η διαδικασία εκκινεί από μία διαμέριση $P_1 = \{\{v\} \in V\}$ του γράφου σε n κοινότητες, αποτελούμενες από μία μόνο κορυφή (δηλαδή κάθε κορυφή αποτελεί μία κοινότητα). Αρχικά, υπολογίζονται οι αποστάσεις μεταξύ όλων των γειτονικών κορυφών. Έπειτα, η διαμέριση εξελίσσεται επαναλαμβάνοντας τις ακόλουθες διαδικασίες. Σε κάθε βήμα k:

- Επιλέγει δύο κοινότητες C_1, C_2 στην P_k, σύμφωνα με ένα κριτήριο που βασίζεται στην απόσταση μεταξύ κοινοτήτων, το οποίο θα αναλύσουμε στη συνέχεια.
- Συγχωνεύει αυτές τις δύο κοινότητες σε μία νέα $C_3 = C_1 \cup C_2$ και δημιουργεί μία νέα διαμέριση $P_{k+1} = (P_k \setminus \{C_1, C_2\} \cup \{C_3\})$.
- Ανανεώνει τις αποστάσεις μεταξύ των κοινοτήτων (κάτι το οποίο γίνεται μόνο για γειτονικές κοινότητες).

Ύστερα από n-1 βήματα, ο αλγόριθμος σταματά και εν τέλει $P_n = \{V\}$. Κάθε βήμα του ορίζει και μία διαμέριση P_k του γράφου σε κοινότητες, η οποία παράγει μία ιεραρχική δομή που

αναπαρίσταται μέσω δενδρογράμματος. Τα σημαντικά σημεία του αλγορίθμου είναι ο τρόπος με τον οποίο επιλέγονται οι κοινότητες προς συγχώνευση, η ανανέωση των αποστάσεων και το πως εκτιμάται η ποιότητα της εκάστοτε διαμέρισης P_k.

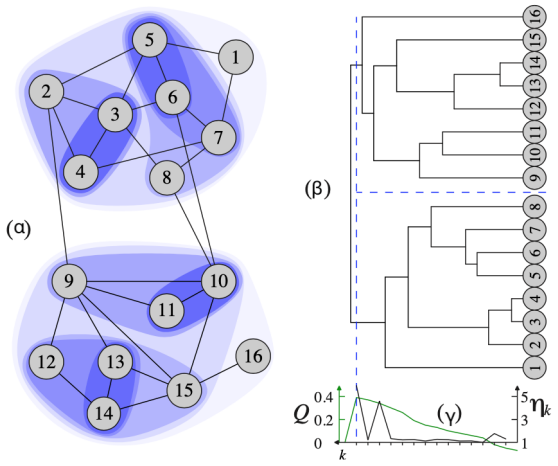
Η επιλογή των δύο κοινοτήτων, που θα συγχωνευτούν, γίνεται σύμφωνα με τη μέθοδο του Ward. Σε κάθε βήμα k, συγχωνεύουμε δύο κοινότητες που ελαχιστοποιούν το μέσο όρο σ_k των τετραγώνων των αποστάσεων, ανάμεσα σε κάθε κορυφή και την κοινότητά της:

$$\sigma_k = \frac{1}{n} \sum_{C \in P_k} \sum_{i \in C} r_{iC}^2$$

Ωστόσο το πρόβλημα αυτό είναι NP-πλήρες και για το λόγο αυτό, για κάθε ζεύγος γειτονικών κοινοτήτων C_1, C_2, υπολογίζεται η απόκλιση $\Delta\sigma(C_1, C_2)$ του σ που προκαλείται αν συγχωνεύσουμε τις C_1, C_2 σε μία νέα κοινότητα $C_3 = C_1 \cup C_2$. Η ποσότητα αυτή εξαρτάται μόνον από τις κορυφές των C_1 και C_2 και όχι από τις άλλες κοινότητες, ή το βήμα k του αλγορίθμου:

$$\Delta\sigma(C_1, C_2) = \frac{1}{n}\left(\sum_{i \in C_3} r_{iC_3}^2 - \sum_{i \in C_1} r_{iC_1}^2 - \sum_{i \in C_2} r_{iC_2}^2\right)$$

Εν τέλει, συγχωνεύονται οι δύο κοινότητες που δίνουν τη μικρότερη τιμή του $\Delta\sigma$[1].



Εικόνα 6.2: *Σχηματική Αναπαράσταση του* ***Walktrap***.

(α) Ένα παράδειγμα κοινοτικής δομής, όπως αυτή εντοπίστηκε από τον αλγόριθμο χρησιμοποιώντας τυχαίους περιπάτους μήκους $t = 3$. (β) Οι φάσεις του αλγορίθμου κωδικοποιούνται στο εν λόγω δενδρόγραμμα. (γ) Σύμφωνα με το κριτήριο της Αρθρωτότητας (modularity) Q και του λόγου αύξησης(increase ratio)η_k, η βέλτιστη διαμέριση αντιστοιχεί σε δύο κοινότητες. Ο λόγος αύξησης η_k είναι και αυτός ένα κριτήριο αξιολόγησης της ποιότητας των διαμερίσεων. Συγκεκριμένα, προτιμάται από την Αρθρωτότητα, για τον εντοπισμό κοινοτήτων διαφορετικής κλίμακας.[25]

Ο αλγόριθμος παράγει μία ακολουθία $(P_k)_{1 \le k \le n}$ διαμερίσεων. Όπως αναφέραμε και στην εισαγωγική περιγραφή της μεθόδου, η Αρθρωτότητα Q χρησιμοποιείται ευρέως ως κριτήριο ποιότητας των εν λόγω διαμερίσεων. Η καλύτερη διαμέριση είναι αυτή που μεγιστοποιεί την Q.

[1] Ο υπολογισμός των Δσ και η ανανέωση των αποστάσεων δύναται να υπολογιστεί αναλυτικά, λόγω λημμάτων που προκύπτουν, από το γεγονός ότι η απόσταση που όρισαν οι συγγραφείς είναι Ευκλείδεια απόσταση.

Ωστόσο, όταν οι κοινότητες είναι διαφορετικής κλίμακας μεγέθους, η Αρθρωτότητα δεν είναι κατάλληλο κριτήριο αξιολόγησης της ποιότητας των παραγόμενων διαμερίσεων. Για το λόγο αυτό, οι συγγραφείς εισήγαγαν το *λόγο αύξησης* η_k (increase ratio) της ποσότητας σ_k:

$$\eta_k = \frac{\Delta_{\sigma_k}}{\Delta_{\sigma_{k-1}}} = \frac{\sigma_{k+1} - \sigma_k}{\sigma_k - \sigma_{k-1}}$$

Οι μεγαλύτερες τιμές του η_k είναι αυτές που αντιστοιχούν και στις καλύτερες διαμερίσεις. Στην παρούσα εργασία, θα χρησιμοποιηθεί το κριτήριο της Αρθρωτότητας, προκειμένου να αξιολογηθούν τα πειραματικά αποτέλεσμα που θα προκύψουν από τη σύγκριση των δύο προτεινόμενων αλγορίθμων, για τον εντοπισμό κοινοτήτων, **Walktrap** και **Hyperbolic Girvan-Newman**.

Κεφάλαιο 7

Πειραματική Αξιολόγηση

Στο κεφάλαιο αυτό, γίνεται αρχικά μία συνοπτική παρουσίαση των τοπολογιών που χρησιμοποιήθηκαν για τη συλλογή αποτελεσμάτων. Οι τοπολογίες αφορούν Πραγματικά και Τεχνητά Σύνθετα Δίκτυα. Στη συνέχεια, στην ενότητα 7.2 επιχειρείται σύγκριση μεταξύ των μεθόδων HGN και Walktrap, όπου συγκεκριμένα παρατίθενται συγκριτικά αποτελέσματα για το χρόνο εκτέλεσης των αλγορίθμων, καθώς και για ποιότητα τον διαμερίσεων που προκύπτουν, μέσω της μετρικής της Αρθρωτότητας (modularity). Οι πραγματικές και συνθετικές τοπολογίες που χρησιμοποιήθηκαν για την διεξαγωγή των πειραμάτων, χωρίζονται σε δύο κατηγορίες. Η πρώτη αφορά δίκτυα, που γνωρίζουμε από πριν ποια είναι η σωστή διαμέριση τους σε κοινότητες. Η δεύτερη αφορά δίκτυα, όπου δεν είναι γνωστή η πληροφορία της βέλτιστης διαμέρισης τους και για αυτό το λόγο, για τις ανάγκες τις εργασίας, η συγκεκριμένη πληροφορία παρήχθη με τη βοήθεια της μεθόδου του Blondel [43] για τη Μεγιστοποίηση της ΑρΘρωτότητας (Ενότητα 4). Τέλος, παρουσιάζονται τα αποτελέσματα για δίκτυα με γνωστή ομαδοποίηση σε κοινότητες (πραγματικά και σύνθετα), τα οποία συγκρίνονται ως προς τη μετρική της Κανονικοποιημένης Αμοιβαίας Πληροφορίας (NMI).

Η εκτέλεση των πειραμάτων που παρουσιάζονται παρακάτω, έγινε σε απλό προσωπικό υπολογιστή με τα εξής χαρακτηριστικά: Dual-Core Intel Core i5 2,6 GHz, 8 GB RAM και λειτουργικό σύστημα macOS (64 bit).

7.1 Τοπολογίες Δικτύων Πειραμάτων

Η πειραματική αξιολόγηση πραγματοποιήθηκε σε ένα πλήθος δικτύων, ορισμένα συνθετικά και κάποια πραγματικά, με σκοπό την όσο το δυνατόν καλύτερη εξαγωγή συμπερασμάτων. Στη παρούσα ενότητα, θα γίνει παρουσίαση συγκεκριμένων τοπολογιών που ανήκουν στις εν λόγω κατηγορίες δικτύων, με σκοπό την καλύτερη κατανόηση των πειραμάτων. Κατόπιν, τα αποτελέσματα που προκύπτουν από αυτά, παρουσιάζονται στη συνέχεια.

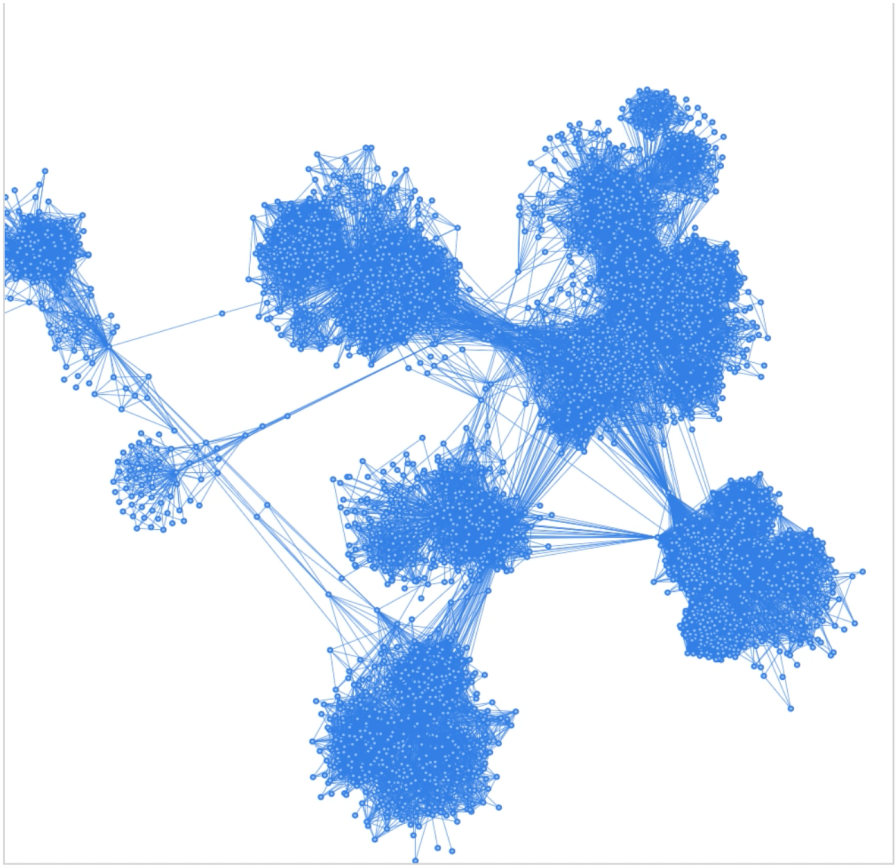
7.1.1 Πραγματικά Δίκτυα

Σε αυτή τη κατηγορία, τα δίκτυα που χρησιμοποιήθηκαν ανακτήθηκαν από την ιστοσελίδα [47], όπου παρατίθεται εκτενής αναφορά για τις αρχικές πηγές τους, καθώς και λεπτομέρειες

για τη φύση τους. Στην παρούσα εργασία χρησιμοποιούνται τα παρακάτω:

- Social circles: Facebook
  Το δίκτυο αυτό αποτελείται από ομάδες χρηστών του Facebook που αλληλεπιδρούν μεταξύ τους και μοιράζονται κοινά ενδιαφέροντα (ανήκουν στην ίδια λίστα φιλίας). Το δίκτυο αυτό αποτελείται από 4039 κόμβους (χρήστες), όπου ενώνονται μεταξύ τους με 88234 ακμές. Μια σχηματική αναπαράσταση του δικτύου δίνεται στην Εικόνα 7.1.

- email-Eu-core:
  Το παρόν δίκτυο αφορά δεδομένα ηλεκτρονικής αλληλογραφίας, προερχόμενα από ένα μεγάλο Ερευνητικό Ινστιτούτο στην Ευρώπη. Αποτελείται από 1005 χρήστες (κόμβους) που επικοινωνούν μεταξύ τους, ανταλλάσσοντας μηνύματα ηλεκτρονικής αλληλογραφίας.



Εικόνα 7.1: *Απεικόνιση του δικτύου Social circles: Facebook.*

7.1.2 Τεχνητά Σύνθετα Δίκτυα

Για τις ανάγκες της παρούσας εργασίας, εκτός από τα προαναφερθέντα πραγματικά δίκτυα, δημιουργήθηκαν και ορισμένα τεχνητά δίκτυα. Συγκεκριμένα, χρησιμοποιήθηκαν το μοντέλο Τυχαίων Γεωμετρικών Γράφων για την κατασκευή χωρικών δικτύων, το μοντέλο των Barabasi-Albert για την παραγωγή δικτύων Ελεύθερης Κλίμακας, καθώς και το μοντέλο των Watts-Strogatz για την κατασκευή δικτύων Μικρού Κόσμου. Στους παρακάτω πίνακες παρουσιάζονται ξεχωριστά ιδιότητες και βασικά χαρακτηριστικά για κάθε είδος γράφου.

Πίνακας 7.1: *Χαρακτηριστικά Τυχαίων Γεωμετρικών Γράφων*

Δίκτυο	Αριθμός Κόμβων	Κατώφλι	Μέγιστη Αρθρωτότητα (Αριθμός Κοινοτήτων)
rgg1	100	0,2	0,61 (5)
rgg2	100	0,3	0,42 (4)
rgg3	100	0,4	0,35 (3)
rgg4	100	0,5	0,27 (2)

Πίνακας 7.2: *Χαρακτηριστικά Δικτύων Ελεύθερης Κλίμακας*

Δίκτυο	Αριθμος Κόμβων	Αριθμός Ακμών	Ελάχιστος Βαθμός	Μέγιστη Αρθρωτότητα (Αριθμός Κοινοτήτων)
scf1	1000	5964	6	0,24 (10)
scf2	500	2964	6	0,25 (8)
scf3	100	564	6	0,22 (7)
scf4	100	651	7	0,21 (6)
scf5	100	475	5	0,25 (6)
scf6	100	384	4	0,32 (7)

Πίνακας 7.3: *Χαρακτηριστικά Δικτύων Μικρού Κόσμου*

Δίκτυο	Αριθμος Κόμβων	Αριθμός Ακμών	Αριθμός Κοντινότερων Γειτόνων	Πιθανότητα Επανασύνδεσης	Μέγιστη Αρθρωτότητα (Αριθμός Κοινοτήτων)
smw1	1000	2000	4	0,3	0,7 (26)
smw2	1000	4000	8	0,3	0,58 (11)
smw3	100	200	4	0,3	0,59 (9)
smw4	100	100	3	0,2	0,8 (10)
smw5	100	200	5	0,3	0,64 (10)
smw6	100	200	5	0,1	0,71 (8)

Πίνακας 7.4: *Χαρακτηριστικά Δικτύων Με Γνωστές Κοινότητες*

Δίκτυο	Αριθμος Κόμβων	Αριθμός Ακμών	Μέσος Βαθμός Κορυφής	Ελάχιστος Αριθμός Κόμβων/Κοινότητα	Αριθμός Κοινοτήτων
s100	100	104	3	5	7
m200	200	414	5	20	9
s250	250	1056	5	20	3
s500	500	499	3	100	4

Επίσης, ακολουθώντας τη μέθοδο που περιγράφεται στο [41], παράχθηκαν τέσσερα δίκτυα με κοινότητες που παρουσιάζονται παρακάτω στον Πίνακα 7.4.

7.2 Σύγκριση των Μεθόδων Hyperbolic Girvan-Newman και Walktrap

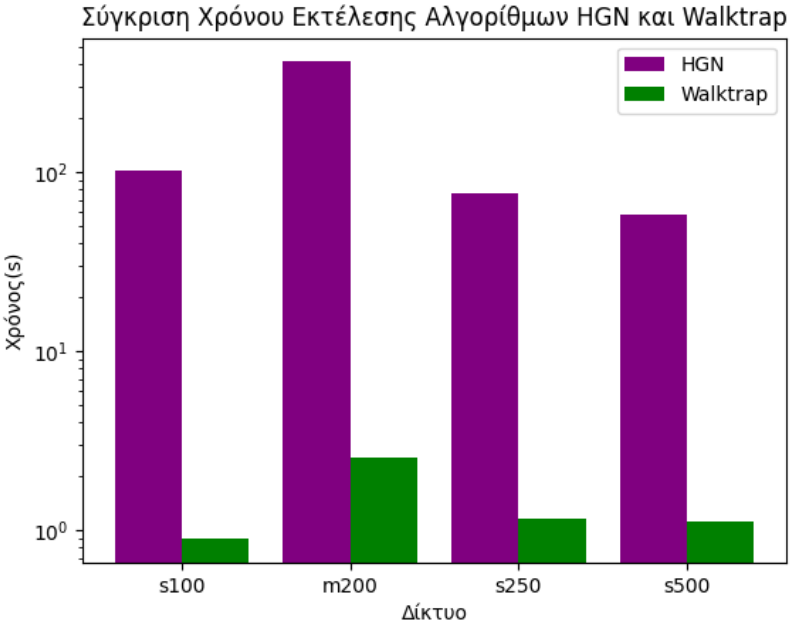
Στην ενότητα αυτή, συγκρίνονται οι αλγόριθμοι HGN (Hyperbolic Girvan-Newman) και Walktrap. Παρακάτω, η πειραματική αξιολόγηση εφαρμόζεται σε δύο κατηγορίες δικτύων. Στα δικτύα των οποίων η βέλτιστη ομαδοποίηση τους είναι γνωστή και στα δίκτυα για τα οποία δεν είναι γνωστό εξαρχής, ποια είναι η βέλτιστη διαμέριση τους σε κοινότητες. Στην πρώτη κατηγορία ανήκουν ορισμένα πραγματικά, αλλά και τεχνητά δίκτυα που δημιουργήθηκαν με τη μέθοδο του [41] και γνωρίζουμε τις κοινότητες διαμέρισης τους. Για τη δεύτερη κατηγορία χρησιμοποιείται η μέθοδος της Μεγιστοποίησης Αρθρωτότητας [4] για την εύρεση κοινοτήτων, ως μεθόδου αναφοράς.

- Δίκτυα με Γνωστές Κοινότητες

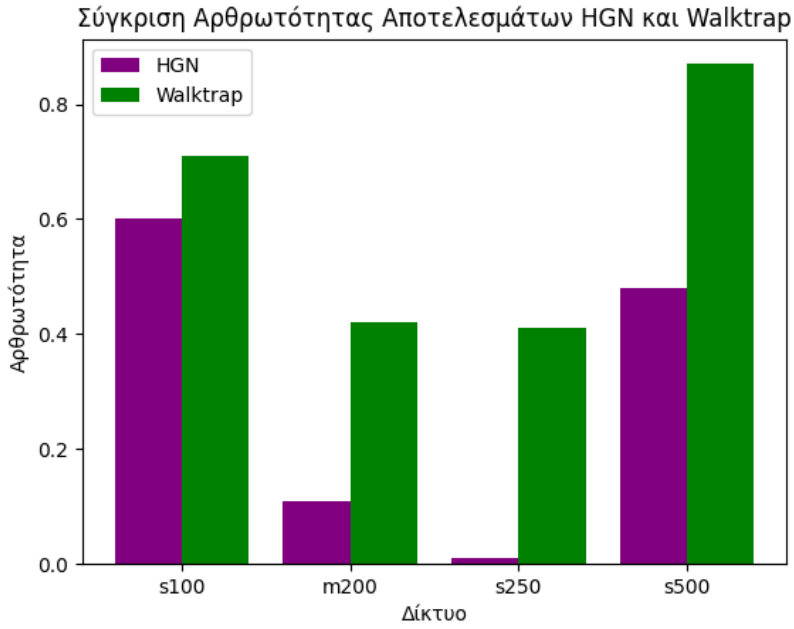
Πίνακας 7.5: *Σύγκριση Χρόνου Εκτέλεσης Αλγορίθμων HGN και Walktrap για Δίκτυα με Άγνωστες Κοινότητες.*

Δίκτυο	Χρόνος HGN (s)	Αρθρωτότητα HGN	Χρόνος Walktrap (s)	Αρθρωτότητα Walktrap
email-EU-core	1516,3	0,0014	0,47	0,38
s100	102,9	0,6	0,89	0,71
m200	414,7	0,11	2,52	0,42
s250	76,34	0,01	1,16	0,41
s500	58,02	0,48	1,12	0,87

Για τα δίκτυα του Πίνακα 7.5, προκύπτει ότι ο χρόνος εκτέλεσης του αλγορίθμου Walktrap είναι σημαντικά ταχύτερος από το χρόνο εκτέλεσης του αλγορίθμου HGN. Όσον αφορά τις τιμές της Αρθρωτότητας, τα συμπεράσματα διαφέρουν ανά δίκτυο. Για στο s100 οι αλγόριθμοι Walktrap και HGN δίνουν εξίσου καλά αποτελέσματα, καθώς πρόκειται για ένα αραιό δίκτυο, χωρίς πυκνές συνδέσεις. Αντίθετα, στο δίκτυο email-EU-core, ο HGN αποτυγχάνει να εντοπίσει σωστά κάποια κοινοτική διαμέριση, σε αντίθεση με τον Walktrap, ο οποίος αποδίδει πολύ καλύτερα, ακόμα και σε πολύ πυκνά δίκτυα, χωρίς καθαρή δομή, που απαρτίζονται από μεγάλες συνεκτικές συνιστώσες. Παρακάτω, δίνονται τα αντίστοιχα γραφήματα στις Εικόνες 7.2 και 7.3 για τη σύγκριση του χρόνου εκτέλεσης και της Αρθρωτότητας αντίστοιχα, για όλα τα δίκτυα που παράχθηκαν με τη μέθοδο [41].



Εικόνα 7.2: *Σύγκριση Χρόνου Εκτέλεσης των δύο αλγορίθμων.*



Εικόνα 7.3: *Αρθρωτότητα που προκύπτει από την εφαρμογή των HGN και Walktrap.*

- Δίκτυα με Άγνωστες Κοινότητες

Στο Πίνακα 7.6 παρουσιάζονται οι τιμές για τον χρόνο εκτέλεσης και την Αρθρωτότητα για τους αλγορίθμους HGN και Walktrap. Ανάλογα με το τύπο του δικτύου, προκύπτουν τα παρακάτω αποτελέσματα:

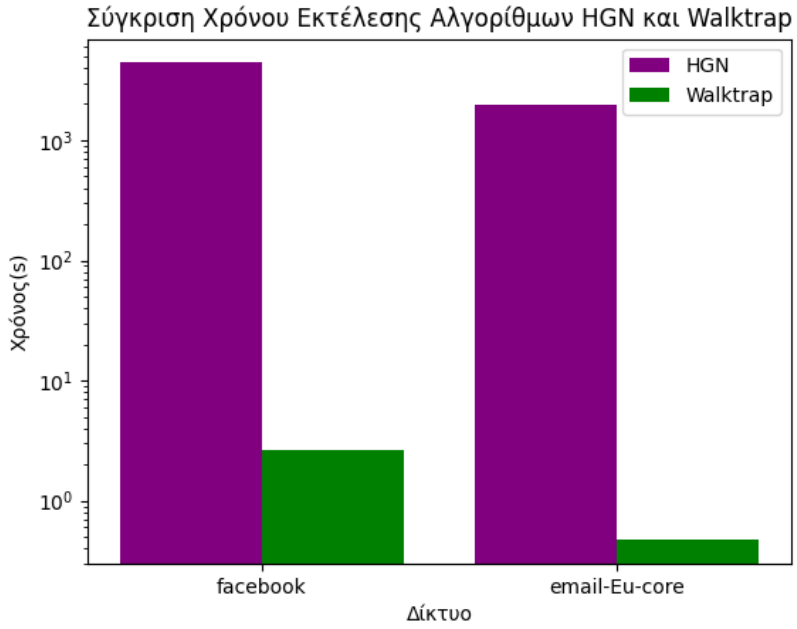
Ο χρόνος εκτέλεσης του αλγορίθμου Walktrap είναι σημαντικά ταχύτερος από το χρόνο εκτέλεσης του αλγορίθμου HGN, για όλους τους τύπους δικτύων που παρουσιάζονται παρα-

Πίνακας 7.6: *Σύγκριση Χρόνου Εκτέλεσης Αλγορίθμων HGN και Walktrap για Δίκτυα με Άγνωστες Κοινότητες.*

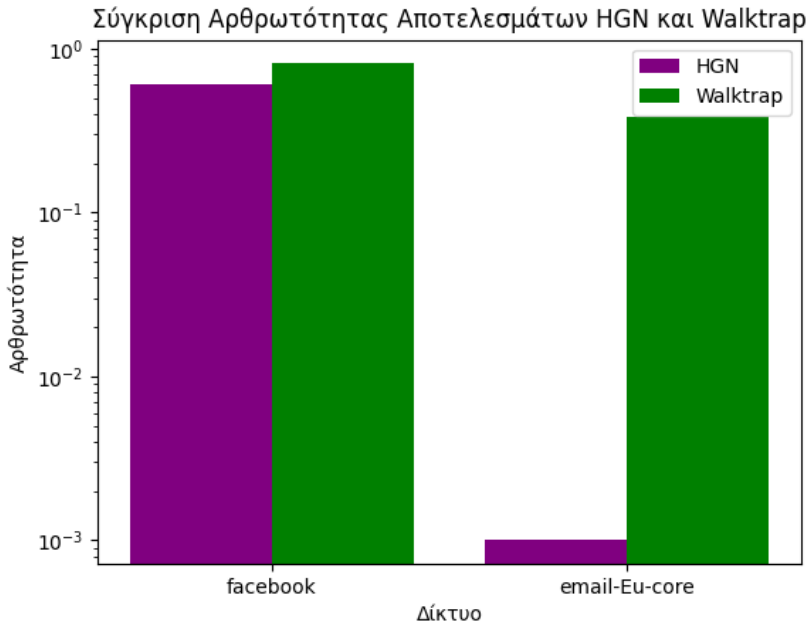
Δίκτυο	Χρόνος HGN (s)	Αρθρωτότητα HGN	Χρόνος Walktrap (s)	Αρθρωτότητα Walktrap
facebook_combined	4416,3	0,6	2,63	0,81
scf1	944,96	0,002	0,4	0,16
scf2	1343,15	0,001	0,6	0,16
scf3	572	0,027	0,73	0,18
scf4	401	0,04	0,88	0,14
scf5	82,98	0,027	0,79	0,19
scf6	1002,7	0,08	0,76	0,27
smw1	898,05	0,11	0,88	0,65
smw2	2074	0,13	0,85	0,61
smw3	173,32	0,61	0,80	0,69
smw4	103,4	0,76	0,84	0,79
smw5	126,25	0,50	0,31	0,62
smw6	78,27	0,61	0,30	0,72
rgg1	156,23	0,67	0,79	0,63
rgg2	322,18	0,5	0,32	0,49
rgg3	1170,42	0,41	0,32	0,39
rgg4	507,08	0,28	0,34	0,23

πάνω. Η ταχύτερη εκτέλεση του αλγορίθμου Walktrap έγκειται στο μειωμένο υπολογισμό των αποστάσεων μεταξύ των κορυφών, που βασίζεται στην μέθοδο του Ward [46]. Επίσης, η απόσταση μεταξύ κορυφών (μέτρο δομικής ομοιότητας) που ανήκουν στην ίδια κοινότητα, υπολογίζεται εύκολα, καθώς εξαρτάται από το μήκος των τυχαίων περιπάτων. Το συγκεκριμένο μήκος είναι σύντομο, όταν οι τυχαίοι περίπατοι παραμένουν εντός της ίδιας κοινότητας. Αντιθέτως, στην περίπτωση του HGN, ο συνολικός χρόνος εκτέλεσης αυξάνεται, λόγω του χρόνου που απαιτείται κάθε φορά για την διαδικασία της Ενσωμάτωσης Rigel. Όσο αφορά την τιμή της Αρθρωτότητας, για τα πραγματικά και τα δίκτυα Ελεύθερης Κλίμακας, καθώς τα πραγματικά κοινωνικά προσεγγίζουν το μοντέλο των δικτύων Ελεύθερης Κλίμακας, προκύπτουν λίγοι κόμβοι στο δίκτυο με υψηλό βαθμό (highly connected), οι οποίοι βρίσκονται σε πολλά συντομότερα μονοπάτια μεταξύ κόμβων. Αυτό συνεπάγεται την αφαίρεση των ακμών που συνδέουν υψηλά συνδεδεμένους κόμβους με τους γειτονικούς τους, διότι παρουσιάζουν τις μεγαλύτερες τιμές Κεντρικότητας. Έτσι, προκύπτουν κοινότητες που απαρτίζονται από μεγάλες συνεκτικές συνιστώσες, μεμονωμένους κόμβους ή μικρές ομάδες κόμβων χαμηλού βαθμού. Μία τέτοια διαμέριση οδηγεί σε χαμηλές τιμές Αρθρωτότητας και συνεπώς, η λύση του αλγόριθμου Walktrap είναι προτιμότερη. Συγκεκριμένα προτιμάται, διότι δεν οδηγεί σε μεμονωμένες κορυφές, καθώς συγχωνεύει κάθε φορά δύο κοινότητες, οδηγώντας σε μια διαμέριση που θα αποτελείται από πιο πυκνά συνεκτικούς υπογράφους. Στην Εικόνα 7.4 και στην Εικόνα 7.4 φαίνονται οι διαφορές στον χρόνο εκτέλεσης και στην παραγόμενη Αρθρωτότητα για τα πραγματικά δίκτυα Social circles: Facebook, email-Eu-core. Αντίστοιχα,

στις Εικόνες 7.6 και 7.7 παρουσιάζονται τα αποτελέσματα που αφορούν δίκτυα Ελεύθερης Κλίμακας του Πίνακα 7.6.

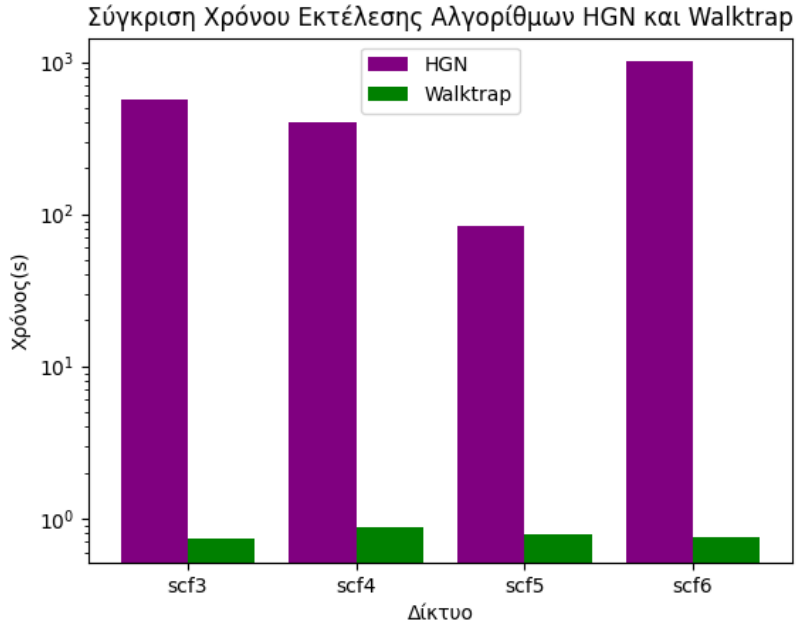


Εικόνα 7.4: *Σύγκριση Χρόνου Εκτέλεσης των δύο αλγορίθμων.*

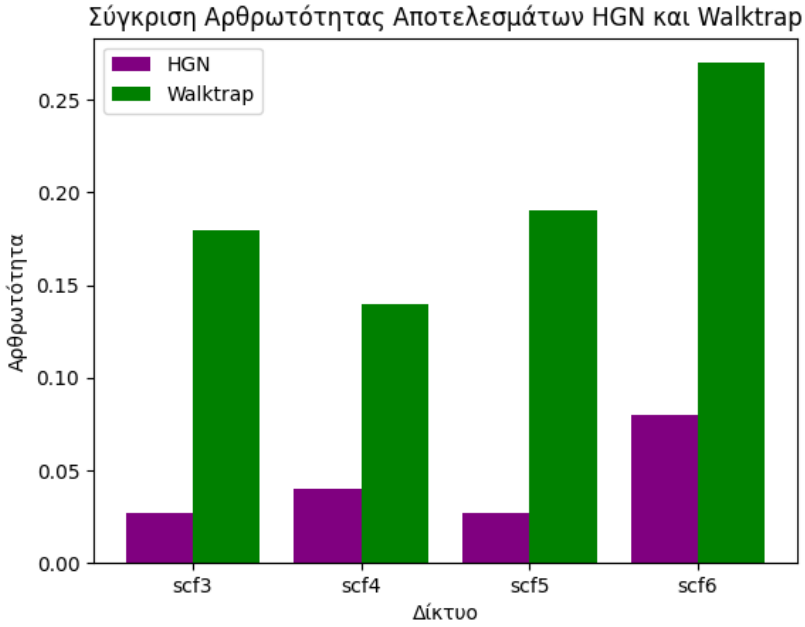


Εικόνα 7.5: *Αρθρωτότητα που προκύπτει από την εφαρμογή των HGN και Walktrap.*

Για τα δίκτυα Μικρού Κόσμου, όσον αφορά τις Αρθρωτότητες των διαμερίσεων που παράγονται, ο αλγόριθμος Walktrap δίνει ελαφρώς καλύτερες τιμές από τον HGN, οι οποίες μάλιστα είναι κοντά στις τιμές που προκύπτουν από την εφαρμογή της Μεγιστοποίησης Αρθρωτότητας. Παρακάτω, δίνονται τα αντίστοιχα γραφήματα στις Εικόνες 7.8 και 7.9 για τη σύγκριση

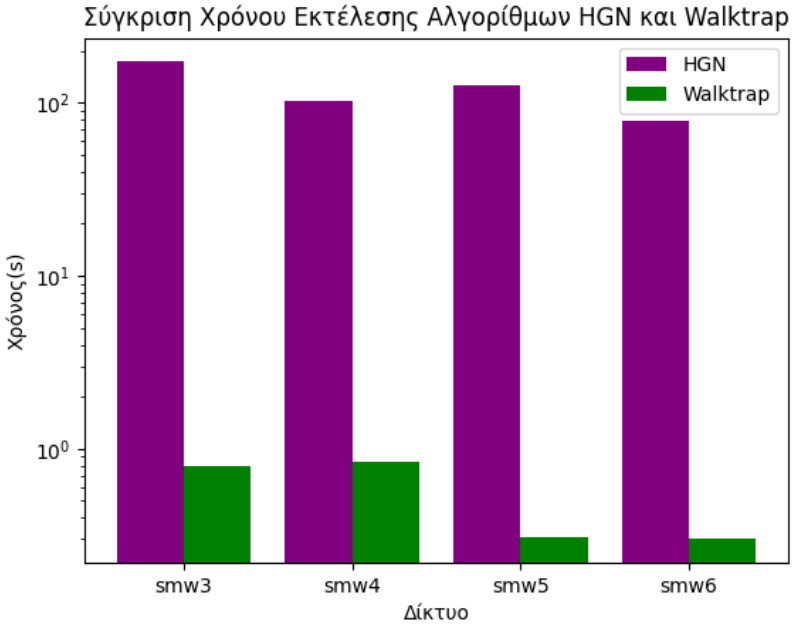


Εικόνα 7.6: *Σύγκριση Χρόνου Εκτέλεσης για 4 Δίκτυα Ελεύθερης Κλίμακας.*

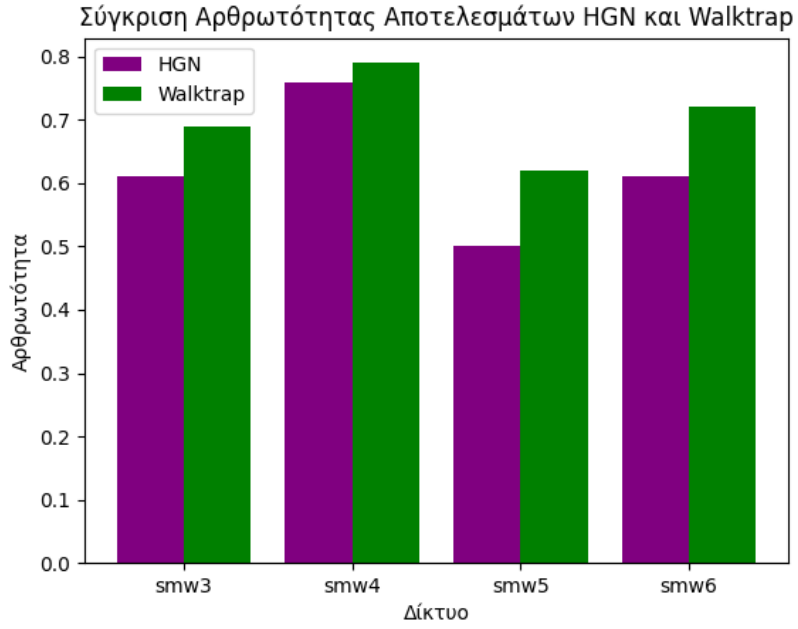


Εικόνα 7.7: *Παραγόμενη Αρθρωτότητα για τα 4 Δίκτυα Ελεύθερης Κλίμακας του Πίνακα 7.6.*

του χρόνου εκτέλεσης και της Αρθρωτότητας, αντίστοιχα.

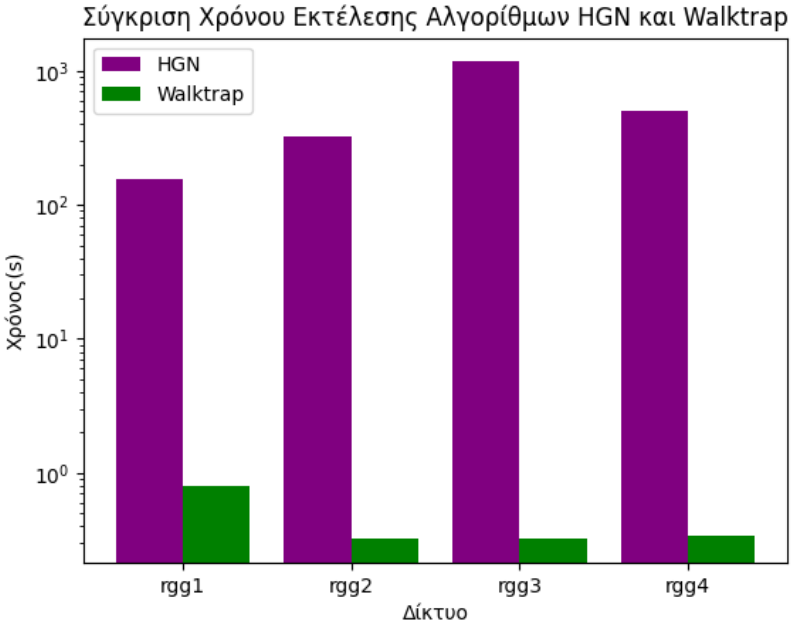


Εικόνα 7.8: *Σύγκριση Χρόνου Εκτέλεσης για 4 Δίκτυα Μικρού Κόσμου.*

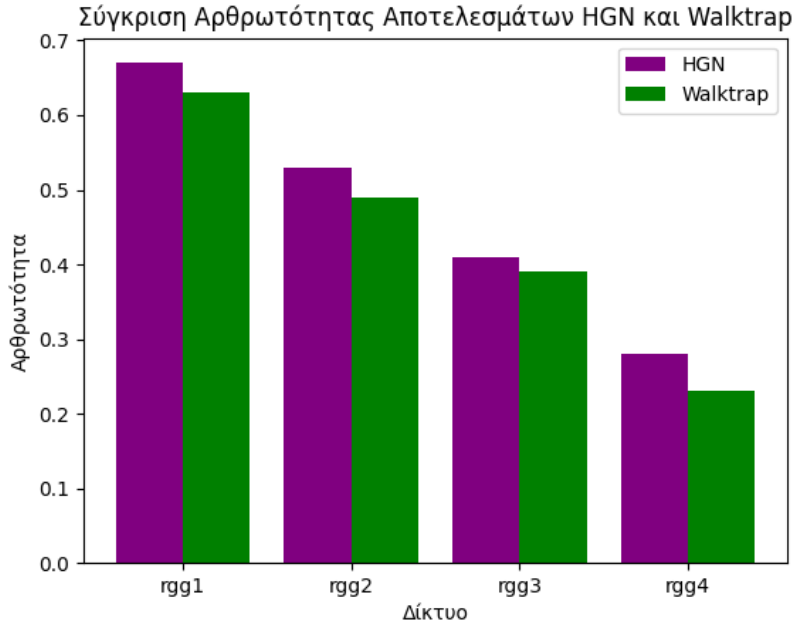


Εικόνα 7.9: *Παραγόμενη Αρθρωτότητα για τα 4 Δίκτυα Μικρού Κόσμου του Πίνακα 7.6.*

Τέλος, όσον αφορά τους Τυχαίους Γεωμετρικούς Γράφους, οι Αρθρωτότητες των διαμερίσεων που παράγονται από τους δύο αλγορίθμους, έχουν παραπλήσιες τιμές, όπως φαίνεται στην Εικόνα 7.11.



Εικόνα 7.10: *Σύγκριση Χρόνου Εκτέλεσης για 4 Τυχαίους Γεωμετρικούς Γράφους.*



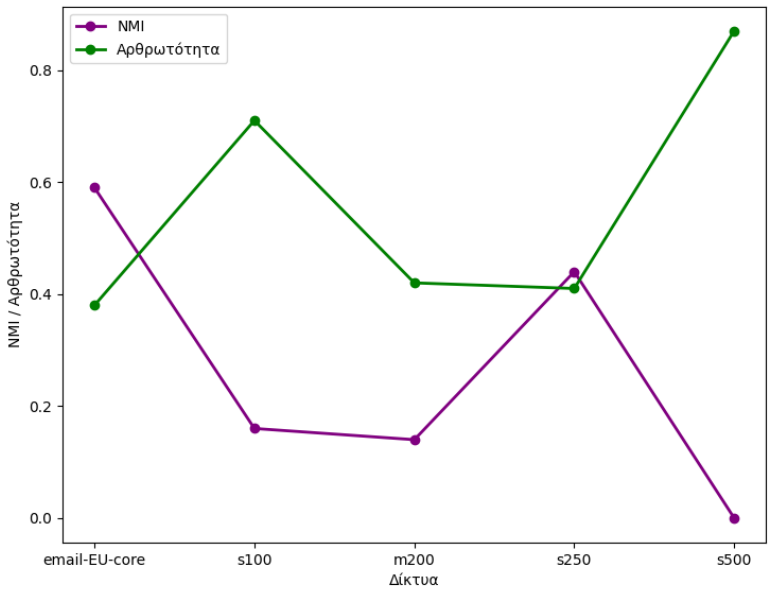
Εικόνα 7.11: *Παραγόμενη Αρθρωτότητα για τα 4 Δίκτυα Τυχαίου Γεωμετρικού Γράφου του Πίνακα 7.6.*

Άλλη μία μετρική, ως προς την οποία εκτιμήθηκε η απόδοση των δύο αλγόριθμων HGN και Walktrap και αξιολογήθηκαν οι παραγόμενες διαμερίσεις, είναι η Κανονικοποιημένη Αμοιβαία Πληροφορία (Normalized Mutual Information - NMI) [3.2.3]. Η αξιολόγηση αφορά δίκτυα για τα οποία γνωρίζουμε ποια είναι βέλτιστη διαμέριση τους. Παρατηρώντας τον Πίνακα 7.7, προκύπτει ότι ο αλγόριθμος Walktrap δίνει καλύτερα αποτελέσματα ως προς τη μετρική (NMI). Ανιχνεύει όμως, σε μικρό βαθμό την πραγματική δομή των κοινοτήτων.

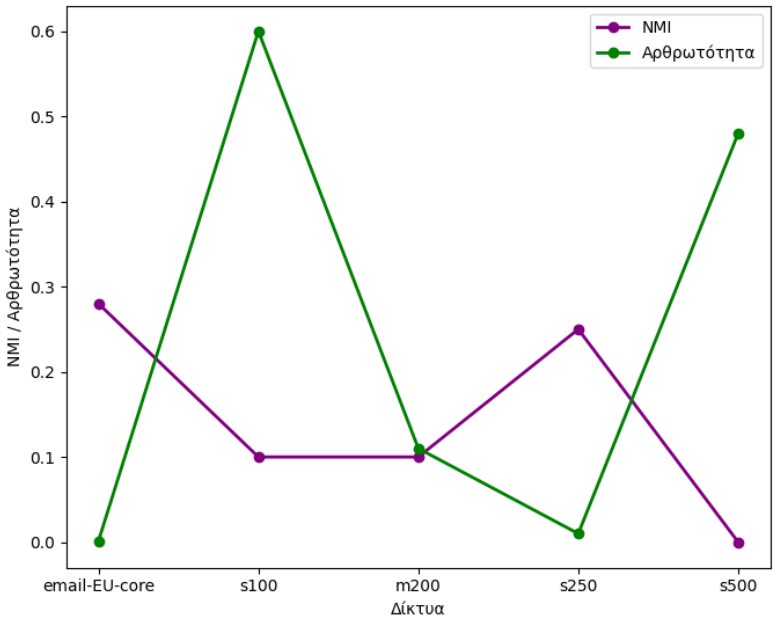
Πίνακας 7.7: *Απόδοση Αλγορίθμων HGN και Walktrap ως προς τη μετρική NMI*

Δίκτυο	NMI HGN	NMI Walktrap
email-EU-core	0,28	0,59
s100	0,1	0,16
m200	0,1	0,14
s250	0,25	0,44
s500	0	0

Ωστόσο, όπως φαίνεται στα διαγράμματα των Εικόνων 7.12 και 7.13, οι τιμές της μετρικής (NMI) είναι δυσανάλογες με τις τιμές της μετρικής της Αρθρωτότητας. Συγκεκριμένα, αυτή η παρατήρηση αφορά δίκτυα όπως το s100 και s500. Λαμβάνοντας λοιπόν υπόψιν, ότι η Αρθρωτότητα είναι μια μετρική που ποσοτικοποιεί την ποιότητα μιας δομής κοινοτήτων, με ένα "τυφλό" τρόπο, δηλαδή χωρίς τη χρήση της δομής κοινοτήτων αναφοράς, μπορεί να προκύψει το εξής αποτέλεσμα:
Δύο δομές κοινοτήτων, μπορούν ταυτόχρονα να φτάσουν σε πολύ παρόμοια αποτελέσματα και να είναι τοπολογικά πολύ διαφορετικές. Υπάρχουν δύο σημαντικές ερμηνείες σε αυτό το αποτέλεσμα. Πρώτον, μία εκτιμώμενη δομή κοινοτήτων μπορεί να φτάσει σε υψηλή τιμή Αρθρωτότητας, χωρίς κατ' ανάγκη να είναι τοπολογικά παρόμοια με τη πραγματική δομή κοινοτήτων. Δεύτερον, δύο κατ' εκτίμηση διαχωρισμοί μπορεί να φτάσουν το ίδιο σκορ, χωρίς να έχουν αυτόματα τις ίδιες τοπολογικές ιδιότητες.



Εικόνα 7.12: *Σύγκριση τιμών Αρθρωτότητας και μετρικής NMI του αλγορίθμου Walktrap, για τα 5 δίκτυα του Πίνακα 7.7*



Εικόνα 7.13: *Σύγκριση τιμών Αρθρωτότητας και μετρικής NMI του αλγορίθμου HGN, για τα 5 δίκτυα του Πίνακα 7.7*

Όπως φάνηκε από τα πειράματα που εκτελέστηκαν, ο αλγόριθμος Walktrap είναι ταχύτερος από τον HGN. Όσον αφορά την Αρθρωτότητα, επιτυγχάνει καλύτερη διαμέριση από τον HGN, για όλους τους τύπους δικτύου που εξετάστηκαν παραπάνω, εκτός που τους Τυχαίους Γεωμετρικούς Γράφους, όπου οι τιμές Αρθρωτότητας, που παράγονται και από τους δύο αλγορίθμους, είναι παραπλήσιες. Τέλος και οι δύο αλγόριθμοι παρουσιάζουν χαμηλά ποσοστά ανίχνευσης της πραγματικής δομής των κοινοτήτων, για δίκτυα των οποίων γνωρίζουμε την αρχική διαμέριση των κόμβων τους σε κοινότητες. Η συγκεκριμένη παρατήρηση έρχεται σε αντίθεση με τις υψηλές τιμές Αρθρωτότητας που επιτυγχάνει ως επί το πλείστον ο αλγόριθμος Walktrap, για τα εν λόγω δίκτυα.

Κεφάλαιο 8

Επίλογος

Στην ενότητα αυτή, επιχειρείται μία σύνοψη των αποτελεσμάτων της Διπλωματικής Εργασίας, καθώς και των συμπερασμάτων που προέκυψαν από την σύγκριση των δύο αλγορίθμων ανίχνευσης κοινοτήτων, **Walktrap** και **Hyperbolic Girvan-Newman**. Στη συνέχεια, αναφέρονται ορισμένες προτάσεις για μελλοντικές επεκτάσεις, με σκοπό τη βελτίωση του προτεινόμενου πλαισίου.

8.1 Σύνοψη και Συμπεράσματα

Στην παρούσα Διπλωματική Εργασία, σε πρώτο στάδιο επιχειρήθηκε μία θεωρητική ανασκόπηση των μοντέλων κατασκευής των τεχνητών δικτύων που χρησιμοποιήθηκαν στην διαδικασία της πειραματικής αξιολόγησης. Έπειτα μελετήθηκε ο αλγόριθμος **Hyperbolic Girvan-Newman**, μία αποδοτικότερη παραλλαγή του κλασικού Girvan-Newman για δίκτυα μεγαλύτερης κλίμακας, η οποία επιτυγχάνεται χάρη στην Ενσωμάτωση του γράφου στον Υπερβολικό Χώρο. Συγκεκριμένα χρησιμοποιεί την Ενσωμάτωση Rigel η οποία αναθέτει στους κόμβους συντεταγμένες ενός Υπερβολικού Χώρου, με διάσταση που καθορίζεται από το χρήστη. Κάνοντας χρήση της ιδιότητας αυτής, υπολογίζει την Υπερβολική Κεντρικότητα Ενδιαμεσικότητα Ακμής (Υ.Κ.Ε.Α) αντί της Κεντρικότητας Ενδιαμεσικότητας Ακμής, για κάθε ακμή του γράφου, με σκοπό την αναζήτηση μίας ταχύτερης λύσης, αφαιρώντας σε κάθε επανάληψη του αλγορίθμου ένα σύνολο ακμών με τη μεγαλύτερη τιμή (Υ.Κ.Ε.Α). Ο αριθμός των ακμών που αφαιρούνται καθορίζεται από το χρήστη και εξαρτάται από το μέγεθος του γράφου. Η επαναληπτική αυτή διαδικασία συνεχίζεται έως ότου υπάρχουν τόσες συνεκτικές συνιστώσες στο δίκτυο όσες και οι κοινότητες που ορίστηκαν από την αρχή του αλγορίθμου.

Στη συνέχεια, ο δεύτερος αλγόριθμος που επιλέχθηκε προς σύγκριση με τον παραπάνω, είναι ο **Walktrap**, ένας αλγόριθμος, επίσης ιεραρχικής συσταδοποίησης (από κάτω προς τα πάνω) που βασίζεται σε τυχαίους περιπάτους. Συγκεκριμένα, με βάση την πληροφορία που παρέχουν οι τυχαίοι περίπατοι σε ένα γράφο, ορίζεται η συνάρτηση απόστασης μεταξύ δύο οποιοδήποτε κόμβων. Ανάλογα με την τιμή αυτής της μετρικής προκύπτει, αν δύο κόμβοι βρίσκονται εντός της ίδιας κοινότητας ή όχι. Υψηλή τιμή αυτής της μετρικής σημαίνει ότι οι κόμβοι βρίσκονται σε διαφορετικές κοινότητες, ενώ χαμηλή σημαίνει ότι βρίσκονται στην

ίδια. Ο αλγόριθμος σε κάθε βήμα του παράγει μία ακολουθία διαμερίσεων, επιστρέφοντας στο τέλος μια ιεραρχική δομή που αναπαρίσταται μέσω δενδρογράμματος. Η επιλογή της βέλτιστης διαμέρισης γίνεται σύμφωνα με το κριτήριο της Αρθρωτότητας και συγκεκριμένα είναι εκείνη που μεγιστοποιεί την τιμή της Αρθρωτότητας (modularity-Q).

Τα πειράματα μας αποδεικνύουν, ότι ο αλγόριθμος **Walktrap** είναι γρηγορότερος για μεγάλα δίκτυα από τον αλγόριθμο **Hyperbolic Girvan-Newman**. Επίσης, όσον αφορά την Αρθρωτότητα των διαμερίσεων που παράγει δίνει καλύτερα αποτέλεσμα σχεδόν σε όλες τις κατηγορίες δικτύων που αναφέρθηκαν. Τέλος, και οι δύο αλγόριθμοι δεν καταφέρνουν ικανοποιητικά να εντοπίσουν τις σωστές κοινότητες, σύμφωνα με τη μετρική NMI, ενώ ταυτόχρονα μπορεί να επιτυγχάνουν μεγάλες τιμές Αρθρωτότητας, υπονοώντας την ύπαρξη μιας καλής κοινοτικής δομής. Συνεπώς, προκύπτει το συμπέρασμα ότι παραδοσιακά μέτρα αξιολόγησης κοινοτικών δομών, όπως τα προαναφερθέντα, δεν είναι απολύτως προσαρμοσμένα ούτε για την αξιολόγηση ενός αλγορίθμου ανίχνευσης κοινοτήτων σε απόλυτους όρους, ούτε για τη σύγκριση των διαφόρων αλγορίθμων.

8.2 Μελλοντικές Επεκτάσεις

Όσον αφορά τις μελλοντικές επεκτάσεις της παρούσας διπλωματικής εργασίας, ιδιαίτερο ενδιαφέρον παρουσιάζουν τα ακόλουθα:

- Στην εργασία αυτή καθ' όλη τη διάρκεια εκτέλεσης του αλγορίθμου HGN, παράμετροι εκτέλεσης, όπως η διάσταση του Υπερβολικού Χώρου, για την ενσωμάτωση Rigel παραμένουν σταθεροί. Παρόλα αυτά, ίσως αποτελεί καλύτερη προσέγγιση η δυναμική ρύθμιση τους, ανάλογα με τα δίκτυα που προκύπτουν κατά την διάρκεια της ομαδοποίησης. Μία τέτοια προσέγγιση ενδεχομένως θα οδηγούσε σε ταχύτερη ολοκλήρωση του αλγορίθμου, αφού οι ακμές με τη μεγαλύτερη Κεντρικότητα θα αφαιρούνταν συντομότερα. Επίσης, βελτίωση του χρόνου εκτέλεσης θα μπορούσε να επιτευχθεί με το βέλτιστο καθορισμό του μεγέθους Δέσμης, ανάλογα με το τύπου του δικτύου.
- Μια παραλλαγή του μέτρου της Αρθρωτότητας, όπου δεν θα λαμβάνει υπόψιν την δομή κοινοτήτων μόνο ως απλά έναν διαχωρισμό κόμβων, αλλά θα κάνει και χρήση της πληροφορία της τοπολογίας του γράφου, θα μπορούσε να οδηγήσει στον εντοπισμό αλληλοεπικαλυπτόμενων κοινοτήτων, επιφέροντας μία καλύτερη διαμέριση.
- Ο αλγόριθμος Walktrap ως έχει, εφαρμόζεται πάνω σε μη-κατευθυνόμενα γραφήματα. Μία ενδιαφέρουσα κατεύθυνση για περαιτέρω έρευνα, θα αποτελούσε η εφαρμογή του σε κατευθυνόμενους γράφους, όπου οι τυχαίοι περίπατοι συμπεριφέρονται εντελώς διαφορετικά.

Παραρτήματα

Παράρτημα Α΄

Πηγαίος Κώδικας

Στα παρακάτω κομμάτια κώδικα έγινε χρήση των βιβλιοθηκών NetworkX (2.7.1) [48], python-igraph (0.9.9) [49] και scikit-learn 1.0.2 [50].

Α΄.0.1 Κώδικας παραγωγής LFR δικτύων και εξαγωγής των κοινοτήτων τους

```
from itertools import combinations, groupby
import random
import networkx as nx
import numpy as np
from igraph import *
import igraph as graph
from networkx.generators.community import LFR_benchmark_graph
from pyvis.network import Network

n = 200
tau1 = 3
tau2 = 2.6
mu = 0.9
G = LFR_benchmark_graph(n, tau1, tau2, mu, average_degree=5, min_community=20, seed=10)

#----Get communities from the node attributes of the graph---#
communities = {frozenset(G.nodes[v]["community"]) for v in G}
clusters = [list(x) for x in communities]
print(len(G.edges()))
print(len(clusters))
g = graph.Graph()
g.add_vertices(G.nodes())
g.add_edges(G.edges)
membership = [None] * g.vcount()
for c, cluster in enumerate(clusters):
    for v in cluster:
        membership[v] = c

clustering = VertexClustering(g, membership=membership)
```

Εικόνα Α΄.1: *Αλγόριθμος παραγωγής LFR δικτύων.*

Α΄.0.2 Η μέθοδος Modularity Maximization

Παρακάτω δίνεται ο κώδικας της μεθόδου Modularity Maximization εφαρμοσμένη σε Τυχαίους Γεωμετρικούς Γράφους, δίκτυα Ελεύθερης Κλίμακας και δίκτυα Μικρού Κόσμου:

```
import networkx as nx
from networkx.algorithms.community import greedy_modularity_communities
import networkx.algorithms.community as nx_com
import numpy as np
import random

# -----Rgg Networks-------#
n = 100
R = 0.3
f = nx.random_geometric_graph(n, R)

# ---- Scale Free networks-----#
n = 100
d = 6
f = nx.barabasi_albert_graph(n, d)

# small world
n = 100
d = 5
p = 0.1
seedno = 5
my = np.random.RandomState(seedno)
f = nx.watts_strogatz_graph(n, d, p, my)

print(nx.is_connected(f))
c = greedy_modularity_communities(f)
mod = nx_com.modularity(f, c)
print("Number of communities: ", len(c))
print("Modularity :", mod)
```

Εικόνα Α΄.2: *Η μέθοδος Modularity Maximization εφαρμοσμένη σε τεχνητά σύνθετα δίκτυα [8].*

Α΄.0.3 Βοηθητικός κώδικας για τον αλγόριθμο HGN

Το παρακάτω κομμάτι κώδικα καθιστά ένα γράφο συνδεδεμένο, πράγμα απαραίτητο για την εκτέλεση του αλγορίθμου HGN:

```
import networkx as nx
import sys
import random
from itertools import combinations, groupby

import numpy as np
from pyvis.network import Network

filename = sys.argv[1]
G = nx.read_edgelist(filename, nodetype=int)
G.remove_edges_from(nx.selfloop_edges(G))

components = dict(enumerate(nx.connected_components(G)))
components_combs = combinations(components.keys(), r=2)

for _, node_edges in groupby(components_combs, key=lambda x: x[0]):
    node_edges = list(node_edges)
    random_comps = random.choice(node_edges)
    source = random.choice(list(components[random_comps[0]]))
    target = random.choice(list(components[random_comps[1]]))
    G.add_edge(source, target)
```

Εικόνα Α΄.3: *Παραγωγή Συνδεδεμένου Γράφου*

Α΄.0.4 Ο αλγόριθμος Walktrap

Παρακάτω δίνεται στιγμιότυπο του κώδικα εκτέλεσης του αλγορίθμου Walktrap, εφαρμοσμένος σε Πραγματικά Δίκτυα και Τυχαίους Γεωμετρικούς Γράφους:

```
import sys
import time

start_time = time.time()
import networkx as nx
import igraph
from sklearn.metrics.cluster import normalized_mutual_info_score
from igraph import *

# ---- Real Datasets ----- #
filename = sys.argv[1]
groundfile = sys.argv[2]
# G = nx.read_edgelist(filename, nodetype=int)
f = Graph.Read_Edgelist(filename, directed=False)
v = f.community_walktrap()
clusters = v.as_clustering()
print(clusters.modularity)

# ------- RGG model ------ #
n = 100
R = 0.2
f = nx.random_geometric_graph(n, R)
g = igraph.Graph(directed=True)
g.add_vertices(f.nodes())
g.add_edges(f.edges())
v = g.community_walktrap()
clusters = v.as_clustering()
print(clusters.modularity)
```

Εικόνα Α΄.4: *Κώδικας Αλγορίθμου Walktrap*

Α΄.0.5 Υπολογισμός Μετρικής NMI

Παρακάτω δίνεται ο κώδικας υπολογισμού της μετρικής NMI, ο οποίος δέχεται ως είσοδο δύο λίστες, εκ των οποίων, η μία περιέχει την εκτιμώμενη δομή κοινοτήτων (cdlist) και η άλλη τη δομή κοινοτήτων αναφοράς (groundl):

```
#-----NMI calculation----#
from sklearn.metrics.cluster import normalized_mutual_info_score

from igraph import *

filecd = sys.argv[1]
groundfile = sys.argv[2]

with open(filecd) as file:
    lines1 = [line.rstrip() for line in file]

with open(groundfile) as file:
    lines2 = [line.rstrip() for line in file]

cdlist = list(map(int, lines1))
groundl = list(map(int, lines2))

NMI = normalized_mutual_info_score(groundl, cdlist)
print("NMI:", NMI)
```

Εικόνα Α΄.5: *Κώδικας Υπολογισμού Μετρικής NMI*

Παράρτημα Β΄

Κατηγορίες Βιβλιογραφικών Αναφορών

Τύπος βιβλιογραφικής πηγής	Αριθμός αναφοράς
Βιβλίο ξενόγλωσσο	[10], [11], [35], [39], [13], [27]
Βιβλίο ελληνικό	[26]
Άρθρο σε επιστημονικό περιοδικό	[14], [9], [15], [17], [18], [19], [20], [21], [22], [31], [46], [7], [42], [43], [3], [32], [36], [34], [24], [37], [6], [41], [38], [30], [2], [28], [5], [1]
Παρουσίαση σε επιστημονικό συνέδριο	[12], [16], [23], [25], [44], [45] [40]
Ιστοσελίδα	[47], [8], [4], [29], [48], [49], [50]

Βιβλιογραφία

[1] V.Nicosia, G.Mangioni, V.Carchiolo και M.Malgeri. *Extending the definition of modularity to directed graphs with overlapping communities*. *Journal of Statistical Mechanics Theory and Experiment*, 4(03), 2008.

[2] Alberto Costa. *Some remarks on modularity density*. *Social and Information Networks (cs.SI)? Physics and Society (physics.soc-ph)*, 1(1409), 2014.

[3] E. N. Gilbert. *Random Graphs*. *Ann. Math. Statist.*, 30(4):1141 – 1144, 1959.

[4] *Random Geometric Graph*. https://commons.wikimedia.org/wiki/File:Random_Geometric_Graph.gif.

[5] ALBERT LÁSZLÓ BARABÁSI. *Scale-Free Networks: A Decade and Beyond*. *SCIENCE*, 325(5939):412–413, 2009.

[6] Brandes. U. *On variants of shortest-path betweenness centrality and their generic computation*. *Social Networks*, 30(2):136–145, 2008.

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

[8] *Networkx modularity maximization*. https://networkx.org/documentation/networkx-2.4/reference/algorithms/generated/networkx.algorithms.community.modularity_max.greedy_modularity_communities.html#networkx-algorithms-community-modularity-max-greedy-modularity-communities.

[9] Vasileios Karyotis, Konstantinos Tsitseklis, Konstantinos Sotiropoulos και Symeon Papavassiliou. *Big Data Clustering via Community Detection and Hyperbolic Network Embedding in IoT Applications*. *Sensor Networks*, 18(4):1205, 2018.

[10] S.Wasserman και K.Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[11] Kadushin Charles. *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford University Press, 2011.

[12] M. E. J. Newman. *The structure and function of complex networks*. *SIAM Review 45, 167-256*, 2003.

[13] Vasileios Karyotis, Eleni Stai και Symeon Papavassiliou. *Evolutionary Dynamics of Complex Communications Networks*. CRC Press, 2014.

[14] Konstantinos Tsitseklis, Maria Krommyda, Vasileios Karyotis, Verena Kantere και Symeon Papavassiliou. *Scalable Community Detection for Complex Data Graphs via Hyperbolic Network Embedding and Graph Databases*. *IEEE Transactions on Network Science and Engineering*, 8(21016248):1269 – 1282, 2020.

[15] K. Krishna και M. N. Murty. *Genetic K-means algorithm*. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3):433–439, 1999.

[16] H. P. Kriegel και M. Pfeifle. *Density-based clustering of uncertain data*. *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 672-677*, 2005.

[17] Z. Wang, Y. Zhao, Z. Chen και Q. Niu. *An Improved Topology-Potential-Based Community Detection Algorithm for Complex Network*. *The Scientific World Journal*, 2014(12):121609, 2014.

[18] Y. Han, D. Li και T. Wang. *Identifying different community members in complex networks based on topology potential*. *Frontiers of Computer Science in China*, 2011(5):87–99, 2011.

[19] J.-P.Zhang, H.-B.Li, J.Yang και J.-B.Bai. *Community discovery method with uncertainty measure of overlapping nodes based on topological potential*. *Journal of Harbin Institute of Technology (New Series)*, 19(2):16–22, 2012.

[20] B.Yang, J.Di, J.Liu και D.Liu. *Hierarchical community detection with applications to real-world network analysis*. *Data Knowledge Engineering*, 83(2013):20–38, 2013.

[21] Andrea Lancichinetti1, Santo Fortunato και János Kertész. *Detecting the overlapping and hierarchical community structure in complex networks*. *New Journal of Physics*, 11(3):033015, 2009.

[22] X.S.Zhang, R.S.Wang, Y.Wang, J.Wang, Y.Qiu, L.Wang και L.Chen. *Modularity optimization in community detection of complex networks*. *EPL (Europhysics Letters)*, 87(3):38002, 2009.

[23] W. Wang και C. Li. *A Core-based Community Detection Algorithm for Networks*. *2010 International Conference on Computational Aspects of Social Networks, 607-610*, 2010.

[24] ALBERT LÁSZLÓ BARABÁSI και RÉKA ALBERT. *Emergence of Scaling in Random Networks*. *SCIENCE*, 286(5439):509–512, 1999.

[25] Pons P. και Latapy M. *Computing communities in large networks using random walks*. *Computer and Information Sciences - ISCIS 2005, 284-293*, 2005.

[26] G. Manolopoulos. *ΜΑΘΗΜΑΤΑ ΘΕΩΡΙΑΣ ΓΡΑΦΩΝ¨*. ΕΚΔΟΣΕΙΣ ΝΕΩΝ ΤΕΧΝΟΛΟΓΙΩΝ, 1996.

[27] Douglas Brent West. *Introduction to Graph Theory*. Pearson College Div? Subsequent edition, 2000.

[28] JOHN W. ESSAM και MICHAEL E. FISHER. *Some Basic Definitions in Graph Theory*. *American Physical Society*, 42(2):271, 1970.

[29] *yEd graph editor*. https://www.yworks.com/products/yed.

[30] Santo Fortunato και Claudio Castellano. *Community Structure in Graphs*. *Physics and Society (physics.soc-ph)? Statistical Mechanics (cond-mat.stat-mech)? Computational Physics (physics.comp-ph)*, 1, 2007.

[31] M. E. J. Newman και M. Girvan. *Finding and evaluating community structure in networks*. *Physical Review E*, 69(026113), 2004.

[32] P. Erdös και A. Rényi. *On Random Graphs I*. *Publicationes Mathematicae*, σελίδες 290 – 297, 1959.

[33] *Complex Network Topologies and Applications*. https://helios.ntua.gr/pluginfile.php/118125/mod_resource/content/4/SNA_Lecture02_complex_topologies.pdf.

[34] Watts DJ και Strogatz SH. *Collective dynamics of 'small-world' networks*. *Nature*, 393(6684):440–442, 1998.

[35] Mathew Penrose. *Random geometric graphs*. Oxford?New York :Oxford University Press, 2003.

[36] A. Antonioni και M. Tomassini. *Degree correlations in random geometric graphs*. *Physical Review E*, 86(3):037101, 2012.

[37] R. D. Luce και A. D. Perry. *A method of matrix analysis of group structure*. *Psychometrika*, 14(1):95–116, 1949.

[38] A. Strehl και J. Ghosh. *Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions*. *Journal of Machine Learning Research*, 3(3):583–617, 2002.

[39] T. M. Cover και J. A. Thomas. *Elements of Information Theory*. John Wiley Sons, 2012.

[40] Ana L. N. Fred και Arjun Jain. *Robust data clustering*. *Conference: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on Volume: 2*, 2003.

[41] A. Lancichinetti1, S. Fortunato και Radicchi F. *Benchmark graphs for testing community detection algorithms*. *Physical review E*, 78(4):046110, 2008.

[42] M. E. J. Newman. *Fast algorithm for detecting community structure in networks*. *Phys. Rev. E 69*, 69(6):066133, 2003.

[43] V.D. Blondel, J. L. Guillaume, R. Lambiotte και E. Lefebvre. *Fast unfolding of communities in large networks*. *Journal of Statistical Mechanics Theory and Experiment*, 10:10008, 2008.

[44] X. Zhao, A. Sala, H. Zheng και B. Y. Zhao. *Efficient shortest paths on massive social graphs*. *7th International Conference on Collaborative Computing: Networking, Applications and Worksharing (Collaborate- Com), pp. 77-86, IEEE*, 2011.

[45] Eleni Stai, Konstantinos Sotiropoulos, Vasileios Karyotis και Symeon Papavassiliou. *Hyperbolic Embedding for Efficient Computation of Path Centralities and Adaptive Routing in Large-Scale Complex Commodity Networks*. *IEEE Transactions on Network Science and Engineering, 140-153*, 2017.

[46] J. H. Ward. *Hierarchical grouping to optimize an objective function*. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[47] *Stanford Large Network Dataset Collection*. http://snap.stanford.edu/data/index.html#communities.

[48] *NetworkX Library*. https://networkx.org/documentation/stable/reference/index.html.

[49] *python-igraph*. https://igraph.org/python/api/latest/.

[50] *scikit-learn*. https://scikit-learn.org/stable/user_guide.html.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

βλπ	βλέπε
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
BPF	Band Pass Filter
HGN	Hyperbolic Girvan-Newman
NMI	Normalized Mutual Information

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
αδερφός	sibling
αμεταβλητότητα	idempotency
ανάκτηση πληροφορίας	information retrieval
αντιμεταθετικότητα	commutativity
απόγονος	descedant
απορρόφηση	absorption
βάση δεδομένων	database
γνώρισμα	attribute
διαπροσωπεία	interface
διαφορά	difference
δικτυακός κατάλογος	portal catalog
δικτυωτή δομή	lattice
δομικές επερωτήσεις	structural queries
δομικές σχέσεις	structural relationships
δομικό σχήμα	schema
εγκυρότητα	validity
ένωση	union

Ευρετήριο ελληνικών όρων

Αμοιβαία, 30
Αρθρωτότητα, 33, 45, 47, 50, 53, 58, 60
ακμή-ων, 15, 17, 19, 20, 22, 23, 26, 27, 30, 34, 39, 41, 52, 59, 60
αλληλοεπικαλυπτόμενες, 24, 60

Ενδιαμεσικότητα, 30, 35, 38

Γράφος, 19

Υπερβολικό Χώρο, 16
Υπερβολικός Χώρος, 1, 17, 18, 36, 37, 39, 59, 60

Κανονικοποιημένη, 32
κεντρικότητα, 17, 25, 30, 37–39, 52, 59, 60
κοινότητες, 15, 22–24, 33, 35, 39, 43–45, 47, 49, 50, 52, 59, 60
κόμβοι, 1, 15, 16, 19, 23, 24, 26–29, 33, 37, 38, 52, 59

Μεγιστοποίηση, 34, 47, 53
μοντέλο, 17, 23, 25–27, 29, 48, 52, 59

ομαδοποίηση, 17, 18, 25, 27, 29, 30, 47, 50, 60
ομοιότητα, 15, 17, 22, 23, 30, 31, 36, 42, 52

πιθανότητα, 23, 26–31, 42–44, 49

Σύνθετα Δίκτυα, 1, 15, 17, 25, 29, 36, 47, 48

Τυχαίοι Περίπατοι, 1, 41, 42, 52, 59, 60

υπογράφος, 20, 22, 52

Ευρετήριο ξενόγλωσσων όρων

Barabasi-Albert, 25, 29
Betweeness, 35, 37

communities, 3, 24

degree, 20
density, 19
directed, 20, 24

graph, 19

Hyperbolic Girvan-Newman, 1, 17, 35, 39, 46, 50, 59

modularity, 16, 33, 34, 45, 47, 60

nodes, 19

RGG, 26
Rigel, 35, 36, 39, 52, 59, 60

scale-free, 28
subgraph, 20

undirected, 20

vertices, 19

Walktrap, 1, 17, 41, 44, 46, 50, 51, 53, 58, 59
Watts-Strogatz, 25, 27