Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΕΘΟΔΟΙ ΕΝΤΟΠΙΣΜΟΥ ΚΟΙΝΩΝΙΩΝ ΣΕ ΔΙΚΤΥΑ

Ναπολέων Κωνσταντινόπουλος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς

Νοέμβριος 2018

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Κούτρας Μάρκος (Επιβλέπων)
- Επίκουρος Καθηγητής Ευαγγελάρας Χαράλαμπος
- Επίκουρος Καθηγητής Πελέκης Νικόλαος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

U N I V E R S I T Y O F P I R A E U S

DEPARTMENT OF STATISTICS

AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

NETWORK COMMUNITY DETECTION TECHNIQUES

By

Napoleon Konstantinopoulos

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

November 2018

Στην οικογένειά μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κ. Μ. Κούτρα, αρχικά διότι δέχθηκε να αναλάβει την επίβλεψη της παρούσας εργασίας αλλά και διότι η καθοδήγηση και οι συμβουλές του, ήταν πολύτιμες κατά τη διάρκεια εκπόνησής της. Ευχαριστώ, επίσης, τα μέλη της τριμελούς επιτροπής, τον Επίκουρο Καθηγητή κ. Χ. Ευαγγελάρα και τον Επίκουρο Καθηγητή κ. Ν. Πελέκη, για το χρόνο που αφιέρωσαν στη μελέτη και διόρθωση της εργασίας αυτής.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου και τον αδελφό μου, η συμπαράσταση των οποίων υπήρξε καθοριστική για μένα καθ' όλη τη διάρκεια των σπουδών μου (προπτυχιακών και μεταπτυχιακών) αλλά και κατά την εκπόνηση της εργασίας αυτής.

Περίληψη

Ο εντοπισμός κοινωνιών σε δίκτυα είναι ένα απο τα πιο δημοφιλή θέματα της σύγχρονης επιστήμης δικτύων. Ένα απο τα πιο βασικά χαρακτηριστικά ενός δικτύου, το οποίο αναπαρίσταται απο έναν γράφο, είναι η κοινοτική του δομή. Ο εντοπισμός της κοινοτικής δομής σε ένα δίκτυο ή συσταδοποίησή του, δηλαδή η οργάνωση των κορυφών του σε ομάδες, με πολλές ακμές που ενώνουν κορυφές της ίδιας ομάδας και συγκριτικά με αυτές λιγότερες ακμές που ενώνουν κορυφές διαφορετικών ομάδων, είναι ένα θέμα που μελετάται και εξελίσσεται διαρκώς μέχρι και σήμερα. Σε επιστήμες όπως κοινωνιολογία, βιολογία ή επιστήμη των υπολογιστών όπου τα δίκτυα αναπαρίστανται ως γράφοι, ο εντοπισμός κοινοτήτων σε δίκτυα είναι μεγάλης σημασίας. Το πρόβλημα της συσταδοποίησης δικτύου ή γράφου, είναι ένα σύγχρονο επιστημονικό αντικείμενο, το οποίο έχει μελετηθεί απο διαφορετικές επιστημονικές σκοπιές. Στην παρούσα εργασία θα προσπαθήσουμε να κάνουμε μια διεξοδική ανάλυση του θέματος, απο τον ορισμό των βασικών συστατικών του, πάνω στα οποία δεν υπάρχει κοινώς αποδεκτό πρωτόκολλο, μέχρι την ανάλυση πολλών μεθόδων που έχουν αναπτυχθεί ανα τα χρόνια, ξεκινώντας απο παραδοσιακές μεθόδους και καταλήγοντας σε σύγχρονες. Στο τέλος θα κάνουμε μια σύγκριση αλγορίθμων απο διάφορες κατηγορίες μεθόδων που αναλύθηκαν, χρησιμοποιώντας συνθετικά δεδομένα (τεχνητά δίκτυα αναφοράς).

Abstract

Community detection in networks is one of the most popular topics of modern network science. One of the most basic features of networks, which are represented by graphs, is their community structure. The underlying of community structure in a network, or clustering, meaning the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively less edges joining vertices of different clusters, is a scientific subject which is thoroughly studied and hasn't stop evolving until today. Network community detection is of great importance in sociology, biology or computer science where networks are represented as graphs. The problem of community detection in networks or graphs, is a modern scientific subject, studied from various scientific aspects. In this thesis, we will attempt an in-depth analysis of the subject, from the definition of its main elements for which there is no universal protocol, to the exposition of many methods for community detection which have been proposed throughout the years. In the end we will compare algorithms coming out of different methods we analyzed, using benchmark graphs.

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1. Εισαγωγή

Ένα σύνθετο δίκτυο είναι ένα μαθηματικό μοντέλο φαινομένων που αλληλεπιδρούν και λαμβάνουν χώρα στον πραγματικό κόσμο, πάνω στο οποίο έχει αναπτυχθεί μια ισχυρή υπολογιστική βάση για την ανάλυση τέτοιων αλληλεπιδρόντων φαινομένων. Ένα κρίσιμο πρόβλημα, το οποίο έχει μελετηθεί ευρέως στη βιβλιογραφία απο την πρώιμη ανάλυση σύνθετων δικτύων, είναι ο εντοπισμός των κοινοτήτων που είναι κρυμμένες στη δομή των δικτύων αυτών. Μια κοινότητα (community), ή συστάδα (cluster), ή τμήμα (module) είναι διαισθητικά αντιληπτή ως ένα σύνολο οντοτήτων, στο οποίο κάθε οντότητα είναι πιο κοντά, υπο την έννοια του δικτύου, με άλλες οντότητες εντός της κοινότητας παρά με εκείνες που είναι εκτος αυτής. Επομένως οι κοινότητες είναι ομάδες οντοτήτων που πιθανόν μοιράζονται ορισμένες κοινές ιδιότητες ή/και παίζουν παρόμοιο ρόλο στο εκάστοτε φαινόμενο αλληλεπίδρασης που αναπαρίσταται. Κοινότητες εμφανίζονται σε πολλά σύνθετα δίκτυα επιστημονικών κλάδων, όπως της βιολογίας, της πληροφορικής, της μηχανικής, των οικονομικών, της πολιτικής και μπορεί να ανιστοιχούν, μεταξύ άλλων, σε ομάδες σελίδων του παγκόσμιου ιστού, σε λειτουργικές ενότητες, όπως κύκλοι και μονοπάτια σε μεταβολικά δίκτυα ή σε ομάδες συνδεδεμένων ατόμων σε κοινωνικά δίκτυα.

Ο εντοπισμός κοινοτήτων είναι πολύ σημαντικός και μπορεί να έχει σαφείς εφαρμογές. Για παράδειγμα, η συσταδοποίηση πελατών με κοινά ενδιαφέροντα, οι οποίοι είναι γεωγραφικά κοντά ο ένας στον άλλο, μπορεί να βελτιώσει την απόδοση υπηρεσιών που παρέχονται στον Παγκόσμιο Ιστό, υπο την έννοια οτι κάθε συστάδα πελατών θα μπορούσε να εξυπηρετείται απο έναν ειδικό διακομιστή «καθρέφτη» (mirror server) (Krishnamurthy & Wang, 2000). Επίσης ο εντοπισμός πελατών με κοινά ενδιαφέροντα σε ένα δίκτυο σχέσεων αγοράς ανάμεσα σε πελάτες και προϊόντα online καταστημάτων (όπως πχ η Amazon.com) μπορεί να ενεργοποιήσει την κατασκευή αποτελεσματικών συστημάτων συστάσεων (recommendation systems) (Reddy et al., 2002), τα οποία καθοδηγούν τους πελάτες μέσω καταλόγων αντικειμένων των λιανοπωλητών, βελτιώνοντας τις επιχειρηματικές ευκαιρίες. Συστάδες μεγάλων γράφων, μπορούν να χρησιμοποιηθούν για την κατασκευή δομών δεδομένων, ωστε να αποθηκεύονται αποτελεσματικά τα δεδομένα του γράφου και τα ερωτήματα πλοήγησης, όπως αναζητήσεις μονοπατιών, να είναι εύκολα διαχειρίσιμα.

Η κοινοτική ανίχνευση είναι σημαντική και για άλλους λόγους. Ο προσδιορισμός τμημάτων και των ορίων τους επιτρέπει μια ταξινόμηση των κορυφών σύμφωνα με τη διαρθρωτική τους

θέση στα τμήματα αυτά. Συνεπώς, κορυφές με κεντρική θέση στις συστάδες τους, δηλαδή κορυφές που μοιράζονται ένα μεγάλο αριθμό ακμών με τις άλλες κορυφές της συστάδας, έχουν σημαντικό ρόλο ελέγχου και σταθερότητας εντος της κοινότητάς τους. Απο την άλλη πλευρά κορυφές που βρίσκονται στα όρια μεταξύ μεταξύ των κοινοτήτων, διαδραματίζουν σημαντικό ρόλο διαμεσολαβητή και καθοδηγούν τις σχέσεις και τις ανταλλαγές μεταξύ διαφορετικών κοινοτήτων. Τέτοιου είδους ταξινόμηση των κορυφών είναι σημαντική σε κοινωνικά και μεταβολικά δίκτυα.

Μια άλλη σημαντική πτυχή που σχετίζεται με την κοινοτική δομή, είναι η ιεραρχική οργάνωση που εμφανίζεται στα περισσότερα δίκτυα στον πραγματικό κόσμο. Τα πραγματικά δίκτυα, αποτελούνται συνήθως απο κοινότητες στις οποίες συμπεριλαμβάνονται άλλες μικρότερου μεγέθους, που με τη σειρά τους συμπεριλαμβάνουν μικρότερες και ούτω καθεξής. Το ανθρώπινο σώμα είναι ένα παράδειγμα τέτοιας ιεραρχικής οργάνωσης. Αποτελείται απο όργανα, τα όργανα αποτελούνται απο ιστούς, οι ιστοί απο κύτταρα. Παρόμοια ιεραρχική οργάνωση συναντάμε στις επιχειρήσεις οι οποίες χαρακτηρίζονται απο οργάνωση προτύπου πυραμίδας, από τους εργαζόμενους στον πρόεδρο, με ενδιάμεσα επίπεδα που αντιστοιχούν σε τμήματα εργασίας και διαχείρισης. Η ιεραρχία παίζει σημαντικό ρόλο στη δομή και εξέλιξη σύνθετων συστημάτων. Η παραγωγή και εξέλιξη ενός συστήματος το οποίο είναι οργανωμένο σε αλληλένδετα σταθερά υποσυστήματα είναι πολύ γρηγορότερη απο το αν το σύστημα ήταν μη δομημένο, διότι είναι πολύ πιο εύκολο να συγκεντρωθούν τα μικρότερα υποτμήματα πρώτα και να χρησιμοποιηθούν ως δομικά στοιχεία για να επιτευχθούν μεγαλύτερες δομές, μέχρι να συναρμολογηθεί όλο. Με αυτόν τον τρόπο είναι επίσης πολύ πιο δύσκολο να συμβούν λάθη κατά την εξέλιξη της διαδικασίας.

Ο εντοπισμός κοινοτήτων βρίσκεται σε αναλογία με το πρόβλημα συσταδοποίησης, το οποίο είναι ένα παραδοσιακό έργο εξόρυξης δεδομένων. Στην εξόρυξη δεδομένων, η συσταδοποίηση είναι μια διαδικασία μη εποπτευόμενης μάθησης, η οποία στοχεύει στο να δημιουργήσει διαμερίσεις μεγάλων συνόλων δεδομένων σε ομοιογενείς ομάδες (συστάδες). Στην πραγματικότητα, ο εντοπισμός κοινοτήτων σε δίκτυα μπορεί θα θεωρηθεί μια διαδικασία εξόρυξης δεδομένων σε γράφους, για αυτό και ονομάζεται συσταδοποίηση δικτύου η γράφου (network/graph clustering). Ο στόχος της κοινοτικής ανίχνευσης σε γράφους είναι να προσδιοριστούν οι κοινότητες και ενδεχομένως η ιεραρχική τους οργάνωση, χρησιμοποιώντας μόνο τις πληροφορίες που κωδικοποιούνται στην τοπολογία του γράφου. Το πρόβλημα έχει μακρά παράδοση και έχει εμφανιστεί σε διάφορες μορφές σε διάφορους κλάδους. Όπως θα δούμε σε

επόμενο κεφάλαιο, υπάρχουν ποικίλοι ορισμοί για την έννοια της κοινότητας, η οποία μπορεί να έχει διαφορετικά χαρακτηριστικά. Κοινότητες μπορεί για παράδειγμα να είναι ιεραρχικά ή αλληλοκαλυπτόμενα διαμορφωμένες εντός του γράφου. Ο γράφος μπορεί να περιλαμβάνει κατευθυνόμενες ακμές μεταξύ των κορυφών του, δίνοντας έτσι σημασία σε αυτή την κατεύθυνση κατά την εξέταση των σχέσεων μεταξύ των κορυφών του. Κοινότητες μπορούν επίσης να είναι δυναμικές, δηλαδή να εξελίσσονται με την πάροδο του χρόνου. Ως αποτέλεσμα αυτής της πληθώρας ορισμών και των χαρακτηριστικών τους, έχουν δημοσιευτεί πολλές και διαφορετικές λύσεις στο εν λόγω πρόβλημα. Στην εργασία αυτή έχουμε κατηγοριοποιήσει μεγάλο μέρος των μεθόδων αυτών σε παραδοσιακές και μοντέρνες, σύμφωνα με τις μεθοδολογικές τους αρχές.

Κεφάλαιο 2. Κοινότητες

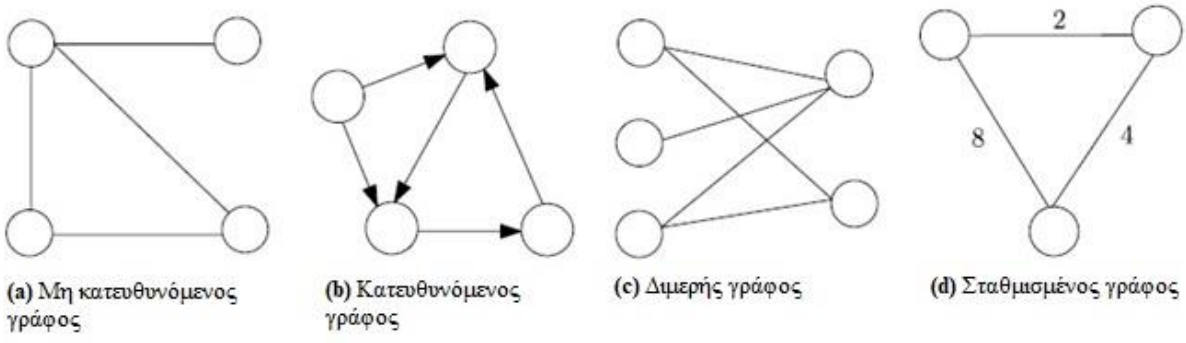
Ο εντοπισμός κοινωνιών σε δίκτυα, ή αλλιώς συσταδοποίηση δικτύου ή γράφου (network or graph clustering), είναι ένα πρόβλημα εγγενώς ασαφές, καθώς δεν υπάρχει καθολικός ορισμός των αντικειμένων που πρέπει να ψάχνει ο ενασχολούμενος. Η ασάφεια αυτή, απο τη μια πλευρά επιτρέπει ελευθερία προσεγγίσεων στο πρόβλημα, οι οποίες κάθε φορά εξαρτώνται απο το συγκεκριμένο ερευνητικό ζητούμενο και το εκάστοτε υπο μελέτη σύστημα, απο την άλλη ωστόσο έχει δημιουργήσει αοριστία στο συγκεκριμένο επιστημονικό πεδίο, καθυστερώντας την πρόοδο του. Τα βασικά συστατικά του προβλήματος, δηλαδή οι έννοιες της κοινότητας (community) και της διαμέρισης (partition) του δικτύου, καθώς δεν είναι αυστηρά ορισμένες, χρήζουν κάποιο βαθμό τυχαιότητας αλλά και κοινής λογικής στην αντιμετώπισή τους. Η ενότητα αυτή έχει να κάνει με την ιδέα της κοινότητας, περιγράφοντας την εξέλιξή της από τις κλασσικές ερμηνείες της, βασιζόμενες στη θεωρία γράφων, στην πιο μοντέρνα στατιστική της- και όχι μόνο- υπόσταση.

Εφόσον ένα δίκτυο αναπαρίσταται απο ένα γράφο, κατα τη διάρκεια της ανάλυσής μας, οι όροι δίκτυο και γράφος θα εναλλάσσονται, υπονοώντας πάντα το ίδιο αντικείμενο. Ξεκινώντας, θα δοθούν διευκρινίσεις των βασικών εννοιών και οι ορισμοί της θεωρίας γράφων, οι οποίες θα αποτελέσουν τη βάση για την κατασκευή της έννοιας της κοινότητας.

2.1 Βασική ορολογία, ορισμοί και προκαταρκτικές γνώσεις.

Ένας γράφος (graph) $G = (V, E)$ είναι ένα σύνολο κόμβων (nodes) V και ένα σύνολο ακμών (edges) $E \subseteq V \times V$ το οποίο συνδέει ζευγάρια κόμβων. Οι κόμβοι και οι ακμές, αναφέρονται συχνά και ως κορυφές (vertices) και σύνδεσμοι (links) αντίστοιχα. Ο αριθμός των κόμβων στο δίκτυο συμβολίζεται με $n = |V|$ και ο αριθμός των ακμών $m = |E|$. Ένας γράφος $G' = (V', E')$ καλείται υπογράφος (subgraph) του $G = (V, E)$ εάν $V' \subset V$ *και* $E' \subset E$. Τύποι γράφων συνοψίζονται στις ακόλουθες κατηγορίες: Κατευθυνόμενος ή μη (directed or undirected), μονομερής ή διμερής (unipartite or bipartite) και σταθμισμένος ή μη (weighted or unweighted). Μια οπτικοποιημένη αναπαράσταση διαφόρων τύπων γράφων (ομάδων σημείων που συνδέονται απο γραμμές), δίνεται στα παρακάτω σχήματα:

Σχήμα 1. Παραδείγματα των διαφόρων τύπων γράφων



Στην περίπτωση του κατευθυνόμενου γράφου (*directed graph*) (b) τα βέλη δηλώνουν την κατεύθυνση του κάθε συνδέσμου. Σε ένα σταθμισμένο γράφο (*weighted graph*) οι τιμές σε κάθε άκρο συμβολίζουν τα βάρη. (Malliaros-Vazirgiannis, 2013)

Εάν κάθε ακμή του γράφου $G = (V, E)$ είναι ένα διατεταγμένο ζεύγος κορυφών, δηλαδή εάν κάθε ακμή $(i\ , j) \in E$ συνδέει τον κόμβο με τον κόμβο j μιλάμε για διατεταγμένο γράφο (directed graph). Στην περίπτωση αυτή το διατεταγμένο ζεύγος (i, j) είναι μια ακμή που κατευθύνεται απο τον κόμβο i στον j (ξεκινάει απο τον i και καταλήγει στον j). Ένας γράφος G, ονομάζεται σταθμισμένος (weighted), όταν ένας πραγματικός αριθμός συνδέεται με κάθε ένα απο τα άκρα του. Επίσης ένας γράφος $G = (V_1, V_2, E)$, καλείται διμερής (bipartite) έαν η ομάδα κόμβων V, διαχωρίζεται σε δύο κομματιασμένες (disjoint) υποομάδες V_1, V_2 και κάθε άκρο συνδέει μια κορυφή του V_1 με μια κορυφή του V_2, 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 $n \times n$ ($|V| \times |V|$) πίνακας που ορίζεται ως εξής:

$$A = (a_{ij})_{n \times n} = \begin{cases} 1, & εαν\ (i,j) \in E, \quad \forall\ i,j \in 1, \dots \dots, n \\ 0, & διαφορετικά \end{cases}$$

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.

Έχοντας εισάγει τις βασικές αυτές έννοιες, προχωρούμε προσσεγγίζοντας την έννοια της κοινότητας ποσοτικοποιώντας την, έννοια η οποία διαισθητικά - και κατευθυντήριο σημείο αναφοράς για όλους τους ορισμούς που θα δούμε στη συνέχεια – σημαίνει, οτι οι ακμές μεταξύ των εσωτερικών της κόμβων είναι περισσότερες απο αυτές που συνδέουν τους κόμβους της με το υπόλοιπο δίκτυο.

Έστω C υπογράφος του γράφου G, με $n = |G|$ και $n_c = |C|$. Ο εσωτερικός (internal degree) k_i^{int} και εξωτερικός (external degree) k_i^{ext} , βαθμός μιας κορυφής $i \in C$, ορίζονται ως ο αριθμός των ακμών που συνδέουν τον κόμβο i με τους υπόλοιπους κόμβους του C και τον G αντίστοιχα, ορισμοί που μπορούν να εκφραστούν εν συντομία και μέσω του πίνακα γειτνίασης A: $k_i^{int} = \sum_{j \in C} A_{ij}$, $k_i^{ext} = \sum_{j \notin C} A_{ij}$. Εξ ορισμού, $k_i = k_i^{int} + k_i^{ext} = \sum_j A_{ij}$. Ορίζονται επίσης οι έννοιες της εσωτερικής πυκνότητας (intra-cluster density), $\delta_{\text{int}}(C)$ και μεταξύ των συστάδων πυκνότητας (inter-cluster density), $\delta_{\text{ext}}(C)$, κατά αντιστοιχία ως ακολούθως:

$$\delta_{int}(C) = \frac{\#\ \varepsilon\sigma\omega\tau\varepsilon\rho\iota\kappa\acute{\omega}\nu\ \sigma\upsilon\nu\delta\acute{\varepsilon}\sigma\mu\omega\nu\ \tau\omega\upsilon\ C}{n_c(n_c-1)/2}$$

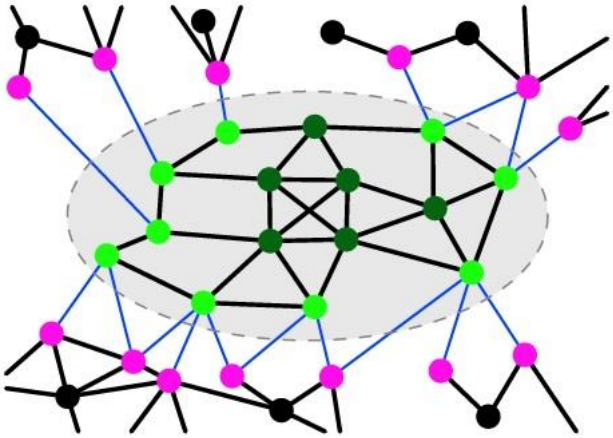
$$\delta_{ext}(C) = \frac{\#\ \mu\varepsilon\tau\alpha\xi\acute{\upsilon}\ \tau\omega\nu\ \sigma\upsilon\sigma\tau\acute{\alpha}\delta\omega\nu,\ \ \sigma\upsilon\nu\delta\acute{\varepsilon}\sigma\mu\omega\nu\ \tau\omega\upsilon\ C}{n_c(n-n_c)}$$

Η $\delta_{\text{int}}(C)$ εκφράζεται ώς ο λόγος του αριθμού των εσωτερικών-του υπογράφου C-συνδέσμων προς τον αριθμό όλων των πιθανών εσωτερικών συνδέσμων και ομοίως η $\delta_{ext}(C)$ ως ο λόγος του αριθμού των συνδέσμων των κόμβων του C προς το υπόλοιπο δίκτυο, προς τον μέγιστο αριθμό συνδέσμων μεταξύ των συστάδων. Τέλος η μέση πυκνότητα συνδέσμων του γράφου G, $\delta(G)$, ορίζεται ως:

$$\delta(G) = \frac{\#\sigma\upsilon\nu\delta\acute{\varepsilon}\sigma\mu\omega\nu\ \sigma\tau o\nu\ G}{n(n-1)/2}$$

Δηλαδή ως ο λόγος του πλήθους των συνδέσμων του γράφου G προς το μέγιστο αριθμό δυνατών συνδέσμων. Για να μιλάμε για κοινότητες, η αναλογική σχέση μεταξύ των προαναφερθέντων ποσοτήτων που πρέπει να πληρείται και ο βασικός στόχος των περισσότερων αλγορίθμων συσταδοποίησης, είναι η $\delta_{int}(C)$ να είναι σημαντικά μεγαλύτερη απο τη $\delta(G)$ καθώς επίσης και η $\delta_{ext}(C)$ να είναι σημαντικά μικρότερη απο τη $\delta(G)$. Ένας απλός τρόπος να επιτευχθεί αυτό είναι μέσω της μεγιστοποίησης των αθροισμάτων των διαφορών $\delta_{int}(C) - \delta_{ext}(C)$ όλων των συστάδων της διαμέρισης (Mancoridis, Mitchell, Rorres, Chen, & Gansner, 1998)

Σχήμα 2. Οπτικοποιημένη αναπαράσταση συνεκτικού υπογράφου

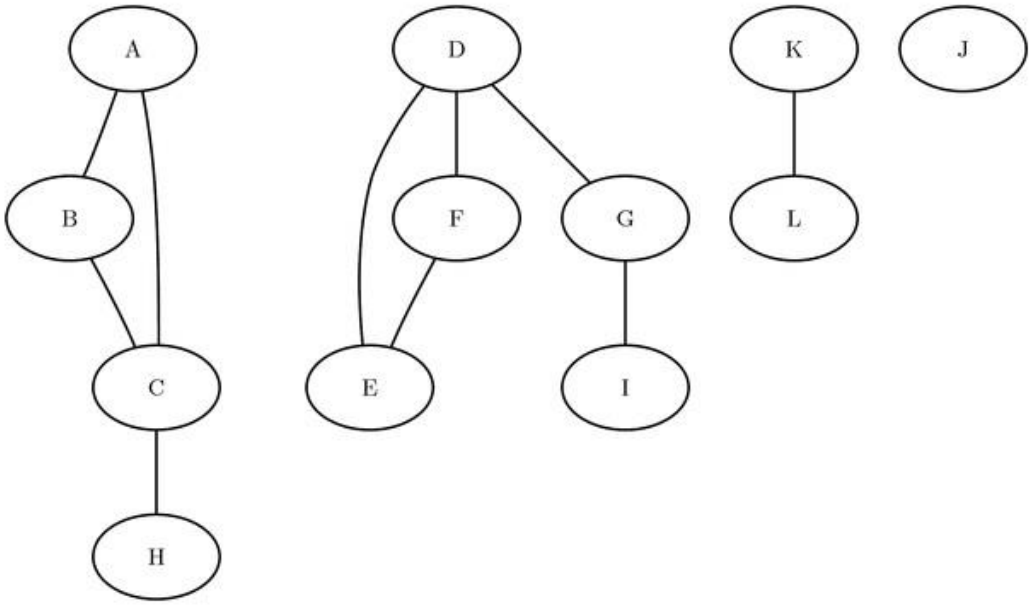


Τέλος, για να μιλάμε για κοινότητες, μια απαραίτητη ιδιότητα που πρέπει να πληρείται είναι αυτή της συνεκτικότητας (connectedness). Αυτή αναφέρεται στο μονοπάτι (path) που πρέπει να υπάρχει μεταξύ κάθε ζεύγους κόμβων (i, j) εκτεινόμενων εντός του υπογράφου C.

Στο παραπάνω σχήμα συνεκτικού υπογράφου, οι μωβ βούλες (dots) αντιπροσωπεύουν τους εξωτερικούς κόμβους που συνδέονται με τον υπογράφο C, ενώ οι μαύρες είναι οι εναπομείναντες κόμβοι του δικτύου. Οι μπλέ γραμμές συμβολίζουν τις ακμές που ενώνουν τον υπογράφο C με το υπόλοιπο δίκτυο. Οι σκούρες πράσινες βούλες στην εικόνα είναι εσωτερικές κορυφές (internal vertexes) του υπογράφου C, καθώς για αυτές ισχύει $k_i^{int} > 0 \; και \; k_i^{ext} = 0$, δηλαδή οι i κόμβοι έχουν γείτονες μόνο στο εσωτερικό του C. Αντίστοιχα με ανοιχτό πράσινο χρώμα συμβολίζονται οι οριακές κορυφές (boundary vertices) για τις οποίες ισχύει $k_i^{int} > 0 \; και \; k_i^{ext} > 0$. Προφανώς εαν $k_i^{int} = 0$, η κορυφή είναι αποκομμένη απο τον C. (Fortunato-Hric, 2016)

Με τον όρο συνεκτική συνιστώσα (connected component) του γράφου, εννοούμε ένα σύνολο κόμβων που είναι όλοι προσβάσιμοι μεταξύ τους. Δηλαδή, εάν δύο κόμβοι ανήκουν στην ίδια συνεκτική συνιστώσα, υπάρχει ένα μονοπάτι μεταξύ τους. Η έννοια θα διασαφηνιστεί καλύτερα με το παρακάτω σχήμα:

Σχήμα 3. Απλός μη κατευθυνόμενος γράφος G με 4 συνεκτικές συνιστώσες.



Στο σχήμα οι κορυφές A, B, C, H αποτελούν μια συνεκτική συνιστώσα καθώς υπάρχει ένα μονοπάτι μεταξύ κάθε ζεύγους τους, οι κορυφές D, E, F, G, I μια άλλη συνεκτική συνιστώσα για τον ίδιο λόγο, οι K, L μια άλλη και ο εναπομείνων κόμβος J θεωρούμε οτι αποτελεί μια συνεκτική συνιστώσα απο μόνος του.

Όλες οι προαναφερθείσες βασικές έννοιες επεκτείνονται ανάλογα με το είδος του γράφου στον οποίο κάθε φορά αναφερόμαστε, κατευθυνόμενος η μη, σταθμισμένος η μη. Οι προεκτάσεις όμως αυτές δεν επηρρεάζουν δομικά και νοηματικά την έννοια της κοινότητας, στην οποία θα μπούμε αμέσως μετά, για αυτό και παραλείφθηκε η αναφορά τους.

Με τις βασικές προαπαιτούμενες γνώσεις υπόψιν, εισάγουμε τους βασικούς ορισμούς της κοινότητας, οι οποίοι προέρχονται απο πολλά ετερογενή επιστημονικά πεδία, όπως κοινωνικές επιστήμες, επιστήμη των υπολογιστών ή φυσική.

2.2 Οπτικές κοινότητας δικτύου.

2.2.1 Παραδοσιακή οπτική κοινότητας δικτύου

Ο κλασσικός ορισμός κοινότητας, εστιάζει -επι το πλείστον- στον αριθμό των ακμών, είτε εσωτερικών είτε εξωτερικών, διαχειριζόμενος τις μεταξύ τους σχέσεις κατάλληλα, αλλά και σε μέτρα προσαρμογής (fitness measures), ωστε να διαμορφώσει την έννοια της.

Εφόσον οι κοινότητες θεωρούνται ξεχωριστές οντότητες εντος του γράφου, έκαστη με τη δική της αυτονομία, οι τοπικοί ορισμοί εστιάζουν στον αντίστοιχο υπογράφο-κοινότητα, παρά στο γράφο καθολικά, προσσέγγιση που αποτελεί την αρχαιότερη προσπάθεια ορισμού κοινότητας και προέρχεται απο αναλυτές κοινωνικών δικτύων. Οι κοινωνικές ομάδες ορίζονται ως υποομάδες των οποίων τα μέλη είναι όλα γειτονικά μεταξύ τους.

Σε όρους γραφημάτων αυτός ο ορισμός αντιστοιχεί στην έννοια της κλίκας (clique), (Luce and Perry, 1949) δηλαδή μιας υποομάδας της οποίας οι κορυφές είναι όλες γειτονικές μεταξύ τους. Μια κλίκα είναι ένας πλήρης γράφος (complete graph), δηλαδή ένας υπογράφος που κάθε μια απο τις κορυφές του συνδέεται με όλες τις υπόλοιπες με μια ακμή. Επιπλέον είναι ένας μέγιστος υπογράφος (maximal subgraph) που σημαίνει οτι δε δύναται να συμπεριληφθεί σε έναν μεγαλύτερο πλήρη υπογράφο, δηλαδή δεν μπορεί να επεκταθεί με την προσθήκη νέων κορυφών και ακμών σε αυτόν χωρίς να χάσει τη χαρακτηριστική του αυτή ιδιότητα. Στη μοντέρνα επιστήμη δικτύων μια κλίκα συνηθίζεται να λέγεται πλήρης γράφος, αλλά όχι απαραίτητα maximal. Τα τρίγωνα είναι η απλούστερη μορφή κλίκας και είναι πολύ συνήθη σε πραγματικά δίκτυα, σε αντίθεση με μεγαλύτερες κλίκες που είναι σπανιότερες. Η έννοια της κλίκας ωστόσο, μολονότι χρήσιμη, δεν είναι ιδανικό μέτρο για τον ορισμό της κοινότητας, για τους εξής λόγους: Αρχικά παρότι η κλίκα έχει τη μεγαλύτερη πιθανή εσωτερική πυκνότητα ακμών, εφόσον όλες οι εσωτερικές ακμές είναι παρούσες, οι κοινότητες κατά βάση δεν είναι πλήρεις γράφοι. Επίσης στην κλίκα, όλες οι κορυφές έχουν πανομοιότυπο ρόλο και είναι απόλυτα συμμετρικές, κάτι που δεν ισχύει σε κοινότητες πραγματικών δικτύων, στις οποίες κάποιες κορυφές είναι σημαντικότερες απο άλλες, εξαιτίας της ετερογένειας που υπάρχει στα πρότυπα σύνδεσής τους αλλά και ιεραρχίας, με κορυφές στον πυρήνα της κοινότητας να συνυπάρχουν με περιφερειακές. Η έννοιά της λοιπόν έχει παραλλαχθεί στην πιο μοντέρνα εκδοχή της καθώς έχουν παραχθεί εναλλακτικές

θεματολογίες στηριζόμενες σε αυτή, όπως η n-clique (Alba, 1973), (Luce, 1950) οι n-clans and n-clubs (Mokken R. J., 1979). Άλλοι εφάμιλλοι (ποσοτικοί) ορισμοί βασίζονται στην ιδέα οτι μια κορυφή πρέπει να είναι γειτονική σε έναν προκαθορισμένο ελάχιστο αριθμό κόμβων εντος του υπογράφου.

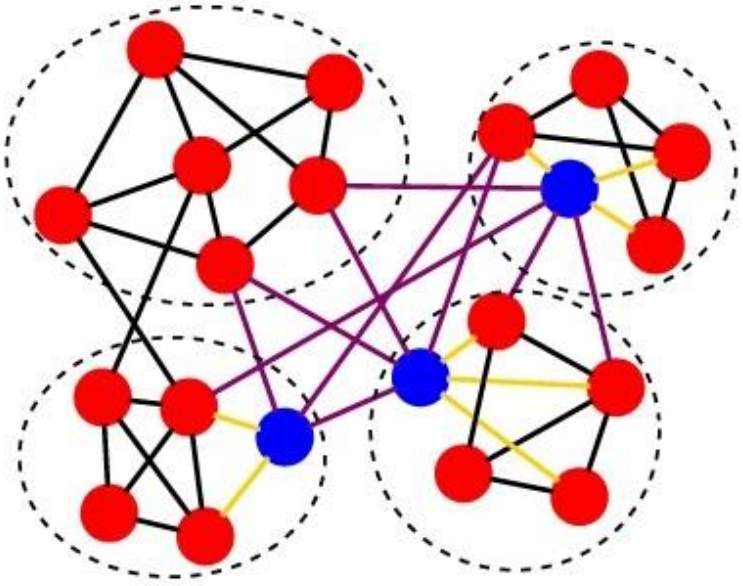
Για να οριστεί όμως η έννοια της κοινότητας, δεν μας αρκεί μόνο η συνεκτικότητα του υπογράφου που εξετάζουμε, αλλά και η συνεκτικότητα μεταξύ του συγκεκριμένου υπογράφου και του υπόλοιπου δικτύου. Πάνω στην ιδέα αυτή, στηρίζονται και οι έννοιες της ισχυρής (strong community) και αδύναμης (weak community) κοινότητας αντίστοιχα. Αυτές ορίζονται με διαφορετικό τρόπο ανάλογα με τον εκάστοτε συγγραφέα:

- Μια ισχυρή κοινότητα *(*Radicchi et al., 2004), ή αλλιώς LS-set (Luccio & Sami, 1969), αυστηρά, είναι ένας υπογράφος για τον οποίο ισχύει $k_i^{int} > k_i^{ext} \ \forall \ i \ \in C$, δηλαδή ο εσωτερικός βαθμός κάθε κόμβου στον υπογράφο είναι μεγαλύτερος απο τον αντίστοιχο εξωτερικό. Κατά τους ίδιους συγγραφείς, αδύναμη κοινότητα είναι ένας υπογράφος που ο συνολικός εσωτερικός του βαθμός είναι μεγαλύτερος απο τον εξωτερικό, δηλαδή: $\sum_{i \in C} k_i^{int} > \sum_{i \in C} k_i^{ext}$. Μια ισχυρή κοινότητα λοιπόν είναι και μια ασθενής κοινότητα, χωρίς να ισχύει όμως το αντίστροφο.
- Κατά τους Hu et al. (Hu, et al., 2008), ένας υπογράφος C είναι μια ισχυρή κοινότητα εαν ο εσωτερικός βαθμός οποιασδήποτε κορυφής που εμπεριέχεται σε αυτόν, είναι μεγαλύτερος απο τον αριθμό των συνδέσμων της συγκεκριμένης κορυφής με οποιαδήποτε άλλη κοινότητα. Αντίστοιχα ένας υπογράφος C είναι μια αδύναμη κοινότητα, εάν ο συνολικός εσωτερικός της βαθμός, ξεπερνά των αριθμό των συνδέσμων μεταξύ αυτής και των άλλων κοινοτήτων.

‘Οποια κοινότητα είναι ισχυρή (ή ασθενής) με βάση τον πρώτο ορισμό είναι και ισχυρή (ή ασθενής) για το δεύτερο, χωρίς όμως να ισχύει το αντίστροφο. Η διαφορά των δύο έγκειται στο οτι ο πρώτος θεωρεί το εκτός του υπογράφου δίκτυο ένα ενιαίο αντικείμενο, ενώ ο δεύτερος το χωρίζει σε κοινότητες.

Οπτικοποιημένη αναπαράσταση γράφων με ισχυρές και αδύναμες κοινότητες δίνεται στα παρακάτω σχήματα:

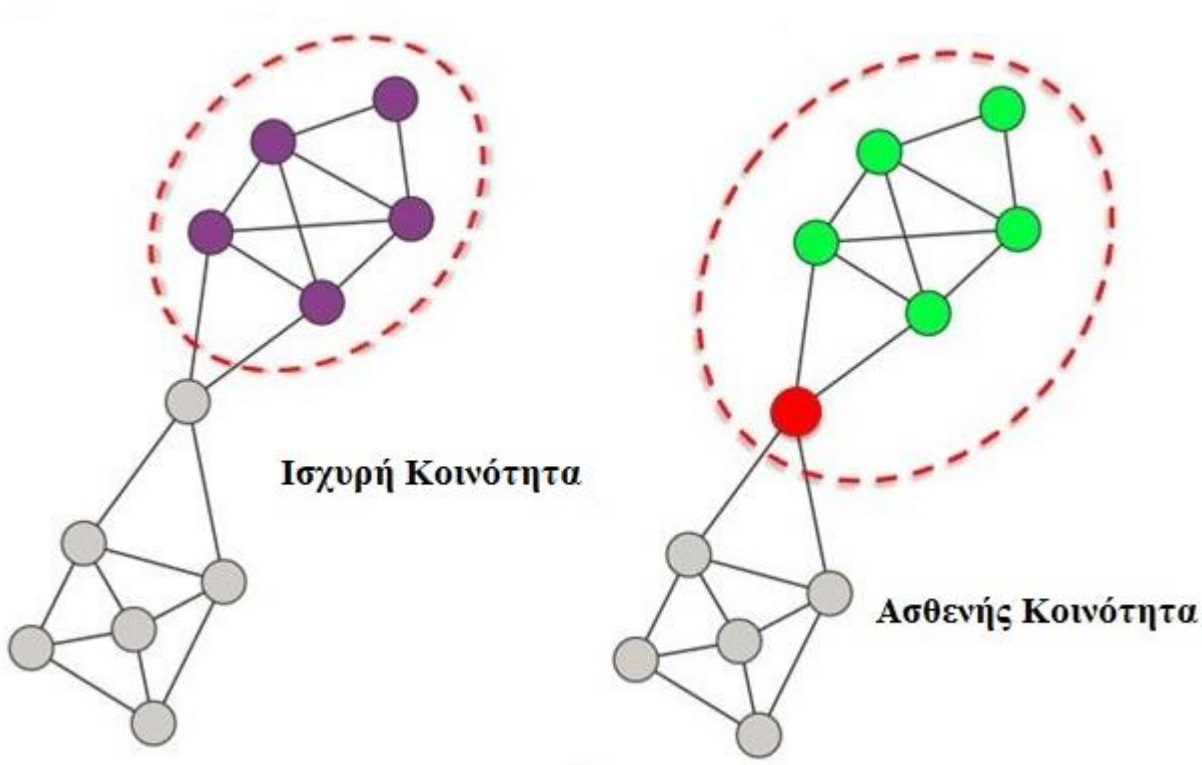
Σχήμα 4. Ισχυρές και Ασθενείς Κοινότητες.



Οι 4 υπογράφοι που περικλείονται από τα διακεκομμένα περιγράμματα είναι ασθενείς κοινότητες σύμφωνα με τους ορισμούς και των Radicchi et al. (2004) αλλά και των Hu et al. (2008). Είναι επίσης ισχυρές κοινότητες με βάση τους Hu et al., εφόσον ο εσωτερικός βαθμός κάθε κορυφής σε αυτές είναι μεγαλύτερος του αριθμού των ακμών που συνδέουν την εν λόγω κορυφή με τις κορυφές οποιασδήποτε άλλης κοινότητας.

Ωστόσο 3 απο τους υπογράφους (αυτοί που περιέχουν τις μπλε κορυφές) δεν είναι ισχυρές κοινότητες με βάση τους Radicchi et al, καθώς οι μπλε κορυφές έχουν εξωτερικό βαθμό μεγαλύτερο απο τον εσωτερικό τους. (οι εσωτερικές και εξωτερικές ακμές είναι χρωματισμένες με κίτρινο και μωβ αντίστοιχα). (Fortunato-Hric, 2016)

Σχήμα 5. Ισχυρή και Ασθενής Κοινότητα.



Στο σχήμα αυτό με βάση τους ορισμούς των Radicchi et al, ο πρώτος υπογράφος που περικλύεται στο διακεκομμένο περίγραμμα (μωβ κορυφές) είναι ισχυρή κοινότητα καθώς ο εσωτερικός βαθμός κάθε κόμβου σε αυτόν είναι μεγαλύτερος απο τον αντίστοιχο εξωτερικό, δηλαδή $k_i^{int} > k_i^{ext} \ \forall \ i \in C$. Εφόσον είναι ισχυρή κοινότητα θα είναι και ασθενής, καθώς $\sum_{i \in C} k_i^{int} > \sum_{i \in C} k_i^{ext}$. Ο δεύτερος υπογράφος που περικλύεται στο διακεκομμένο περίγραμμα, είναι επίσης αδύναμη κοινότητα εφόσον ο συνολικός εσωτερικός του βαθμός είναι μεγαλύτερος απο τον εξωτερικό, αλλά όχι ισχυρή εφόσον η κόκκινη κορυφή σε αυτόν δεν έχει εξωτερικό βαθμό μεγαλύτερο απο εσωτερικό, αλλά τον ίδιο. (Radicchi, Castellano, Cecconi, Loreto, & Parisi, 2004)

Κλείνοντας, εκτός των προαναφερθέντων μεταβλητών (εσωτερικών και εξωτερικών βαθμών) οι οποίες είναι μεταβλητές που εστιάζουν στο μέγεθος της κοινότητας, η έννοια της κοινότητας

επίσης προσσεγγίζεται και από μέτρα προσαρμοργής (fitness measures), όπως η εσωτερική πυκνότητα,

$$\delta_{int}(C) = \frac{\#\ \varepsilon\sigma\omega\tau\varepsilon\rho\iota\kappa\acute{\omega}\nu\ \sigma\upsilon\nu\delta\acute{\varepsilon}\sigma\mu\omega\nu\ \tau\sigma\upsilon\ C}{n_c(n_c-1)/2}.$$

Με βάση αυτή, ένας υπογράφος C, είναι κοινότητα εαν η $\delta_{int}(C)$ είναι μεγαλύτερη απο ένα κατώφλι, έστω ξ.

2.2.2 Μοντέρνα οπτική κοινότητας δικτύου.

Στην προηγούμενη ενότητα είδαμε πως οι παραδοσιακοί ορισμοί της κοινότητας στηρίζονται κατά βάση στη μέτρηση ακμών (εσωτερικών και εξωτερικών) με διάφορους τρόπους. Οι πιο μοντέρνες οπτικές ορισμού κοινότητας στηρίζονται σε μέτρα ομοιότητας και ανομοιότητας των κορυφών, στην πιθανότητα οι κορυφές να έχουν ακμές εντος του υπογράφου, στην έννοια της «φυλής», αλλά και στις ενέργειες των κόμβων εντός του δικτύου.

Είναι φυσικό να θεωρήσουμε τις κοινότητες ως ομάδες κορυφών που είναι όμοιες μεταξύ τους, ομοιότητα η οποία μετριέται ποικιλοτρόπως, με κάποια ιδιότητα αναφοράς, ανεξαρτήτως απο το εάν οι κόμβοι συνδέονται μεταξύ τους με κάποιο σύνδεσμο ή όχι. Κάθε κορυφή καταλήγει στη συστάδα της οποίας οι κορυφές είναι περισσότερο όμοιες με την κορυφή αυτή. Τα μέτρα ομοιότητας (similarity measures) είναι η βάση των παραδοσιακών μεθόδων συσταδοποίησης που θα αναφερθούν σε επόμενο κεφάλαιο, όπως της ιεραρχικής, μη-ιεραρχικής και φασματικής συσταδοποίησης. (Fortunato, Community detection in graphs, 2010)

Μέτρο ομοιότητας (similarity measure) μεταξύ δύο αντικειμένων είναι ένα αριθμητικό μέτρο του βαθμού στον οποίο δύο αντικείμενα είναι παρόμοια. Προφανώς και η ομοιότητα είναι υψηλότερη για ζευγάρια αντικειμένων που ομοιάζουν περισσότερο. Συνήθως είναι μη αρνητική και το εύρος τιμών της είναι το [0,1], όπου 0 συμβολίζει την πλήρη ανομοιότητα (τα αντικείμενα είναι τελείως διαφορετικά) και το 1 την πλήρη ομοιότητα (τα αντικείμενα είναι ίδια).

Μέτρο ανομοιότητας (dissimilarity measure) μεταξύ δύο αντικειμένων είναι ένα αριθμητικό μέτρο του βαθμού στον οποίο δύο αντικείμενα είναι διαφορετικά. Η ανομοιότητα είναι χαμηλότερη για περισσότερο όμοια ζεύγη αντικειμένων. Η απόσταση (distance) είναι το πιο κλασσικό μέτρο

ανομοιότητας. Το εύρος τιμών των μέτρων ανομοιότητας, πολλές φορές είναι το διάστημα [0,1], αλλά είναι επίσης σύνηθες να είναι το [0, +∞), όπου το 0 σημαίνει καθόλου ανομοιότητα (τα αντικείμενα είναι ίδια) και το ∞ συμβολίζει την πλήρη ανομοιότητα (τα αντικείμενα είναι τελείως διαφορετικά).

Τα μέτρα ομοιότητας και ανομοιότητας στη βιβλιογραφία αναφέρονται ως μέτρα εγγύτητας (proximity measures).

Εαν μπορούμε απεικονίσουμε τις κορυφές ενός γράφου σε έναν n-διάστατο Ευκλείδειο χώρο, ορίζοντας μια θέση σε αυτές μέσα σε αυτόν, τότε μπορούμε να χρησιμοποιήσουμε την απόσταση σαν μέτρο ανομοιότητας μεταξύ ζευγών κορυφών. Ενδεικτικά αναφέρουμε κάποια κλασσικά μέτρα απόστασης.

Αν $A = (\alpha_1, \alpha_2, \dots, \alpha_n)$ και $B = (b_1, b_2, \dots, b_n)$ είναι δύο n-διάστατες παρατηρήσεις τότε η απόσταση μεταξύ του A και του B μπορεί να ορισθεί με βάση τους τύπους:

$d_{AB} = \sqrt{\sum_{k=1}^{n}(a_k - \beta_k)^2}$ (Ευκλείδεια απόσταση)

$d_{AB} = \sum_{k=1}^{n} |a_k - \beta_k|$ (απόσταση Manhattan)

Εαν ο γράφος δεν μπορεί να απεικονιστεί στο χώρο, η ανομοιότητα προκύπτει αναγκαστικά απο τις γειτονικές σχέσεις μεταξύ κορυφών. Η απόσταση μεταξύ των κορυφών τότε, μπορεί να οριστεί ως:

$d_{ij} = \sqrt{\sum_{k \neq i,j}(A_{ik} - A_{jk})^2}$ (απόσταση που προκύπτει από τον πίνακα γειτνίασης),

όπου $A = (a_{ij})_{n \times n} = \begin{cases} 1, & \varepsilon\alpha\nu\ (i,j) \in E, \quad \forall\ i,j \in 1, \dots \dots, n \\ 0, & \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$, ο πίνακας γειτνίασης.

Αυτό είναι ένα μέτρο ανομοιότητας που στηρίζεται στην έννοια της δομικής ισοδυναμίας (structural equivalence) (Lorrain & White, 1971). Δύο κορυφές, i, j είναι δομικά ισοδύναμες αν έχουν τους ίδιους γείτονες, ακόμα και αν δεν είναι γειτονικές μεταξύ τους. Εαν οι i, j είναι δομικά ισοδύναμες, τότε $d_{ij} = 0$. Κορυφές με μεγάλο βαθμό και διαφορετικούς γείτονες θεωρούνται πολύ απομακρυσμένες μεταξύ τους. Μέτρο ομοιότητας που στηρίζεται στη δομική ισοδυναμία είναι η

αλληλοκάλυψη (overlap) ανάμεσα σε δύο γειτονιές Γ(i), Γ(j), δύο κορυφών i, j, ως ο λόγος της τομής και της ένωσης των γειτονιών:

$$\omega_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}$$

Άλλο μέτρο ομοιότητας που στηρίζεται στη δομική ισοδυναμία είναι ο συντελεστής συσχέτισης του Pearson, μεταξύ γραμμών ή στηλών του πίνακα γειτνίασης:

$$C_{ij} = \frac{\Sigma_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n \sigma_\iota \sigma_j}, \text{ όπου } \mu_i = \frac{(\Sigma_j A_{ij})}{n}, \sigma_i = \sqrt{\frac{\sum_j (A_{ij} - \mu_i)^2}{n}}.$$

Άλλη σημαντική κατηγορία μέτρων ομοιότητας, πάνω στα οποία υπάρχει έντονη ερευνητική δραστηριότητα, βασίζεται στις ιδιότητες των τυχαίων περιπάτων σε γράφους.

Ενδεικτικά, αναφέρουμε μία απο αυτές, το χρόνο μετακίνησης (commute time) ανάμεσα σε ζεύγη κορυφών, δηλαδή το μέσο αριθμό βημάτων που χρειάζεται ένας τυχαίος περιπατητής (random walker), ξεκινώντας απο οποιαδήποτε κορυφή για να φτάσει σε μια άλλη για πρώτη φορά και να επιστρέψει στην αρχική. Όσο περισσότερος ο χρόνος, τόσο πιο απομακρυσμένες (λιγότερο όμοιες) οι κορυφές.

Με βάση το survey Community Detection in Networks: A user guide (Fortunato & Hric, Community detection in networks: A user guide, 2016), η ύπαρξη κοινοτήτων στη μοντέρνα της οπτική υπονοεί ένα πιθανοτικό μοτίβο προτίμησης συνδεσιμότητας κορυφών της ίδιας κοινότητας (κορυφές της ίδιας κοινότητας έχουν μεγαλύτερη πιθανότητα να διαμορφώσουν συνδέσμους μεταξύ τους αντι με κορυφές του υπόλοιπου δικτύου). Η έννοια αυτή αντιστοιχίζεται ακριβώς στις έννοιες της ισχυρής και αδύναμης κοινότητας που αναφέρθηκαν στην προηγούμενη ενότητα. Στους αντίστοιχους ορισμούς δεν αλλάζει τίποτα πλην του οτι αντικαθίσταται ο βαθμός των κορυφών, εσωτερικός η εξωτερικός (k_i^{int}, k_i^{ext}) με πιθανότητες συνδέσμων μεταξύ κορυφών της ίδιας συστάδας (p_{in}) και κορυφών μεταξύ διαφορετικών συστάδων (p_{out}), αντίστοιχα.

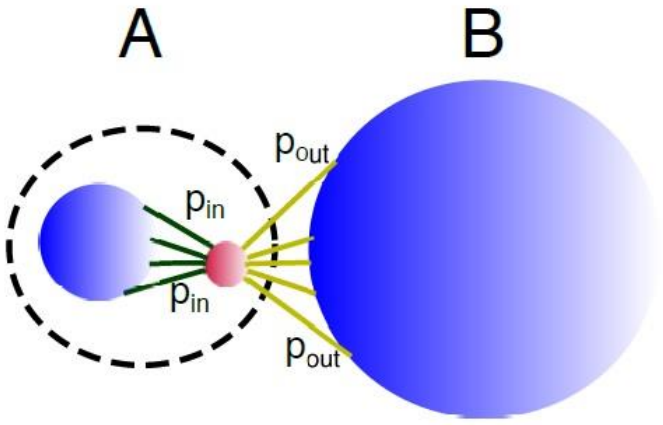
Σύμφωνα με αυτά και δεδομένου οτι έχουμε υπολογίσει τις πιθανότητες των ακμών όλων των ζευγών κορυφών με κάποιο τρόπο, οι ορισμοί των ισχυρών και αδύναμων κοινοτήτων είναι οι εξής:

- Μια ισχυρή κοινότητα, είναι ένας υπογράφος που κάθε μια απο τις κορυφές του έχει μεγαλύτερη πιθανότητα να συνδέεται με κάθε άλλη εντός του υπογράφου, παρά με οποιαδήποτε άλλη του υπόλοιπου γράφου.
- Μια αδύναμη κοινότητα, είναι ένας υπογράφος που η μέση πιθανότητα συνδέσμου κάθε κορυφής με τις υπόλοιπες κορυφές εντος του υπογράφου, είναι μεγαλύτερη απο την μέση πιθανότητα συνδέσμου της κορυφής με τις κορυφές οποιουδήποτε άλλου υπογράφου του υπόλοιπου γράφου.

Η διαφορά ανάμεσα στους δύο ορισμούς έγκειται στο οτι στην μεν ισχυρή κοινότητα η ανισότητα μεταξύ πιθανότητας συνδέσμων κορυφών αναφέρεται σε ζεύγη κορυφών, ενώ στην αδύναμη κοινότητα η ανισότητα αναφέρεται σε μέσους όρους πιθανότητας συνδέσμων κορυφών μεταξύ των υπογράφων. Συνεπώς, μια ισχυρή κοινότητα είναι και αδύναμη, με το αντίστροφο να μην ισχύει γενικά.

Με το επόμενο παράδειγμα θα διασαφηνιστεί γιατί οι έννοιες της ισχυρής και αδύναμης κοινότητας όπως ορίστηκαν στην προηγούμενη ενότητα κατά τους Radicchi et al. και Hu et al., δεν είναι επαρκείς.

Σχήμα 6. Προβλήματα κλασσικών ορισμών ισχυρής και αδύναμης κοινότητας.



(S. Fortunato, D. Hric, (2016))

Για τους δύο υπογράφους A και B του παραπάνω δικτύου, έστω n_A το πλήθος κορυφών του A και n_B το πλήθος κορυφών του B αντίστοιχα, με $n_B \gg n_A$. Έστω επίσης p_{in} η πιθανότητα

συνδέσμου μεταξύ κορυφών του υπογράφου A και p_{out} μεταξύ κορυφών των υπογράφων A και B αντίστοιχα, με $p_{out} < p_{in}$. Η κόκκινη κουκίδα είναι αντιπροσωπευτική κορυφή του υπογράφου A και η μπλέ κουκίδα εντός του A, συμβολίζει τις υπόλοιπες κορυφές του. Ο αναμενόμενος εσωτερικός βαθμός μιας κορυφής στον A θα είναι $k_A^{int} = p_{in} n_A$. (Συμβατικά, θεωρούμε οτι το πλήθος των γειτονικών κορυφών μιας κορυφής εντος του υπογράφου είναι n_A και όχι $n_A - 1$, συμπεριλαμβάνοντας στο πλήθος συνδέσμων και το σύνδεσμο ενός κόμβου με τον εαυτό του. Η σύμβαση αυτή δημιουργεί αμελητέες διαφοροποιήσεις στα αποτελέσματα όταν τα μεγέθη των κοινοτήτων είναι πολύ μεγαλύτερα της μονάδας και υιοθετείται για λόγους απλότητας.) Αντίστοιχα ο αναμενόμενος εξωτερικός βαθμός μιας κορυφής στον A θα είναι $k_A^{ext} = p_{out} n_B$. Ο αναμενόμενος εσωτερικός και εξωτερικός βαθμός του υπογράφου A θα είναι $K_A^{int} = p_{in} n_A^2$ και $K_A^{ext} = p_{out} n_A n_B$, αντίστοιχα. Για οποιεσδήποτε δύο τιμές των p_{in} και $p_{out} < p_{in}$, το n_B μπορεί να είναι επαρκώς μεγαλύτερο του n_A έτσι ώστε $k_A^{int} < k_A^{ext}$, το οποίο υπονοεί επίσης οτι $K_A^{int} < K_A^{ext}$. Σύμφωνα με τα προαναφερθέντα, οι υπογράφοι A και B δεν είναι ούτε ισχυρές αλλά ούτε και αδύναμες κοινότητες με βάση τους ορισμούς των Radicchi et al. και Hu et al. της προηγούμενης ενότητας.

Ο υπολογισμός των p_{in}, p_{out} είναι μια διαδικασία που εξαρτάται απο το μοντέλο διαμόρφωσης των συνδέσμων, το οποίο πρέπει να στηρίζεται στην παρουσία ομάδων κορυφών που συμπεριφέρονται με τον ίδιο τρόπο.

Οι Coscia et al.(2011) συμπληρώνουν τους ορισμούς κοινότητας με δύο κατηγορίες ορισμών, αυτούς που στηρίζονται στη δράση των κόμβων (*action-based definitions*) και τους ορισμούς που στηρίζονται στη διάδοση επιρροής (*influence propagation-based definitions*). Στην πρώτη κατηγορία ορισμών οι κοινότητες διαμορφώνονται με βάση τις κλάσεις ενεργειών που εκτελούν εντός του δικτύου. Σε αυτή την κατηγορία ορισμού η ανακάλυψη κοινοτήτων πραγματοποιείται ανεξάρτητα απο το αν υπάρχει άμεσος σύνδεσμος μεταξύ των οντοτήτων που τις διαμορφώνουν. Οι ορισμοί διάδοσης επιρροής απο την άλλη, στηρίζονται στην έννοια της «φυλής» (tribe), δηλαδή μιας ομάδας οντοτήτων που επηρρεάζονται απο τους ίδιους «αρχηγούς» (leaders). «Αρχηγός» θεωρείται ένας κόμβος εαν ύστερα απο έναν καθορισμένο χρόνο εκτέλεσης μιας ενέργειας, ένας επαρκής αριθμός άλλων κόμβων εκτελούν την ίδια ενέργεια. Με βάση αυτό το μοτίβο, επομένως, διαμορφώνονται και οι αντίστοιχες ομάδες.

2.2.3 Τύποι Κοινοτήτων

Όπως προαναφέρθηκε στην εισαγωγή του κεφαλαίου, ο εντοπισμός κοινωνιών σε δίκτυα είναι μια διαδικασία που περιέχει μια εγγενή ασάφεια, η οποία κάθε φορά αποκρυσταλλώνεται απο διαφορετικούς παράγοντες, διαδικαστικούς (της αλγοριθμικής προσσέγισης που ο ενασχολούμενος προτίθεται να ακολουθήσει ανάλογα με τα δεδομένα του) ή/και δομικούς (μορφολογία του δικτύου).

Στο σημείο αυτό θα δώσουμε μια βασική κατηγοριοποίηση τύπων κοινωνιών, όπως αυτή προκύπτει απο δίκτυα πραγματικού χρόνου:

- Αλληλοκαλυπτόμενες 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 γράφων, πολλά φαινόμενα πραγματικού χρόνου, αναπαρίστανται απο μεταξύ τους συνδέσμους οι οποίοι δεν είναι «αμοιβαίοι», όπως για παράδειγμα ένα hyperlink στον ιστό, που οδηγεί το χρήστη απο 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*): Στην κατηγορία κοινοτήτων αυτή, μιλάμε για ομάδες συνδέσμων που εμφανίζονται και εξαφανίζονται, επομένως και οι αντίστοιχες κοινότητες εξελίσσονται και αναδιαμορφώνονται με την πάροδο του χρόνου.

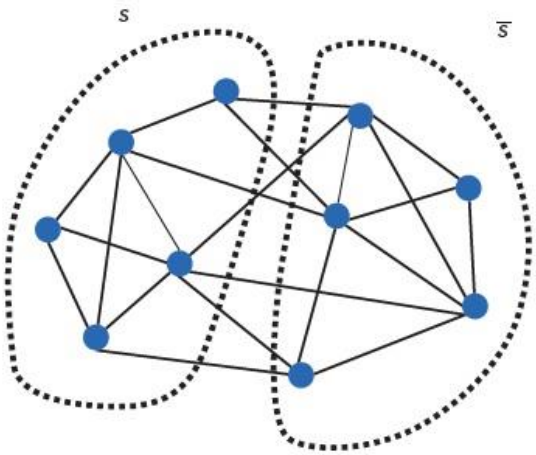
Κεφάλαιο 3. Παραδοσιακές Μέθοδοι Εντοπισμού Κοινωνιών σε Δίκτυα

Οι αλγόριθμοι που χρησιμοποιούνται για τον εντοπισμό κοινοτήτων έχουν κατηγοριοποιηθεί ποικιλοτρόπως στη βιβλιογραφία, με δύο απο τις πιο διάσημες τους κατηγοριοποιήσεις να γίνονται είτε με βάση τις μεθοδολογικές τους αρχές (Fortunato, 2010) είτε με βάση τον αντίστοιχο ορισμό κοινότητας που χρησιμοποείται (Coscia et al, 2011). Το κεφάλαιο αυτό περιέχει μια συνοπτική περιγραφή και κατηγοριοποίηση παραδοσιακών τεχνικών εντοπισμού κοινοτήτων σε δίκτυα με βάση τις μεθοδολογικές τους αρχές. Κάποιες απο τις μεθόδους που θα αναφερθούν στη συνέχεια αποτελούν και κλασσικές μεθόδους συσταδοποίησης Πολυμεταβλητής Ανάλυσης.

3.1 Διαμέριση του Γράφου

Η μέθοδος αυτή ασχολείται με το διαχωρισμό των κορυφών σε g συστάδες προκαθορισμένου μεγέθους, τέτοιες ώστε ο αριθμός των συνδέσμων ανάμεσα στις συστάδες αυτές, δηλαδή το μέγεθος αποκοπής (cut size), να είναι το ελάχιστο. Στην τεχνική της διαμέρισης του γράφου (Graph Partitioning), ο αριθμός των συστάδων που προκύπτουν είναι προκαθορισμένος.

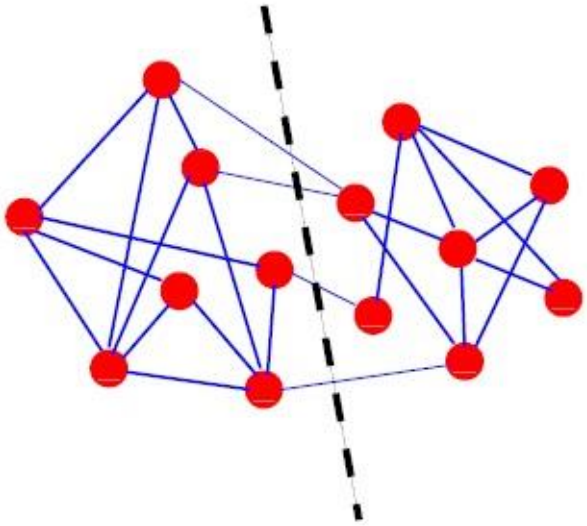
Σχήμα 7. Διχοτόμηση του Γράφου.



(Sanjeev Arora, Satish Rao, Umesh Vazirani, (2008))

Στο σχήμα παρατηρούμε τη διαμέριση του γράφου σε 2 συστάδες $S, \bar{S}$, με αριθμό συνδέσμων μεταξύ των συστάδων, δηλαδή μέγεθος αποκοπής, 7. Οι δύο συστάδες που προκύπτουν έχουν τον ίδιο αριθμό κορυφών, για αυτό και η διαμέριση ονομάζεται διχοτόμηση. (Arora, Rao, & Vazirani, Geometry, Flows, and Graph Partitioning Algorithms, 2008)

Σχήμα 8. Ελάχιστο μέγεθος αποκοπής.



(S. Fortunato, (2010))

Στο εν λόγω σχήμα, η διακεκομμένη γραμμή συμβολίζει το ελάχιστο μέγεθος αποκοπής, δηλαδή τη διαμέριση (διχοτόμηση εδώ) του γράφου σε 2 συστάδες ίδιου μεγέθους με τον ελάχιστο αριθμό συνδέσμων μεταξύ τους.

Εαν κάποιος «επέβαλλε» μια διαμέριση με το ελάχιστο μέγεθος αποκοπής, αγνοώντας όμως τον αριθμό των συστάδων, θα υπήρχαν τα εξής προβλήματα:

- Όλες οι κορυφές θα κατέληγαν να ανήκουν στην ίδια συστάδα, εφόσον το μέγεθος αποκοπής απαιτείται να είναι το ελάχιστο δυνατό. Απο τη στιγμή που δεν υπάρχουν περιορισμοί για το πλήθος των συστάδων που χρειαζόμαστε, το ελάχιστο δυνατό μέγεθος αποκοπής ουσιαστικά ανάγει τη λύση του προβλήματος διαμέρισης του γράφου σε g συστάδες, στον εκφυλισμό του γράφου σε μια συστάδα.

- Η λύση του προβλήματος θα αναγόταν στο διαχωρισμό της κορυφής χαμηλότερου βαθμού απο τον υπόλοιπο γράφο.

Η αποδοτικότητα της διαμέρισης επιτυγχάνεται επιλέγοντας κάποια μέτρα βελτιστοποίησης των μεγεθών των συστάδων, όπως η συνάρτηση οφέλους Q, στην οποία στηρίζεται ο αλγόριθμος Kernighan-Lin (1970), η οποία αντιπροσωπεύει τη διαφορά ανάμεσα στο πλήθος των ακμών εντός των συστάδων και του πλήθους των ακμών μεταξύ τους. Ο αλγόριθμος ξεκινάει με μια αρχική διαμέριση του γράφου σε δύο ομάδες προκαθορισμένου μεγέθους και στη συνέχεια ισομεγέθη υποσύνολα κορυφών ανταλλάσσονται μεταξύ των ομάδων για να επιτευχθεί η μέγιστη δυνατή αύξηση της συνάρτησης κέρδους.

Διάσημες τεχνικές που υπάγονται στην μέθοδο της διαμέρισης του γράφου, είναι επίσης η γεωμετρική τεχνική και η τεχνική που που δρομολογεί ροές στον γράφο. Οι δύο αυτές τεχνικές καταλήγουν στο να παράγουν γρήγορους αλλά και ποιοτικούς όσον αφορά τις αποκοπές, αλγορίθμους. Η περιγραφή τους δίνεται στο παράρτημα 1. Άλλη τεχνική που υπάγεται στην κατηγορία μεθόδων διαμέρισης του γράφου, είναι αυτή της φασματικής διχοτόμησης *(spectral bisection method)* η οποία στηρίζεται στις ιδιοτιμές του πίνακα Laplace (*Laplacian Matrix*), $L = (L_{ij})$, ο οποίος ορίζεται ως εξής:

$$L_{ij} = \begin{cases} k_i, & εαν\ i = j \\ -1, & εαν\ i, j\ γειτονικά, \\ 0, & διαφορετικά \end{cases}$$

όπου k_i είναι ο βαθμός του κόμβου i.

Στην τεχνική αυτή, κάθε διαμέριση ενός γράφου n κορυφών σε 2 ομάδες, αναπαρίσταται απο ένα διάνυσμα-δείκτη $\mathbf{s}$, του οποίου η συνιστώσα s_i, παίρνει τιμές 1 ή -1 αντίστοιχα, ανάλογα με το αν η εν λόγω κορυφή ανήκει στη μια ή στην άλλη ομάδα. Σκοπός είναι η ελαχιστοποίηση του μεγέθους αποκοπής R,

όπου

$R = \frac{1}{4}\mathbf{s}^T\mathbf{L}\mathbf{s}$,

όπου $\mathbf{s}^T$ ο ανάστροφος του διανύσματος $\mathbf{s}$ και $\mathbf{s} = \sum_i a_i \mathbf{v}_i$, όπου $\mathbf{v}_\text{i}$ τα ιδιοδιανύσματα του πίνακα Laplace.

Μέτρα που αφορούν την ελαχιστοποίηση του μεγέθους αποκοπής, είναι:

- Η αγωγιμότητα (*conductance*) ενός υπογράφου C του γράφου G,

$$\Phi(C) = \frac{c(C,G\backslash C)}{\min(k_c,k_{G\backslash C})},$$

όπου $c(C, G\backslash C)$ μέγεθος αποκοπής του C, και $k_c, k_{G\backslash C}$ οι συνολικοί βαθμοί του C και του υπόλοιπου εκτος του C, γράφου, αντίστοιχα. Η αγωγιμότητα του γράφου είναι η τιμή της πιο «αραιής» αποκοπής σε σχέση όλες τις πιθανές. Οι αποκοπές αφορούν μόνον μη κενές ομάδες, καθώς διαφορετικά το κλάσμα δε θα οριζόταν. Η ελαχιστοποίηση της αγωγιμότητας, προφανώς υπονοεί χαμηλές τιμές του αριθμητή και μεγάλες τιμές του παρονομαστή, η μέγιστη των οποίων επιτυγχάνεται όταν οι συνολικοί βαθμοί της συστάδας C και του συμπληρωματικού της συνόλου $G\backslash C$ είναι ίσες.

- Ο λόγος αποκοπής (cut ratio), $\Phi_C(C) = \frac{c(C,G\backslash C)}{n_c n_{G\backslash C}}$, όπου n_c, $n_{G\backslash C}$ ο ο αριθμός κορυφών του C και του εκτός του C δικτύου αντίστοιχα.
- Η κανονικοποιημένη αποκοπή (normalized cut), $\Phi_N(C) = \frac{c(C,G\backslash C)}{k_c}$, όπου k_c ο συνολικός βαθμός του υπογράφου C.

Χαρακτηριστικό των μεθόδων ελαχιστοποίησης των μέτρων αυτών είναι οτι ευνοούν διχοτομήσεις σε συστάδες ίδιου μεγέθους, απο την άποψη αριθμού κόμβων και συνδέσμων μεταξύ τους. Η μέθοδος της διχοτόμησης του γράφου ωστόσο απαιτεί συγκεκριμένες παραδοχές για τα μεγέθη των συστάδων ενώ η ελαχιστοποίηση των προαναφερθέντων μέτρων όχι.

Η διαμέριση γράφου κρίνεται ακατάλληλη για εντοπισμό κοινοτήτων τόσο σε σχέση με την πρακτικότητα της εφαρμογής της αλλά και σε σχέση με τη μεθοδολογία. Απο τη μια πλευρά χρειάζεται πολλές φορές να ξέρουμε και τον αριθμό αλλά και το μέγεθος των συστάδων, για τα οποία συνήθως δε διαθέτουμε στοιχεία. Απο μεθοδολογική σκοπιά, εαν, για παράδειγμα, θέλουμε μια διαίρεση του γράφου σε 3 συστάδες, αυτή αναγκαστικά πραγματοποιείται μέσω διχοτόμησης κάποιας απο τις αρχικές 2 συστάδες. Ωστόσο σε πολλές περιπτώσεις η διαμέριση ελάχιστου μεγέθους αποκοπής επιτυγχάνεται εαν η τρίτη συστάδα είναι συγχώνευση μερών των δύο αρχικών.

3.2 Ιεραρχική Συσταδοποίηση

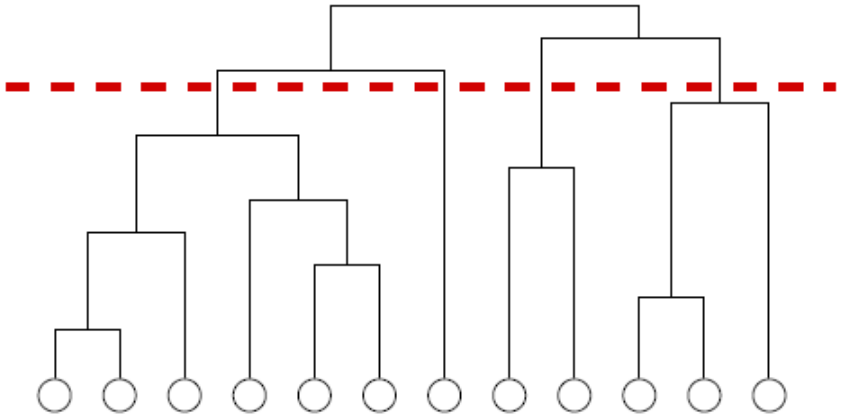
Η ιεραρχική συσταδοποίηση (hierarchical clustering) χρησιμοποιείται κατα κόρον όταν ο γράφος έχει ιεραρχική δομή, δηλαδή διαφορετικά επίπεδα συσταδοποίησης των κορυφών, με μικρές συστάδες να εμπεριέχονται σε μεγαλύτερες, οι μεγαλύτερες σε ακόμα μεγαλύτερες και ούτω καθεξής. Οι αλγόριθμοι που υπάγονται στην τεχνική αυτή, δεν προαπαιτούν γνώση για τον αριθμό των συστάδων και παρέχουν μια ιεραρχία «δενδροειδούς μορφής» όπου στα διάφορα στάδια, το πλήθος k των ομάδων παίρνει όλες τις δυνατές τιμές απο το 1 έως το n. Στο ένα άκρο της ιεραρχίας αυτής υπάρχει μόνο μια ομάδα που περιέχει n άτομα και στο άλλο άκρο n ομάδες οπου η κάθε μια περιέχει μόνο ένα άτομο. Η λογική τους συνοπτικά, στηρίζεται σε ένα μέτρο ομοιότητας μεταξύ των κορυφών και στον υπολογισμό του για κάθε ζεύγος κορυφών, ανεξάρτητα απο το εαν αυτές συνδέονται η όχι. Στο τέλος της εκάστοτε διαδικασίας καταλήγουμε σε έναν $n \times n$ πίνακα X, τον πίνακα ομοιότητας. Οι τεχνικές αυτές εντοπίζουν ομάδες κορυφών υψηλής ομοιότητας-όπως αυτή κάθε φορά ορίζεται. Υπάρχουν δύο βασικές κατηγορίες ιεραρχικών μεθόδων:

- Οι συσσωρευτικές μέθοδοι (*agglomerative methods*), οι οποίες ξεκινούν με n ομάδες και με διαδοχικές συγχωνεύσεις καταλήγουν σε μια ομάδα που περιέχει όλους τους κόμβους του δικτύου, δηλαδή οι συστάδες επαναληπτικά συγχωνεύονται αν η ομοιότητα τους είναι επαρκώς υψηλή. (*bottom-up algorithms*)
- Οι διαιρετικές μέθοδοι (*divisive methods*), οι οποίες εκτελούν την αντίθετη διεργασία δηλαδή ξεκινούν με μια μόνο ομάδα που περιέχει n κόμβους και διαιρούν τα δεδομένα σε όλο και μικρότερες ομάδες, δηλαδή αφαιρούν συνδέσμους που συνδέουν κορυφές με χαμηλή ομοιότητα. (*top down algorithms*)

Η ομοιότητα μεταξύ συστάδων υπολογίζεται διαφορετικά ανάλογα με τον εκάστοτε αλγόριθμο. Παραδείγματος χάρη, στη μέθοδο απλής συνένωσης (single linkage clustering) η ομοιότητα μεταξύ δύο συστάδων υπολογίζεται απο το ελάχιστο στοιχείο x_{ij} του πίνακα απόστασης με τον κόμβο i να ανήκει στη μία και ο j στην άλλη, ενώ στη μέθοδο της πλήρους συνένωσης (complete linkage clustering) απο το μέγιστο στοιχείο x_{ij}.

Οι διαδικασίες αναπαριστώνται μέσω δενδροδιαγραμμάτων όπως στο παρακάτω σχήμα:

Σχήμα 9. Δενδροδιάγραμμα ή ιεραρχικό δένδρο.



((Newman and Girvan), 2004)

Η οριζόντια διακεκομμένη κόκκινη γραμμή αντιπροσωπεύει τις επιθυμητές διαμερίσεις του γράφου σε συστάδες, στην περίπτωση του παραδείγματος, 4.

Μειονεκτήματα της ιεραρχικής συσταδοποίησης έγκεινται αρχικά στο οτι τα αποτελέσματα της επηρρεάζονται απο το μέτρο ομοιότητας που χρησιμοποιούμε καθώς επίσης στο οτι δεν παρέχει κάποιο τρόπο επιλογής της συσταδοποίησης που αντιπροσωπεύει καλύτερα την κοινοτική δομή του δικτύου.

Επίσης οδηγεί στη διαμόρφωση μιας ιεραρχίας του δικτύου που μπορεί να είναι εικονική τις περισσότερες φορές, καθώς ο γράφος μπορεί να μην έχει καθόλου ιεραρχική δομή εξαρχής. Ακόμα, οι ομάδες που φτιάχνονται σε αρχικά βήματα δεν μπορούν να χωρίσουν και μένουν μαζί για πάντα. Συχνά ένας αλγόριθμος ιεραρχικής ομαδοποίησης καταλήγει στη δημιουργία ενός μικρού πλήθους ομάδων με πολλές παρατηρήσεις και αφήνει αρκετές παρατηρήσεις να είναι απο μόνες τους ανεξάρτητες ομάδες. Βασικό μειονέκτημα και των συσσωρευτικών αλλα και των διαιρετικών μεθόδων είναι επίσης οτι είναι ασύμφορες απο την άποψη υπολογιστικού φόρτου σε μεγάλα σύνολα δεδομένων, που είναι και το ζητούμενό μας.

Οι πιθανές ενώσεις δύο απο τους n διαθέσιμους κόμβους στο πρώτο στάδιο ενός συσσωρευτικού αλγορίθμου είναι πολυωνυμικής τάξης $n(n-1)/2$ ενώ στο πρώτο στάδιο ενός διαιρετικού αλγορίθμου οι πιθανοί διαμερισμοί του συνόλου των n κόμβων σε δύο μη κενά σύνολα είναι εκθετικής τάξης $2^{n-1}-1$, επομένως στην πρώτη περίπτωση η αύξηση του n οδηγεί το πλήθος των δυνατών επιλογών σε αύξηση ανάλογη του τετραγώνου του n, ενώ στη δεύτερη εκθετική αύξηση του πλήθους, γεγονότα που επιβαρύνουν τον πίνακα ομοιότητας που πρέπει να ανανεώνεται κάθε φορά.

3.3 Μη Ιεραρχική Συσταδοποίηση

Στόχος των μη ιεραρχικών μεθόδων (non hierarchical clustering) είναι η ομαδοποίηση των n κόμβων σε k ομάδες, με το k να είναι καθορισμένο απο την αρχή. Ο τρόπος λειτουργίας τους έχει ως εξής:

- Θεωρούν k συγκεκριμένα μητρικά σημεία (seed points) και γύρω απο αυτά ταξινομούν τους υπόλοιπους κόμβους εως ότου διαμορφωθούν οι επιθυμητές ομάδες
- Ξεκινούν με έναν αρχικό διαμερισμό των κόμβων και στη συνέχεια μετακινούν τους κόμβους μεταξύ των ομάδων εως ότου πετύχουν την καλύτερη διαμέριση.

Οι μέθοδοι που υπάγονται σε αυτή την κατηγορία λειτοργούν επαναληπτικά και χρησιμοποιούν την έννοια του κέντρου της ομάδας, δηλαδή του διανύσματος των μέσων ανα μεταβλητή για όλες τις παρατηρήσεις της ομάδας. Κατά την εφαρμογή των μεθόδων, οι παρατηρήσεις κατατάσσονται σε ομάδες ανάλογα με την απόσταση τους απο τα κέντρα βάρους (centroids) όλων των ομάδων . Για κάθε παρατήρηση υπολογίζεται η απόστασή της απο τα κέντρα των διαθέσιμων ομάδων και στη συνέχεια κατατάσσεται στην ομάδα της οποίας το κέντρο είναι πιο κοντά στην παρατήρηση. Η συνάρτηση κόστους (*cost function*) είναι μια συνάρτηση η οποία στηρίζεται στις αποστάσεις που χρησιμοποιούνται για την κατάταξη των παρατηρήσεων. Διαφέρει ανάλογα με τη μέθοδο και την ιδέα πίσω απο τον εκάστοτε αλγόριθμο και μπορεί ενδεικτικά να είναι: η διάμετρος της συστάδας, δηλαδή η μεγαλύτερη απόσταση μεταξύ δυο κόμβων μιας συστάδας (*Minimum k-clustering*), ο μέσος όρος αποστάσεων μεταξύ όλων των ζευγών παρατηρήσεων της συστάδας (*k-clustering sum*), η μέγιστη απόσταση μεταξύ κάθε κόμβου της συστάδας και του κέντρου βάρους της (*k-center*), ο μέσος όρος της απόστασης μεταξύ κάθε κόμβου της συστάδας και του κέντρου βάρους της (*k-median*). Η διαφοροποίηση των

μεθόδων έγκειται στο σημείο στο οποίο γίνεται η ανανέωση των κέντρων των ομάδων και η ταξινόμηση των υπόλοιπων παρατηρήσεων σε αυτές.

Η πιο διάσημη τεχνική μη ιεραρχικής συσταδοποίησης είναι η μέθοδος MacQueen ή k-means method (MacQueen, 1967). Για την ταξινόμηση n ατόμων σε k ομάδες ο αλγόριθμος αυτός αποτελείται απο τα ακόλουθα βήματα:

1. Καθόρισε αρχικά ένα αρχικό σύνολο απο k μητρικά σημεία χρησιμοποιώντας k απο τα n διαθέσιμα άτομα.
2. Κατάταξε καθένα απο τα εναπομείναντα $n - k$ άτομα στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση απο το άτομο. Μετά απο κάθε τοποθέτηση υπολόγισε ξανά το κέντρο βάρους της αλλαγμένης πλεον ομάδας.
3. Όταν όλα τα άτομα έχουν τοποθετηθεί σε ομάδες μέσω του βήματος 2, θεώρησε τα δημιουργηθέντα κέντρα βάρους ως μητρικά σημεία και εκτέλεσε μια ακόμη σάρωση στα δεδομένα τοποθετώντας κάθε άτομο των δεδομένων στο πλησιέστερο μητρικό σημείο.

Η όλη διαδικασία εκτελεί $k(2n - k)$ υπολογισμούς αποστάσεων, $(k - 1)(2n - k)$ συγκρίσεις αποστάσεων και $n - k$ ανανεώσεις κέντρων βάρους.

Το μεγάλο μειονέκτημα του αλγορίθμου αυτού είναι οτι εξαρτάται απο τα αρχικά μητρικά σημεία ή αρχικές διαμερίσεις τα οποία αν δεν είναι σωστά επιλεγμένα μπορεί να οδηγήσουν σε μια τελείως διαφορετική ομαδοποίηση απο τη φυσική ομαδοποίηση που υπάρχει στα δεδομένα. Επίσης η ύπαρξη έκτροπων παρατηρήσεων (*outliers*) μπορεί να οδηγήσει στη δημιουργία ομάδων με πολύ διεσπαρμένες παρατηρήσεις. Επίσης, αν είναι γνωστό εκ των προτέρων οτι ο πληθυσμός που μελετάμε αποτελείται απο k ομάδες, και συμβεί στο δείγμα μας να μην αντιπροσωπεύεται κάποια απο αυτές, απαιτώντας το διαχωρισμό σε k ομάδες, θα οδηγηθούμε σε αφύσικες/παραπλανητικές ομαδοποιήσεις. Για το λόγο αυτό καλό είναι ο αλγόριθμος (και κατ'επέκταση οι αλγόριθμοι) να εφαρμόζεται για διαφορετικές επιλογές του k και να συγκρίνουμε τα αποτελέσματα χρησιμοποιώντας και τη διαίσθησή μας για να πετύχουμε την καλύτερη ομαδοποίηση.

3.4 Φασματική Συσταδοποίηση

Έστω μια $n-$διάστατη παρατήρηση $X = (x_1, x_2, \dots, x_n)$ και μια συνάρτηση (μέτρο) ομοιότητας, s, η οποία είναι συμμετρική και μη αρνητική: $s_{ij} = s(x_i, x_j) = s(x_j, x_i) \geq 0, \forall\, i, j = 1, \dots, n$ με τον αντίστοιχο πίνακα ομοιότητας $S = (s_{ij})_{i,j=1,\dots.n}$. Η φασματική συσταδοποίηση (Spectral Clustering) περιλαμβάνει όλες εκείνες τις τεχνικές που διαμερίζουν την $n-$διάστατη παρατήρηση σε συστάδες, χρησιμοποιώντας ιδιοδιανύσματα πινάκων ομοιότητας όπως του πίνακα Laplace ή του πίνακα Γειτνίασης (ή πίνακα Βαρών για σταθμισμένους γράφους). Έστω οτι η $n-$διάστατη παρατήρηση $X = (x_1, x_2, \dots, x_n)$ αναπαριστά τις κορυφές ενός γράφου. Κατά τη φασματική συσταδοποίηση, οι κορυφές αυτές προβάλλονται σε σημεία ενός $l-$διάστατου Ευκλείδειου χώρου, $\mathbb{R}^l$, τα οποία αντιστοιχούν σε κάποια ιδιοδιανύσματα του πίνακα ομοιότητας. Τα σημεία αυτά, στη συνέχεια συσταδοποιούνται με κάποια παραδοσιακή μη ιεραρχική μέθοδο, όπως η μέθοδος k-means.

Το βασικό εργαλείο της φασματικής συσταδοποίησης και ο πίνακας που χρησίμοποείται κατά κόρον σε αυτή είναι ο πίνακας Laplace και οι παραλλαγές του.

Στο σημείο αυτό θα αναφερθούμε στις διάφορες παραλλαγές των πινάκων Laplace και σε κάποιες βασικές ιδιότητες τους, οι οποίες θα βοηθήσουν στην κατανόηση του τρόπου λειτουργίας της φασματικής συσταδοποίησης. Στη βιβλιογραφία δεν υπάρχει καθολική σύμβαση για το ποιος πίνακας καλείται πίνακας Laplace. Εμείς έχουμε επιλέξει τη σύμβαση που ακολουθεί ο Ulrike Von Luxburg (2006).

Για τους ορισμούς που θα ακολουθήσουν, ισχύουν οι εξής υποθέσεις: Ο γράφος G είναι ένας μη κατευθυνόμενος & σταθμισμένος γράφος, με πίνακα βαρών $W = (w_{ij})_{n\times n}$, όπου $w_{ij} = w_{ji} \geq 0$. Τα ιδιοδιανύσματα ενός πίνακα δεν είναι απαραίτητα κανονικοποιημένα. Για παράδειγμα το σταθερό διάνυσμα $\mathbb{1} = (1,1, \dots,1)$ και ένα πολλαπλάσιο του $\alpha\mathbb{1}$, για κάποιο $\alpha \neq 0$, θεωρούνται τα ίδια ιδιοδιανύσματα. Οι ιδιοτιμές πάντα θα ταξινομούνται κατά αύξοντα τρόπο, κατά την πολλαπλότητα τους. Με τον όρο «τα πρώτα k ιδιοδιανύσματα», αναφερόμαστε στα ιδιοδιανύσματα που αντιστοιχούν στις k μικρότερες ιδιοτιμές. Κατά τον Ulrike von Luxburg (Luxburg, 2006) :

- Ο μη κανονικοποιημένος πίνακας Laplace ορίζεται ως: $L = D - W$, όπου $D = (d_{ij})_{n\times n}$ διαγώνιος πίνακας που το d_{ii} στοιχείο του, αντιστοιχεί στο βαθμό της κορυφής i.

  Ο L πληρεί τις ακόλουθες ιδιότητες, οι οποίες θα δοθούν χωρίς τις αποδείξεις τους:

  1. *Για κάθε διάνυσμα* $f \in \mathbb{R}^n$*, ισχύει :* $f'Lf = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$
  2. *Ο L είναι συμμετρικός και θετικά ημιορισμένος.*
  3. *Η μικρότερη ιδιοτιμή του L είναι* 0 *και το ιδιοδιάνυσμα που αντιστοιχεί στην ιδιοτιμή αυτή είναι το σταθερό διάνυσμα* $\mathbb{1} = (1,1,\dots,1)$.
  4. *Ο L έχει* n *πραγματικές, μη αρνητικές ιδιοτιμές* $0 = \lambda_1 \leq \lambda_2 \dots\dots\dots \leq \lambda_n$.

Οι ιδιοτιμές και τα ιδιοδιανύσματα του μη κανονικοποιημένου πίνακα Laplace μπορούν να χρησιμοποιηθούν για να περιγράψουν πολλές απο τις φασματικές ιδιότητες του γράφου, δηλαδή τις ιδιότητες του φάσματος του αντίστοιχου πίνακα Laplace του γράφου. Μια βασική απο αυτές είναι η εξής:

- *Εστω* G *ένας μη κατευθυνόμενος γράφος με μη αρνητικά βάρη. Τότε η πολλαπλότητα* k *της ιδιοτιμής 0 του L ισούται με τον αριθμό των συνεκτικών συνιστωσών* $A_1, \dots\dots, A_k$ *στο γράφο. Ο ιδιοχώρος της ιδιοτιμής* 0 *παράγεται απο τα χαρακτηριστικά διανύσματα-δείκτες* $\mathbb{1}_{A_1}, \mathbb{1}_{A_2}, \dots\dots, \mathbb{1}_{A_k}$ *των συνιστωσών αυτών.*

Έστω οτι έχουμε μια συνεκτική συνιστώσα στο γράφο, δηλαδή $k = 1$ και έστω f ιδιοδιάνυσμα με ιδιοτιμή 0. Σύμφωνα με την πρώτη ιδιότητα , $0 = f'Lf = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$. Το βάρος μεταξύ δύο συνεκτικών κορυφών v_i *και* v_j είναι θετικό ($w_{ij} > 0$), συνεπώς πρέπει $f_i = f_j$. Συνεπώς το ιδιοδιάνυσμα f πρέπει να είναι σταθερό σε όλη την εν λόγω συνεκτική συνιστώσα, A_k. Στην περίπτωση αυτή, μόνο το σταθερό διάνυσμα $\mathbb{1} = (1,1,\dots\dots,1)$ είναι το ιδιοδιάνυσμα με ιδιοτιμή 0, το οποίο θα είναι χαρακτηριστικό διάνυσμα δείκτης της συνεκτικής συνιστώσας.

Γενικεύοντας την παραπάνω πρόταση, έστω οτι έχουμε k συνεκτικές συνιστώσες. Χωρίς βλάβη της γενικότητας υποθέτουμε οτι οι κορυφές είναι διατεταγμένες ανάλογα με τις συνεκτικές συνιστώσες στις οποίες ανήκουν. Στην περίπτωση αυτή, ο πίνακας βαρών W έχει μια block-

διαγώνια μορφή (δηλαδή τα διαγώνια στοιχεία του είναι και αυτά πίνακες, ενώ τα υπόλοιπα μηδενικά).

Το ίδιο θα ισχύει και για τον πίνακα Laplace:

$$L = \begin{bmatrix} L_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & L_k \end{bmatrix}$$

Κάθε ένα απο τα blocks, L_i, είναι ένας πίνακας Laplace, ο οποίος αντιστοιχεί στον υπογράφο της i −οστής συνεκτικής συνιστώσας. Όπως σε όλους τους block-διαγώνιους πίνακες, το φάσμα του L, δίνεται απο την ένωση των φασμάτων των L_i και τα αντίστοιχα ιδιοδιανύσματα του L, είναι τα ιδιοδιανύσματα των L_i. Εφόσον κάθε L_i είναι ένας πίνακας Laplace ενός συνεκτικού γράφου, κάθε L_i έχει ιδιοτιμή 0 με πολλαπλότητα 1, και το ιδιοδιάνυσμα που αντιστοιχεί στην ιδιοτιμή αυτή είναι το σταθερό διάνυσμα $\mathbb{1}$ της i-οστής συνεκτικής συνιστώσας. Επομένως ο πίνακας L, έχει τόσες μηδενικές ιδιοτιμές όσες συνεκτικές συνιστώσες και τα ιδιοδιανύσματα που αντιστοιχούν στις ιδιοτιμές αυτές είναι τα χαρακτηριστικά διανύσματα δείκτες των συνεκτικών αυτών συνιστωσών.

- Υπάρχουν δύο πίνακες που στη βιβλιογραφία αναφέρονται ως κανονικοποιημένοι πίνακες Laplace. Ορίζονται ως:
  a. $L_{sym} \coloneqq D^{-\frac{1}{2}} L D^{\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{\frac{1}{2}}$
  b. $L_{rw} \coloneqq D^{-1} L = I - D^{-1} W$

Οι L_{sym}, L_{rw} πληρούν τις ακόλουθες ιδιότητες, οι οποίες θα δοθούν χωρίς τις αποδείξεις τους:

1. *Για κάθε διάνυσμα* $f \in \mathbb{R}^n$*, ισχύει :* $f' L_{sym} f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}})^2$*, όπου* d_i, d_j *ο βαθμός των κορυφών* i, j *αντίστοιχα.*
2. *Η* λ *είναι ιδιοτιμή του* L_{rw} *που αντιστοιχεί στο ιδιοδιάνυσμα* v *αν και μόνο αν η* λ *είναι ιδιοτιμή του* L_{sym} *που αντιστοιχεί στο ιδιοδιάνυσμα* $w = D^{\frac{1}{2}} v$.

3. *Η λ είναι ιδιοτιμή του L_{rw} που αντιστοιχεί στο ιδιοδιάνυσμα v αν και μόνο αν τα λ και v αποτελούν λύσεις του γενικευμένου προβλήματος ιδιοτιμών: $Lv = \lambda Dv$.*
4. *Η ιδιοτιμή 0 του L_{rw}, είναι η ιδιοτιμή που αντιστοιχεί στο σταθερό ιδιοδιάνυσμα $\mathbb{1}$. Η ιδιοτιμή 0 του L_{sym}, είναι η ιδιοτιμή αντιστοιχεί στο ιδιοδιάνυσμα $D^{\frac{1}{2}}\mathbb{1}$.*
5. *Οι L_{sym}, L_{rw} είναι θετικά ημιορισμένοι και έχουν n το πλήθος μη αρνητικές, πραγματικές ιδιοτιμές $0 = \lambda_1 \leq \lambda_2 \ldots\ldots \leq \lambda_n$.*

Όπως και στην περίπτωση του μη κανονικοποιημένου πίνακα Laplace, η πολλαπλότητα της ιδιοτιμής 0 των κανονικοποιημένων πινάκων Laplace σχετίζεται με τον αριθμό των συνεκτικών συνιστωσών του γράφου. Πιο συγκεκριμένα:

- Έστω G ένας μη κατευθυνόμενος γράφος με μη αρνητικά βάρη. Τότε η πολλαπλότητα k της ιδιοτιμής 0 των L_{sym}, L_{rw} ισούται με τον αριθμό των συνεκτικών συνιστωσών $A_1, \ldots\ldots, A_k$ στο γράφο. Ο ιδιοχώρος της ιδιοτιμής 0 παράγεται απο τα χαρακτηριστικά ιδιοδιανύσματα δείκτες $\mathbb{1}_{A_1}, \mathbb{1}_{A_2}, \ldots, \mathbb{1}_{A_k}$ των συνιστωσών αυτών.

Στο σημείο αυτό θα παρουσιάσουμε 3 βασικούς αλγορίθμους φασματικής συσταδοποίησης, οι οποίοι χρησιμοποιούν τις διαφορετικές μορφές των πινάκων Laplace, όπως αυτές αναλύθηκαν πιο πάνω. Ξεκινώντας θα αναφερθούμε στη μη κανονικοποιημένη φασματική συσταδοποίηση, η οποία όπως υπονοείται και απο την ονομασία της χρησιμοποιεί τον μη κανονικοποιημένο πίνακα Laplace.

Θεωρούμε, όπως αναφέρθηκε και στην αρχή του κεφαλαίου, μια $n-$διάστατη παρατήρηση $X = (x_1, x_2, \ldots, x_n)$ και μια συνάρτηση ομοιότητας, s και τον αντίστοιχο πίνακα ομοιότητας $S = (s_{ij})_{i,j=1,\ldots\ldots n}$.

3.4.1 Μη κανονικοποιημένη φασματική συσταδοποίηση

Τα δεδομένα εισόδου του αλγορίθμου είναι ο πίνακας βαρών W (εφόσον στην ανάλυση που προηγήθηκε αναφερθήκαμε σε σταθμισμένο γράφο) και ο αριθμός των συστάδων k. Τα βήματα υλοποίησης της μεθόδου είναι τα ακόλουθα:

1. Κατασκευή του μη κανονικοποιημένου πίνακα Laplace, L.
2. **Υπολογισμός των πρώτων k ιδιοδιανυσμάτων $v_1, \dots, v_k$ του L.**
3. Κατασκευή του πίνακα $V \in \mathbb{R}^{n \times k}$ του οποίου οι στήλες είναι τα $v_1, .., v_k$ ιδιοδιανύσματα.
4. Για $i = 1, \dots, n$, έστω $y_i \in \mathbb{R}^l$ το διάνυσμα που αντιστοιχεί στην i −οστή γραμμή του V (η οποία αναπαριστά την i κορυφή του γράφου).
5. Συσταδοποίηση των σημείων $(y_i)_{i=1,\dots,n}$ με τον αλγόριθμο k −means στις συστάδες $C_1, \dots, C_k$.

Τα δεδομένα εξόδου του αλγορίθμου είναι οι συστάδες $A_1, \dots, A_k$, με $A_i = \{j | y_j \in C_i\}$

3.4.2 Κανονικοποιημένη φασματική συσταδοποίηση

Σε αυτή την κατηγορία μεθόδων ανήκουν δύο διαφορετικοί αλγόριθμοι, ανάλογα με το ποιος κανονικοποιημένος πίνακας Laplace χρησιμοποιείται, ο L_{sym} ή ο L_{rw}.

- Ο αλγόριθμος που έχει προταθεί απο τους Shi και Malik (Shi & Malik, 2000) χρησιμοποιεί τον L_{rw} ενώ
- Ο αλγόριθμος που έχει προταθεί απο τους Ng, Jordan και Weiss (Ng, Jordan, & Weiss, 2001) χρησιμοποιεί τον L_{sym}.

Τα δεδομένα εισόδου του αλγορίθμου των Shi και Malik (2000) είναι ο πίνακας βαρών W (εφόσον στην ανάλυση που προηγήθηκε αναφερθήκαμε σε σταθμισμένο γράφο) και ο αριθμός των συστάδων k. Τα βήματα υλοποίησης της μεθόδου είναι τα ακόλουθα:

1. Κατασκευή του μη κανονικοποιημένου πίνακα Laplace, L.
2. **Υπολογισμός των πρώτων k ιδιοδιανυσμάτων $v_1, \dots, v_k$ που αποτελούν λύση του γενικευμένου προβλήματος ιδιοτιμών: $Lv = \lambda Dv$.**
3. Κατασκευή του πίνακα $V \in \mathbb{R}^{n \times k}$ του οποίου οι στήλες είναι τα $v_1, \dots, v_k$ ιδιοδιανύσματα.
4. Για $i = 1, \dots, n$, έστω $y_i \in \mathbb{R}^l$ το διάνυσμα που αντιστοιχεί στην i −οστή γραμμή του V (η οποία αναπαριστά την i κορυφή του γράφου).

5. Συσταδοποίηση των σημείων $(y_i)_{i=1,\dots,n}$ με τον αλγόριθμο $k-$means στις συστάδες $C_1, \dots, C_k$.

Τα δεδομένα εξόδου του αλγορίθμου είναι οι συστάδες $A_1, \dots, A_k$, με $A_i = \{j | y_j \in C_i\}$.

Η διαφορά του αλγορίθμου αυτού σε σχέση με τον αλγόριθμο της μη-κανονικοποιημένης φασματικής συσταδοποίησης είναι οτι ο αλγόριθμος των Shi and Malik χρησιμοποιεί τα γενικευμένα ιδιοδιανύσματα του L, που αντιστοιχούν στα ιδιοδιανύσματα του πίνακα L_{rw}.

Τέλος, ο τρίτος αλγόριθμος που θα παρουσιάσουμε, αυτός που προτάθηκε απο τους Ng, Jordan and Weiss, χρησιμοποιεί τα ιδιοδιανύσματα του L_{sym} και έχει ένα επιπλέον βήμα σε σχέση με τους υπόλοιπους 2. Τα δεδομένα εισόδου του αλγορίθμου των Ng. et al (2002) είναι ο πίνακας βαρών W (εφόσον στην ανάλυση που προηγήθηκε αναφερθήκαμε σε σταθμισμένο γράφο) και ο αριθμός των συστάδων k. Τα βήματα υλοποίησης της μεθόδου είναι τα ακόλουθα:

1. Κατασκευή του κανονικοποιημένου πίνακα Laplace L_{sym}.
2. **Υπολογισμός των πρώτων k ιδιοδιανυσμάτων $v_1, \dots, v_k$ του L_{sym}.**
3. Κατασκευή του πίνακα $V \in \mathbb{R}^{n \times k}$ του οποίου οι στήλες είναι τα $v_1, \dots, v_k$ ιδιοδιανύσματα.
4. **Κατασκευή του πίνακα $U \in \mathbb{R}^{n \times k}$ μέσω του V, κανονικοποιώντας τα αθροίσματα των γραμμών στη νόρμα 1**, ώστε $u_{ij} = \frac{v_{ij}}{(\sum_k v_{ik}^2)^{\frac{1}{2}}}$.
5. Για $i = 1, \dots, n$, έστω $y_i \in \mathbb{R}^l$ το διάνυσμα που αντιστοιχεί στην $i-$οστή γραμμή του U. (η οποία αναπαριστά την i κορυφή του γράφου)
6. Συσταδοποίηση των σημείων $(y_i)_{i=1,\dots,n}$ με τον αλγόριθμο $k-$means στις συστάδες $C_1, \dots, C_k$

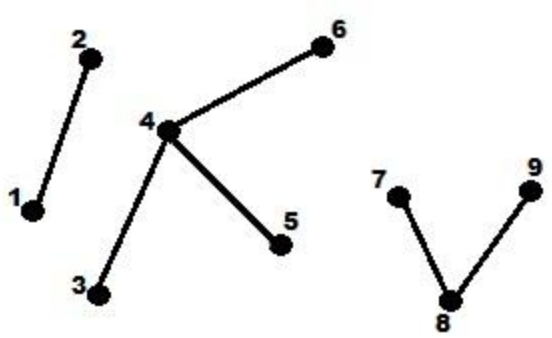
Τα δεδομένα εξόδου του αλγορίθμου είναι οι συστάδες $A_1, \dots, A_k$, με $A_i = \{j | y_j \in C_i\}$.

Ο βασικός στόχος των τριών αλγορίθμων που παρατέθηκαν, είναι η αλλαγή της αναπαράστασης των συνιστωσών της $n-$διάστατης παρατήρησης $X = (x_1, x_2, \dots, x_n)$ σε σημεία y_i του $l-$διάστατου Ευκλείδειου χώρου $\mathbb{R}^l$. Λόγω των ιδιοτήτων των πινάκων Laplace, η αλλαγή

στην αναπαράσταση των σημείων καθιστά τα χαρακτηριστικά των συστάδων πιο εμφανή και τη συσταδοποίηση ευκολότερη.

Για την καλύτερη κατανόηση του τρόπου λειτουργίας της μεθόδου φασματικής συσταδοποίησης παραθέτουμε το ακόλουθο παράδειγμα:

Σχήμα 10. Γράφος με 3 συνεκτικές συνιστώσες



Θεωρούμε το γράφο G του σχήματος, με τρεις συνεκτικές συνιστώσες: η πρώτη που αποτελείται απο τις κορυφές 1 και 2, η δεύτερη που αποτελείται τις κορυφές 3,4,5 και 6 και η τρίτη που αποτελείται απο τις κορυφές 7, 8 και 9.

Κατά τη φασματική συσταδοποίηση κάθε συνεκτική συνιστώσα του γράφου προβάλλεται σε ένα σημείο του $\mathbb{R}^l$. Στο παράδειγμά μας, εφόσον η n −διάστατη παρατήρηση συμβολίζει τις κορυφές του γράφου, $n = 9$ και εφόσον ο γράφος αποτελείται απο 3 συνεκτικές συνιστώσες, $l = 3$. Τα σημεία αυτά του $\mathbb{R}^3$ θα είναι τα: $y_1 = (1,0,0),\ y_2 = (0,1,0),\ y_3 = (0,0,1)$ όπως θα δούμε στη συνέχεια.

Ο πίνακας Laplace που αντιστοιχεί στο γράφο του σχήματος είναι ο 9×9 πίνακας:

$$L = \begin{pmatrix} \mathbf{1} & \mathbf{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{-1} & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & \mathbf{-1} & \mathbf{0} & \mathbf{0} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & \mathbf{-1} & \mathbf{0} & \mathbf{0} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{-1} & \mathbf{3} & \mathbf{-1} & \mathbf{-1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{0} & \mathbf{-1} & \mathbf{0} & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{-1} & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{-1} & \mathbf{2} & \mathbf{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & \mathbf{-1} & \mathbf{1} \end{pmatrix}$$

Κάθε ένα απο τα 3 blocks, L_i, χρωματισμένα με κόκκινο, μωβ και πράσινο χρώμα αντίστοιχα είναι ένας πίνακας Laplace, ο οποίος αντιστοιχεί στον υπογράφο της $i-$οστής συνεκτικής συνιστώσας, $i = 1,2,3$. Οι ιδιοτιμές του πίνακα Laplace είναι οι εξής: $\lambda_1 = 0$ με πολλαπλότητα 3, $\lambda_2 = 1$ με πολλαπλότητα 3, $\lambda_3 = 2$ με πολλαπλότητα 1, $\lambda_4 = 3$ με πολλαπλότητα 1 και $\lambda_5 = 4$ με πολλαπλότητα 1. Παρατηρούμε οτι η ιδιοτιμή $\lambda_1 = 0$ έχει πολλαπλότητα 3, όσες και οι συνεκτικές συνιστώσες στο γράφο. Τα πρώτα 3 ιδιοδιανύσματα του πίνακα Laplace είναι τα $v_1 = (1,1,0,0,0,0,0,0,0), v_2 = (0,0,1,1,1,1,0,0,0),\ v_3 = (0,0,0,0,0,0,1,1,1)$. Κάθε ιδιοδιάνυσμα αντιστοιχεί σε ένα block κορυφών στο γράφο. Το $v_1 = (1,1,0,0,0,0,0,0,0)$ για την πρώτη συνεκτική συνιστώσα με τις κορυφές 1,2 το , $v_2 = (0,0,1,1,1,1,0,0,0)$ για τη συνεκτική συνιστώσα με τις κορυφές 3,4,5,6 και το $v_3 = (0,0,0,0,0,0,1,1,1)$ για τη συνεκτική συνιστώσα με τις κορυφές 7,8,9. Κατασκευάζουμε τον πίνακα $V \in \mathbb{R}^{9\times 3}$ με στήλες τα 3 ιδιοδιανύσματα:

$$V = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Παρατηρούμε οτι οι 2 πρώτες γραμμές του πίνακα αντιστοιχούν στο διάνυσμα $y_1 = (1,0,0)$. Οι επόμενες 4 γραμμές του πίνακα αντιστοιχούν στο $y_2 = (0,1,0)$ και οι 3 τελευταίες στο $y_3 = (0,0,1)$. Η $i-$οστή γραμμή του πίνακα αυτού λοιπόν είναι ένα διάνυσμα με 3 συνιστώσες που αναπαριστούν την $i-$οστή κορυφή του γράφου. Τα διανύσματα που αναπαριστούν κορυφές που ανήκουν στην ίδια συνεκτική συνιστώσα ταυτίζονται και έτσι το $y_1 = (1,0,0)$ εμφανίζεται 2 φορές στον πίνακα όσες και οι κορυφές της πρώτης συνεκτικής συνιστώσας του γράφου με κορυφές τις 1 και 2, το $y_2 = (0,1,0)$ εμφανίζεται 4 φορές όσες και οι κορυφές της δεύτερης συνεκτικής συνιστώσας με κορυφές τις 3,4,5,6 και το $y_3 = (0,0,1)$ εμφανίζεται 3 φορές όσες και οι κορυφές της τρίτης συνεκτικής συνιστώσας με κορυφές τις 7,8,9. Για αυτό και αναφέραμε οτι η αλλαγή στην αναπαράσταση των σημείων καθιστά τα χαρακτηριστικά των συστάδων πιο εμφανή και τη συσταδοποίηση πλέον μέσω μιας παραδοσιακής μεθόδου όπως η k-means, ευκολότερη.

Ένα βασικό μειονέκτημα της φασματικής συσταδοποίησης είναι το οτι απαιτεί τον υπολογισμό ιδιοδιανυσμάτων του πίνακα Laplace, κάτι που σημαίνει οτι όσο μεγαλύτερος είναι ο γράφος, τόσο πιο δύσκολος και χρονοβόρος γίνεται ο ακριβής υπολογισμός των ιδιοδιανυσμάτων αυτών. Για το πρόβλημα αυτό έχουν προταθεί διάφορες προσσεγιστικές τεχνικές, όπως η μέθοδος Lanczos (Golub & Van Loan, 1989). Η ταχύτητα σύγκλισης της μεθόδου αυτής εξαρτάται απο το μέγεθος της απόστασης μεταξύ δύο διαδοχικών ιδιοτιμών (eigengap), $\gamma_k = |\lambda_k - \lambda_{k+1}|$. Όσο μεγαλύτερο είναι αυτό, τόσο πιο γρήγορα συγκλίνουν οι αλγόριθμοι.

Αναφορικά με το ποιον πίνακα Laplace θα χρησιμοποιούμε κάθε φορά αξίζει να αναφερθούν τα εξής: Αν οι κορυφές του γράφου έχουν τους ίδιους ή παρεμφερείς βαθμούς δεν υπάρχει ουσιαστική διαφορά ανάμεσα στον μη-κανονικοποιημένο και κανονικοποιημένους πίνακες Laplace. Αν όμως υπάρχει μεγάλη ανομοιογένεια στους βαθμούς του γράφου διαφορετική επιλογή πίνακα Laplace δίνει και διαφορετικά αποτελέσματα. Οι μέθοδοι φασματικής συσταδοποίησης που χρησιμοποιούν τους κανονικοποιημένους πίνακες Laplace είναι προτιμότεροι διότι κατά την εφαρμογή τους εκτελούν διαμερίσεις κατά τις οποίες η εσωτερική πυκνότητα των συστάδων, $\delta_{int}(C)$ είναι υψηλή και η μεταξύ των συστάδων πυκνότητα, $\delta_{ext}(C)$, είναι χαμηλή. Ο μη κανονικοποιημένος πίνακας Laplace ωστόσο σχετίζεται μόνο με τη $\delta_{int}(C)$. Επίσης η μη κανονικοποιημένη φασματική συσταδοποίηση δεν συγκλίνει πάντα και πολλές φορές παράγει διαμερίσεις κατά τις οποίες μια η περισσότερες συστάδες αποτελούνται απο μια μόνο κορυφή.

Μεταξύ των κανονικοποιημένων πινάκων Laplace, ο L_{rw} είναι προτιμότερος απο τον L_{sym}. Αυτό συμβαίνει διότι τα ιδιοδιανύσματα του L_{rw} που αντιστοιχούν στις χαμηλότερες (μηδενικές) ιδιοτιμές είναι τα χαρακτηριστικά διανύσματα δείκτες $\mathbb{1}_{A_i}$ των συστάδων, ενώ τα ιδιοδιανύσματα του L_{sym} που αντιστοιχούν σε μηδενικές ιδιοτιμές παράγονται μέσω πολλαπλασιασμού των ιδιοδιανυσμάτων του L_{rw}, με $D^{\frac{1}{2}}$. Συνεπώς, οι συντεταγμένες των ιδιοδιανυσμάτων που αντιστοιχούν σε κορυφές της ίδιας συνεκτικής συνιστώσας δεν είναι πλέον ίσες, κάτι που μπορεί να οδηγήσει σε ανεπιθύμητα αποτελέσματα.

Κεφάλαιο 4. Μοντέρνες Μέθοδοι Διαμερίσεων Δικτύων.

4.1 Modularity

Πριν αναφερθούμε στη συνάρτηση ποιότητας modularity (διάρθρωση/τμηματικότητα), θα δώσουμε τη διευκρίνιση δύο όρων: διαμέριση και κάλυμμα, οι οποίοι αφορούν διαιρέσεις του γράφου σε συστάδες. Με τον όρο διαμέριση (partition) εννοούμε τη διαίρεση του γράφου σε συστάδες, ώστε κάθε κορυφή να ανήκει σε μία μόνον συστάδα. Με τον όρο κάλυμμα (cover) εννοούμε τη διαίρεση του γράφου σε αλληλοκαλυπτόμενες κοινότητες. Στο κεφάλαιο αυτό θα ασχοληθούμε με μοντέρνες τεχνικές που παράγουν διαμερίσεις του δικτύου.

Συνάρτηση ποιότητας είναι ένα ποσοτικό κριτήριο αξιολόγησης του πόσο «καλή» είναι η διαμέριση του γράφου που επιτυγχάνεται από την εκάστοτε μέθοδο. Με τον όρο «καλή» διαμέριση εννοούμε οτι πληρεί τις ιδιότητες, του πλούτου (richness) και της σταθερότητας (consistency). Η ιδιότητα του πλούτου σημαίνει οτι δεδομένης μιας διαμέρισης, κάποιος μπορεί να τοποθετήσει τις ακμές μεταξύ των κορυφών του γράφου με τέτοιο τρόπο, ώστε η εν λόγω διαμέριση να είναι είναι μια φυσική συνέπεια της δομής του γράφου (αυτό μπορεί να επιτευχθεί τοποθετώντας ακμές μόνο μεταξύ κορυφών της ίδιας συστάδας). Η ιδιότητα της σταθερότητας σημαίνει οτι δεδομένης μιας διαμέρισης, διαγράφοντας εσωτερικές ακμές και προσθέτοντας μεταξύ των συστάδων ακμές, η διαμέριση που προκύπτει είναι ίδια με την αρχική. Η εκάστοτε συνάρτηση ποιότητας αντιστοιχίζει ένα «score» σε κάθε διαμέριση του γράφου, το οποίο όσο μεγαλύτερο είναι, τόσο καλύτερη είναι και αυτή. Ωστόσο το επίθετο «καλύτερη» χαρακτηρίζεται απο υποκειμενικότητα, διότι εξαρτάται απο τον ορισμό της κοινότητας που υιοθετείται και απο τη συνάρτηση ποιότητας που χρησιμοποιείται. Η πιο διάσημη συνάρτηση ποιότητας είναι η modularity των Newman και Girvan (Newman & Girvan, Finding and evaluating community structure in networks, 2004).

Η modularity αρχικά εισήχθηκε ως ένα μέτρο αξιολόγησης της ισχύος των διαμερίσεων που παρήχθεισαν απο έναν αλγορίθμο ιεραρχικής συσταδοποίησης (υποδηλώνοντας ποια διαμέριση πρέπει να κρατηθεί). Το μέτρο αυτό στηρίζεται στην ιδέα οτι γράφοι με εγγενή κοινοτική δομή συνήθως διαφέρουν απο τυχαίους γράφους: Απο τη στιγμή που οι τυχαίοι γράφοι δεν αναμένεται να έχουν κοινοτική δομή, μετρώντας την απόκλιση ανάμεσα στην πυκνότητα των ακμών του προς

εξέταση γράφου και την αναμενόμενη πυκνότητα τυχαία κατανεμημένων ακμών, θα έχουμε μια ένδειξη παρουσίας (ή έλλειψης) κοινοτικής δομής. Μεγαλύτερες τιμές της modularity υπονοούν και καλύτερη κοινοτική δομή, δεδομένου οτι το πλήθος ακμών εντός των κοινοτήτων είναι μεγαλύτερο απο το αναμενόμενο πλήθος ακμών αν αυτές ήταν τοποθετημένες τυχαία. Αυτή η αναμενόμενη πυκνότητα του τυχαίου γράφου εξαρτάται απο το μηδενικό μοντέλο (null model), δηλαδή ένα αντίγραφο του προς εξέταση γράφου, το οποίο διατηρεί κάποιες απο τις δομικές του ιδιότητες, αλλά δεν παρουσιάζει κοινοτική δομή. Οι τύποι της modularity ποικίλουν στη βιβλιογραφία. Εμείς θα παρουσιάσουμε ορισμένους απο τους πιο διαδεδομένους, οι οποίοι αφορούν διαμερίσεις μη κατευθυνόμενων αλλά και κατευθυνόμενων γράφων.

Με βάση τα προαναφερθέντα, ξεκινάμε απο τον παρακάτω τύπο για μη κατευθυνόμενους γράφους.

$$Q_u = \frac{1}{2m}\sum_{ij}(A_{ij} - P_{ij})\delta(C_i, C_j)$$

όπου το άθροισμα, όπως φαίνεται στον τύπο, αφορά όλα τα ζεύγη κορυφών, A είναι ο πίνακας γειτνίασης, $m = \frac{1}{2}\sum_{i \in V} k_i$, είναι ο συνολικός αριθμός ακμών στο γράφο και η ποσότητα P_{ij} συμβολίζει τον αναμενόμενο αριθμό ακμών ανάμεσα στις κορυφές i και j του μηδενικού μοντέλου.

Τέλος,

$$\delta(C_i, C_j) = \begin{cases} 1, \ εαν\ C_i = C_j\ (δηλαδή\ οι\ \ i\ και\ j\ ανήκουν\ στην\ ίδια\ κοινότητα) \\ 0, \ διαφορετικά \end{cases}$$

Το μηδενικό μοντέλο επιλέγεται αυθαίρετα και μπορούν να υπάρξουν πολλές παραλλαγές του. Κάποιος, για παράδειγμα, μπορεί να απαιτήσει το μηδενικό μοντέλο να διατηρεί τον ίδιο αριθμό ακμών με τον προς εξέταση γράφο και οι ακμές αυτές να είναι τοποθετημένες με την ίδια πιθανότητα μεταξύ κάθε ζεύγους κορυφών. (τυχαίος γράφος Bernoulli). Στην περίπτωση αυτή ο όρος του μηδενικού μοντέλου στην παραπάνω εξίσωση, θα είναι μια σταθερά: $P_{ij} = p = \frac{2m}{[n(n-1)]}, \forall\, i, j.$ Ωστόσο, η επιλογή του συγκεκριμένου μηδενικού μοντέλου δεν είναι αντιπροσωπευτική πραγματικών δικτύων, διότι οι βαθμοί των κορυφών του ακολουθούν την κατανομή Poisson, η οποία διαφέρει πολύ απο τις λοξές κατανομές που συναντάμε σε πραγματικά δίκτυα. Στο σύνηθες μηδενικό μοντέλο της modularity, η αναμενόμενη ακολουθία βαθμών είναι

ίδια με την ακολουθία βαθμών (δηλαδή την ακολουθία βαθμών των κορυφών του, με τους αριθμούς τοποθετημένους σε αύξουσα σειρά, με επαναλήψεις, όπως απαιτείται κάθε φορά) του γράφου που εξετάζεται. Μια κορυφή μπορεί να συνδέεται με οποιαδήποτε άλλη κορυφή του γράφου και η πιθανότητα οι κορυφές i και j, με βαθμούς k_i, k_j αντίστοιχα, να συνδέονται, υπολογίζεται ως εξής. Για να σχηματιστεί μια ακμή μεταξύ των κορυφών i και j, αρκεί να ενώσουμε δύο ημι-ακμές (stubs/half-edges) συνηφασμένες (incident) με τις κορυφές i και j αντίστοιχα. Η πιθανότητα, p_i, να επιλέξουμε τυχαία μια ημι-ακμή συνηφασμένη με την κορυφή i, είναι $\frac{k_i}{2m}$, εφόσον υπάρχουν k_i ημι-ακμές, απο τις 2m στο σύνολο, συνηφασμένες με την κορυφή i. Η πιθανότητα δύο κορυφές i κ j να συνδέονται, θα δίνεται ως $p_i p_j$, εφόσον οι ακμές είναι τοποθετημένες ανεξάρτητα η μία απο την άλλη. Συνεπώς, $p_i p_j = \frac{k_i k_j}{4m^2}$ και το πλήθος των ακμών μεταξύ των κορυφών i κ j τελικά θα είναι: $P_{ij} = 2m p_i p_j = \frac{k_i k_j}{2m}$.

Με βάση τα προαναφερθέντα, η modularity θα δίνεται ως εξής:

$$Q_u = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j)$$

Στην περίπτωση κατευθυνόμενων γράφων, (Malliaros-Vazirgiannis, 2013) έχουν προταθεί πολλές προεκτάσεις για το μέτρο της modularity. Οι Arenas et al. (Arenas, Duch, Fernandez, & Gomez, 2007) έχοντας σαν απώτερο στόχο να μειώσουν το μέγεθος του αρχικού δικτύου (κατευθυνόμενου η μή), διατηρώντας την τιμή της modularity (αυτό είναι πολύ κρίσιμο σημείο καθώς η βελτιστοποίηση της modularity-στην οποία θα αναφερθούμε εκτενώς αργότερα-είναι δύσκολο έργο), πρότειναν μια γενίκευση του τύπου της για κατευθυνόμενους γράφους. Αυτή η γενίκευση, στηρίζεται στην εξής παρατήρηση: Η ύπαρξη μιας κατευθυνόμενης ακμής (i, j) μεταξύ των κόμβων i κ j, εξαρτάται απο τον εξωτερικό και εσωτερικό βαθμό, k_i^{ext}, k_j^{in} των κορυφών i και j αντίστοιχα. Ας θεωρήσουμε οτι ο κόμβος i έχει υψηλό εξωτερικό βαθμό αλλά χαμηλό εσωτερικό βαθμό, ενώ ο κόμβος j έχει υψηλό εσωτερικό βαθμό αλλά χαμηλό εξωτερικό βαθμό. Τότε είναι πιθανότερο να παρατηρήσουμε την κατευθυνόμενη ακμή (i, j) απο τον κόμβο i στον κόμβο j, παρά την ακμή (j, i). Με βάση την υπόθεση αυτή, το μοντέλο διαμόρφωσης (configuration/null model) μπορεί να επεκταθεί στην περίπτωση κατευθυνόμενου γράφου, όπου η ακμή (i, j) απο τον κόμβο

i στον κόμβο j, θα υπάρχει με πιθανότητα $\frac{k_i^{ext} k_j^{in}}{m}$. Τότε η modularity για κατευθυνόμενους γράφους μπορεί να εκφραστεί ως:

$$Q_d = \frac{1}{m}\sum_{ij}(A_{ij} - \frac{k_i^{ext} k_j^{in}}{m})\delta(C_i, C_j)$$

Ο λόγος που το πλήθος ακμών, m, δεν πολλαπλασιάζεται με 2 στον παρονομαστή του κλάσματος, είναι οτι το άθροισμα των εξωτερικών (αντίστοιχα και εσωτερικών) βαθμών των κορυφών στην περίπτωση του κατευθυνόμενου γράφου, ισούται με m.

Οι Arenas et al., επιπλέον έδωσαν και τη σχέση που συνδέει την modularity για κατευθυνόμενους και μη κατευθυνόμενους γράφους.

$$Q_d = Q_u + \frac{1}{4m^2}\sum_{i,j}\left(k_i^{ext} - k_i^{in}\right)\left(k_j^{ext} - k_j^{in}\right)\delta(C_i, C_j).$$

Γενικότερα, η modularity μπορεί να χρησιμοποιηθεί όπως προαναφέρθηκε και ως συνάρτηση ποιότητας για μια συγκεκριμένη διαμέριση του γράφου, αλλά και ως το βασικό συστατικό εξόρυξης κοινοτικής δομής σε ένα δίκτυο. Η διαδικασία αυτή, η οποία ονομάζεται βελτιστοποίηση της modularity (modularity optimization), είναι μια απο τις κυρίαρχες προσεγγίσεις για εντοπισμό κοινωνιών σε δίκτυα.

4.2 Βελτιστοποίηση της Modularity

Η modularity είναι η πιο γνωστή και η πιο ευρέως χρησιμοποιημένη συνάρτηση ποιότητας και αντιπροσωπεύει μια απο τις πρώτες σύγχρονες προσπάθειες για την κατανόηση των θεμελιωδών αρχών του προβλήματος της συσταδοποίησης. Στη συμπαγή της μορφή, ενσωματώνονται όλα τα βασικά συστατικά και ερωτήματα που αφορούν τις κοινότητες και τις διαμερίσεις τους, όπως ο ορισμός της κοινότητας, η επιλογή ενός μοντέλου διαμόρφωσης και η ισχύς των κοινοτήτων και των διαμερίσεων τους.

Στο κεφάλαιο αυτό θα ασχοληθούμε με τις σημαντικότερες τεχνικές εντοπισμού κοινωνιών σε δίκτυα οι οποίες χρησιμοποιούν τη modularity άμεσα η έμμεσα.

Ξεκινάμε απο την παραδοχή οτι υψηλές τιμές της modularity υποδηλώνουν και καλές διαμερίσεις (κάτι που δεν ισχύει πάντοτε, αλλά το κομμάτι αυτό θα το διαπραγματευτούμε

αργότερα στην ανάλυσή μας). Συνεπώς, η διαμέριση του γράφου που αντιστοιχίζεται στη μεγαλύτερη αυτή τιμή, θα πρέπει να είναι και η καλύτερη δυνατή-ή έστω-μια πολύ καλή διαμέριση. Αυτή είναι και η βασική ιδέα για τη βελτιστοποίηση της modularity, η οποία είναι και η πιο ευρέως διαδεδομένη και διάσημη μέθοδος εντοπισμού κοινοτήτων σε γράφους.

Το ολικό μέγιστο της modularity ωστόσο, δε δύναται να επιτευχθεί, καθώς έχει αποδειχθεί (Brandes, et al., 2006) οτι η βελτιστοποίησή της είναι ένα NP-πλήρες (Non deterministic Polynomial Time) πρόβλημα. Δηλαδή ένα πρόβλημα για το οποίο δε γνωρίζουμε αλγόριθμο πολυωνυμικού χρόνου και ταυτόχρονα δεν μπορούμε να αποδείξουμε οτι τέτοιος αλγόριθμος δεν υπάρχει. Ένας αλγόριθμος πολυωνυμικού χρόνου για οποιοδήποτε NP-πλήρες πρόβλημα, θα σήμαινε την ύπαρξη αλγορίθμου πολυωνυμικού χρόνου για όλα. Αλγόριθμος πολυωνυμικού χρόνου, είναι ένας αλγόριθμος που πληρεί την εξής ιδιότητα: Υπάρχουν απόλυτες σταθερές $c > 0$ και $d > 0$ έτσι ώστε για κάθε στιγμιότυπο εισόδου μεγέθους n, ο χρόνος εκτέλεσης του φράσσεται απο cn^d στοιχειώδη υπολογιστικά βήματα. Η παραπάνω έκφραση ουσιαστικά αποτελεί μια ποσοτικοποίηση του δικαιολογημένου χρόνου εκτέλεσης ενός αλγορίθμου. Ο χώρος των λύσεων ενός προβλήματος, τείνει να αυξάνεται εκθετικά σε συνάρτηση με το μέγεθος της εισόδου του αλγορίθμου, έστω n. Εαν το μέγεθος της εισόδου αυξηθεί κατά ένα, ο αριθμός των δυνατών λύσεων αυξάνεται πολλαπλάσια. Θα θέλαμε λοιπόν, όταν το μέγεθος εισόδου αυξάνεται κατά μια σταθερή ποσότητα, για παράδειγμα κατά 2, ο αλγόριθμος να επιβραδύνεται μόνο κατά ένα σταθερό συντελεστή c. Έχουν προταθεί, ωστόσο, πολλοί αλγόριθμοι οι οποίοι βρίσκουν αρκετά καλές προσσεγίσεις του μεγίστου της modularity σε δικαιολογημένο χρονικό διάστημα. Στο σημείο αυτό θα κατηγοριοποιήσουμε και θα παραθέσουμε ορισμένες βασικές τεχνικές βελτιστοποίησης της modularity για εντοπισμό κοινωνιών σε δίκτυα.

4.2.1 Λαίμαργες Τεχνικές.

Στην ανάλυσή μας θα αναφερθούμε στη λαίμαργη μέθοδο που προτάθηκε απο τον Newman (2004), η οποία είναι η πρώτη προσπάθεια βελτιστοποίησης της modularity. Πάνω σε αυτή στηρίχθηκαν πολλές εκλεπτύνσεις και βελτιστοποιήσεις της, μερικές απο τις οποίες θα δούμε και στην εφαρμογή μας. Λαίμαργος αλγόριθμος (greedy algorithm) είναι ένας αλγόριθμος που κατασκευάζει μια λύση τμηματικά, επιλέγοντας πάντοτε το επόμενο βήμα που προσφέρει το πιο άμεσο και προφανές όφελος, δηλαδή προχωράει στο επόμενο βήμα με την απόφαση που εκείνη

τη στιγμή φαίνεται να είναι η καλύτερη για την επίλυση του προβλήματος (χωρίς αυτό να σημαίνει οτι αποδίδει πάντα τη βέλτιστη λύση). Ο αλγόριθμος του Newman ανήκει στην κατηγορία των συσσωρευτικών μεθόδων ιεραρχικής συσταδοποίησης, όπου συστάδες κορυφών συγχωνεύονται διαδοχικά για να σχηματίσουν μεγαλύτερες κοινότητες με τέτοιο τρόπο ώστε και η modularity να αυξάνεται μετά τη συγχώνευση.

Ξεκινάμε απο n συστάδες, κάθε μια εκ των οποίων περιέχει μια μοναδική κορυφή. Οι ακμές μεταξύ των κορυφών δεν είναι εξ'αρχής παρούσες, αλλά προστίθενται μια προς μία κατά τη διάρκεια της διαδικασίας. Η modularity των διαμερίσεων που διερευνάται κατά τη διάρκεια της διαδικασίας, ωστόσο, υπολογίζεται πάντα σύμφωνα με όλη την τοπολογία του γράφου, καθώς επιζητούμε να βρούμε το μέγιστο της στο χώρο των διαμερίσεων του πλήρη γράφου. Προσθέτοντας μια πρώτη ακμή στο σύνολο των μη συνδεδεμένων κορυφών, ο αριθμός των συστάδων μειώνεται απο n σε $n-1$ και παράγει μια νέα διαμέριση του γράφου. Η ακμή επιλέγεται κατά τέτοιον τρόπο ώστε η διαμέριση να δίνει τη μέγιστη αύξηση (ή ελάχιστη μείωση) της modularity, σύμφωνα με την προηγούμενη σύμβαση. Όλες οι υπόλοιπες ακμές που προστίθενται, βασίζονται στην ίδια αρχή. Εαν η εισαγωγή μιας ακμής δεν αλλάζει τη διαμέριση, δηλαδή η ακμή είναι εσωτερική μιας απο τις συστάδες που προηγουμένως δημιουργήθηκαν, η modularity παραμένει ίδια. Ο συνολικός αριθμός των διαμερίσεων που παράγονται είναι n και κάθε μια διαμέριση αποτελείται απο διαφορετικό αριθμό συστάδων, απο n μέχρι 1. Η μεγαλύτερη τιμή της modularity από το σύνολο των διαμερίσεων είναι η προσέγγιση του μεγίστου της modularity που δίνεται απο τον αλγόριθμο. Σε κάθε επαναληπτικό βήμα του αλγορίθμου, υπολογίζεται η απόκλιση ΔQ της modularity που προκύπτει αν συγχωνευτούν οποιεσδήποτε δύο κοινότητες της τρέχουσας διαμέρισης, έτσι ώστε η συγχώνευση αυτή που θα επιλεχθεί να είναι η καλύτερη δυνατή. Η συγχώνευση κοινοτήτων μεταξύ των οποίων δεν υπάρχουν ακμές, ωστόσο, δεν μπορεί ποτέ να οδηγήσει στην αύξηση της τιμής της modularity, άρα τα ζεύγη κοινοτήτων που πρέπει να ελέγχονται είναι αυτά που συνδέονται με ακμές, οι οποίες δεν μπορούν να ξεπερνούν το συνολικό πλήθος ακμών του γράφου, m. Ο υπολογισμός κάθε απόκλισης ΔQ μπορεί να γίνει σε σταθερό χρόνο, δηλαδή μπορεί να δοθεί άνω όριο στο χρόνο που χρειάζεται ο αλγόριθμος για να την υπολογίσει, το οποίο δεν επηρρεάζεται απο τις παραμέτρους εισόδου του αλγορίθμου. Αυτό σημαίνει οτι το βήμα του αλγορίθμου αυτό, απαιτεί $O(m)$ χρόνο. Η μαθηματική αυτή έκφραση ή οποία ονομάζεται και ασυμπτωτικό άνω όριο του αλγορίθμου (Big-O notation), είναι μια ποσότητα που συνήθως αναφέρεται στον απαιτούμενο χρόνο ή στην απαιτούμενη μνήμη για να

εκτελεστεί ένας αλγόριθμος. Πιο συγκεκριμένα, έστω $T(n)$ μια συνάρτηση, στην περίπτωση μας, ο χρόνος εκτέλεσης χειρότερης περίπτωσης ενός αλγορίθμου για μια είσοδο μεγέθους n. Δεδομένου μιας άλλης συνάρτησης $f(n)$ λέμε οτι η $T(n)$ είναι $O(f(n))$, δηλαδή η $T(n)$ είναι τάξης $f(n)$, αν για αρκετά μεγάλο n, η συνάρτηση $T(n)$ φράσσεται εκ των άνω απο ένα σταθερό πολλαπλάσιο της $f(n)$. Τότε γράφουμε $T(n) = O(f(n))$. Δηλαδή η $T(n)$ είναι $O(f(n))$ αν υπάρχουν σταθερές c και $n_o \geq 0$ έτσι ώστε, για όλα τα $n \geq n_o$, να ισχύει $0 \leq T(n) \leq c \cdot f(n)$.

Η απόκλιση $\Delta(Q)$, ύστερα απο τη συγχώνευση δύο κοινοτήτων, δίνεται ως εξής:

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j)$$

Για να εξηγήσουμε τους όρους, θα πρέπει πρώτα να δώσουμε τους ακόλουθους ορισμούς:

Έστω ο $n \times n$ πίνακας **e** (όπου n το συνολικό πλήθος κοινοτήτων στο δίκτυο) του οποίου το e_{ij} στοιχείο είναι το κλάσμα των ακμών στο δίκτυο που συνδέουν τις κορυφές της κοινότητας i με τις κορυφές της κοινότητας j. Δηλαδή εάν υπάρχουν m ακμές συνολικά στο δίκτυο, και k ακμές που συνδέουν κορυφές της κοινότητας i με αυτές της κοινότητας j, τότε $e_{ij} = \frac{k}{m}$. Είναι σημαντικό να βεβαιωθούμε οτι κάθε ακμή μετριέται μόνο μια φορά στον πίνακα $\mathbf{e} = (e_{ij})_{n \times n}$ δηλαδή η ίδια ακμή δεν πρέπει να εμφανίζεται εκατέρωθεν της διαγωνίου. Στην περίπτωση μας, μια ακμή που συνδέει τις κοινότητες i και j χωρίζεται σε δύο μισά, μεταξύ των ij και ji στοιχείων, κάτι που ανάγει τον πίνακα σε συμμετρικό. Αρα το e_{ij} στοιχείο, είναι το πρώτο-μισό του κλάσματος των ακμών στο δίκτυο που συνδέει τις κορυφές της συστάδας i με αυτές της συστάδας j, ώστε το συνολικό κλάσμα των ακμών αυτών να δίνεται ως $e_{ij} + e_{ji}$. Με e_{ii} συμβολίζουμε το κλάσμα των ακμών που βρίσκονται εντός της κοινότητας i. Συνεπώς, το ίχνος του πίνακα $\mathrm{Tr}\,\mathbf{e} = \sum_i e_{ii}$ θα είναι το συνολικό κλάσμα των ακμών στο δίκτυο που βρίσκονται εντός των κοινοτήτων. Η μέγιστη τιμή του αθροίσματος αυτού είναι 1 και για μια καλή διαμέριση του δικτύου αναμένουμε το ίχνος αυτό να έχει μεγάλη τιμή, χωρίς όμως αυτό να αποτελεί πληροφορία για την κοινοτική δομή. (αν τοποθετούσαμε όλες τις ακμές σε μια μοναδική κοινότητα θα είχαμε $\mathrm{Tr}\,\mathbf{e} = 1$)

Με α_i συμβολίζουμε το κλάσμα όλων των άκρων των ακμών (ends of edges) που συνδέονται με τις κορυφές της συστάδας i. Μπορούμε να υπολογίσουμε το α_i άμεσα ως εξής: $\alpha_i = \sum_j e_{ij}$. Εάν οι άκρες των ακμών συνδέονται τυχαία, τότε το κλάσμα των ακμών που συνδέοουν κορυφές εντος

της συστάδας i θα είναι $\alpha_i{}^2$. Η modularity με βάση τα προαναφερθέντα ορίζεται απο τον Newman ως:

$$Q = \sum_i (e_{ii} - a_i{}^2)$$

(Εαν μια συγκεκριμένη διαμέριση του δικτύου δε δίνει περισσότερες ακμές εντος των κοινοτήτων απο αυτές που αναμένεται να υπάρχουν τυχαία, τότε ο παραπάνω τύπος θα δίνει $Q = 0$. Τιμές διαφορετικές του 0 σημαίνουν και αποκλίσεις απο την τυχαιότητα και στην πράξη τιμές μεγαλύτερες του 0.3 υποδεικνύουν κοινοτική δομή. Εφόσον μια υψηλή τιμή της Q είναι ενδεικτική και μιας καλής διαμέρισης του δικτύου, η βελτιστοποίηση της Q σε όλες τις πιθανές διαμερίσεις ήταν το ερώτημα που αποτέλεσε τη βάση του αλγορίθμου του Newman.)

Οι ποσότητες e_{ij} είναι αρχικά ίσες με τα μισά των αντίστοιχων στοιχείων του πίνακα γειτνίασης, δηλαδή $\frac{1}{2}$ για ζεύγη κορυφών που συνδέονται με μια ακμή και 0 για αυτά που δε συνδέονται. Ύστερα απο μια συγχώνευση κοινοτήτων, τα στοιχεία του πίνακα e_{ij} που αντιστοιχούν στις συγχωνευμένες κοινότητες πρέπει να ανανεώνονται, διαδικασία που έχει ως χρόνο εκτέλεσης χειρότερης περίπτωσης $O(n)$. Συνεπώς κάθε βήμα του αλγορίθμου, έχει ως χρόνο εκτέλεσης χειρότερης περίπτωσης $O(m + n)$. Υπάρχουν $n - 1$ στο σύνολο διαδικασίες συγχώνευσης και άρα $n - 1$ επαναλήψεις για να ολοκληρωθεί ο αλγόριθμος (και να κατασκευαστεί το πλήρες δενδρόγραμμα) , συνεπώς ο χρόνος εκτέλεσης του αλγορίθμου είναι $O(m(m + n))$ ή $O(n^2)$ για αραιούς γράφους. Στη βιβλιογραφία μέχρι σήμερα, όπως αναφέραμε έχουν προταθεί βελτιστοποιήσεις και εκλεπτύνσεις του αλγορίθμου του Newman. Οι Clauset et al (Clauset, Newman, & Moore, 2004) βελτίωσαν την πολυπλοκότητα του αλγορίθμου χρησιμοποιώντας δομές δεδομένων για αραιούς γράφους, όπως σωρούς μεγίστων (max heaps), που αναδιατάσσουν τα δεδομένα σε μορφή δυαδικών δέντρων (binary trees). Τέλος, ένας άλλος αλγόριθμος λαίμαργης βελτιστοποίησης της modularity είναι αυτός των Blondel et al (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008), που εισήχθηκε για την γενική περίπτωση σταθμισμένων γράφων. Παρόλα αυτά όμως, η ακρίβεια της λαίμαργης βελτιστοποίησης δεν είναι τόσο καλή, συγκρινόμενη με άλλες τεχνικές που θα αναφερθούν στη συνέχεια.

4.2.2 Φασματική Βελτιστοποίηση

Η βελτιστοποίηση της modularity μπορεί επίσης να επιτευχθεί μέσω της χρήσης ιδιοτιμών και ιδιοδιανυσμάτων του πίνακά της (modularity matrix) **B**, του οποίου το B_{ij} στοιχείο δίνεται ως

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

Ο πίνακας της modularity στη διαδικασία βελτιστοποίησης της, έχει τον ίδιο ρόλο που έχει ο πίνακας Laplace στη φασματική διχοτόμηση και ο τύπος της modularity όπως θα δοθεί παρακάτω, είναι αντίστοιχος του ανάλογου τύπου μεγέθους αποκοπής.

Αν συμβολίσουμε με **s** το διάνυσμα που αντιπροσωπεύει οποιαδήποτε διαμέριση του γράφου σε 2 συστάδες, J και K, τότε $s_i = \begin{cases} +1, \text{εαν η i κορυφή ανήκει στη J} \\ -1, \text{εαν η i κορυφή ανήκει στη K} \end{cases}$. Σύμφωνα με αυτά η modularity μπορεί να γραφεί ώς:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j) = \frac{1}{4m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m})(s_i s_j + 1) =$$

$\frac{1}{4m} \sum_{ij} B_{ij} s_i s_j = \frac{1}{4m} s^T \boldsymbol{B} \boldsymbol{s}$. (Στη φασματική διχοτόμηση είχαμε μέγεθος αποκοπής $R = \frac{1}{4} \mathbf{s}^T \mathbf{L} \mathbf{s}$)

Η τελευταία έκφραση υποδηλώνει συνήθη γινόμενα πινάκων. Το διάνυσμα **s**, μπορεί να εκφραστεί ως προς τη βάση που αποτελείται απο ιδιοδιανύσματα $\mathbf{u}_i$ $(i = 1, \dots, n)$ του πίνακα της modularity **B**, ως $\mathbf{s} = \sum_i a_i \mathbf{u}_i$, με $a_i = \mathbf{u}_i^{\mathrm{T}} \cdot \mathbf{s}$. Εισάγοντας τα αμέσως προαναφερθέντα στοιχεία στον παραπάνω τύπο της modularity τελικά θα έχουμε:

$$Q = \frac{1}{4m} \sum_i a_i \, \boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{B} \sum_{\boldsymbol{j}} a_{\boldsymbol{j}} \boldsymbol{u}_{\boldsymbol{j}} = \frac{1}{4m} \sum_{i=1}^{n} (\boldsymbol{u}_i^{\mathrm{T}} \cdot \boldsymbol{s})^2 \, \beta_i = \frac{1}{4m} \sum_i a_i^2 \beta_i$$

Όπου $\mathbf{u}_i$ το ιδιοδιάνυσμα του πίνακα **B** που αντιστοιχεί στην ιδιοτιμή β_i . Υποθέτουμε οτι οι ιδιοτιμές διατάσσονται κατά φθίνουσα σειρά $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$. Η διαδικασία μεγιστοποίησης της modularity, επιτυγχάνεται επιλέγοντας τις ποσότητες a_i^2 έτσι ωστε να πέφτει όσο το δυνατόν μεγαλύτερο δυνατό βάρος απο το παραπάνω άθροισμα στους όρους που αντιστοιχούν στις μεγαλύτερες (πιο θετικές) ιδιοτιμές.

Όπως και στη φασματική διχοτόμηση, αυτό θα ήταν απλός στόχος αν η επιλογή του **s** δεν είχε περιορισμούς, πλην της κανονικοποίησης του: θα επιλέγαμε το **s** ένα πολλαπλάσιο του

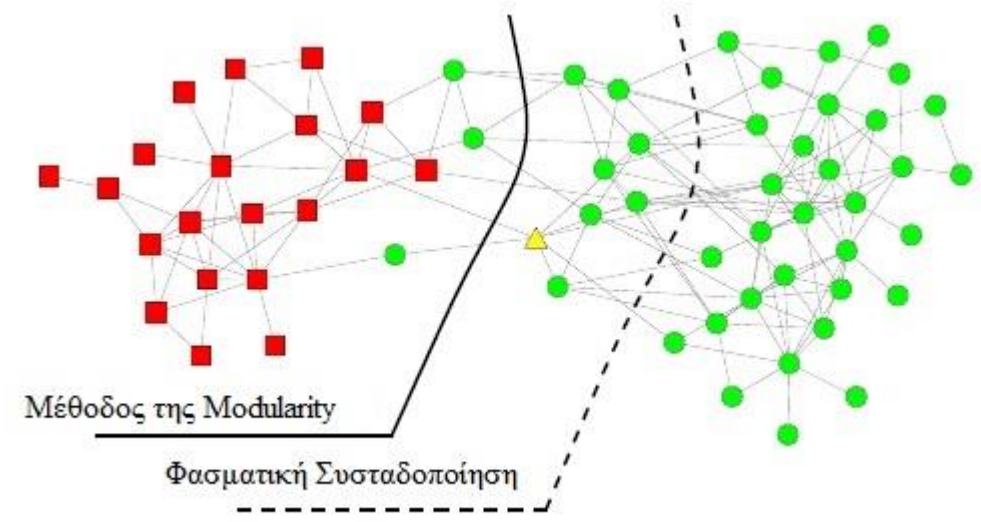
μεγαλύτερου ιδιοδιανύσματος $\mathbf{u}_1$ του πίνακα της Modularity **B**. Ωστόσο οι συνιστώσες του **s** περιορίζονται στις τιμές $s_j = \pm 1$, κάτι που σημαίνει οτι το **s** δε μπορεί επιλεγεί παράλληλο στο $\mathbf{u}_1$. Για το λόγο αυτό, λειτουργούμε προσσεγγιστικά, επιλέγοντας το **s** να είναι όσο το δυνατόν πιο «κοντά» στο να είναι παράλληλο με το $\mathbf{u}_1$. Αυτό το καταφέρνουμε θέτοντας:

$$s_j = \begin{cases} +1, & \varepsilon\alpha\nu\ u_i^{(1)} \geq 0 \\ -1, & \varepsilon\alpha\nu\ u_i^{(1)} \leq 0 \end{cases}$$

Με τη σύμβαση αυτή κατασκευάζουμε τον απλούστερο αλγόριθμο βελτιστοποίησης της Modularity για εντοπισμό κοινοτήτων στο δίκτυο: Εντοπίζουμε το ιδιοδιάνυσμα **s** που αντιστοιχεί στην πιο θετική ιδιοτιμή β_i του πίνακα της Modularity και διαμερίζουμε το δίκτυο σε δύο ομάδες, ανάλογα με το πρόσημο του διανύσματος. Ο αλγόριθμος δίνει καλύτερα αποτελέσματα συγκριτικά με αυτόν της φασματικής διχοτόμησης, όπως θα δούμε παρακάτω.

Τα αποτελέσματα του αλγορίθμου, ο οποίος εφαρμόστηκε σε ένα κοινωνικό δίκτυο δελφινιών που κατασκευάστηκε και μελετήθηκε απο τον Lusseau (Lusseau, 2003) συγκρινόμενα με αυτά της φασματικής διχοτόμησης, συνοψίζονται στο παρακάτω σχήμα. Οι κορυφές αναπαριστούν δελφίνια και οι ακμές τις σχέσεις μεταξύ ζευγών δελφινιών, όπως αυτές διαμορφώθηκαν ύστερα απο πολυετή παρατήρηση.

Σχήμα 11. Το κοινωνικό δίκτυο δελφινιών του Lusseau.



Η διακεκομμένη καμπύλη αναπαριστά τη διαίρεση του δικτύου σε δύο ισομεγέθη μέρη, όπως αυτά παρήχθεισαν μέσω της τυπικής φασματικής διχοτόμησης. Η ενιαία καμπύλη αναπαριστά τη διαίρεση του δικτύου, όπως αυτή βρέθηκε απο τον προαναφερθέντα αλγόριθμο φασματικής βελτιστοποίησης. Τα κόκκινα τετράγωνα και οι πράσινοι κύκλοι αναπαριστούν τη φυσική διαίρεση του δικτύου, όπως αυτή παρατηρήθηκε όταν η κοινότητα των δελφινιών χωρίστηκε στη μέση λόγω της αποχώρησης ενός κομβικού για αυτήν, μέλους (το μέλος αυτό αναπαρίσταται στο σχήμα, απο το κίτρινο τρίγωνο). (Newman, Finding community structure in networks using the eigenvectors of matrices, 2006)

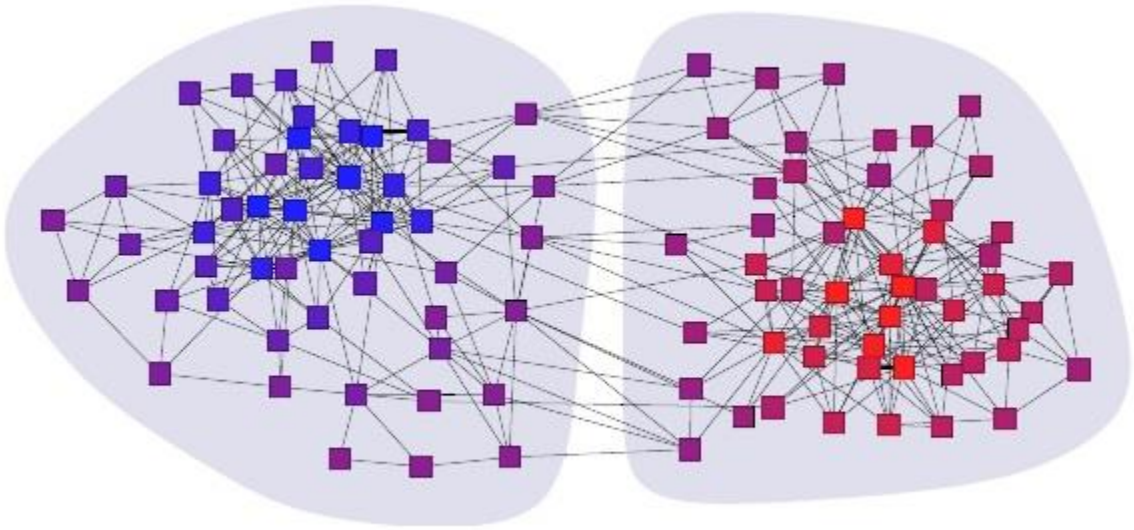
Όπως είναι εμφανές, η μέθοδος φασματικής βελτιστοποίησης της modularity που συζητήθηκε, παράγει πιο ρεαλιστικά -και κοντινά στη φυσική ομαδοποίηση των δεδομένων- αποτελέσματα, συγκριτικά με αυτά της τυπικής φασματικής διχοτόμησης. Το πλεονέκτημά της, έγκειται στο γεγονός οτι ενώ η τυπική φασματική διχοτόμηση, περιορίζεται απο το γεγονός οτι προαπαιτεί τα μεγέθη των κοινοτήτων, η φασματική βελτιστοποίηση μεγαλύτερου ιδιοδιανύσματος, δεν έχει τέτοιο περιορισμό.

Επιπλέον, οι τιμές των συνιστωσών του μεγαλύτερου ιδιοδιανύσματος $\mathbf{u}_1$, του πίνακα της Modularity, περιέχουν χρήσιμη πληροφορία αναφορικά με το πόσο «ισχυρά» συμμετέχουν οι κορυφές του δικτύου στις κοινότητες που ανήκουν. Πιο συγκεκριμένα, συνιστώσες των οποίων οι τιμές είναι κοντά στο 0, βρίσκονται στο σύνορο μεταξύ των δύο κοινοτήτων και θα μπορούσε να θεωρηθεί οτι ανήκουν και στις δύο. Η διχοτόμηση που προκύπτει μέσω της μεθόδου, μπορεί να

βελτιστοποιηθεί περαιτέρω, μετακινώντας «αυτόνομες» κορυφές απο τη μία κοινότητα στην άλλη, ώστε να έχουμε τη μέγιστη άυξηση (η ελάχιστη μείωση) της modularity. Αυτή η τεχνική εκλέπτυνσης, μπορεί επίσης να εφαρμοστεί για να τελειοποιηθούν τα αποτελέσματα και άλλων τεχνικών βελτιστοποίησης της modularity (όπως π.χ των «άπληστων» αλγορίθμων).

Στο παράδειγμα που ακολουθεί, οι κορυφές του δικτύου αναπαριστούν βιβλία πολιτικής των Ηνωμένων Πολιτειών, με ακμές να συνδέουν ζεύγη βιβλίων όπως συνήθως αγοράζονται απο τους ίδιους πελάτες μέσω της Amazon.com. Εφαρμόζοντας τον αλγόριθμο φασματικής βελτιστοποίησης, το δίκτυο διαιρείται όπως δείχνεται στην εικόνα.

Σχήμα 12. Το δίκτυο βιβλίων πολιτικής των Ηνωμένων Πολιτειών του Krebs.



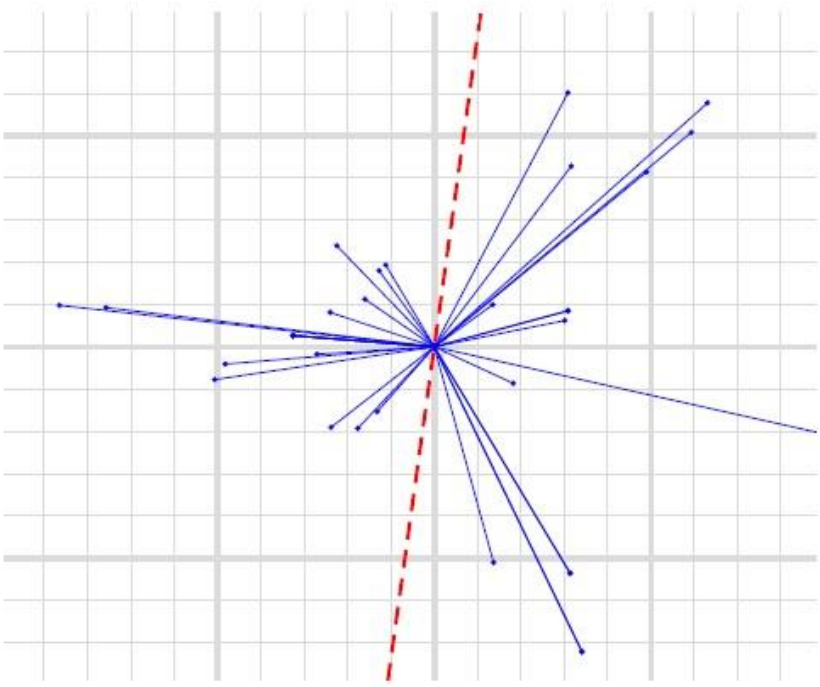
Τα χρώματα των κορυφών είναι δοσμένα σε αντιστοιχία με τις τιμές των συνιστωσών του μεγαλύτερου ιδιοδιανύσματος του πίνακα της Modularity. Οι μπλέ και κόκκινες κορυφές αναπαριστούν βιβλία ακραίων αριστερών και ακραίων δεξιών πεποιθήσεων αντίστοιχα. (Newman, Finding community structure in networks using the eigenvectors of matrices, 2006)

Η διαδικασία φασματικής βελτιστοποίησης, όπως παρουσιάστηκε προηγουμένως, έχει δύο εμφανή προβλήματα. Αρχικά, διαμερίζει το δίκτυο σε δύο κοινότητες μόνον, ενώ τα δίκτυα πραγματικού χρόνου σίγουρα περιέχουν περισσότερες κοινότητες. Επιπλέον χρησιμοποιεί μόνο

το μεγαλύτερο ιδιοδιάνυσμα του πίνακα της Modularity και αγνοεί όλα τα υπόλοιπα. Κατά αυτόν τον τρόπο όμως, αγνοεί και τη χρήσιμη πληροφορία που εμπεριέχεται στα ιδιοδιανύσματα αυτά.

Γενικεύοντας την παραπάνω διαδικασία, μπορούμε να κρατήσουμε τα κυρίαρχα p ιδιοδιανύσματα και να κατασκευάσουμε n το πλήθος p-διάστατα διανύσματα, κάθε ένα απο τα οποία θα αντιστοιχίζεται σε μια κορυφή του δικτύου, όπως και στη φασματική συσταδοποίηση που διαπραγματευτήκαμε σε προηγούμενο κεφάλαιο. Οι συνιστώσες του διανύσματος της κορυφής i είναι ανάλογες με τις p εισόδους των των ιδιοδιανυσμάτων στη θέση i. Αθροίζοντας τα διανύσματα κορυφών της ίδιας κοινότητας, ορίζονται διανύσματα κοινοτήτων. Μπορεί να δειχθεί οτι εαν τα διανύσματα δύο κοινοτήτων δημιουργούν γωνία μεγαλύτερη του $\pi/2$, κρατώντας τις κοινότητες αποκομμένες, πετυχαίνουμε μεγαλύτερες τιμές της modularity συγκριτικά με το να τις συγχωνεύαμε. Κατά αυτόν τον τρόπο, το μέγιστο της modularity σε έναν $\mathrm{p}-$διάστατο χώρο, αντιστοιχεί σε μια διαμέριση $\mathrm{p}+1$ κοινοτήτων.

Σχήμα 13. Φασματική Βελτιστοποίηση της Modularity



Χρησιμοποιώντας τα δύο πρώτα ιδιοδιανύσματα του πίνακα της modularity, οι κορυφές του δικτύου αναπαρίστανται ως σημεία ενός επιπέδου. Κόβοντας το επίπεδο με μια γραμμή που περνάει απο την πηγή (όπως η

διακεκομμένη κόκκινη γραμμή στο σχήμα) μπορούμε να έχουμε διχοτομήσεις του γράφου με πιθανόν μεγάλες τιμές της modularity. (Newman, Finding community structure in networks using the eigenvectors of matrices, 2006)

Η φασματική βελτιστοποίηση της modularity είναι πολύ γρήγορη. Το κυρίαρχο ιδιοδιάνυσμα του πίνακα της, μπορεί να υπολογιστεί με τον επαναληπτικό πολλαπλασιασμό του πίνακα **B** με ένα αυθαίρετο διάνυσμα, το οποίο επιλέγεται έτσι ωστε να μην είναι ορθογώνιο με το $\mathbf{u}_1$. Ο αριθμός των απαιτούμενων επαναλήψεων για να συγκλίνει η μέθοδος είναι $O(n)$. Κάθε πολλαπλασιασμός του πίνακα **B** απαιτεί $O(n^2)$ χρόνο, αλλά η ιδιαίτερη μορφή του επιτρέπει τον υπολογισμό του σε συντομότερο χρόνο, πολυπλοκότητας $O(m+n)$. Μια διχοτόμηση του γράφου απαιτεί $O[n(m+n)]$ χρόνο ή $O(n^2)$ για αραιούς γράφους. Η εύρεση της βέλτιστης modularity απαιτεί τη γνώση του αριθμού των επακόλουθων διχοτομήσεων του δικτύου που ισούται με το βάθος d του προκύπτοντος ιεραρχικού δέντρου. Στη χειρότερη περίπτωση $d = O(n)$, αλλά σε πρακτικές περιπτώσεις η διαδικασία σταματάει πολύ πριν να φτάσει στα φύλλα του δενδρογράμματος. Συνεπώς, για την εύρεση της βέλτιστης Q μπορούμε συμβατικά να θεωρήσουμε $(d) \sim logn,$ με $O(n^2 logn)$ πολυπλοκότητα.

4.2.3 Άλλες Τεχνικές Βελτιστοποίησης-Προσομοιωμένη Ανόπτηση.

Σύμφωνα με τη σχετική προς το θέμα βιβλιογραφία, έχουν καταγραφεί και προταθεί στρατηγικές βελτιστοποίησης της modularity οι οποίες προέρχονται απο ετερογενή επιστημονικά πεδία. Για παράδειγμα, η τεχνική της προσομοιωμένης ανόπτησης (simulated annealing) (Kirkpatrick, Gelatt, & Vecchi, 1983) η οποία χρησιμοποιεί αρχές στατιστικής μηχανικής. Η στατιστική μηχανική είναι ένας κλάδος που επικεντρώνεται στην ανάλυση συγκεντρωτικών ιδιοτήτων μεγάλου πλήθους ατόμων που βρίσκονται σε δείγματα υγρής η στερεής ύλης. Η προσομοιωμένη ανόπτηση είναι μια διαδικασία βελτιστοποίησης η οποία χρησιμοποιείται σε μεγάλο εύρος προβλημάτων. Εκτελεί μια εξερεύνηση του χώρου των δυνατών καταστάσεων, αναζητώντας το ολικό βέλτιστο (έστω το μέγιστο) μιας συνάρτησης F. Οι μεταβάσεις απο τη μια κατάσταση στην άλλη, συμβαίνουν με πιθανότητα 1, εάν η μετάβαση είναι αποδεκτή (δηλαδή η F αυξάνεται) αλλιώς γίνεται αποδεκτή με κάποια πιθανότητα μικρότερη του 1, η οποία μειώνεται εκθετικά ως προς την ακαταλληλότητα της μετάβασης, σύμφωνα με μια παράμετρο θερμοκρασίας T. Η διαδικασία ξεκινά με μια μεγάλη τιμή της T και στη συνέχεια η T μειώνεται σταδιακά. Σε κάποιο στάδιο της διαδικασίας, το σύστημα συγκλίνει σε μια σταθερή κατάσταση που μπορεί να

αποτελέσει μια αυθαίρετα καλή προσέγγιση του μεγίστου της F, ανάλογα με το πόσες καταστάσεις διερευνήθηκαν και πόσο αργά μεταβάλλεται η T. Η προσομοιωμένη ανόπτηση χρησιμοποιήθηκε για πρώτη φορά για τη βελτιστοποίηση της modularity απο τους Guimerà και Amaral (2005), των οποίων τον αλγόριθμο θα εξετάσουμε και στην εφαρμογή μας. Συνδυάζει δύο τύπους «κινήσεων»: τοπικές κινήσεις (local moves), κατά τις οποίες μια μόνο κορυφή που επιλέγεται τυχαία, μετακινείται απο τη μια συστάδα στην άλλη και καθολικές κινήσεις (global moves), οι οποίες αποτελούνται απο συγχωνεύσεις και διαχωρισμούς κοινοτήτων. Κατά την εφαρμογή της διαδικασίας, μέσω των τοπικών κινήσεων, μεμονωμένες κορυφές μετακινούνται απο τη μία συστάδα στην άλλη και η θερμοκρασία T μειώνεται μέχρι να φτάσει στην τρέχουσα τιμή ολικής βελτιστοποίησης της modularity. Οι καθολικές κινήσεις βοηθούν, μειώνοντας τον κίνδυνο του να παγιδευτεί η διαδικασία σε τοπικά ελάχιστα. Στην πρακτική της εφαρμογή, συνδυάζονται n^2 τοπικές κινήσεις με n καθολικές κινήσεις σε μια επανάληψη. Η μέθοδος της προσομοιωμένης ανόπτησης μπορεί να φτάσει πολύ κοντά στο πραγματικό μέγιστο της modularity, αλλά το μειονέκτημα της είναι οτι είναι αργή. Η πολυπλοκότητα της δεν μπορεί να υπολογιστεί ακριβώς, καθώς εξαρτάται και απο άλλες παραμέτρους πλην του μεγέθους του γράφου, όπως για παράδειγμα τη θερμοκρασία. Η μέθοδος δυνητικά να προσεγγίσει το πραγματικό μέγιστο της modularity, αλλά είναι αργή. Χρησιμοποιείται κατά κόρον για μικρούς γράφους, της τάξης των 10^4 κορυφών. (Guimerà & Amaral, 2005)

4.3 Αξιοπιστία της Modularity

Στην παράγραφο αυτή θα αναφερθούμε σε κάποια χαρακτηριστικά της modularity που κρίνονται βασικά για την εφαρμογή της και θα αξιολογηθεί η αξιοπιστία της στο πρόβλημα της συσταδοποίησης γράφου.

Μια σημαντική ερώτηση που αρχικά πρέπει να θέσουμε αφορά τη μέγιστη τιμή της modularity του δικτύου, Q_{max}. Η Q_{max} μπορεί να εκφραστεί μέσω και των εσωτερικών και των μεταξύ των συστάδων ακμών του γράφου, ως εξής:

$$Q_{max} = maxp\left\{\sum_{c=1}^{n_c}\left[\frac{l_c}{m} - (\frac{d_c}{2m})^2\right]\right\} = \frac{1}{m}\ maxp\{\sum_{c=1}^{n_c}[l_c - Ex(l_c)]\} =$$

$$-\frac{1}{m}\ minp\{\sum_{c=1}^{n_c}[l_c - Ex(l_c)]\}$$

όπου $maxp$ και $minp$ υποδηλώνουν τη μέγιστη και ελάχιστη απο όλες τις πιθανές διαμερίσεις P και $Ex(l_c) = {d_c^2}/{4m}$ ο αναμενόμενος αριθμός ακμών στη συστάδα C στο μηδενικό μοντέλο της modularity. Συνεχίζοντας,

$$Q_{max} = -\frac{1}{m}\, minp\{\sum_{c=1}^{n_c}[l_c - Ex(l_c)]\} = -\frac{1}{m}\, minp\big[(m - \sum_{c=1}^{n_c} l_c) - (m - \sum_{c=1}^{n_c} Ex(l_c))\big] = -\frac{1}{m} minp(|Cutp| - ExCutp).$$

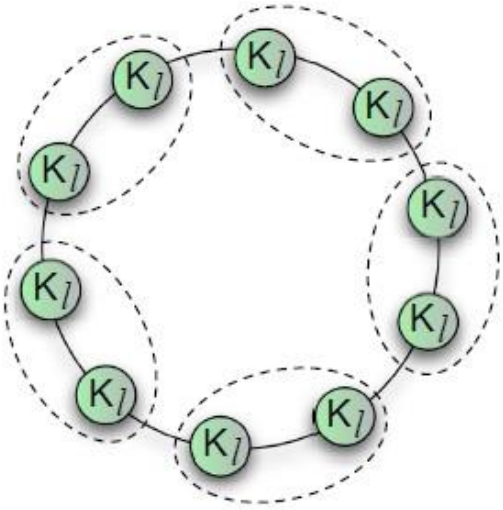
Στην τελευταία έκφραση της Q_{max}, ο όρος $|Cutp| = m - \sum_{c=1}^{n_c} l_c$ είναι ο αριθμός των εντός των συστάδων ακμών της διαμέρισης P και ο όρος $ExCutp = m - \sum_{c=1}^{n_c} Ex(l_c)$ είναι ο αναμενόμενος αριθμός των μεταξύ των συστάδων ακμών της διαμέρισης στο μηδενικό μοντέλο της modularity. Μια μεγάλη τιμή της Q_{max} δε σημαίνει απαραίτητα οτι ο γράφος έχει κοινοτική δομή. Οι τυχαίοι γράφοι θεωρείται οτι δεν έχουν κοινοτική δομή διότι η πιθανότητα σύνδεσης κορυφών σε αυτούς, είτε είναι μια σταθερά είτε είναι μια συνάρτηση του βαθμού των εν λόγω κορυφών, συνεπώς δεν υπάρχει a priori μεροληψία για σύνολα κορυφών. Ωστόσο, είναι δυνατόν να υπάρχουν τυχαίοι γράφοι με μεγάλες τιμές της Q. Σύμφωνα με τον ορισμό της modularity, η κοινοτική δομή ενός γράφου βρίσκεται σε συνάρτηση με έναν τυχαίο γράφο ίδιου μεγέθους και μια αναμενόμενη ακολουθία βαθμών. Ως εκ τούτου, η Q_{max} σε ένα γράφο αποκαλύπτει σημαντική κοινοτική δομή μόνο εάν είναι αισθητά μεγαλύτερη απο την Q_{max} τυχαίων γράφων του ίδιου μεγέθους και αναμενόμενης ακολουθίας βαθμών. Έχει αποδειχθεί (Reichardt & Bornholdt, 2006) οτι η αναμενόμενη μέγιστη τιμή της modularity για έναν τυχαίο γράφο αυξάνεται όταν ο γράφος γίνεται πιο «αραιός». Επομένως η μέθοδος της βελτιστοποίησης της modularity για εντοπισμό κοινοτήτων σε «αραιούς» γράφους κρίνεται ακατάλληλη.

Ένα πιο δομικό ζήτημα, αφορά την ικανότητα της modularity να εντοπίζει «καλές» διαμερίσεις. Εάν ο γράφος που εξετάζεται έχει ξεκάθαρη κοινοτική δομή, τότε αναμένουμε τη μέγιστη τιμή της modularity να την αποκαλύψει. Το μηδενικό μοντέλο της modularity υποθέτει οτι οποιαδήποτε κορυφή i «βλέπει» οποιαδήποτε κορυφή j και ο αναμενόμενος αριθμός ακμών μεταξύ τους είναι $p_{ij} = \frac{k_i k_j}{2m}$. Ομοίως, ο αναμενόμενος αριθμός ακμών μεταξύ δύο συστάδων A και B με συνολικούς βαθμούς K_A και K_B αντίστοιχα, θα είναι $P_{AB} = \frac{K_A K_B}{2m}$. Η μεταβολή της modularity που καθορίζεται απο την τιμή της στη διαμέριση όπου οι συστάδες A και B συγχωνεύονται και τη διαμέριση που οι συστάδες A και B είναι ξεχωριστές δίνεται ως:

$$\Delta Q_{AB} = \frac{l_{AB}}{m} - \frac{K_A K_B}{2m^2}$$

όπου l_{AB} ο αριθμός των ακμών που συνδέουν τις συστάδες A και B. Εάν $l_{AB} = 1$, δηλαδή υπάρχει μια μόνον ακμή που τις συνδέει, αναμένουμε οτι οι δύο υπογράφοι θα έπρεπε να κρατηθούν ξεχωριστά. Ωστόσο έαν $\frac{K_A K_B}{2m} < 1$, $\Delta Q_{AB} > 0$. Εαν υποθέσουμε οτι $K_A \sim K_B = K$, δηλαδή οι υπογράφοι είναι ισομεγέθεις υπο την οπτική ακμών σε αυτούς, έχει αποδειχθεί οτι εαν $K < \sim\sqrt{2m}$ και οι υπογράφοι A και B συνδέονται, η modularity είναι μεγαλύτερη εάν οι υπογράφοι A και B είναι συγχωνευμένοι (βρίσκονται στην ίδια συστάδα). Ο λόγος είναι διαισθητικός: έαν υπάρχουν περισσότερες ακμές μεταξύ των A και B απο τις αναμενόμενες, υπάρχει ισχυρή τοπολογική συσχέτιση μεταξύ των υπογράφων αυτών. Εαν οι υπογράφοι είναι επαρκώς μικροί σε μέγεθος (όσον αφορά το βαθμό τους), ο αναμενόμενος αριθμός ακμών στο μηδενικό μοντέλο μπορεί να είναι μικροτερος της μονάδος, συνεπώς ακόμα και η πιο αδύναμη δυνατή σύνδεση μεταξύ τους (δηλαδή μια μόνον ακμή) είναι αρκετή ωστε να τους κρατήσει συνδεδεμένους. Αυτό ισχύει ακόμα και εαν οι υπογράφοι είναι κλίκες, δηλαδή υπογράφοι με την μεγαλύτερη δυνατή πυκνότητα εσωτερικών ακμών, που αναπαριστούν και τις πιο ισχυρές δυνατές κοινότητες.

Σχήμα 14. Όριο ανάλυσης της Βελτιστοποίησης της Modularity.



Όπως φαίνεται και στο σχήμα, η φυσική κοινοτική δομή του γράφου που αναπαρίσταται απο τις μεμονωμένες κλίκες, δεν αναγνωρίζεται μέσω της βελτιστοποίησης της modularity, εάν οι κλίκες είναι μικρότερες απο μια κλίμακα που εξαρτάται απο το μέγεθος του γράφου. Στην περίπτωση αυτή η Q_{max} αντιστοιχεί σε μια διαμέριση του γράφου

σε συστάδες οι οποίες εμπεριέχουν δύο η περισσότερες κλίκες (όπως αυτές που αναπαρίστανται απο τα διακεκομμένα περιγράμματα). (Fortunato & Barthélemy, Resolution limit in community detection, 2007).

Δηλαδή η βελτιστοποίηση της modularity έχει ένα όριο ανάλυσης (resolution limit) το οποίο την εμποδίζει απο το να εντοπίζει κοινότητες που είναι μικρές συγκριτικά με το μέγεθος του γράφου, ακόμα και όταν αυτές είναι καλά ορισμένες, όπως οι κλίκες. Αρα, εαν η διαμέριση στην οποία αντιστοιχίζεται η Q_{max} συμπεριλαμβάνει κοινότητες με βαθμό τάξης $\sqrt{m}$ ή μικρότερο, δεν μπορούμε εκ των προτέρων να γνωρίζουμε εαν οι συστάδες είναι μεμονωμένες κοινότητες ή συνδυασμοί μικρότερων κοινοτήτων έστω και αν αυτές είναι ασθενώς συνδεδεμένες μεταξύ τους. Το πρόβλημα του ορίου ανάλυσης της διαδικασίας βελτιστοποίησης της modularity έχει μεγάλη επιρροή σε πρακτικές εφαρμογές, διότι πραγματικοί γράφοι με κοινοτική δομή, συνήθως περιέχουν κοινότητες με μεγέθη που ποικίλουν και πολλές μικρές κοινότητες μπορεί να μην εντοπίζονται. Έχουν προταθεί ανα τη βιβλιογραφία διάφοροι τρόποι αντιμετώπισης του προβλήματος που γεννά το όριο ανάλυσης, όπως η εκτέλεση επιπλέον υποδιαιρέσεων των κοινοτήτων που αποκτήθηκαν απο την τεχνική βελτιστοποίησης με σκοπό να εξαλειφθούν ενδεχόμενες τεχνητές τους συγχωνεύσεις. Σε δίκτυα πραγματικού χρόνου, έχει αποδειχθεί (Good, Montjoye, & Clauset, 2009) οτι το ολικό μέγιστο της modularity είναι αδύνατον να βρεθεί. Οι διαμερίσεις που προκύπτουν απο προσσεγγίσεις της βέλτιστης τιμής της modularity στα δίκτυα αυτά, χαρακτηρίζονται απο μεγάλη δομική ανομοιογένεια και χωρίς την ύπαρξη επιπλέον πληροφορίας, δεν μπορούμε να τις εμπιστευτούμε.

4.4 Μέθοδοι στηριζόμενοι σε Στατιστική Συμπερασματολογία

Όταν το σύνολο δεδομένων μας είναι ένας γράφος, οι μέθοδοι Στατιστικής Συμπερασματολογίας (*Statistical Inference*), χρησιμοποιώντας υποθέσεις για το πως οι κορυφές ενός γράφου συνδέονται μεταξύ τους, ενσωματώνουν ένα μοντέλο σε αυτόν το οποίο θα πρέπει να προσαρμόζεται με τον καλύτερο δυνατό τρόπο στην πραγματκή τοπολογία του γράφου. Στο κεφάλαιο αυτό, θα αναπτύξουμε κάποιες βασικές τεχνικές συσταδοποίησης που προσπαθούν να εντοπίσουν μοντέλα με την καλύτερη δυνατή προσαρμογή στο γράφο. Στις τεχνικές αυτές το μοντέλο «υποθέτει» οτι οι κορυφές του γράφου έχουν κάποιου είδους κατηγοριοποίηση, σύμφωνα με τα πρότυπα συνδεσιμότητάς τους. Θα εστιάσουμε σε μεθόδους που υϊοθετούν Μπεϋζιανή συμπερασματολογία (Bayesian inference) (Winkler, 2003) στις οποίες η καλύτερη προσαρμογή

επιτυγχάνεται μέσω μεγιστοποίησης μιας πιθανοφάνειας, τα λεγόμενα παραγωγικά μοντέλα (generative models), αλλά και σε άλλες τεχνικές βασιζόμενες στη μοντελοποίηση σε blocks (blockmodeling) (Doreian, Batagelj, & Ferligoj, 2005) και στη θεωρία πληροφορίας (information theory) (Mackay D. J., 2003).

4.4.1 Παραγωγικά Μοντέλα

Η Μπεϋζιανή συμπερασματολογία χρησιμοποιεί παρατηρήσεις για να εκτιμήσει την πιθανότητα μια δεδομένη υπόθεση να είναι αληθής. Απαρτίζεται απο δύο συστατικά: Τις ενδείξεις/δεδομένα (evidence) οι οποίες εκφράζονται μέσω της πληροφορίας D που έχουμε για το σύστημα (την οποία την αποκτούμε μέσω μετρήσεων) και ένα στατιστικό μοντέλο με παραμέτρους $\{\theta\}$. Αρχικά θεωρούμε την πιθανοφάνεια $P(D|\{\theta\})$ οτι οι παρατηρηθείσες ενδείξεις παράγονται απο ένα δοθέν σύνολο παραμέτρων $\{\theta\}$. Σκοπός είναι ο προσδιορισμός του συνόλου των παραμέτρων $\{\theta\}$ που μεγιστοποιούν την εκ των υστέρων κατανομή $P(\{\theta\}|D)$ των παραμέτρων, δοθέντος του μοντέλου και των ενδείξεων. Χρησιμοποιώντας το θεώρημα του Bayes, έχουμε:

$$P(\{\theta\}|D) = \frac{1}{Z}\, P(D|\{\theta\})P(\{\theta\})$$

όπου $P(\{\theta\})$ η εκ των προτέρων κατανομή των παραμέτρων του μοντέλου και

$$Z = \int P(D|\{\theta\})P(\{\theta\})\, d\theta.$$

Τα παραγωγικά μοντέλα διαφέρουν μεταξύ τους στο πως υπολογίζουν το εν λόγω ολοκλήρωμα και στο πως επιλέγεται η εκ των προτέρων κατανομή $P(\{\theta\})$ σε κάθε ένα απο αυτά.

Η Μπεϋζιανή συμπερασματολογιά χρησιμοποιείται συχνά στην ανάλυση και μοντελοποίηση πραγματικών δικτύων, όπως κοινωνικών και βιολογικών. Η συσταδοποίηση του γράφου μπορεί να θεωρηθεί ένα πρόβλημα συμπερασματολογίας: Οι ενδείξεις/δεδομένα στην περίπτωση μας αναπαρίστανται απο τη δομή του γράφου (πίνακας γειτνίασης ή πίνακας βαρών) και υπάρχει και ένα επιπλέον χαρακτηριστικό το οποίο αναπαρίσταται απο την κατηγοριοποίηση των κορυφών σε ομάδες. Σε όλες τις τεχνικές στις οποίες θα αναφερθούμε, μεγιστοποιείται η πιθανοφάνεια $P(D|\{\theta\})$ οτι το μοντέλο είναι σε αρμονία με την παρατηρηθείσα δομή του γραφήματος με

διαφορετικούς, κάθε φορά, περιορισμούς. Το σύνολο των παραμέτρων $\{\theta\}$, ορίζεται ώς $(\{q\},\{\pi\},k)$, όπου η παράμετρος $\{q\}$ δηλώνει τις κοινοτικές αναθέσεις των κορυφών, $\{\pi\}$ είναι το σύνολο παραμέτρων του μοντέλου και k ο αριθμός των συστάδων. Στην ανάλυση μας θα αναφερθούμε σε 2 μεθόδους: Τη μέθοδο που προτάθηκε απο τον Hastings (Hastings, 2006), και τη μέθοδο των Newman και Leicht (Newman & Leicht, Mixture Models and Exploratory analysis in Networks, 2007).

Η τεχνική του Hastings επιλέγει ως κοινοτικό μοντέλο του δικτύου, το μοντέλο *εμφωλευμένης διαμέρισης* (*planted partition model*). Σε αυτό, n κορυφές ανατίθενται σε q ομάδες: Οι κορυφές της ίδιας ομάδας συνδέονται με μια πιθανότητα p_{in} ενώ οι κορυφές διαφορετικών ομάδων συνδέονται με πιθανότητα p_{out}. Εαν $p_{in} > p_{out}$ το μοντέλο έχει ενσωματωμένη κοινοτική δομή. Η κατηγοριοποίηση των κορυφών αναπαρίσταται απο το σύνολο ετικετών $\{q_i\}$. Η πιθανότητα οτι, δοθέντος ενός γράφου, η κατηγοριοποίηση $\{q_i\}$ των κορυφών σε αυτόν είναι η σωστή σύμφωνα με το μοντέλο, είναι αναλογική με ($y \propto x$ σημαίνει οτι $y = kx$ για κάποια σταθερά k):

$$p(\{q_i\}) \propto \left\{exp\left[-\sum_{<ij>} J\delta_{q_i q_j} - \sum_{i\#j} \acute{J}\delta_{q_i q_j}/2\right]\right\}^{-1} ,$$

όπου $J = log\{[p_{in}(1-p_{out})]/[p_{out}(1-p_{in})]\}$, $\acute{J} = \log[(1-p_{in})/(1-p_{out})]$ και το πρώτο άθροισμα τρέχει για όλες τις πλησιέστερες γειτονικές κορυφές. Η μεγιστοποίηση της $p(\{q_i\})$ είναι ισοδύναμη με την ελαχιστοποίηση του ορίσματος στο εκθετικό, το οποίο ονομάζεται *Hamiltonian of a Potts* μοντέλο και ανάγεται σε πρόβλημα στατιστικής μηχανικής. Προσσεγγίστηκε απο τον Hastings μέσω *belief propagation*, αλγορίθμου «διάδοσης μηνυμάτων» (Gallager, 1963). Σε αραιούς γράφους, η πολυπλοκότητα του αλγορίθμου αναμένεται να είναι $O(nlog^{a}n)$, όπου το α προσεγγίζεται αριθμητικά. Σε γενικές γραμμές, κάποιος πρέπει να εισάγει τις παραμέτρους p_{in} και p_{out}, οι οποίες συνήθως είναι άγνωστες σε πρακτικές εφαρμογές. Ωστόσο αποδεικνύεται, οτι μπορούν να επιλεγούν μάλλον αυθαίρετα και κακές επιλογές μπορούν να αναγνωριστούν και να διορθωθούν.

Οι Newman και Leicht, πρότειναν μια προσσέγγιση για εντοπισμό κοινωνιών σε κατευθυνόμενους γράφους, βασιζόμενοι σε ένα μικτό μοντέλο και την τεχνική προσδοκίας-μεγιστοποίησης (expectation-maximization technique) (Dempster, Laird, & Rubin, 1977). Πιο συγκεκριμένα, έστω c ο αριθμός των κοινοτήτων στο δίκτυο και g_i η κοινότητα στην οποία ανήκει ο κόμβος i. Σε ποιες κοινότητες ανήκουν αρχικά οι κορυφές, τα μέλη των κοινοτήτων δηλαδή,

είναι αρχικά άγνωστα. Σκοπός του αλγορίθμου είναι να τα εξάγει απο την παρατηρηθείσα κοινοτική δομή. Για το λόγο αυτό οι συγγραφείς έχουν προτείνει τη χρήση ενός μικτού μοντέλου για τις υποκείμενες κοινότητες και τις ιδιότητές τους, στο οποίο οι παράμετροι προσαρμόζονται έτσι ωστε να εντοπιστεί η καλύτερη προσαρμογή στο δίκτυο. Το μικτό μοντέλο (mixture model) είναι ενα πιθανοτικό μοντέλο για την αναπαράσταση της παρουσίας υποπληθυσμών εντός ενός συνολικού πληθυσμού, χωρίς να απαιτεί τον προσδιορισμό του υπο-πληθυσμού στον οποίο ανήκει μια μεμονωμένη παρατήρηση ενός παρατηρηθέντος σύνολου δεδομένων. Αυτό το σημείο είναι ιδιαίτερης σημασίας διότι η μέθοδος δεν υποθέτει καμία εκ των προτέρων πληροφορία για τη δομή του δικτύου.

Με π_r συμβολίζουμε τη μεταβλητή που αναπαριστά το μέρος (fraction) των κορυφών στην κοινότητα r, (εάν υπάρχουν n κορυφές συνολικά στο δίκτυο, και k κορυφές στην κοινότητα r τότε $\pi_r = \frac{k}{n}$), με θ_{ri} την πιθανότητα της ύπαρξης κατευθυνόμενης ακμής απο μια συγκεκριμένη κορυφή στην κοινότητα r σε μια κορυφή i. (δηλαδή οι «προτιμήσεις» των κόμβων της κοινότητας r για τους κόμβους με τους οποίους συνδέονται) και με g_i την κοινότητα της κορυφής i. Το μοντέλο ορίζεται απο τις παρακάτω ποσότητες: Τα δεδομένα του δικτύου $\{A_{ij}\}$, ($\boldsymbol{A}$ ο πίνακας γειτνίασης του γράφου), τα δεδομένα που λείπουν $\{g_i\}$ και τα σύνολα των παραμέτρων του μοντέλου $\{\pi_r\}$ και $\{\theta_{ri}\}$. Εξ ορισμού τα σύνολα $\{\pi_r\}$ και $\{\theta_{ri}\}$ ικανοποιούν τις συνθήκες κανονικοποίησης $\sum_{r=1}^{c} \pi_r = 1$ και $\sum_{i=1}^{n} \theta_{ri} = 1$. Όπως έχουμε προαναφέρει, δεδομένου του ορισμού της κοινότητας ως ένα σύνολο κορυφών που έχουν παρόμοια πρότυπα σύνδεσης μεταξύ τους, ο εντοπισμός κοινοτήτων στο δίκτυο μπορεί να διατυπωθεί ως ένα πρόβλημα μεγιστοποίησης πιθανοφάνειας. Στην περίπτωση μας, ο σκοπός είναι η μεγιστοποίηση της πιθανότητας $P(\boldsymbol{A}, g|\pi, \theta)$ δηλαδή της πιθανότητας οτι τα δεδομένα δημιουργήθηκαν απο το δεδομένο μοντέλο, σύμφωνα με τις παραμέτρους του.

Μια συνήθης προσέγγιση είναι η μεγιστοποίηση της συνάρτησης του λογαρίθμου της πιθανοφάνειας αντι για την πιθανοφάνεια καθεαυτή. Η καλύτερη κατάταξη των κορυφών ανιστοιχίζεται στη μέγιστη τιμή του μέσου λογαρίθμου της πιθανοφάνειας, $\bar{\mathcal{L}}$, οτι το μοντέλο, που περιγράφεται απο τις τιμές των παραμέτρων $\{\pi_r\}, \{\theta_{ri}\}$ προσαρμόζεται στον πίνακα γειτνίασης $\boldsymbol{A}$ του γράφου. Σύμφωνα με το θεώρημα του Bayes, οι αναμενόμενες πιθανότητες q_{ir} οτι ο κόμβος i ανήκει στην κοινότητα r, μπορεί να εκφραστεί μέσω των $\{\pi_r\}$ και $\{\theta_{ri}\}$ ως:

$$q_{ir} = \frac{\pi_r \prod_j \theta_{rj}^{A_{ij}}}{\sum_s \pi_s \prod_j \theta_{sj}^{A_{ij}}}.$$

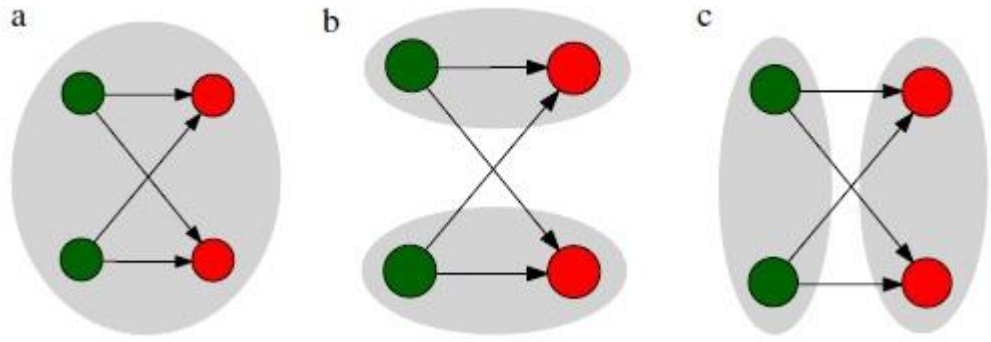
Σύμφωνα με τους συγγραφείς, η μεγιστοποίηση του μέσου λογαρίθμου πιθανοφάνειας, $\bar{\mathcal{L}}$, συμβαίνει όταν

$$\pi_r = \frac{1}{n}\sum_i q_{ir}, \qquad \theta_{rj} = \frac{\sum_i A_{ij} q_{ir}}{\sum_i k_i^{out} q_{ir}},$$

Όπου k_i^{out} ο εξωτερικός βαθμός της κορυφής i. Συνδυάζοντας τις παραπάνω εξισώσεις, ένας αλγόριθμος προσδοκίας-μεγιστοποίησης μπορεί να εφαρμοστεί για να παράγει τις πιθανότητες q_{ir}. Η σύγκλιση του αλγορίθμου σύμφωνα με τους συγγραφείς είναι γρήγορη και μπορεί να εφαρμοστεί σε αρκετά μεγάλα γραφήματα, της τάξης των 10^6 κορυφών.

Το βασικό πλεονέκτημα αυτής της μεθόδου είναι οτι είναι ανεξάρτητη της υποκείμενης κοινοτικής δομής του δικτύου, κάτι που την καθιστά ικανή να αποκαλύπτει διαφορετικούς τύπους κοινοτικής δομής. Ωστόσο, ο αριθμός των κοινοτήτων είναι μια παράμετρος που πρέπει να προσδιορίζεται εκ των προτέρων, αλλά οι συγγραφείς υποστηρίζουν οτι ο μπορεί να εξαχθεί απο τα δεδομένα. Το μειονέκτημα της μεθόδου (Ramasco & Mungan, 2008) είναι το εξής: Παρατηρήθηκε οτι στο μοντέλο των Newman και Leicht η πιθανότητα θ_{ri} οτι ένας κόμβος i έχει μια εισερχόμενη ακμή κατευθυνόμενη απο έναν κόμβο στην κοινότητα r, υποδηλώνει οτι κάθε κοινότητα r, θα πρέπει να έχει έναν τουλάχιστον κόμβο με μη μηδενικό εξωτερικό βαθμό. Αυτός ο περιορισμός όμως μπορεί να επηρεάσει τις κοινότητες που παράγει ο αλγόριθμος όπως στο σχήμα

Σχήμα 15. Το πρόβλημα της μεθόδου των Newman και Leicht.



Στο σχήμα παρατηρούμε οτι το μικτό μοντέλο που προτάθηκε απο τους Newman και Leicht, έχει πρόβλημα στο να κατανήμει τους κόμβους του δικτύου σε κοινότητες, όπως φαίνεται και στο διμερή γράφο του σχήματος. Τα χρώματα εδώ δηλώνουν τις κλάσεις των κορυφών. Οι πιθανές ομαδοποιήσεις της μεθόδου είναι αυτές που παρουσιάζονται στις σκιασμένες περιοχές των (a) και (b), ενώ η φυσική ομαδοποίηση που παρουσιάζεται στο σχήμα (c) δεν μπορεί να ανιχνευθεί. Για το λόγο αυτό προτάθηκε μια γενίκευση της τεχνικης προσδοκίας-μεγιστοποίησης (EM), με τέτοιο τρόπο που η κατεύθυνση των ακμών δεν περιορίζει την πιθανή ανάθεση των κόμβων σε γκρουπ. Αυτό μπορεί να επιτευχθεί αντικαθιστώντας τις πιθανότητες των ακμών θ_{ri} με τρεις νέους τύπους πιθανοτήτων:

i. Η $\overrightarrow{\theta_{ri}}$ αναπαριστά την πιθανότητα μια ακμή να κατευθύνεται απο έναν κόμβο στην κοινότητα r σε έναν κόμβο i.
ii. Η $\overleftarrow{\theta_{ri}}$ αναπαριστά την πιθανότητα μια ακμή να κατευθύνεται απο τον κόμβο i σε έναν κόμβο εντός της κοινότητας r.
iii. Η $\overleftrightarrow{\theta_{ri}}$ αναπαριστά την πιθανότητα η ακμή να είναι αμφίδρομη ανάμεσα σε έναν κόμβο i και τον κόμβο εντός μιας κοινότητας r.

Κατά τον τρόπο αυτό, το πρόβλημα αναδιαμορφώνεται με νέες παραμέτρους αυτή τη φορά και η γενικευμένη τεχνική προσδοκίας-μεγιστοποίησης είναι ικανή να εντοπίσει ευρύ φάσμα διαφορετικών τύπων κοινοτήτων.

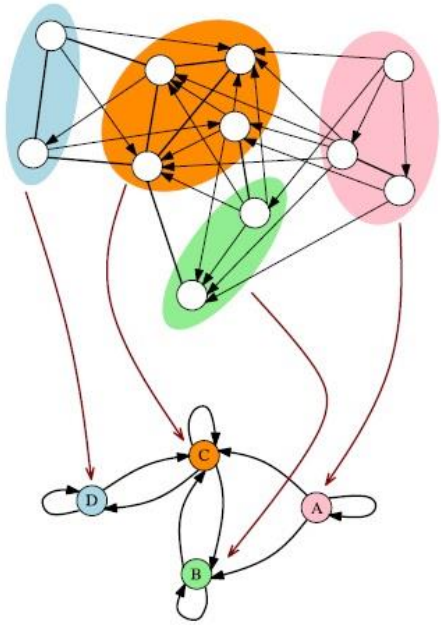
Συνοψίζοντας τα παραπάνω, στη μέθοδο του Hastings μεγιστοποιείται η πιθανοφάνεια $P(D|\{q\},\{\pi\},k)$ σε όλο το σύνολο των πιθανών κοινοτικών εκχωρήσεων των κορυφών, δοθέντος του αριθμού των συστάδων k και των παραμέτρων του μοντέλου (δηλαδή των πιθανοτήτων p_{in} και p_{out}). Στη μέθοδο των Newman και Leicht μεγιστοποιείται η πιθανοφάνεια $P(D|\{q\},\{\pi\},k)$ δοθέντος του αριθμού των συστάδων, για όλες τις δυνατές επιλογές των παραμέτρων του μοντέλου και των κοινοτικών εκχωρήσεων των κορυφών, με την παραγωγή των βέλτιστων επιλογών και για τις δύο μεταβλητές μέσω της διαδικασίας προσδοκίας-μεγιστοποίησης.

4.4.2 Μοντελοποίηση σε Μπλοκ και Θεωρία Πληροφορίας

Η μοντελοποίηση σε μπλοκς (Blockmodeling) είναι μια προσέγγιση που έχει χρησιμοποιηθεί εκτενώς για την ανάλυση και περιγραφή της δομής κοινωνικών δικτύων και κατ'επέκταση

σχεσιακών δεδομένων. Ο στόχος της, είναι να αναπαραστήσει ένα μεγάλο και πιθανόν «ασυναφές» (incoherent) δίκτυο μέσω μιας μικρότερης δομής, η οποία μπορεί να ερμηνευθεί ευκολότερα δηλαδή να αποσυνθέσει ένα γράφο σε κλάσεις κορυφών με κοινές ιδιότητες για να διευκολύνει την ερμηνεία του. Κατά τη μοντελοποίηση αυτή, οι κόμβοι του δικτύου ομαδοποιούνται σε κλάσεις ισοδυναμίας σύμφωνα με το πόσο «ισοδύναμοι» (equivalent) είναι. Ανά τη βιβλιογραφία υπάρχουν δύο βασικοί ορισμοί τοπολογικής ισοδυναμίας κορυφών: (i) Η διαρθρωτική ισοδυναμία (structural equivalence) (Lorrain & White, 1971) σύμφωνα με την οποία οι κόμβοι είναι ισοδύναμοι εαν έχουν τα ίδια μοτίβα σύνδεσης με τους ίδιους γείτονες (δηλαδή αν έχουν τους ίδιους συνδέσμους/ακμές με τους ίδιους γειτονικούς κόμβους. Σε ένα κοινωνικό δίκτυο για παράδειγμα, μπορεί να υπάρχουν διαφορετικοί τύποι συνδέσμων/ακμών) και (ii) Η κανονική ισοδυναμία (regular equivalence), σύμφωνα με την οποία οι κόμβοι είναι ισοδύναμοι εαν έχουν τα ίδια ή παρεμφερή μοτίβα σύνδεσης με διαφορετικούς γείτονες. (δηλαδή κόμβοι μιας κλάσης έχουν παρεμφερή μοτίβα σύνδεσης με κόμβους των άλλων κλάσεων, όπως για παράδειγμα γονείς/παιδιά). Η κανονική ισοδυναμία δεν απαιτεί οι ακμές/σύνδεσμοι να περιορίζονται σε συγκεκριμένους κόμβους-στόχους, δηλαδή είναι μια έννοια γενικότερη της διαρθρωτικής ισοδυναμίας. Κόμβοι που είναι διαρθωτικά ισοδύναμοι είναι και κανονικά ισοδύναμοι, αλλά το αντίστροφο δεν ισχύει. Η διαρθρωτική ισοδυναμία μπορεί να επεκταθεί σε πιθανοτικά μοντέλα, όπου εισάγεται η έννοια της στοχαστικής ισοδυναμίας (stochastic equivalence): οι κόμβοι της ίδιας ομάδας λέγονται στοχαστικά ισοδύναμοι, εαν οι πιθανότητες σύνδεσης τους με οποιαδήποτε άλλο κόμβο στο γράφο είναι ίδιες. Η στοχαστική ισοδυναμία περιγράφεται ως εξής (Holland, Laskey, & Leinhardt, 1983) : "*Θα λέμε οτι δύο κόμβοι a και b είναι στοχαστικά ισοδύναμοι, άν και μόνο αν η πιθανότητα οποιουδήποτε συμβάντος για τα δίκτυα παραμένει αναλλοίωτη απο την εναλλαγή των κόμβων a και b*". Ο ορισμός αυτός, είναι και η βάση των στοχαστικών μεθόδων μοντελοποίησης σε μπλοκ (stochastic blockmodeling methods), στις οποίες κάθε ζεύγος κορυφών που ανήκει στην ίδια κοινότητα είναι και στοχαστικά ισοδύναμο. Στην εν λόγω περίπτωση κορυφές τοποθετούνται σε κλάσεις, έτσι ώστε οι πιθανότητες συνδέσμου μιας κορυφής με τις κορυφές του υπόλοιπου γράφου να είναι ίδιες για κορυφές που ανήκουν στην ίδια κλάση. Τα μοντέλα στοχαστικών μπλοκς μπορούν να θεωρηθούν παραγωγικά μοντέλα για κοινότητες ή μπλοκς σε δίκτυα και το πρόβλημα ανάγεται σε ένα πρόβλημα εκτίμησης μέγιστης πιθανοφάνειας.

Σχήμα 16. Αναπαράσταση της μοντελοποίησης σε μπλοκς.



Ο γράφος στο επάνω μέρος του σχήματος αντιστοιχεί στον αρχικό κατευθυνόμενο γράφο και ο γράφος στο κάτω μέρος του σχήματος, στην αναπαράστασή του, μέσω της μοντελοποίησης σε μπλοκς. ((Batagelj, Mirvar, & Pajek, 2002)

Στην ανάλυση μας θα αναφερθούμε στο στοχαστικό μπλοκ-μοντέλο που πρότειναν οι Yang et al. (2010), το οποίο ονομάζεται μοντέλο συνδέσμων δημοτικότητας και παραγωγικότητας. (PPL-Popularity and Productivity link Model). Σκοπός του είναι η μοντελοποίηση εισερχόμενων και εξερχόμενων συνδέσμων, ταυτόχρονα. Οι συγγραφείς, για να επιτύχουν το στόχο αυτό, εισήγαγαν δύο λανθάνουσες (latent) μεταβλητές, που ονομάζονται παραγωγικότητα και δημοτικότητα, ωστε να τιθασσεύσουν λεπτομερώς (explicitly capture) εισερχόμενους και εξερχόμενους συνδέσμους αντίστοιχα. Στη γενική του μορφή, το PPL μοντελοποιεί την απο κοινού πιθανότητα $\Pr(i,j) = \Pr(i\rightarrow, j\leftarrow)$ δηλαδή την πιθανότητα οτι υπάρχει μια κατευθυνόμενη ακμή απο τον κόμβο i στον κόμβο j. Προκειμένου να δοθεί έμφαση στους διαφορετικούς ρόλους των κόμβων i και j ο συμβολισμός $\Pr(i\rightarrow, j\leftarrow)$ υποδηλώνει οτι ο i έχει το ρόλο του κόμβου που παράγει το σύνδεσμο και ο j παίζει το ρόλο του κόμβου που λαμβάνει το σύνδεσμο. Η πιθανότητα αυτή μοντελοποιείται ως ακολούθως:

$$\Pr(i\rightarrow, j\leftarrow) = \sum_c \Pr(i\rightarrow|k)\Pr(j\leftarrow|k)\Pr(k)$$

$$= \sum_k \left(\frac{\gamma_{ik} a_i}{\sum_i \gamma_{ic} a_i} \frac{\gamma_{jk} \beta_j}{\sum_i \gamma_{ik} \beta_i} \sum_i \gamma_{ik} c_i \right),$$

όπου γ_{ik} αναπαριστά την πιθανοτήτα ο κόμβος i να ανήκει στην κοινότητα k, a_i την παραγωγικότητα του κόμβου i (δηλαδή, πόσο πιθανόν είναι μια ακμή να ξεκινάει απο τον i), β_j τη δημοτικότητα του κόμβου j (δηλαδή πόσο πιθανόν είναι μια ακμή να λαμβάνεται απο τον j) και c_i το βάρος του κόμβου i για την απόφαση της πιθανότητας $\Pr(k)$ με την οποία ο κόμβος i ανήκει στην κοινότητα k. Για την παραπάνω εξίσωση κατασκευάζεται μια παραγωγική διαδικασία και μέσω ενός αλγορίθμου προσδοκίας μεγιστοποίησης παράγεται η προσσέγιση της μέγιστης πιθανοφάνειας. Για έναν κατευθυνόμενο γράφο $G = (V, E)$, θα χρησιμοποιήσουμε τους ακόλουθους συμβολισμούς: $V = \{1, \dots, N\}$ για τους κόμβους και $E = \{(i, j) | s_{ij} \neq 0\}$ για τις κατευθυνόμενες ακμές, όπου η ποσότητα s_{ij} καταγράφει την τιμή που σχετίζεται με την ακμή απο τον κόμβο i στον κόμβο j. Η s_{ij} ποσότητα μπορεί να είναι είτε δυαδική, και να υποδηλώνει το αν υπάρχει ακμή απο τον κόμβο i στον κόμβο j, είτε να έχει μη αρνητικές τιμές, και να υποδηλώνει το βάρος του συνδέσμου. Με K θα συμβολίζουμε τον αριθμό των κοινοτήτων, $z_i \in \{1, \dots, K\}$ τη μεταβλητή της κοινότητας του κόμβου i και με $\gamma_i = (\gamma_{i1}, \dots, \gamma_{iK})$ συμβολίζουμε τις κοινοτικές συμμετοχές (community memberships) του κόμβου i. Με άλλα λόγια γ_{ik} είναι η πιθανότητα ο κόμβος i να ανήκει στην κοινότητα k.

Τα βήματα της παραγωγικής διαδικασίας του PPL είναι τα ακόλουθα:

- Κατασκευή μιας κοινότητας z, σύμφωνα με την εκ των προτέρων κατανομή $\pi_1, \pi_2, \dots, \pi_K$, όπου π_k υπολογίζεται ως εξής : $\pi_k = \sum_{i=1}^{N} \gamma_{ik} c_i$.
- Δοθέντος της κοινότητας z, η δεσμευμένη πιθανότητα συνδέσμων, δίνεται ως:

$$\Pr(i\rightarrow, j\leftarrow | \mathrm{z}) = \Pr(i\rightarrow | \mathrm{z}) \Pr(j\leftarrow | \mathrm{z}) =$$

$$= \frac{\gamma_{iz} a_i}{\sum_i \gamma_{iz} a_i} \frac{\gamma_{jz} \beta_j}{\sum_i \gamma_{iz} \beta_i}$$

Υπάρχουν δύο μοναδικά χαρακτηριστικά στην παραπάνω παραγωγική διαδικασία:

- Η εκ των προτέρων πιθανότητα $\pi_k = \sum_{i=1}^{N} \gamma_{ik} c_i$ για έναν σύνδεσμο να παράγεται στην κοινότητα k, κατασκευάζεται ως το σταθμισμένο άθροισμα κοινοτικής συμμετοχής των κόμβων γ_{ik}, όπου c_i το βάρος του κόμβου i στο συνδυασμό. Αυτή η

κατασκευή ενισχύει τη σταθερότητα μεταξύ των πιθανοτήτων κοινοτικής συμετοχής γ_{ik} και της πιθανότητας $\{\pi_k\}_{k=1}^{K}$.

- Στην έκφραση της δεσμευμένης πιθανότητας $\Pr(i\rightarrow, j\leftarrow|z)$ που δόθηκε πιο πάνω, τα δύο άκρα της ακμής $i\rightarrow j$ αντιμετωπίζονται με διαφορετικό τρόπο όταν αυτή μοντελοποιείται: Εκτός απο την εξάρτησή τους απο τις πιθανότητες κοινοτικών συμμετοχών γ_{ik} και γ_{jk}, οι $\Pr(i\rightarrow|z)$ και $\Pr(j\leftarrow|z)$ μοντελοποιούνται επίσης απο τις ποσότητες α_i (παραγωγικότητα του κόμβου i) και b_j(δημοτικότητα του κόμβου j), αντίστοιχα, κάτι που οδηγεί στη διαφοροποίηση των ρόλων των δύο κόμβων.

Μέσω της σχέσης για την απο κοινού πιθανότητα συνδέσμων $\Pr(i\rightarrow, j\leftarrow)$ που δόθηκε πιο πάνω, ο λογάριθμος της πιθανοφάνειας των συνδέσμων μπορεί να γραφεί ως:

$$\mathcal{L}(\alpha, b, c, \gamma) = \sum_{(i,j)\in E} s_{ij} log \frac{\gamma_{ik}a_i}{\sum_i \gamma_{ik}a_i} \frac{\gamma_{jk}\beta_j}{\sum_i \gamma_{ik}\beta_i} \sum_i \gamma_{ik}c_i$$

Οι παράμετροι γ, α, b μπορούν να βρεθούν μέσω της μεγιστοποίησης του λογαρίθμου της πιθανοφάνειας $\mathcal{L}(\alpha, b, c, \gamma)$.

Το PPL μοντέλο επιδέχεται παραλλαγές, ανάλογα με τους περιορισμούς των παραμέτρων α, b, c. Οι τρεις βασικές του παραλλαγές, στις οποίες θα αναφερθούμε μόνο ονομαστικά είναι το Popularity Link Model (PoL), το Productivity Link Model (Prl) και το Regularized PPL Model (PPL-D). Έχει αποδειχθεί οτι υπο τη βέλτιστη λύση, η απο κοινού πιθανότητα συνδέσμων $\Pr(i\rightarrow, j\leftarrow)$ και για τα 3 μοντέλα είναι η ίδια. Οι αντίστοιχοι αλγόριθμοι προσδοκίας-μεγιστοποίησης συγκλίνουν στις προσεγγίσεις μέγιστης πιθανοφάνειας (MLE-Maximum Likelihood Estimation) για το PoL και PrL μοντέλα, και τη προσέγγιση μέγιστης εκ των προτέρων πιθανότητας (MAP-Maximum A-posteriori Probability) για το PPL-D μοντέλο. Πιο αξιόπιστα αποτελέσματα εκ των τριών σύμφωνα με εφαρμογή τους, έχει αποδειχθεί οτι δίνει το PPL-D.

Μια εξέχουσα μεθοδολογία για την εξαγωγή της κοινοτικής δομής ενός δικτύου, είναι αυτή που εφαρμόζει τις αρχές της θεωρίας πληροφορίας (information theory) και συμπίεσης (compression). Γενικότερα, η ύπαρξη κοινοτήτων σε ένα δίκτυο αντιπροσωπεύει κάποια δομικά πρότυπα τα οποία μπορούν να χρησιμοποιηθούν για την αποτελεσματική συμπίεση του. Στην ανάλυσή μας, θα αναφερθούμε εκτενώς στο πλαίσιο θεωρίας πληροφορίας για τον εντοπισμό κοινοτικής δομής που εισήχθηκε απο τους Rosvall και Bergstrom (2007) και μέσω της μεθόδου

αυτής θα εξηγήσουμε πως λειτουργεί η θεωρία πληροφορίας για τον εντοπισμό κοινωνιών σε δίκτυα. (Rosvall & Bergstrom, An information-theoretic framework for resolving community structure in complex networks, 2007)

Πολλά αντικείμενα στη φύση, απο πρωτεϊνες μέχρι άτομα, αλληλεπιδρούν σε ομάδες που συνθέτουν κοινωνικά, τεχνολογικά και βιολογικά συστήματα. Οι ομάδες αυτές, διαμορφώνουν ένα διακριτό ενδιάμεσο πλαίσιο, ανάμεσα στις μικροσκοπικές και μακροσκοπικές δυνατές περιγραφές του συστήματος στο οποίο ανήκουν και η δομή τους συχνά μπορεί να συνδέεται με πτυχές της λειτουργίας του συστήματος, όπως η ευρωστία (robustness) και η σταθερότητα (stability). Όταν απεικονίζουμε τις αλληλεπιδράσεις μεταξύ των συνθετικών στοιχειών ενός πολύπλοκου συστήματος σε ένα δίκτυο με κόμβους που συνδέονται με ακμές, αυτές οι ομάδες των αντικειμένων που αλληλεπιδρούν, διαμορφώνουν τμήματα/ενότητες (modules) υψηλής εσωτερικής σύνδεσης, τα οποία είναι ασθενώς συνδεδεμένα μεταξύ τους. Μπορούμε λοιπόν να κατανοήσουμε τη δομή ενός χαοτικού και περίπλοκου δικτύου, με το να εντοπίσουμε τα τμήματα ή κοινότητες απο τις οποίες αυτό αποτελείται. Όταν περιγράφουμε ένα δίκτυο ως ένα σύνολο διασυνδεδεμένων τμημάτων, τονίζουμε ορισμένες «ομαλότητες» (regularities) της δομής του δικτύου ενώ αγνοούμε τις σχετικά ασήμαντες λεπτομέρειες. Επομένως, μια περιγραφή του δικτύου σε τμήματα, μπορεί να θεωρηθεί συμπίεση της τοπολογίας του και το πρόβλημα προσδιορισμού των κοινοτήτων σε αυτό, ένα πρόβλημα εύρεσης μιας αποδοτικής συμπίεσης της δομής του.

Η άποψη αυτή υποδηλώνει οτι πρόβλημα προσδιορισμού της κοινοτικής δομής ενός σύνθετου δικτύου, μπορούμε να το προσεγγίσουμε ως ένα θεμελιώδες πρόβλημα θεωρίας πληροφορίας. Στο σημείο αυτό, θα παραθέσουμε τις βάσεις της θεωρίας πληροφορίας που απαιτούνται για την προσέγγιση του προβλήματος εντοπισμού κοινοτήτων και θα διερευνήσουμε τα πλεονεκτήματά της συγκριτικά με άλλες μεθόδους. Θεωρούμε τη διαδικασία περιγραφής ενός σύνθετου δικτύου απο μια απλουστευμένη σύνοψη της τμηματικής του δομής, ως μια διαδικασία επικοινωνίας. Η πλήρης δομή του σύνθετου δικτύου είναι μια τυχαία μεταβλητή X. Ενας διαβιβαστής (signaler) γνωρίζει την πλήρη μορφή του δικτύου, X και έχει στόχο να μεταφέρει όση περισσότερη απο την πληροφορία αυτή κατα έναν μειωμένο τρόπο, σε έναν δέκτη του σήματος (signal receiver). Για να το κάνει αυτό, ο διαβιβαστής κωδικοποιεί την πληροφορία για το X σε μια απλοποιημένη περιγραφή του, Y. Στέλνει το κωδικοποιημένο μήνυμα μέσω ενός αθόρυβου καναλιού

επικοινωνίας. Ο δέκτης του σήματος, παρατηρεί το μήνυμα Y και στη συνέχεια το «αποκωδικοποιεί», χρησιμοποιώντας το για να κάνει υποθέσεις Z για τη δομή του αρχικού δικτύου, X.

Υπάρχουν πολλοί διαφορετικοί τρόποι να περιγράψουμε ένα δίκτυο X, μέσω μιας απλούστερης περιγραφής Y. Η απάντηση στο ποιος απο τους τρόπους είναι ο καλύτερος εξαρτάται απο το τι θέλουμε κάθε φορά να κάνουμε με την περιγραφή. Ωστόσο, η θεωρία πληροφορίας δίνει μια ενδιαφέρουσα γενική απάντηση στο ερώτημα αυτό: Δοθέντος ενός συνόλου υποψήφιων περιγραφών Y_i, η καλύτερη περιγραφή Y της τυχαίας μεταβλητής X, είναι αυτή που μεγιστοποιεί την κοινή πληροφορία $I(X:Y)$ μεταξύ της περιγραφής και του δικτύου.

Απο τη στιγμή που στόχος μας είναι ο εντοπισμός κοινοτικής δομής του X, θα διερευνήσουμε περιγραφές Y που συνοψίζουν τη δομή του δικτύου X, απαριθμώντας τις κοινότητες/τμήματα μέσα σε αυτό, περιγράφοντας τις μεταξύ τους σχέσεις. Η μέθοδος κωδικοποίησης του δικτύου που ακολουθούν οι συγγραφείς, ξεκινά με τις ακόλουθες υποθέσεις.

Θεωρούμε ένα μη σταθμισμένο και μη κατευθυνόμενο δίκτυο X, μεγέθους n, με l ακμές, το οποίο μπορεί να περιγραφεί απο τον πίνακα γειτνίασης:

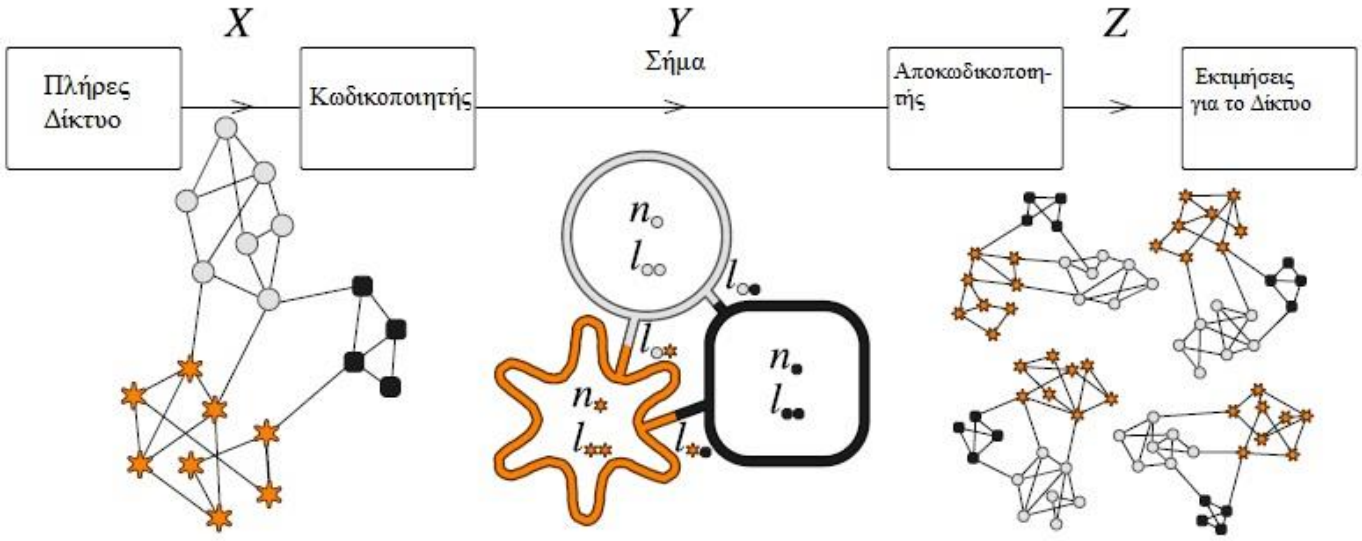
$$\boldsymbol{A} = (a_{ij})_{n\times n} = \begin{cases} 1, & \textit{εαν υπάρχει ακμή μεταξύ των κόμβων } i \textit{ και } j \\ 0, & \textit{διαφορετικά} \end{cases}.$$

Επιλέγουμε την περιγραφή:

$$\Upsilon = \left\{ \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}, \boldsymbol{M} = \begin{pmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \ddots & \vdots \\ l_{m1} & \cdots & l_{\boldsymbol{mm}} \end{pmatrix} \right\}$$

για m τμήματα, όπου $\boldsymbol{\alpha}$ το διάνυσμα τμηματικών εκχωρήσεων, $\alpha_i \in \{1,2,\dots,m\}$ και $\boldsymbol{M}$ ο πίνακας τμημάτων (module matrix). Ο πίνακας τμημάτων $\boldsymbol{M} = \boldsymbol{M}(X, \boldsymbol{\alpha})$ περιγράφει πως τα m τμήματα, που δίνονται απο το διάνυσμα τμηματικών εκχωρήσεων, συνδέονται στο πραγματικό δίκτυο. Το τμήμα i έχει n_i το πλήθος κόμβους και συνδέεται με το τμήμα j με l_{ij} ακμές. Σχηματική αναπαράσταση της παραπάνω διαδικασίας δίνεται στο παρακάτω σχήμα:

Σχήμα 17. Βασικό πλαίσιο εντοπισμού κοινοτήτων ως διαδικασία επικοινωνίας.



Ο διαβιβαστής, γνωρίζοντας την πλήρη κοινοτική δομή, θέλει να στείλει όσο το δυνατόν περισσότερη πληροφορία για το δίκτυο στον δέκτη του σήματος, κατα μήκος ενός καναλιού επικοινωνίας με περιορισμένη χωρητικότητα. Ο διαβιβαστής λοιπόν κωδικοποιεί το δίκτυο σε τμήματα με τέτοιο τρόπο ωστε να μεγιστοποιείται η ποσότητα πληροφορίας για το αρχικό δίκτυο. Το σχήμα απεικονίζει έναν κωδικοποιητή που συμπιέζει το δίκτυο σε τρία τμήματα, i =κύκλος, τετράγωνο, αστέρι, με n_i κόμβους και l_{ii} ακμές μεταξύ των τμημάτων. Ο δέκτης του σήματος μπορεί να αποκωδικοποιήσει το μήνυμα και να κατασκευάσει ένα σύνολο πιθανών υποψηφίων κοινοτήτων για το αρχικό δίκτυο. Όσο μικρότερο το σύνολο των υποψηφίων κοινοτήτων, τόσο περισσότερη η πληροφορία που κατάφερε να μεταφέρει ο κωδικοποιητής. (Rosvall & Bergstrom, An information-theoretic framework for resolving community structure in complex networks, 2007)

Για να βρούμε την καλύτερη εκχώρηση $\boldsymbol{\alpha}^*$, μεγιστοποιούμε την κοινή πληροφορία για όλες τις πιθανές εκχωρήσεις των κόμβων σε m τμήματα

$$\boldsymbol{\alpha}^* = \arg max_{\boldsymbol{\alpha}} I(X:Y)$$

Εξ'ορισμού, η κοινή πληροφορία $I(X:Y) = H(X) - H(X|Y) = H(X) - H(Z)$, όπου H(X) η πληροφορία που απαιτείται για να περιγράψουμε το X και η δεσμευμένη πληροφορία $H(X|Y) = H(Z)$ είναι η πληροφορία που απαιτείται για να περιγράψουμε το X δοθέντος του Y. Ως εκ τούτου, επιδιώκουμε την ελαχιστοποίηση της ποσότητας $H(Z)$. Αυτό ισοδυναμεί με την κατασκευή ενός διανύσματος εκχώρησης έτσι ώστε το σύνολο των εκτιμήσεων Z της εικόνας να είναι όσο το δυνατόν μικρότερο. Δοθέοντος οτι η περιγραφή Y εκχωρεί κόμβους σε m το πλήθος τμήματα

$$\mathrm{H(Z)} = \log\left[\prod_{i=1}^{m}\binom{n_i(n_i-1)/2}{l_{ii}}\prod_{i>j}\binom{n_i n_j}{l_{ij}}\right],$$

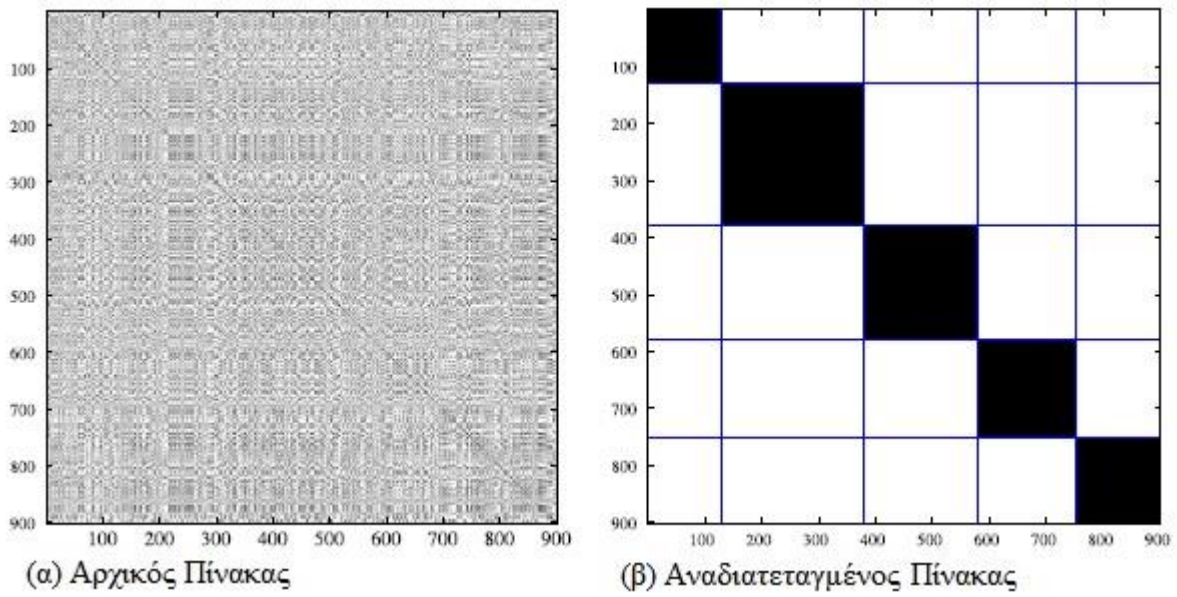
όπου οι παρενθέσεις υποδηλώνουν τους δυωνυμικούς συντελεστές και ο λογάριθμος υπολογίζεται με βάση το 2. Κάθε ένας απο τους m δυωνυμικούς συντελεστές στο πρώτο γινόμενο δίνει τον αριθμό των διαφορετικών τμημάτων που μπορούν να κατασκευαστούν με n_i κόμβους και l_{ii} ακμές. Κάθε ένας απο τους $\mathrm{m(m-1)/2}$ το πλήθος συντελεστές στο δεύτερο γινόμενο μας δίνει τους διαφορετικούς τρόπους κατα τους οποίους το τμήμα i και j μπορούν να συνδεθούν μεταξύ τους. Επειδή είναι υπολογιστικά ανέφικτό να ελεγχθούν όλες οι δυνατές διαμερίσεις ακόμα και σε μικρού μεγέθους δίκτυα, η διαμέριση που μεγιστοποιεί την κοινή πληροφορία μεταξύ της περιγραφής Y και του αρχικού δικτύου X, $I(X:Y)$, βρίσκεται μέσω ενός αλγορίθμου προσομειωμένης ανόπτησης. Η μέθοδος αυτή, έχει κριθεί ανώτερη, σύμφωνα με ελέγχους, απο τη βελτιστοποίηση της Modularity, ειδικά όταν οι κοινότητες είναι διαφορετικών μεγεθών. Ωστόσο είναι σχετικά αργή και μπορεί να εφαρμοστεί σε γράφους της τάξης των 10^4 κορυφών.

Οι ίδιοι συγγραφείς, (Rosvall & Bergstrom, Maps of random walks on complex networks reveal community structure, 2008), στηρίχθηκαν πάλι στην ιδέα της περιγραφής του δικτύου μέσω συμπίεσης της πλήρους πληροφορίας που περιέχεται στον πίνακα γειτνίασης και πρότειναν μια μέθοδο που ονομάζεται Infomap, για τον εντοπισμό κοινοτήτων σε δίκτυα. Στόχος της είναι η βέλτιστη συμπίεση της πληροφορίας που χρειάζεται ώστε να περιγραφεί η διαδικασία της διάχυσης πληροφορίας (information diffusion) κατα μήκος του γράφου. Στη μέθοδο αυτή χρησιμοποιείται η έννοια του τυχαίου περιπάτου- στην οποία θα ανεφερθούμε εκτενώς σε επόμενο κεφάλαιο- για να περιγράψει τη διαδικασία ροής της πληροφορίας στο δίκτυο. Οι κοινότητες μπορούν να εξαχθούν μέσω της συμπίεσης της περιγραφής που προκύπτει μέσω αυτού. Τη μέθοδο αυτή θα την δούμε και εκτενώς στην εφαρμογή μας.

Μια σχετικά διαφορετική διατύπωση των αρχών της θεωρίας πληροφορίας στο πρόβλημα εντοπισμού κοινοτήτων σε δίκτυα έχει παρουσιαστεί απο τον Chakrabarti (2004). Ο αλγόριθμος (ονομάζεται Autopart) που προτείνεται απο τον συγγραφέα, μπορεί να θεωρηθεί ένα εργαλείο ταυτόχρονης συσταδοποίησης (co-clustering) γραμμών και στηλών πινάκων με δυαδικά στοιχεία (όπως ο πίνακας γειτνίασης στην περίπτωσή μας), στο οποίο εφαρμόζονται έννοιες συμπίεσης ωστε να προσδιοριστεί η υποκείμενη κοινοτική δομή του δικτύου. Στόχος του αλγορίθμου είναι να ομαδοποιήσει τους κόμβους του δικτύου με τέτοιο τρόπο ωστε ο πίνακας γειτνίασης να

χωρίζεται σε ορθογώνια ομοιογενή μπλοκ υψηλής ή χαμηλής πυκνότητας. Τα μπλοκς αυτά υποδηλώνουν οτι οι συγκεκριμένες ομάδες κόμβων έχουν περισσότερες (ή λιγότερες) συνδέσεις με άλλες ομάδες. Αυτό επιτυγχάνεται μέσω μιας διαδικασίας αναδιάταξης του πίνακα γειτνίασης, όπου οι γραμμές και οι στήλες του αναδιατάσσονται ωστε να επιτευχθεί αυτή η δομή.

Σχήμα 18. Παράδειγμα αλγορίθμου εντοπισμού κοινοτήτων μέσω ταυτόχρονης συσταδοποίησης



Εδώ ο πίνακας (α) είναι ο αρχικός πίνακας γειτνίασης και ο πίνακας (β) είναι ο αρχικός πίνακας γειτνίασης με αναδιάταξη των γραμμών και των στηλών με τέτοιο τρόπο ωστε να σχηματιστούν ομογενή μπλοκ (Chakrabarti, 2004).

Η ποιότητα των διαφορετικών πιθανών δομών των ομαδοποιήσεων που προκύπτουν, αξιολογείται μέσω του συνολικού κόστους συμπίεσης Τ. Δεδομένου του κόστους αυτού, το καλύτερο σχήμα συμπίεσης θα πρέπει να επιτύχει μια αντιστάθμιση μεταξύ του αριθμού των παραγόμενων μπλοκς (δηλαδή συστάδων) και της ομοιογένειάς τους. Σε δύο ακραίες περιπτώσεις, κάποιος θα μπορούσε να επιλέξει είτε μόνο ένα μπλοκ (δηλαδή όλο τον πίνακα), αλλά όχι πολύ ομοιογενές, ή n^2 το πλήθος μπλοκς τέλειας ομοιογένειας, μεγέθους 1 (δηλαδή κάθε κελί του πίνακα). Αυτή η αντιστάθμιση επιτυγχάνεται εφαρμόζοντας την αρχή του Ελάχιστου Μήκους Περιγραφής (Minimum Description Length principle) για την επιλογή μοντέλου. Σύμφωνα με αυτή, η καλύτερη ομαδοποίηση (μοντέλο) είναι αυτή που ελαχιστοποιεί τόσο το κόστος συμπίεσης των δεδομένων, όσο και το κόστος για τη «σύνοψη» («summary») των ομάδων των κόμβων.

Το συνολικό κόστος συμπίεσης, T, εμπεριέχει πληροφορία για το συνολικό αριθμό κόμβων στο γράφο, τον αριθμό των μπλοκ, τον αριθμό των κόμβων και ακμών σε κάθε μπλόκ, αλλά και τους πίνακες γειτνίασης για κάθε μπλοκ. Η ελαχιστοποίηση του T, εφαρμόζεται η εξής επαναληπτική διαδικασία. Ξεκινάμε απο μια διαμέριση του γράφου, στην οποία ο γράφος αποτελεί μια συστάδα μόνος του. Σε κάθε βήμα, εφαρμόζεται μια διάσπαση (split) της συστάδας που προέκυψε απο τη διαμέριση, συστάδα η οποία έχει τη μέγιστη Shannon εντροπία ανα κόμβο (Shannon εντροπία ή εντροπία πληροφορίας ορίζεται ως η μέση ποσότητα πληροφορίας που παρήχθη απο μια στοχαστική πηγή δεδομένων- (Shannon C. E., 1948)). Η διάσπαση αυτή εφαρμόζεται με σκοπό να μετακινήσει απο την αρχική συστάδα, τους κόμβους εκείνους με τη μεγαλύτερη συνεισφορά στην εντροπία ανά κόμβο. Έπειτα, ξεκινώντας απο τη διαμέριση που προέκυψε, η οποία έχει μια συστάδα περισσότερη απο την προηγούμενη, το κόστος συμπίεσης T, βελτιστοποιείται ανάμεσα στις διαμερίσεις με τον ίδιο αριθμό συστάδων. Η διαδικασία επαναλαμβάνεται μέχρι να φτάσει σε έναν αριθμό συστάδων k^*, για τις οποίες το T δεν μπορεί να μειωθεί άλλο. Η πολυπλοκότητα της μεθόδου του Chakrabarti είναι $O[l(k^*)m]$, όπου l είναι ο αριθμός των επαναλήψεων που απαιτούνται ωστε να συγκλίνει η βελτιστοποίηση για έναν δεδομένο αριθμό συστάδων. Στα πειράματα που εφαρμόστηκαν απο τον συγγραφέα συνήθως $l \leq 20$ συνεπώς η μέθοδος μπορεί να εφαρμοστεί σε ικανοποιητικά μεγάλους γράφους.

4.5 Μέθοδοι στηριζόμενοι σε Τυχαίους Περιπάτους.

Στο κεφάλαιο αυτό θα ασχοληθούμε με την έννοια του τυχαίου περιπάτου σε γράφους και πως αυτή χρησιμοποείται για τον εντοπισμό κοινοτήτων σε αυτούς. Θα αναλύσουμε και θα περιγράψουμε δύο διάσημους αλγορίθμους που εντάσσονται σε αυτή την κατηγορία τεχνικών, τον Walktrap (Pons & Latapy, 2005) και τον Markov Cluster Algorithm (MCL), (Dongen S. v., 2000). Τη συμπεριφορά των διάσημων 2 αυτών αλγορίθμων, θα την δούμε και στην εφαρμογή μας.

4.5.1 Ο αλγόριθμος Walktrap.

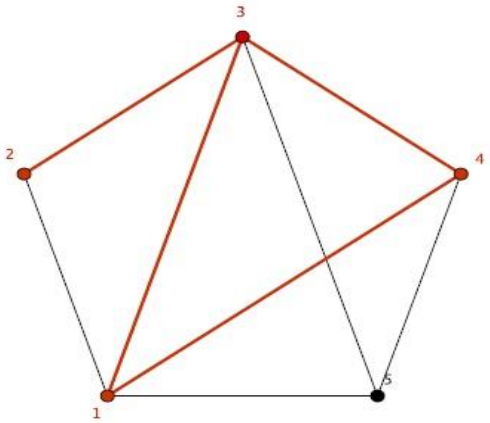
Θα ξεκινήσουμε την ανάλυσή μας απο τον αλγόριθμο των Pons & Latapy και κατά την περιγραφή του, θα αναφερθούμε σε βασικές αρχές και ιδιότητες των τυχαίων περιπάτων σε γράφους και πως αυτές χρησιμοποιήθηκαν απο τους συγγραφείς στην κατασκευή του αλγορίθμου τους.

Η ανάλυση που θα ακολουθήσει, αφορά ένα συνεκτικό μη προσανατολισμένο γράφο $\mathrm{G} = (\mathrm{V}, \mathrm{E})$, με $\mathrm{n} = |V|$ κορυφές και $m = |E|$ ακμές. Ο γράφος G, όπως έχουμε αναφέρει και σε προηγούμενο κεφάλαιο, συνδέεται με τον πίνακα γειτνίασής του A: $A_{ij} = 1$, αν οι κορυφές i και j συνδέονται με κάποια ακμή και $A_{ij} = 0$, διαφορετικά. Ο βαθμός $d(i) = \sum_j A_{ij}$ της κορυφής i είναι ο αριθμός των κορυφών με τις οποίες αυτή συνδέεται (συμπεριλαμβανομένης της ίδιας).

Η ιδέα στην οποία στηρίχθηκε ο αλγόριθμος Walktrap είναι η εξής: Οι τυχαίοι περίπατοι σε γράφους τείνουν να «παγιδεύονται» σε «πυκνά» συνδεδεμένα μέρη που αντιστοιχούν σε κοινότητες. 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 απο τις πιθανότητες μετάβασης ενός τυχαίου περιπατητή απο τη μια κορυφή στην άλλη, για έναν δεδομένο αριθμό βημάτων. Η απόσταση αυτή έχει το πλεονέκτημα οτι υπολογίζεται εύκολα και μπορεί να χρησιμοποιηθεί σε έναν συσσωρευτικό αλγόριθμο ιεραρχικής συσταδοποίησης (επαναληπτική συγχώνευση των κορυφών σε κοινότητες). Κατά τον τρόπο αυτό αποκτάται μια ιεραρχική κοινοτική δομή που αναπαρίσταται μέσω δενδροδιαγράμματος. Ανάλογα με το πλαίσιο στο οποίο χρησιμοποιείται ο αλγόριθμος, μπορούν να χρησιμοποιηθούν διαφορετικά κριτήρια για να επιλεγεί η καλύτερη ή οι καλύτερες διαμερίσεις. Ένα απο τα κριτήρια αυτά, το οποίο χρησιμοποίησαν και οι συγγραφείς στα πειράματά τους, είναι η modularity, λόγω αξιοπιστίας των αποτελεσμάτων της, αλλά και λόγω της διευκόλυνσης που τους παρείχε στη σύγκριση του αλγορίθμου τους με άλλους. Η πολυπλοκότητα του αλγορίθμου για τον υπολογισμό της κοινοτικής δομής είναι $O(mnH)$, όπου H το ύψος του αντίστοιχου δενδροδιαγράμματος. Ο χρόνος εκτέλεσης χειρότερης περίπτωσης είναι $O(mn^2)$. Ωστόσο, τα περισσότερα δίκτυα πραγματικού χρόνου είναι αραιά ($m = O(n)$) και το H είναι γενικά μικρό. Ιδανικά, για να είναι το δενδροδιάγραμμα ισορροπημένο, αρκεί $H = O(logn)$. Στην περίπτωση αυτή η πολυπλοκότητα του αλγορίθμου είναι $O(n^2 logn)$.

Έστω $\{X_n\}_n$, ένας τυχαίος περίπατος, δηλαδή μια στοχαστική διαδικασία διακριτού χρόνου, στις κορυφές του συνεκτικού, μη προσανατολισμένου γράφου G. Σε κάθε βήμα, ο περιπατητής επιλέγει τυχαία μια απο τις ακμές της κορυφής που βρίσκεται και μεταβαίνει στην άλλη κορυφή αυτής της ακμής.

Σχήμα 19. Σχηματική Αναπαράσταση Τυχαίου Περιπάτου σε Γράφο



$$3 \to 4 \to 1 \to 3 \to 2 \to \cdots$$

Η ακολουθία των κορυφών $v_o, v_1, v_2, \dots, v_k, \dots$ επιλεγμένη κατά αυτόν τον τρόπο είναι ένας απλός τυχαίος περίπατος στον γράφο G. Σε κάθε βήμα k, έχουμε μια τυχαία μεταβλητή X_k που παίρνει τιμές στο χώρο V. Η τυχαία ακολουθία $\{X_n\}_n = X_0, X_1, X_2, \dots, X_k, \dots$ λοιπόν, είναι μια στοχαστική διαδικασία διακριτού χρόνου που ορίζεται στον χώρο καταστάσεων V. Δεδομένου οτι ένας περιπατητής βρίσκεται σε μια κορυφή i με βαθμό $d(i)$, όλες οι $d(i)$ κορυφές που συνδέονται με την κορυφή i είναι ισοπίθανες ως επόμενος προορισμός. Η ακολουθία των κορυφών που επισκέφθηκε ο περιπατητής επομένως, είναι μια Μαρκοβιανή αλυσίδα, οι καταστάσεις της οποίας είναι οι κορυφές του γράφου. Σε κάθε βήμα, η πιθανότητα μετάβασης απο την κορυφή i στην κορυφή j, είναι $P_{ij} = \frac{A_{ij}}{d(i)}$. Η πιθανότητα αυτή, ορίζει τον πίνακα μετάβασης P των διαδικασιών τυχαίου περιπάτου. Μπορούμε επίσης να γράψουμε $P = D^{-1}A$, όπου D, ο διαγώνιος πίνακας των βαθμών ($D_{ii} = d(i)$ και $D_{ij} = 0$ για $i \neq j, \forall i, j$).

Η διαδικασία οδηγείται απο τις δυνάμεις του πίνακα P: Η πιθανότητα μετάβασης απο μια κορυφή i σε μια κορυφή j, μέσω ενός τυχαίου περιπάτου μήκους t (δηλαδή αριθμού βημάτων), είναι $(P^t)_{ij}$. Η πιθανότητα αυτή, την οποία θα συμβολίζουμε με P_{ij}^t πληρεί δύο βασικές ιδιότητες

της διαδικασίας τυχαίου περιπάτου, οι οποίες θα βοηθήσουν στην κατασκευή της απόστασης r για τη μέτρηση της ομοιότητας μεταξύ των κορυφών. Οι ιδιότητες αυτές θα δοθούν χωρίς αποδείξεις και είναι οι ακόλουθες:

I. *Όταν το μήκος t ενός τυχαίου περιπάτου που ξεκινά απο μια κορυφή i, τείνει στο άπειρο, η πιθανότητα να βρεθούμε σε σε μια κορυφή j, εξαρτάται μόνο απο το βαθμό της κορυφής j (και όχι απο την αρχική κορυφή i):*

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

II. *Ο λόγος των πιθανοτήτων να μεταβούμε απο την κορυφή i στην j και απο την j στην i μέσω ενός τυχαίου περιπάτου δεδομένου μήκους t, εξαρτάται μόνον απο τους βαθμούς $d(i)$ και $d(j)$:*

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

Για να ομαδοποιηθούν οι κορυφές σε κοινότητες, εισάγεται η απόσταση r μεταξύ των κορυφών, η οποία αποτυπώνει την κοινοτική δομή του γράφου. Η απόσταση αυτή θα πρέπει να είναι μεγάλη, εαν δυο κορυφές ανήκουν σε διαφορετικές κοινότητες και μικρή αν ανήκουν στην ίδια και υπολογίζεται απο την πληροφορία που παίρνουμε απο τους τυχαίους περιπάτους στο γράφο. Η πληροφορία για την κορυφή i που κωδικοποιείται μέσω του πίνακα μεταβάσεων P^t, βρίσκεται στις n πιθανότητες $(P_{ij}^{t})_{1\le k\le n}$, οι οποίες ουσιαστικά είναι η $i-$οστή γραμμή του πίνακα μεταβάσεων P^t και συμβολίζεται με $P_{i\bullet}^{t}$. Δεδομένων των προαναφερθέντων ιδιοτήτων, δίνουμε τους ακόλουθους δύο ορισμούς 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:*

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

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

Θεωρώντας τώρα, τυχαίους περιπάτους που ξεκινούν απο μια κοινότητα, επιλέγοντας τον αρχικό κόμβο τυχαία μεταξύ των κορυφών της, ορίζουμε την πιθανότητα μετάβασης απο την 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 μας επιτρέπει να γενικεύσουμε την απόσταση μεταξύ κοινοτήτων μέσω του ακόλουθου ορισμού.

Ορισμός 2: *Έστω* $\mathrm{C}_1, \mathrm{C}_2 \subset V$*, δύο κοινότητες. Τότε η απόσταση τους,* $r_{\mathrm{C}_1\mathrm{C}_2}$*, ορίζεται ως:*

$$r_{C_1C_2} = \left\| D^{-\frac{1}{2}}P_{C_1\bullet}^{t} - D^{-\frac{1}{2}}P_{C_2\bullet}^{t} \right\| = \sqrt{\sum_{\mathrm{k}=1}^{n} \frac{\left(P_{C_1\mathrm{k}}^{t} - P_{C_2\mathrm{k}}^{t}\right)^2}{d(k)}}$$

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 έναν συσσωρευτικό αλγόριθμο ιεραρχικής συσταδοποίησης που βασίζεται στη μέθοδο του Ward (1963), ο οποίος δίνει πολύ καλά αποτελέσματα, ενώ μειώνει τον υπολογισμό των αποστάσεων.

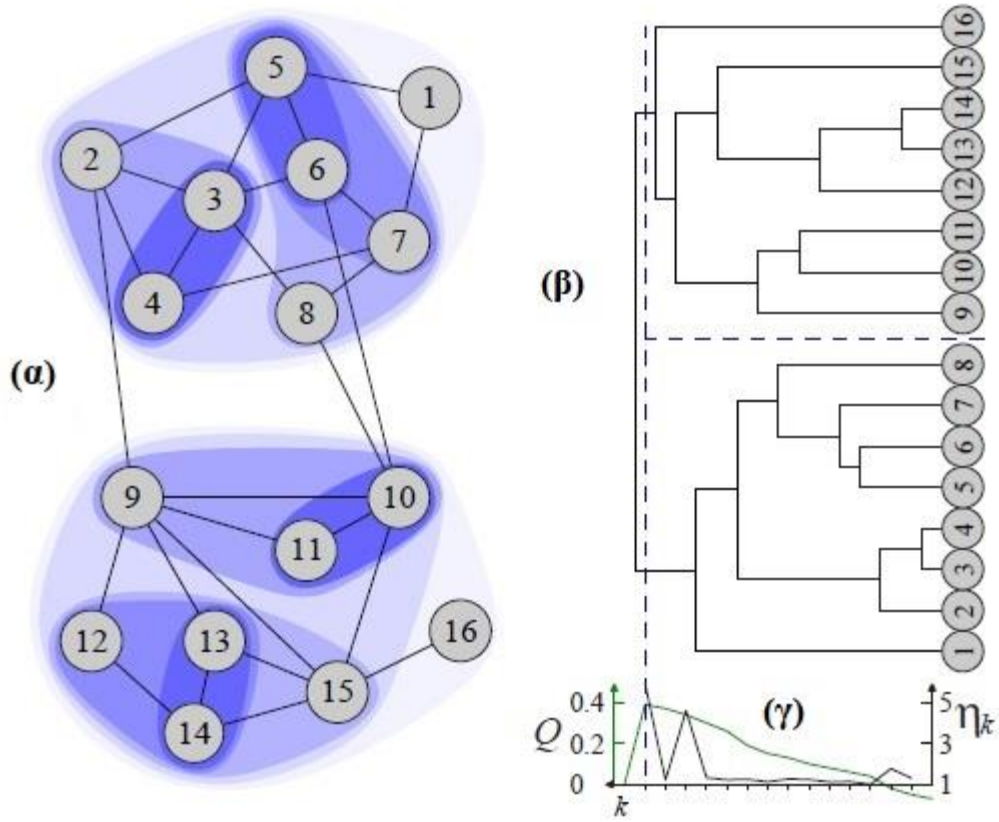
Ξεκινάμε απο μια διαμέριση $\mathrm{P}_1 = \{\{v\} \in V\}$ του γράφου σε n κοινότητες, αποτελούμενες απο μια μόνον κορυφή (δηλαδή κάθε κορυφή αποτελεί μια κοινότητα). Αρχικά υπολογίζονται οι αποστάσεις μεταξύ όλων των γειτονικών κορυφών. Έπειτα η διαμέριση εξελίσσεται επαναλαμβάνοντας τις ακόλουθες 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 μια νέα $C_3 = C_1 \cup C_2$ και δημιουργεί μια νέα διαμέριση $\mathrm{P}_{k+1} = (\mathrm{P}_k \backslash \{C_1, C_2\} \cup \{C_3\})$
- Ανανεώνει τις αποστάσεις μεταξύ κοινοτήτων (κάτι το οποίο γίνεται μόνο για γειτονικές κοινότητες).

Ύστερα απο $n-1$ βήματα ο αλγόριθμος σταματά και εν τέλει $\mathrm{P}_n = \{V\}$. Κάθε βήμα του, ορίζει και μια διαμέριση P_k του γράφου σε κοινότητες, η οποία παράγει μια ιεραρχική δομή που αναπαρίσταται μέσω δενδρογράμματος. Τα σημαντικά σημεία του αλγορίθμου είναι ο τρόπος με

τον οποίο επιλέγονται οι κοινότητες που θα συγχωνευθούν, η ανανέωση των αποστάσεων και το πως εκτιμάται η ποιότητα της εκάστοστε διαμέρισης $\mathcal{P}_k$.

Σχήμα 20. Σχηματική Αναπαράσταση του Walktrap



(α) Ένα παράδειγμα κοινοτικής δομής, όπως αυτή εντοπίστηκε απο τον αλγόριθμο χρησιμοποιώντας τυχαίους περιπάτους μήκους $t = 3$. (β) Οι φάσεις του αλγορίθμου κωδικοποιούνται στο εν λόγω δενδρόγραμμα. (γ) Σύμφωνα με τα κριτήρια, της modularity, Q, και του λόγου αύξησης (increase ratio) η_k, η βέλτιστη διαμέριση αντιστοιχεί σε 2 κοινότητες. Ο λόγος αύξησης, η_k, είναι και αυτός ένα κριτήριο αξιολόγισης της ποιότητας των διαμερίσεων και προτιμάται απο τη modularity για εντοπισμό κοινοτήτων διαφορετικής κλίμακας. (Pons & Latapy, 2005)

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:

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

Ωστόσο το πρόβλημα αυτό είναι NP-πλήρες πρόβλημα 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 μια νέα κοινότητα $C_3 = C_1 \cup C_2$. Η ποσότητα αυτή εξαρτάται μόνον απο τις κορυφές των C_1 και C_2 και όχι απο 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 $\Delta\sigma$.

Ο υπολογισμός των $\Delta\sigma$ και η ανανέωση των αποστάσεων, δύναται να υπολογιστεί αποδοτικά, λόγω λημμάτων που προκύπτουν απο το γεγονός ότι η απόσταση που όρισαν οι συγγραφείς είναι Ευκλείδεια απόσταση.

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

Όταν οι κοινότητες είναι διαφορετικής κλίμακας μεγέθους ωστόσο, η modularity 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.

4.5.2 Ο αλγόριθμος MCL.

Ο δεύτερος αλγόριθμος που θα περιγράψουμε, ονομάζεται Markov Cluster Algorithm (MCL) και εφευρέθηκε απο τον Van Dongen (2000). Οι φυσικές συστάδες (natural clusters) ενός γράφου, χαρακτηρίζονται απο την παρουσία πολλών ακμών μεταξύ των κόμβων μιας συστάδας. Επίσης ο αριθμός των μεγαλύτερου μήκους μονοπατιών ανάμεσα σε δύο αυθαίρετους κόμβους στη

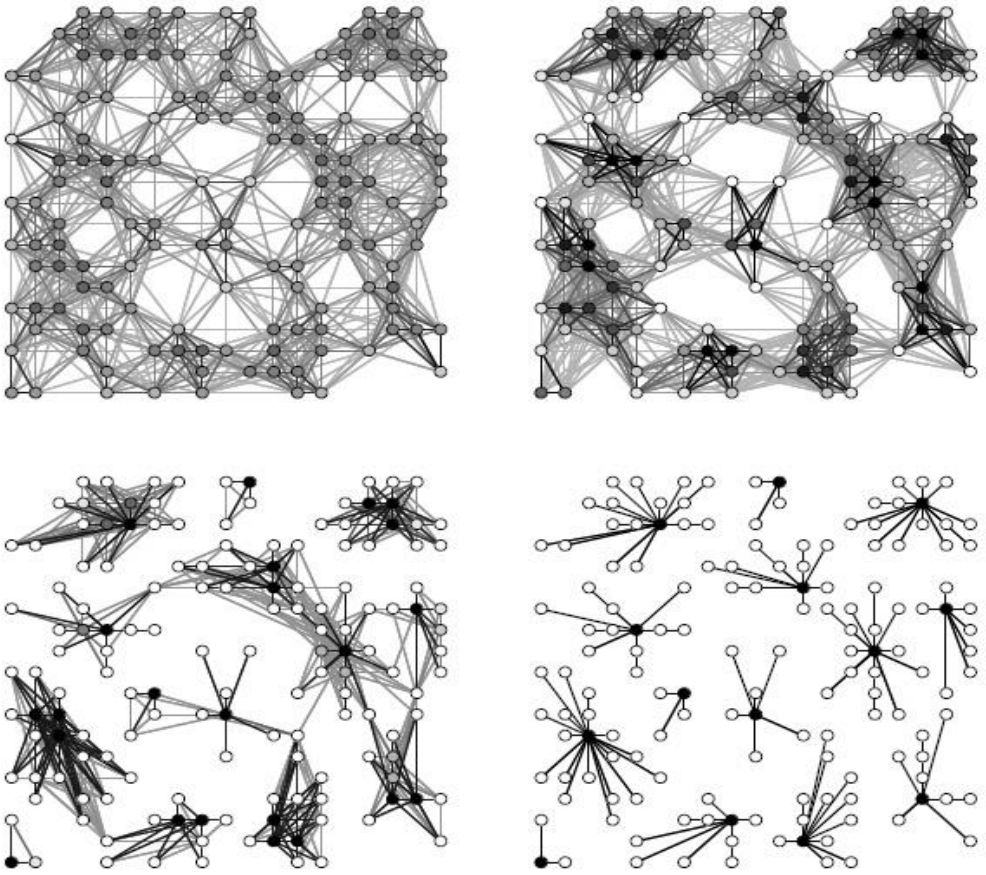
συστάδα αυτή, αναμένεται υψηλός. Πιο συγκεκριμένα ο αριθμός αυτός θα πρέπει να είναι μεγάλος όταν συγκρίνεται με ζεύγη κόμβων που ανήκουν σε διαφορετικές φυσικές συστάδες. Η ιδέα αυτή πηγάζει απο τη βασική ιδιότητα των τυχαίων περιπάτων σε γράφους (σύμφωνα με την οποία κατασκευάστηκαν και πολλοί άλλοι αλγόριθμοι συσταδοποίησης), οτι δηλαδή ένας τυχαίος περίπατος σε έναν γράφο G, όταν επισκέπτεται μια πυκνή συστάδα είναι πολύ πιθανόν να μην την αφήσει μέχρι να επισκεφθεί πολλές κορυφές της.

Ο αλγόριθμος MCL εντοπίζει την κοινοτική δομή ενός γράφου, μέσω της παρακάτω διαδικασίας: Υπολογίζει τις πιθανότητες τυχαίων περιπάτων στο γράφο και χρησιμοποιεί δύο τελεστές οι οποίοι μετατρέπουν ένα σύνολο πιθανοτήτων σε ένα άλλο. Αυτό το καταφέρνει με τη χρήση στοχαστικών πινάκων (Μαρκοβιανών πινάκων), στους οποίους εμπεριέχεται η μαθηματική έννοια των τυχαίων περιπάτων σε γράφους.

Ο αλγόριθμος MCL προσομειώνει τυχαίους περιπάτους σε γράφους με την εναλλαγή δύο τελεστών που ονομάζονται επέκταση (expansion) και εμφύσηση (inflation). Η επέκταση αναπαρίσταται υψώνοντας ένα στοχαστικό πίνακα σε δυνάμεις, χρησιμοποιώντας το σύνηθες γινομένο πινάκων και είναι υπεύθυνη στο να επιτρέπει στη ροή (δηλαδή την πιθανότητα μετάβασης απο έναν κόμβο σε έναν άλλο) να συνδέει διαφορετικές περιοχές του γράφου. Η εμφύσηση, αναπαραρίσταται υψώνοντας έναν στοχαστικό πίνακα σε δυνάμεις Hadamard (Hadamard γινόμενο ως προς τα στοιχεία των πινάκων: δηλαδή για δύο πίνακες A, B του ίδιου τύπου, το γινόμενο Hadamard είναι ο πίνακας $A \circ B$, με στοιχείο $(A \circ B)_{ij} = A_{ij}B_{ij}$) ακολουθούμενη απο ένα βήμα κλιμάκωσης (scaling step), έτσι ώστε ο πίνακας που προκύπτει να είναι επίσης στοχαστικός, δηλαδή τα στοιχεία των στηλών του να αντιστοιχούν σε τιμές πιθανοτήτων. Η εμφύσηση είναι υπεύθυνη για την ενδυνάμωση και την εξασθένισή της ροής στο γράφο.

Η προκύπτουσα αλγεβρική διαδικασία ονομάζεται Μαρκοβιανή Διαδικασία Συσταδοποίησης (Markov Cluster Process) ή διαδικασία MCL και μια σχηματική της αναπαράσταση δίνεται παρακάτω.

Σχήμα 21. Διαδοχικά Στάδια της προσομείωσης της ροής στο γράφο μέσω της διαδικασίας MCL.



Στο σχήμα αναπαρίστανται 4 επαναληπτικά στάδια της διαδικασίας MCL, κοιτάζοντας τα απο αριστερά προς τα δεξιά και απο πάνω προς τα κάτω. Για κάθε κόμβο, υπάρχουν το πολύ δεκαέξι γειτονικοί. Ο κάτω δεξιά γράφος αντιστοιχεί στο όριο της MCL διαδικασίας. Η βαθμίδα σκίασης μιας ακμής μεταξύ δύο κόμβων, υποδηλώνει τη μέγιστη ποσότητα ροής που περνάει απο τις δύο κατευθύνσεις: Όσο πιο σκούρα η ακμή τόσο μεγαλύτερη η ποσότητα της ροής. Η βαθμίδα σκίασης των κόμβων, υποδηλώνει τη συνολική ποσότητα εισερχόμενης ροής. Επομένως, μια σκούρα ακμή ανάμεσα σε έναν λευκό και έναν μαύρο κόμβο, υποδηλώνει οτι η μέγιστη τιμή ροής βρίσκεται στην κατεύθυνση του σκούρου κόμβου και μετά βίας κάποια ποσότητα ροής πηγαίνει προς την αντίθετη κατεύθυνση. Ο κάτω δεξία γράφος αντιπροσωπεύει μια κατάσταση στην οποία η ροή είναι σταθερή: Οι σκούροι κόμβοι είναι οι έλκοντες και οι ακμές δείχνουν ποιοι κόμβοι έλκονται απο αυτούς. Το όριο της MCL διαδικασίας παρουσιάζει όπως παρατηρούμε, ισχυρή κοινοτική δομή. Ο κάτω δεξιά γράφος παράγει μια συσταδοποίηση του αρχικού γράφου, στην οποία συστάδες θεωρούνται όλες οι ανίσχυρα συνδεδεμένες συνιστώσες. (weakly connected components) (Dongen S. v., 2000).

Ένας κατά στήλες στοχαστικός πίνακας (column stochastic) είναι ένας μη αρνητικός πίνακας με την ιδιότητα κάθε μια απο τις στήλες του να αθροίζει στη μονάδα. Δοθέντος ενός τέτοιου πίνακα M και ενός πραγματικού αριθμού, $r > 1$, ο κατα στήλες στοχαστικός πίνακας που προκύπτει απο την εμφύσηση των στηλών του M με το συντελεστή δύναμης r, γράφεται ως $\Gamma_r(M)$ και ο Γ_r ονομάζεται τελεστής εμφύσησης με συντελεστή δύναμης (power coefficient) r. Συμβολίζουμε με $\sum_{\mathrm{r,j}}(M)$ το άθροισμα των στοιχείων της j στήλης του πίνακα M που έχουν υψωθεί στη δύναμη r (το άθροισμα λαμβάνεται αφού πρώτα τα στοιχεία της στήλης έχουν υψωθεί στην εν λόγω δύναμη). Τότε ο $\Gamma_r(M)$ ορίζεται ως:

$$\Gamma_r(M_{ij}) = {M_{ij}}^r / \sum_{\mathrm{r,j}}(M)$$

Κάθε στήλη j ενός στοχαστικού πίνακα M, αντιστοιχεί στον κόμβο j του στοχαστικού γράφου που σχετίζεται με τον πίνακα M. Το στοιχείο της i γραμμής που αντιστοιχεί στη στήλη j (δηλαδή το M_{ij} στοιχείο του πίνακα) αντιστοιχεί στην πιθανότητα μετάβασης απο τον κόμβο j στον κόμβο i. Έχει παρατηρηθεί οτι για τιμές του συντελεστή δύναμης $r > 1$, η εμφύσηση τροποποιεί τις πιθανότητες που σχετίζονται με το σύνολο/συλλογή (collection) των τυχαίων περιπάτων που ξεκινάνε απο έναν συγκεκριμένο κόμβο (αυτό το σύνολο αντιστοιχεί σε μια στήλη του πίνακα), ευνοώντας πιο πιθανούς περιπάτους απο λιγότερο πιθανούς.

Η επέκταση αντιστοιχεί στον υπολογισμό τυχαίων περιπάτων υψηλότερου μήκους, δηλαδή τυχαίων περιπάτων πολλών βημάτων. Συνδέει νέες πιθανότητες με όλα τα ζεύγη κόμβων, όπου ο ένας κόμβος είναι το σημείο αναχώρησης και ο άλλος ο προορισμός. Δεδομένου οτι μονοπάτια υψηλότερου μήκους είναι πιο συνηθισμένα εντός των συστάδων απο οτι μεταξύ διαφορετικών συστάδων, οι πιθανότητες που σχετίζονται με ζεύγη κόμβων που βρίσκονται στην ίδια συστάδα, θα είναι σχετικά μεγάλες, καθώς υπάρχουν πολλοί τρόποι μετάβασης απο τον έναν κόμβο στον άλλο. Η εμφύσηση τότε, θα παίζει τον εξής ρόλο: Θα ενδυναμώνει τις πιθανότητες των τυχαίων περιπάτων εντός των συστάδων (intra-cluster walks) και θα αποδυναμώνει τους περιπάτους μεταξύ των συστάδων (inter-cluster walks). Αυτό επιτυγχάνεται χωρίς καμία εκ των προτέρων γνώση της δομής των συστάδων. Είναι απλά το αποτέλεσμα της δομής που είναι παρούσα.

Τελικά, η επανάληψη της επέκτασης και της εμφύσησης, έχει ως αποτέλεσμα το διαχωρισμό του γράφου σε διαφορετικά τμήματα, μεταξύ των οποίων δεν υπάρχουν πλεον μονοπάτια. Η συλλογή των τμημάτων που προκύπτουν θα είναι και η συσταδοποίηση του γράφου μέσω της

διαδικασίας. Ο τελεστής εμφύσησης μπορεί να τροποποιηθεί, χρησιμοποιώντας την παράμετρο r. Αυξάνοντας την παράμετρο αυτή, ενδυναμώνουμε τον τελεστή εμφύσησης κάτι που αυξάνει το πόσο δεμένες είναι οι συστάδες (tightness of clusters) ή το βαθμό «πιστότητάς» τους (granularity).

Με αυτά, ο MCL αλγόριθμος μπορεί να γραφεί ώς:

```
G is a graph
  add loops to G                          # εξήγηση παρακάτω
  set Γ to some value                     # επηρρεάζει το βαθμό «πιστότητας»
  set M_1 to be the matrix of random walks on G

  while (change) {
    M_2 = M_1 * M_1                       # επέκταση
    M_1 = Γ(M_2)                          # εμφύσηση
    change = difference (M_1, M_2)
  }

  set CLUSTERING as the components of M_1     #εξήγηση παρακάτω
```

Η μεταβλητή change, δύναται να υπολογιστεί απευθείας απο τον πίνακα **M_2** χρησιμοποιώντας τα χαρακτηριστικά των ορίων της MCL διαδικασίας, ωστόσο δε θα αναφερθούμε σε αυτά στην ανάλυσή μας.

Ανεπίσημα, μπορούμε να ισχυριστούμε οτι η επέκταση κάνει τη στοχαστική ροή να διαχέεται εντός των συστάδων, ενώ η εμφύσηση εξαλείφει τη στοχαστική ροή μεταξύ των συστάδων. Η επέκταση και η εμφύσηση αναπαριστούν διαφορετικές παλιρροιακές (tidal) δυνάμεις, οι οποίες τροποποιούνται εως ότου επιτευχθεί μια κατάσταση ισορροπίας. Η κατάσταση ισορροπίας αυτή, ταυτίζεται με έναν «διπλά ταυτοδύναμο» (doubly idempotent) πίνακα, δηλαδή έναν πίνακα που δε δύναται να τροποποιηθεί περαιτέρω απο την προσθήκη βημάτων επέκτασης ή εμφύσησης. Ο γράφος που σχετίζεται με έναν τέτοιο πίνακα αποτελείται απο διαφορετικές συνεκτικές κατευθυνόμενες συνιστώσες. Κάθε μια απο τις συνιστώσες αυτές, ερμηνεύεται ως μια συστάδα, και έχει τη μορφή αστεριού, με έναν έλκοντα κόμβο στο κέντρο και τόξα που κατευθύνονται απο όλους τους κόμβους της εν λόγω συνιστώσας στον έλκοντα κόμβο. Θεωρητικά, μπορεί να υπάρξουν συνιστώσες με περισσότερους απο έναν έλκοντες κόμβους, γεγονός που δεν αλλάζει ωστόσο την ερμηνεία που ήδη δόθηκε. Επίσης, είναι δυνατόν να υπάρξουν κόμβοι που συνδέονται με διαφορετικά αστέρια, κάτι που σημαίνει οτι οι συστάδες που παράγει η διαδικασία είναι αλληλοκαλυπτόμενες.

Αναφορικά με τη σύγκλιση, έχει αποδειχθεί οτι η διαδικασία που προσομειώνει ο αλγόριθμος συγκλίνει τετραγωνικά γύρω απο τις καταστάσεις ισορροπίας της. Στην πράξη, ο αλγόριθμος αρχίζει να συγκλίνει αισθητά απο 3 εως 10 επαναλήψεις. Η ολική σύγκλισή του είναι πολύ δύσκολο να αποδειχθεί. Εικάζεται οτι η διαδικασία συγκλίνει πάντα, εαν ο αρχικός γράφος είναι συμμετρικός.

Ο MCL αλγόριθμος επίσης συνδέει πιθανότητες επιστροφής (ή βρόχους/loops) με κάθε κόμβο του αρχικού γράφου εισόδου. Το πρότυπο ροής(flow paradigm) στο οποίο στηρίζεται ο αλγόριθμος, το απαιτεί λόγω των φασματικών και δομικών ιδιοτήτων των πινάκων που προκύπτουν κατά τις επαναλήψεις του. Τα βάρη των βρόχων που επιλέγονται είναι καλό να είναι ουδέτερων τιμών. Είναι δυνατόν να επιλεχθούν μεγάλα βάρη και κατά αυτόν τον τρόπο να ενισχυθεί ο βαθμός «πιστότητας» των παραγόμενων συστάδων. Ο αλγόριθμος ωστόσο γενικά, δεν είναι πολύ ευαίσθητος σε αλλαγές των βαρών των βρόχων.

Ένα πολύ σημαντικό προτέρημα του αλγορίθμου είναι η «bootstrapping» φύση του: Η δομή των συστάδων ανακτάται απο το αποτύπωμά της, το οποίο δημιουργείται κατά τη διαδικασία της ροής. Περαιτέρω βασικά οφέλη του αλγορίθμου είναι αρχικά οτι δεν «παραπλανάται» απο ακμές που συνδέουν διαφορετικές συστάδες. Είναι γρήγορος και εύκολα επεκτάσιμος (easily scalable). Επιπλέον τα μαθηματικά που σχετίζονται με αυτόν δείχνουν οτι υπάρχει εγγενής σχέση ανάμεσα στη διαδικασία που προσομειώνει και την κοινοτική δομή του αρχικού γράφου. Τέλος, η εφαρμογή του είναι απλή και κομψή. Ο MCL αν και στηρίζεται σε ομοιότητες μεταξύ ζευγών κορυφών, παρουσιάζει την εξής ιδιαιτερότητα: Ανασυνδυάζει (recombines) αυτές τις ομοιότητες (μέσω της επέκτασης) και επομένως επηρρεάζεται απο ομοιότητες σε επίπεδο συνόλων (ως γενίκευση ζευγών). Η εναλλαγή στην επέκταση μέσω της εμφύσησης, αποδεικνύεται οτι είναι ένας κατάλληλος τρόπος εκμετάλλευσης αυτής της ιδιότητας ανασυνδυασμού.

Κεφάλαιο 5. Μέθοδοι Εντοπισμού Αλληλοκαλυπτόμενων Κοινοτήτων σε Δίκτυα.

Οι μέθοδοι που συζητήθηκαν στην προηγούμενη ενότητα, στόχευαν στην ανίχνευση διαμερίσεων του δικτύου, δηλαδή κοινοτήτων στις οποίες κάθε κόμβος ανήκει σε μια μόνον κοινότητα. Οι κοινότητες που εντόπιζαν λοιπόν ήταν «κομματιασμένες»(disjoint). Ωστόσο, όπως έχουμε ήδη αναφέρει, σε πραγματικά δίκτυα οι κορυφές συχνά μοιράζονται ανάμεσα σε κοινότητες και το ζήτημα εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων έχει αναπτυχθεί ιδιαίτερα στη βιβλιογραφία. Ο εντοπισμός αλληλοκαλυπτόμενων κοινοτήτων προκύπτει απο το γεγονός οτι τα πολύπλοκα δίκτυα πραγματικού χρόνου, συνήθως δε χωρίζονται σε «αιχμηρά» υποδίκτυα, αλλά οι κόμβοι σε αυτά εκ φύσεως μπορεί να ανήκουν σε περισσότερες απο μια κοινότητες. Για παράδειγμα, σε ένα κοινωνικό δίκτυο, ένα άτομο μπορεί να ανήκει σε διαφορετικές κοινότητες ταυτόχρονα (αυτή της οικογένειας του, αυτή των φίλων του, αυτή του επαγγέλματός του). Επομένως, το να είμαστε σε θέση να εντοπίζουμε αλληλοκαλυπτόμενες κοινότητες σε δίκτυα μπορεί να προσφέρει προσοδοφόρες γνώσεις για τη δομή του δικτύου.

Στην ενότητα αυτή θα αναφερθούμε σε βασικές τεχνικές εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων σε δίκτυα.

5.1 Μέθοδος Διήθησης της Κλίκας

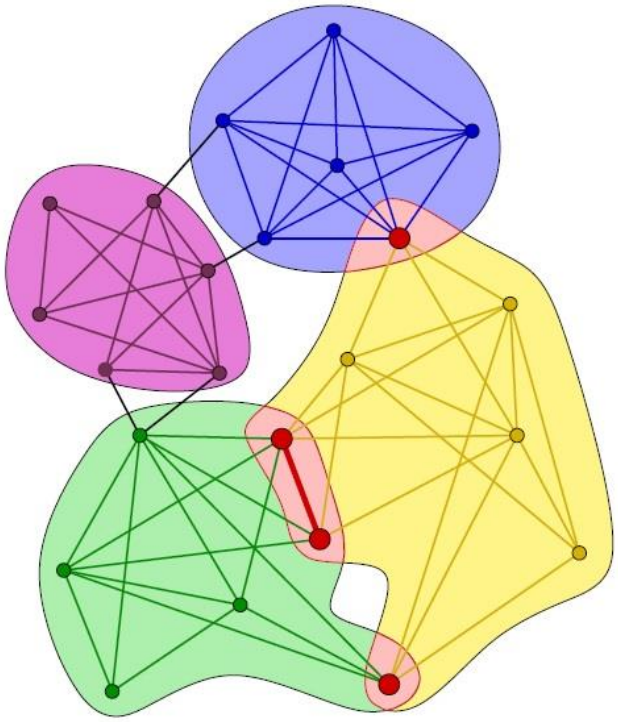
Η πιο διάσημη τεχνική εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων σε δίκτυο, είναι η μέθοδος διήθησης της κλίκας (Clique Percolation Method-CPM) απο τους Palla et al. (2005).

Οι συγγραφείς υιοθετούν έναν ορισμό κοινότητας ο οποίος στηρίζεται στην παρατήρηση οτι ένα μέλος της ενώ συνδέεται με άλλα μέλη της, δε συνδέεται απαραίτητα με όλους της τους κόμβους. Δηλαδή με άλλα λόγια, μια κοινότητα μπορεί να ερμηνευθεί ως μια ένωση μικρότερων πλήρων (πλήρως συνεκτικών) υπογράφων, οι οποίοι μοιράζονται κάποιους κόμβους μεταξύ τους. Στη μαθηματική βιβλιογραφία, τέτοιοι πλήρεις υπογράφοι ονομάζονται k-κλίκες (k-cliques), όπου το k αναφέρεται στον αριθμό των κόμβων στον υπογράφο. Ως εκ τούτου, ορίζεται η κοινότητα k-κλικών (k-clique community) ως η ένωση όλων των k-κλικών που φτάνουν η μια στην άλλη μέσω μιας σειράς γειτονικών k-κλικών (adjacent k-cliques). Δύο k-κλίκες ονομάζονται

γειτονικές εαν μοιράζονται $k-1$ κόμβους. Χρησιμοποιώντας τη γειτνίαση των k-κλικών, ορίζεται η αλυσίδα k-κλικών (k-clique chain) ως η ένωση μιας ακολουθίας γειτονικών k-κλικών. Μέσω της αλυσίδας αυτής ορίζεται και η συνεκτικότητα k-κλίκας (k-clique connectedness): Δύο k-κλίκες είναι συνεκτικές εαν είναι μέλη μιας αλυσίδας k-κλικών. Δηλαδή, εν τέλει μια κοινότητα k-κλικών είναι ο μεγαλύτερος συνεκτικός υπογράφος που προκύπτει απο την ένωση μιας k-κλίκας και όλων των συνεκτικών k-κλικών με αυτήν.

Με βάση τα προαναφερθέντα, μια 2-κλίκα είναι μια ακμή, και μια κοινότητα 2-κλικών είναι η ένωση αυτών των ακμών, οι οποίες φτάνουν η μια στην άλλη μέσω μιας ακολουθίας κοινών κόμβων. Ομοίως, μια κοινότητα 3-κλικών είναι η ένωση τριγώνων, τα οποία φτάνουν το ένα στο άλλο μέσω μιας ακολουθίας κοινών ακμών. Όσο αυξάνεται το k, οι κοινότητες k-κλικών συρρικνώνονται (χάνονται), αλλά απο την άλλη πλευρά γίνονται πιο συνεκτικές (cohesive) απο τη στιγμή που οι κόμβοι τους πρέπει να ανήκουν σε τουλάχιστον μια k-κλίκα.

Σχήμα 22. Αλληλοκαλυπτόμενες κοινότητες 4-κλικών.



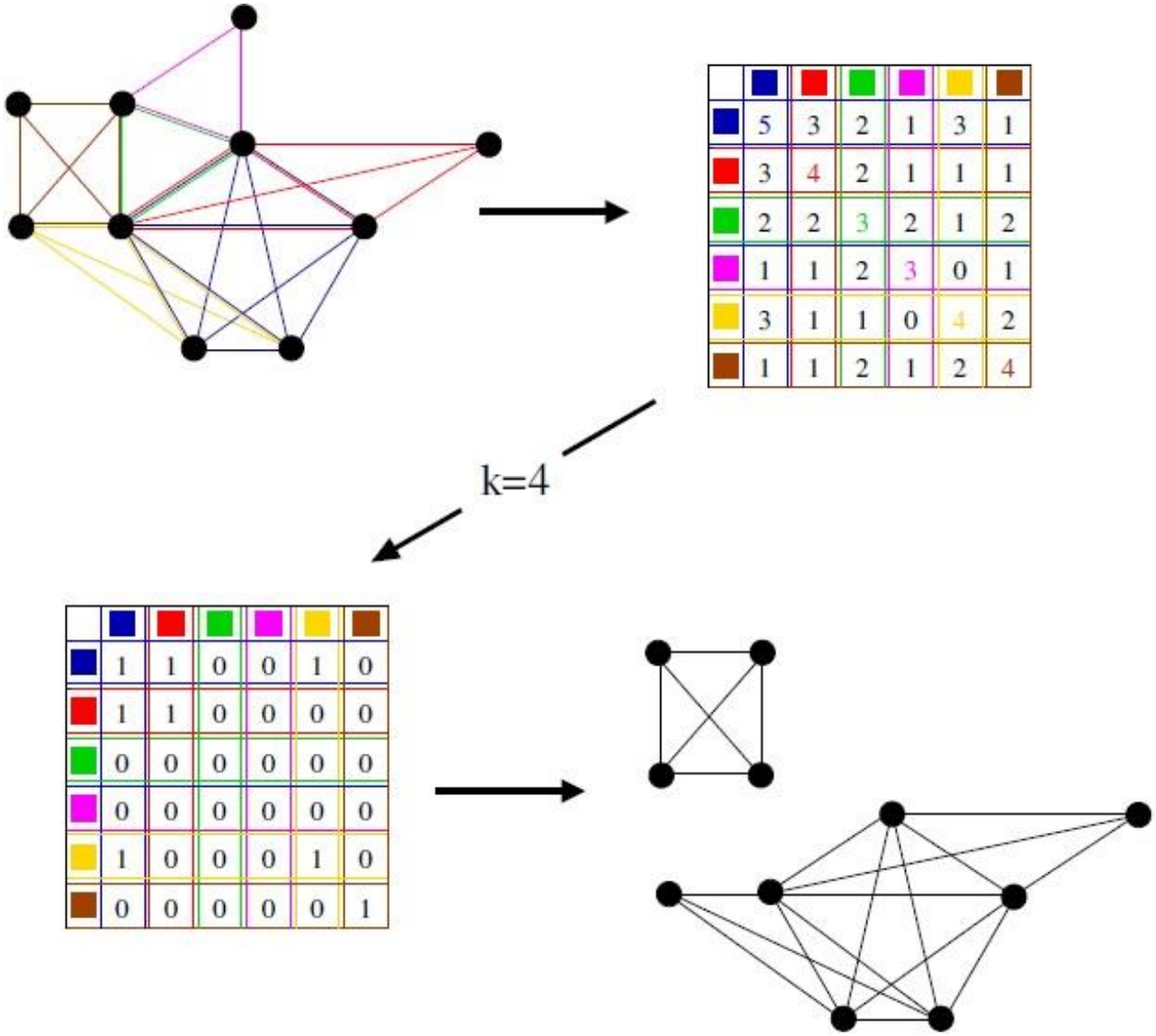
Στο εν λόγω σχήμα έχουμε παρατηρούμε τέσσερις κοινότητες k-κλικών, για $k = 4$, δηλαδή τέσσερις μεγαλύτερους δυνατούς συνεκτικούς υπογράφους, που προκύπτουν απο την ένωση συνεκτικών 4-κλικών, μεταξύ των οποίων υπάρχει αλληλοκάλυψη. Οποιαδήποτε 4-κλικά (δηλαδή πλήρης υπογράφος μεγέθους $k = 4$), φτάνεται μόνο απο τις 4-κλίκες της ίδιας κοινότητας μέσω μιας ακολουθίας γειτονικών 4-κλικών, εφόσον δύο 4-κλίκες είναι γειτονικές αν μοιράζονται $k - 1 = 4 - 1 = 3$ κόμβους στην περίπτωση μας. Παρατηρούμε οτι η κοινότητα χρωματισμένη με κίτρινο αλληλεπικαλύπτεται με την κοινότητα χρωματισμένη με μπλε σε έναν μόνο κόμβο, ενώ με την κοινότητα χρωματισμένη με πράσινο, σε έναν κόμβο και μια ακμή. Οι περιοχές αλληλεπικάλυψης είναι χρωματισμένες με κόκκινο χρώμα. (Palla, Derényi, Farkas, & Vicsek, 2005)

Σε δίκτυα πραγματικού χρόνου συχνά συναντάμε πλήρεις υπογράφους μεγέθους μεταξύ 10 και 100 κόμβων. Ένας τέτοιος μεγάλος πλήρης υπογράφος μεγέθους s, περιέχει $\binom{s}{k}$ διαφορετικές k-κλίκες, επομένως ένας αλγόριθμος που προσπαθεί να εντοπίσει τις k-κλίκες αυτές και να εξετάσει τη γειτνίαση τους θα ήταν αργός. Ωστόσο ένας πλήρης υπογράφος μεγέθους s, είναι προφανώς και ένα συνεκτικό υποσύνολο k-κλικών για οποιοδήποτε $k \leq s$, διότι για κάθε ζεύγος μικρότερων k-κλικών που περιλαμβάνονται σε αυτόν, μπορεί να βρεθεί μια τετριμμένη ακολουθία γειτονικών k-κλικών που τις συνδέει. Επιπλέον, δύο πλήρεις υπογράφοι μεγάλου μεγέθους οι οποίοι μοιράζονται τουλάχιστον $k - 1$ κόμβους, διαμορφώνουν μια συνεκτική συνιστώσα k-κλικών. Αυτό σημαίνει οτι αντί να αναζητάμε k-κλίκες, θα ήταν προτιμότερο να εντοπίζουμε πρώτα τους μεγάλους πλήρεις υπογράφους στο δίκτυο και ύστερα να αναζητάμε τα συνετικά υποσύνολα k-κλικών (δηλαδή τις κοινότητες k-κλικών), για δεδομένο k και να μελετάμε την αλληλεπικάλυψη μεταξύ τους.

Ο αλγόριθμος που πρότειναν οι συγγραφείς, αρχικά εξάγει όλους τους πλήρεις υπογράφους του δικτύου οι οποίοι δεν είναι μέρη μεγαλύτερων πλήρων υπογράφων. Αυτοί οι μέγιστοι πλήρεις υπογράφοι ονομάζονται απλά κλίκες (cliques) και η διαφορά ανάμεσα στις k-κλίκες και τις κλίκες είναι οτι οι k-κλίκες μπορούν να αποτελέσουν υποσύνολα μεγαλύτερων πλήρων υπογράφων. Αφού εντοπιστούν οι κλίκες, κατασκευάζεται ο πίνακας αλληλοκαλυπτόμενων κλικών (clique-clique overlap matrix) O, ο οποίος είναι ένας $n_c \times n_c$ πίνακας, όπου n_c ο αριθμός των κλικών στο γράφο. Κάθε γραμμή και κάθε στήλη του πίνακα αντιστοιχούν σε μια κλίκα. Κάθε στοιχείο του πίνακα, O_{ij}, αντιστοιχεί στον αριθμό των κόμβων που οι αντίστοιχες κλίκες μοιράζονται, ενώ τα διαγώνια στοιχεία του, O_{ii}, αντιστοιχούν στο μέγεθος της κάθε κλίκας. (Η τομή δύο κλικών είναι πάντοτε ένας πλήρης υπογράφος). Οι κοινότητες k-κλικών, για ένα δεδομένο k, είναι ισοδύναμες με τις συνεκτικές συνιστώσες κλικών στις οποίες οι γειτονικές κλίκες μοιράζονται τουλάχιστον

$k-1$ κοινούς κόμβους. Οι συνεκτικές αυτές συνιστώσες μπορούν να βρεθούν διαγράφοντας κάθε μη διαγώνιο O_{ij} $(i \neq j)$ στοιχείο του πίνακα το οποίο είναι μικρότερο του $k-1$ και κάθε διαγώνιο O_{ii} στοιχείο του πίνακα το οποίο είναι μικρότερο του k. Τα εναπομείναντα στοιχεία του πίνακα τα αντικαθιστούμε με 1. Οι συνεκτικές συνιστώσες που αντιστοιχούν στον πίνακα που προκύπτει είναι οι διαφορετικές κοινότητες k −κλικών.

Σχήμα 23. Παράδειγμα CPM



Στο εν λόγω σχήμα έχουμε μια απλή αναπαράσταση του CPM αλγορίθμου για τον εντοπισμό των κοινοτήτων k −κλικών για $k = 4$. Το επάνω αριστερά σχήμα δείχνει ένα γράφο στον οποίο οι διαφορετικές κλίκες δίνονται με διαφορετικά χρώματα. Ο πίνακας αλληλοκαλυπτόμενων κλικών φαίνεται στο επάνω δεξιά σχήμα. Για να εντοπιστούν οι κοινότητες k −κλικών για $k = 4$, διαγράφονται τα μη διαγώνια στοιχεία του πίνακα που είναι μικρότερα του 3 και τα διαγώνια στοιχεία του πίνακα που είναι μικρότερα του 4. Ο πίνακας στον οποίο καταλήγουμε φαίνεται κάτω αριστερά. Οι συνεκτικές συνιστώσες (κοινότητες 4-κλικών) που αντιστοιχούν στον πίνακα αυτόν φαίνονται κάτω δεξιά. (Palla, Derényi, Farkas, & Vicsek, 2005)

Στο σημείο αυτό, θα αναλύσουμε τον τρόπο με τον οποίο ο αλγόριθμος εντοπίζει τις κλίκες, για να συνεχίσει στην κατασκευή του πίνακα O. Όπως προαναφέραμε, σε αντίθεση με τις k −κλίκες, οι κλίκες δεν μπορούν να αποτελέσουν υποσύνολα μεγαλύτερων κλικών, επομένως πρέπει να εντοπίζονται κατα φθίνουσα σειρά μεγέθους. Η μεγαλύτερου δυνατού μεγέθους κλίκα στον υπο μελέτη γράφο, καθορίζεται απο την ακολουθία βαθμών των κόμβων. Ξεκινώντας με την κλίκα του μεγαλύτερου μεγέθους, ο αλγόριθμος επιλέγει επαναλαμβανόμενα έναν κόμβο, εξάγει κάθε κλίκα του εν λόγω μεγέθους που περιέχει τον κόμβο αυτό και έπειτα διαγράφει τον εν λόγω κόμβο και τις ακμές του. Η διαγραφή των κόμβων που έχουν ήδη εξεταστεί εμποδίζει τον αλγόριθμο να εντοπίσει την ίδια κλίκα πολλές φορές. Απο τη στιγμή που δεν έχουν απομείνει κόμβοι προς εξέταση, το μέγεθος της κλίκας μειώνεται κατά ένα και η διαδικασία εύρεσης κλικών επανακκινείται στον αρχικό γράφο. Οι κλίκες που έχουν ήδη βρεθεί επηρρεάζουν την περαιτέρω αναζήτηση απο τη στιγμή που οι μικρότερες κλίκες που δεν έχουν ακόμα βρεθεί δεν μπορούν να αποτελέσουν υποσύνολα τους.

Οι κλίκες μεγέθους s που περιέχουν έναν κόμβο v, μπορούν να βρεθούν εξετάζοντας τις σχέσεις μεταξύ των γειτονικών κόμβων του v. Αυτή η διαδικασία υλοποιείται στον αλγόριθμο ως εξής: Αρχικά κατασκευάζεται ένα σύνολο A, που περιέχει κόμβους που συνδέονται όλοι μεταξύ τους. Αρχικά το A αποτελείται μόνο απο τον κόμβο v και ο στόχος μας είναι να μεγενθύνουμε αυτό το σύνολο στο πραγματικό μέγεθος κλίκας, s. Κατασκευάζεται επίσης ένα σύνολο B, το οποίο αποτελείται από το σύνολο των κόμβων που συνδέονται με κάθε κόμβο του A. Οι κόμβοι του B, ωστόσο, δεν συνδέονται απαραίτητα μεταξύ τους. Αρχικά το B αποτελείται απο τους γειτονικούς κόμβους του v. Το σύνολο A μπορεί να διευρυνθεί με τη μεταφορά κόμβων σε αυτό απο το B. Αυτό επιτυγχάνεται αναδρομικά, προκειμένου να ελεγχθεί κάθε δυνατός συνδυασμός των κόμβων που μεταφέρονται. Για να αποφευχθεί η εύρεση της ίδιας κλίκας πολλές φορές, οι κόμβοι μεταφέρονται απο το σύνολο B στο A κατά αύξουσα ή φθίνουσα διάταξη των δεικτών

τους). Όταν ένας κόμβος w, τοποθετείται απο το σύνολο B στο A, οι κόμβοι που δεν είναι γειτονικοί με αυτόν αφαιρούνται απο το B. (Αυτό γίνεται με σκοπό να διατηρηθεί η ιδιότητα οτι οι κόμβοι του B συνδέονται όλοι με κάθε κόμβο του A). Κάθε φορά που το σύνολο A φτάνει στο μέγεθος s, βρίσκεται και μια νέα κλίκα. Μετά την καταγραφή της ο αλγόριθμος, γυρνάει πάλι πίσω για να ελέγξει τους εναπομείναντες δυνατούς συνδυασμούς των δεικτών των γειτονικών κόμβων. Εαν το σύνολο B εξαντληθεί απο κόμβους πριν το σύνολο A φτάσει στο μέγεθος s ή εαν η ένωση των δύο συνόλων μπορεί να συμπεριληφθεί σε μια μεγαλύτερη κλίκα που έχει ήδη βρεθεί, η αναδρομική διαδικασία γυρνάει πίσω για να ελέγξει άλλες δυνατότητες.

Ο εντοπισμός κλικών στο γράφο απαιτεί χρόνο ο οποίος αυξάνεται εκθετικά με το μέγεθος του γράφου. Ωστόσο οι συγγραφείς, έχοντας κάνει εφαρμογές του αλγορίθμου σε δίκτυα πραγματικού χρόνου, έχουν βρει οτι η διαδικασία είναι πολύ γρήγορη λόγω του αρκετά περιορισμένου αριθμού κλικών που υπάρχουν στα δίκτυα αυτά, όταν είναι αραιά. Αραιά δίκτυα της τάξης των 10^5 κορυφών μπορούν να αναλυθούν σε αρκετά σύντομο χρονικό διάστημα. Η πολυπλοκότητα (actual scalability) του αλγορίθμου εξαρτάται απο πολλούς παράγοντες και δεν μπορεί να εκφραστεί σε κλειστό τύπο. Ο αλγόριθμος CPM έχει επεκταθεί για την ανάλυση σταθμισμένων, κατευθυνόμενων και διμερών γράφων, (Farkas, Abel, Palla, & Vicsek, 2007) (Lehmann, Schwartz, & Hansen, 2008) αλλά έχουν υπάρξει και τροποποιήσεις του για ταχύτερη εφαρμογή του, ο λεγόμενος Sequential Clique Percolation algorithm (SCP) (Kumpula, Kivela, Kaski, & Saramaki, 2008). Ένας άλλος αλγόριθμος που στηρίζεται στη μέθοδο της διήθησης κλίκας, ονομάζεται EAGLE (agglomerative hierarchical clustering based on maximal clique), (Shen, Cheng, Cai, & Hu, 2008) στον οποίο η ιεραρχία των αλληλοκαλυπτόμενων κοινοτήτων εντοπίζεται μέσω μιας διαδικασίας συσσωρευτικής ιεραρχικής συσταδοποίησης.

Το βασικό μειονέκτημα του CPM είναι το εξής: Υποθέτει οτι ο γράφος αποτελείται απο μεγάλο αριθμό κλικών και έτσι μπορεί να αποτύχει στο να παράγει καλύμματα με νόημα σε γράφους με λίγες κλίκες, όπως τεχνολογικά δίκτυα η κάποια κοινωνικά δίκτυα. Απο την άλλη πλευρά, εαν υπάρχουν πολλές κλίκες στο γράφο, η μέθοδος μπορεί να παράγει ασήμαντη κοινοτική δομή του δικτύου, όπως το να θεωρήσει όλο το γράφο ένα κάλυμα με μια ενιαία κοινότητα. Ένα πιο θεμελιώδες ζήτημα είναι το γεγονός οτι η μέθοδος δεν ψάχνει για πραγματικές κοινότητες με την έννοια των πυκνών υπογράφων, αλλά για υπογράφους που περιέχουν πολλές κλίκες, οι οποίοι μπορεί να είναι πολύ διαφορετικά αντικείμενα απο κοινότητες (για παράδειγμα

θα μπορούσαν να είναι αλυσίδες κλικών με χαμηλή εσωτερική πυκνότητα ακμών). Ένα άλλο μεγάλο πρόβλημα είναι οτι σε δίκτυα πραγματικού χρόνου κάποιο κομμάτι κορυφών μπορεί να μένει εκτός των κοινοτήτων, τα λεγόμενα «φύλλα». Ενώ θα μπορούσε να κατασκευαστεί μια διαδικασία ωστε να συμπεριληφθούν και αυτά στις παραγόμενες κοινότητες, για να επιτευχθεί αυτό, θα πρέπει να εισαχθεί ένα νέο κριτήριο έξω απο το πλαίσιο απο το οποίο εμπνεύστηκε η μέθοδος. Τέλος, δεν είναι ξεκάθαρο εκ των προτέρων ποια τιμή του k πρέπει να επιλεχθεί για τον εντοπισμό κοινοτικών δομών με νόημα.

5.2 Modularity για αλληλοκαλυπτόμενες κοινότητες.

Στη μέχρι τώρα ανάλυση της modularity, έχουμε αναφερθεί σε τεχνικές εντοπισμού κοινοτήτων στο δίκτυο στις οποίες κάθε κόμβος ανήκει σε μία μόνον κοινότητα, δηλαδή σε διαμερίσεις του δικτύου χωρίς επικαλύψεις μεταξύ των παραγόμενων κοινοτήτων. Ωστόσο, μια συμπληρωματική προσέγγιση του προβλήματος εντοπισμού κοινοτήτων είναι αυτή στην οποία κόμβοι ανήκουν σε περισσότερες απο μία κοινότητες.

Οι Nicosia et al. (2009), επέκτειναν τη modularity στη γενικότερη περίπτωση των κατευθυνόμενων δικτύων με αλληλοκαλυπτόμενες κοινότητες. Η βασική ιδέα πίσω απο την προσέγγισή τους, είναι η επέκταση του μοντέλου διαμόρφωσης που χρησιμοποιείται στον ορισμό της modularity, επιτρέποντας κόμβους να ανήκουν σε πολλές κοινότητες ταυτόχρονα.

Τυπικά, οι κόμβοι ανήκουν σε κάθε κοινότητα με μια συγκεκριμένη ισχύ και κάθε κόμβος $i \in V$ συνδέεται με έναν συντελεστή $\alpha_{i,c}$ ο οποίος δείχνει πόσο «ισχυρά» ο κόμβος αυτός ανήκει στην κοινότητα c. Πιο συγκεκριμένα, κάθε κόμβος $i \in V$ σχετίζεται με ένα διάνυσμα $[a_{i,1}, \alpha_{i,2}, \dots, \alpha_{i,|c|})]^T$, όπου $|c|$ ο συνολικός αριθμός κοινοτήτων στο δίκτυο. Κατόπιν, ορίζεται ένας παρεμφερής συντελεστής ο οποίος σχετίζεται με τη συμμετοχή των ακμών στις κοινότητες: Για κάθε κατευθυνόμενη ακμή $e = (i, j)$, ο παράγοντας συμμετοχής της στην κοινότητα c αναπαρίσταται απο μια συνάρτηση των αντίστοιχων συντελεστών των κόμβων i, j, $\beta_{e,c} = f(\alpha_{i,c}, a_{j,c})$. Έτσι η $\delta(c_i, c_j)$ συνάρτηση που χρησιμοποείται στον τύπο της modularity, μπορεί να αντικατασταθεί απο δύο διαφορετικούς συντελεστές, r_{ijc} και s_{ijc}, που αφορούν τη συνεισφορά της

ακμής (i, j) στη modularity του δικτύου και στο μοντέλο διαμόρφωσης αντίστοιχα. Τελικά, ο τύπος της modularity για αλληλοκαλυπτόμενες κοινότητες εκφράζεται ως εξής:

$$Q_{ov} = \frac{1}{m}\sum_{\forall c}\sum_{i,j}[r_{ijc}A_{ij} - s_{ijc}\frac{k_i^{out}k_j^{in}}{m}]$$

Εαν δεν υπάρχει επικάλυψη μεταξύ κοινοτήτων, τότε $r_{ijc} = s_{ijc} = \delta(c_i, c_j)$, όπου η ακμή (i, j) συνεισφέρει στη modularity μόνο εάν $c_i = c_j$. Η τιμή του συντελεστή r_{ijc} μπορεί να θεωρηθεί η συνεισφορά μιας ακμής $e = (i, j)$ στη modularity της κοινότητας c, $r_{ijc} = \beta_{e,c} = \mathcal{F}(\alpha_{i,c}, a_{j,c})$. Για τον παράγοντα s_{ijc} που σχετίζεται με το μοντέλο διαμόρφωσης, αν θεωρήσουμε οτι το πόσο ανήκει ένας κόμβος σε μια κοινότητα είναι ανεξάρτητο του πόσο ανήκει οποιοσδήποτε άλλος κόμβος στην κοινότητα αυτή (δηλαδή, η πιθανότητα ένας κόμβος i να ανήκει στην κοινότητα c με ισχύ $\alpha_{i,c}$ δε σχετίζεται με την πιθανότητα οποιοσδήποτε άλλος κόμβος j να ανήκει στην ίδια κοινότητα με ισχύ $\alpha_{j,c}$) τότε η modularity μπορεί να οριστεί ώς:

$$Q_{ov} = \frac{1}{m}\sum_{\forall c}\sum_{i,j}\left[r_{ijc}A_{ij} - s_{ijc}\frac{\beta_{e,c}^{out}k_i^{out}\beta_{e,c}^{in}k_j^{in}}{m}\right] = \frac{1}{m}\sum_{\forall c}\sum_{i,j}\left[r_{ijc}A_{ij} - s_{ijc}\frac{\beta_{i\rightarrow,c}^{out}k_i^{out}\beta_{\leftarrow,c}^{in}k_j^{in}}{m}\right]$$

Όπου

$$\beta_{e,c}^{out} = \beta_{i\rightarrow,c}^{out} = \frac{\sum_{j\epsilon V}\mathcal{F}(\alpha_{i,c},a_{j,c})}{|V|} \kappa\alpha\iota \;\; \beta_{e,c}^{in} = \beta_{\leftarrow,c}^{in} = \frac{\sum_{j\epsilon V}\mathcal{F}(\alpha_{i,c},a_{j,c})}{|V|},$$

οι αναμενόμενοι συντελεστές «συμμετοχής» οποιασδήποτε ακμής $e = (i, j)$ με τον κόμβο i να ανήκει στην κοινότητα c. (δηλαδή ο μέσος όρος συμμετοχής για όλες τις ακμές).

Κάτι που πρέπει να διασαφηνιστεί επιπλέον για να οριστεί η modularity για κατευθυνόμενους γράφους με αλληλοκαλυπτόμενες κοινότητες, αφορά την επιλογή της συνάρτησης $\mathcal{F}(\alpha_{i,c}, a_{j,c})$. Η συνάρτηση αυτή καθορίζει τη συμμετοχή μιας ακμής (i, j) σε μια κοινότητα c σύμφωνα με τους συντελεστές συμμετοχής των κόμβων i και j. Οι Nicosia et al, πρότειναν η επιλογή της συνάρτησης $\mathcal{F}(\cdot)$ πρέπει να πληρεί τις παρακάτω ιδιότητες:

- **i.** Η Q_{ov} θα πρέπει να ισούται με μηδέν όταν δεν μπορεί να εντοπιστεί κοινοτική δομή και όλοι οι κόμβοι ανήκουν στην ίδια κοινότητα, και
- **ii.** Υψηλότερες τιμές της Q_{ov} θα πρέπει να συνεπάγονται και καλύτερη κοινοτική δομή.

Κάθε πιθανή συνάρτηση που διατηρεί αυτές τις ιδιότητες μπορεί να εφαρμοστεί στη modularity. Τέλος, οι συγγραφείς παρουσιάζουν έναν γενετικό αλγόριθμο για τη βελτιστοποίηση του προτεινόμενου κριτηρίου της modularity, ο οποίος μπορεί να χρησιμοποιηθεί για τον εντοπισμό αλληλοκαλυπτόμενης κοινοτικής δομής σε κατευθυνόμενους γράφους. Περιγραφή του δίνεται στο παράρτημα 2.

5.3 Άλλες τεχνικές εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων

Στη βιβλιογραφία μέχρι στιγμής, έχουν προταθεί πολλές και διαφορετικές προσσεγγίσεις εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων σε δίκτυα, οι οποίες μπορούν να χωριστούν χοντρικά σε δύο μεγάλες κατηγορίες: (i) Αλγορίθμους που στηρίζονται σε κόμβους και (ii) Αλγορίθμους που στηρίζονται σε ακμές. Οι αλγόριθμοι εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων που στηρίζονται σε κόμβους, συσταδοποιούν τους κόμβους απευθείας, χρησιμοποιώντας τη δομική τους πληροφορία. Σε αυτή την κατηγορία μεθόδων, ανήκουν πολλοί γνωστοί και εδραιωμένοι αλγόριθμοι, εκτός αυτών που αναλύθηκαν εκτενώς προηγουμένως. Θα τους κατηγοριοποιήσουμε συνοπτικά, σύμφωνα με τον τρόπο με τον οποίο εντοπίζουν τις κοινότητες. Οι αλγόριθμοι εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων που στηρίζονται σε ακμές, πρώτα συσταδοποιούν τις ακμές του δικτύου και ύστερα κάνουν μια αντιστοίχιση των κοινοτήτων ακμών σε κοινότητες κόμβων, συλλέγοντας κόμβους που είναι «συναφείς» με όλες τις ακμές των εν λόγω κοινοτήτων ακμών. Η δεύτερη κατηγορία αλγορίθμων υπερέχει της πρώτης στην ανίχνευση πολύπλοκων και διαφορετικής κλίμακας κοινοτήτων. Ωστόσο οι αλγόριθμοί της έχουν υψηλή υπολογιστική πολυπλοκότητα. Στο σημείο αυτό, θα αναφερθούμε περιληπτικά σε κάποιους διάσημους αλγορίθμους εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων που στηρίζονται σε κόμβους.

Θα ξεκινήσουμε την αναφορά μας απο τους αλγορίθμους που χρησιμοποιούν τις έννοιες της τοπικής επέκτασης και βελτιστοποίησης. Οι αλγόριθμοι αυτοί, βασίζονται κατά κόρον στην ανάπτυξη φυσικών κοινοτήτων (natural communities) (Lancichinetti, Fortunato, & Kertesz, Detecting the overlapping and hierarchical community structure of complex networks, 2009) που αλληλοκαλύπτονται μεταξύ τους. Μια φυσική κοινότητα ενός κόμβου, είναι μια κοινότητα που αναπτύχθηκε απο τον κόμβο αυτό, μέσω ενός άπληστου αλγορίθμου που μεγιστοποιεί μια συνάρτηση προσαρμογής ή οφέλους (local benefit-local fitness function), η οποία εξαρτάται απο

μια παράμετρο ανάλυσης. Η συνάρτηση αυτή, χαρακτηρίζει την ποιότητα μιας πυκνά συνδεδεμένης ομάδας κόμβων.

Οι Baumes et al (2005) πρότειναν μια διαδικασία δύο βημάτων. Το πρώτο βήμα ονομάζεται RankRemoval και χρησιμοποιείται για να κατατάξει τους κόμβους του δικτύου σύμφωνα με κάποιο κριτήριο σημαντικότητας. Η σημαντικότητα των κορυφών καθορίζεται απο τις κεντρικές τους βαθμολογίες (centrality scores) (όπως για παράδειγμα ο βαθμός τους ή η ενδιάμεση κεντρικότητα (betweenness centrality) (Freeman, 1977). Έπειτα η διαδικασία αφαιρεί επαναληπτικά κόμβους υψηλής κατάταξης εως ότου διαμορφωθούν μικροί, διακεκομμένοι πυρήνες συστάδων. Αυτοί οι πυρήνες χρησιμεύουν ως seed κοινότητες για το δεύτερο βήμα της διαδικασίας που ονομάζεται Iterative Scan (IS), το οποίο επεκτείνει τους πυρήνες προσθέτοντας η αφαιρώντας κόμβους έναν προς έναν μέχρις ότου μια συνάρτηση τοπικής πυκνότητας να μη μπορεί να βελτιωθεί περαιτέρω. Η συνάρτηση πυκνότητας αυτή ορίζεται ως εξής:

$$f(c) = \frac{w_{in}^c}{w_{in}^c + w_{out}^c}$$

όπου w_{in}^c και w_{out}^c είναι το συνολικό εσωτερικό και εξωτερικό βάρος της κοινότητας c. Ο χρόνος εκτέλεσης χειρότερης περίπτωσης της όλης διαδικασίας είναι $O(n^2)$. Οι ποιότητα των κοινοτήτων που εντοπίστηκαν, εξαρτάται απο την ποιότητα των seeds. Απο τη στιγμή που ο αλγόριθμος αφαιρεί κορυφές κατα τη διάρκεια της επέκτασης, έχει δειχθεί μέσω πειραμάτων οτι ο IS παράγει μη συνεκτικές συνιστώσες σε κάποιες περιπτώσεις. Για το λόγο αυτό, εισήχθηκε μια τροποποιημένη έκδοσή του, που ονομάζεται CIS, στην οποία η συνεκτικότητα ελέγχεται σε κάθε επανάληψη. Στην περίπτωση που μια κοινότητα είναι «σπασμένη» σε περισσότερα απο ένα μέρη, διατηρείται το μέρος αυτό με τη μεγαλύτερη πυκνότητα. Με τον CIS, αναπτύσσεται επίσης μια καινούρια συνάρτηση πυκνότητας

$$f(c) = \frac{w_{in}^c}{w_{in}^c + w_{out}^c} + \lambda e_p$$

η οποία ενσωματώνει την πιθανότητα ακμών e_p. Η παράμετρος λ, ελέγχει το πως συμπεριφέρεται ο αλγόριθμος σε αραιές περιοχές του δικτύου.

Ο LFM (Lancichinetti, Fortunato, & Kertesz, Detecting the overlapping and hierarchical community structure of complex networks, 2009), στηρίζεται και αυτός στην υπόθεση οτι οι κοινότητες είναι ουσιαστικά τοπικές δομές, που αφορούν κόμβους που ανήκουν σε τμήματα συν

το πολύ μια εκτεταμένη γειτονία τους. Ο LFM επεκτείνει μια κοινότητα από έναν τυχαίο seed κόμβο σε μια φυσική κοινότητα, μέχρι η συνάρτηση προσαρμογής

$$f_G = \frac{k_{in}^G}{\left(k_{in}^G + k_{out}^G\right)^a},$$

να είναι τοπικά μέγιστη, όπου k_{in}^G και k_{out}^G είναι οι συνολικοί εσωτερικοί και εξωτερικοί βαθμοί του τμήματος G και η παράμετρος ανάλυσης (resolution parameter) a είναι μια παράμετρος που παίρνει θετικές πραγματικές τιμές και ελέγχει το μέγεθος των κοινοτήτων. Αφού βρει μια κοινότητα, ο LFM επιλέγει τυχαία έναν κόμβο που δεν έχει ακόμα ανατεθεί σε καμία κοινότητα και αναπτύσσει μια νέα. Ο LFM εξαρτάται σημαντικά απο την παράμετρο ανάλυσης α. Η υπολογιστική του πολυπλοκότητα για μια δεδομένη τιμή της παραμέτρου a, είναι κατα προσέγγιση $O(n_c s^2)$, όπου n_c ο αριθμός των κοινοτήτων και s το μέσο μέγεθός τους. Ο χρόνος εκτέλεσης χειρότερης περίπτωσης είναι $O(n^2)$.

Μια τροποποίηση του LFM, είναι ο αλγόριθμος MONC (Merging of Overlapping Natural Communities) (Havemann, Heinz, & Struck, 2011) ο οποίος χρησιμοποιεί την τροποποιημένη συνάρτηση προσαρμογής του LFM, $f(c) = \frac{k_{in}^c + 1}{\left(k_{in}^c + k_{out}^c\right)^{\alpha}}$, η οποία επιτρέπει στους κόμβους του δικτύου να θεωρούνται κοινότητες απο μόνοι τους. Η προτεινόμενη συνάρτηση προσαρμογής, επιτρέπει στον αλγόριθμο MONC να βρίσκει το εύρος των α για τα οποία ένα σύνολο κόμβων είναι τοπικά βέλτιστο. Ο MONC, σε αντίθεση με τον LFM, δε διερευνά αριθμητικά αυτές τις τιμές της παραμέτρου α. Με την αριθμητική διερεύνηση τιμών, η διαδικασία πρέπει να επαναλαμβάνεται για κάθε τιμή της παραμέτρου ανάλυσης, γεγονός χρονοβόρο αλλά και ανεπαρκές όσον αφορά την ακρίβεια του αλγορίθμου, καθώς παρέχει μόνο εκτιμήσεις των επιπέδων ανάλυσης στα οποία μεταβάλλονται τα καλύμματα του δικτύου. Αντι αυτού, ο MONC υπολογίζει με ακρίβεια τα επίπεδα ανάλυσης στα οποία τα μέλη των φυσικών κοινοτήτων αλλάζουν και ως εκ τούτου καθορίζει τις φυσικές κοινότητες για όλα τα επίπεδα ανάλυσης σε ένα τρέξιμό του. Αυτό το καταφέρνει υπολογίζοντας την επόμενη χαμηλότερη τιμή της παραμέτρου α που έχει ως αποτέλεσμα την περαιτέρω επέκταση μιας κοινότητας. Στην περίπτωση που η φυσική κοινότητα ενός κόμβου i είναι υποσύνολο ενός άλλου κόμβου η ανάλυση του i σταματά. Κατά τον τρόπο αυτό ο MONC συγχωνεύει κοινότητες κατα την επεξεργασία τους και ως εκ τούτου, παράγει καλύματα γρηγορότερα απο τον LFM.

Ένας άλλος αλγόριθμος που ανήκει στην κατηγορία τεχνικών βελτιστοποίησης είναι ο OSLOM (Order Statistics Local Optimization Method) (Lancichinetti et al, 2011). Είναι μια μέθοδος που βελτιστοποιεί τοπικά τη στατιστική σημαντικότητα των συστάδων, η οποία ορίζεται σύμφωνα με ένα καθολικό μηδενικό μοντέλο. Ο OSLOM χρησιμοποιεί τη στατιστική σημαντικότητα των συστάδων ως μέτρο προσαρμογής (fitness measure) τους με σκοπό να τις εκτιμήσει. Η στατιστική σημαντικότητα μιας συστάδας ορίζεται απο τους συγγραφείς ως η πιθανότητα να βρεθεί η συστάδα αυτή σε ένα τυχαίο μηδενικό μοντέλο, δηλαδή σε μια κλάση γράφων χωρίς κοινοτική δομή. Οι συγγραφείς επιλέγουν το μοντέλο διαμόρφωσης (configuration model) των Molloy και Reed (1995) ως μηδενικό μοντέλο.

Στη βιβλιογραφία έχουν προταθεί και πολλοί άλλοι αλγόριθμοι εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων που χρησιμοποιούν τις έννοιες της τοπικής επέκτασης μιας κοινότητας, βελτιστοποιώντας μια τοπική συνάρτηση οφέλους. Ενδεικτικά θα αναφέρουμε ονομαστικά και σε κάποιους άλλους, όπως UEOC (Jin et al, 2011), OCA (Padrol-Sureda et al., 2010), iLCD (Cazabet et al., 2010).

Μια άλλη κατηγορία αλγορίθμων εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων, ονομάζονται ασαφείς αλγόριθμοι (fuzzy community detection algorithms). Οι αλγόριθμοι αυτοί ποσοτικοποιούν την ισχύ της συσχέτισης ανάμεσα σε όλα τα ζεύγη κόμβων και κοινοτήτων. Σε αυτούς, υπολογίζεται ένα διάνυσμα «μαλακής» συμμετοχής (soft membership) ή παράγοντας ένταξης (belonging factor) (Gregory, 2010) για κάθε κόμβο. Ένα μειονέκτημα των αλγορίθμων αυτών, είναι η ανάγκη να προσδιοριστεί η διάσταση k του διανύσματος συμμετοχής. Αυτή η τιμή είτε υπάρχει ως παράμετρος στον αλγόριθμο είτε υπολογίζεται απο τα δεδομένα. Αλγόριθμοι που ανήκουν σε αυτή την κατηγορία είναι ο αλγόριθμος των Nepusz et al. (2008), ο αλγόριθμος των Zhang et al. (2007), ο SPAEM των Newman και Leicht (2007), ο SSDE των Magdon-ismail και Purnel (2011), ο OSBM των Latouche et al. 2011 και ο MOSES των McDaid και Hurley (2010).

Μια τελευταία κατηγορία αλγορίθμων εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων, είναι αυτή της διάδοσης ετικέτας και οι πιο διάσημοι αλγόριθμοι της κατηγορίας αυτής είναι ο SLPA των Xie et al. (2011), ο COPRA του Gregory (2010) και τα πολυμορφικά μοντέλα περιστροφής (Multi-state Spin Models) των Reichard και Bornholdt (2004) και των Lu et al. (2009).

Παρά την πολλή δουλειά και έρευνα που έχει αφιερωθεί στην ανάπτυξη αλγορίθμων εντοπισμού αλληλοκαλυπτόμενων κοινοτήτων, υπάρχουν ορισμένα θεμελιώδη ερωτημάτα που

δεν έχουν ακόμη αντιμετωπιστεί πλήρως. Δύο απο τα πιο σημαντικά είναι: (i) πότε πρέπει να εφαρμόζονται μέθοδοι αλληλοκάλυψης και (ii) πόσο σημαντική είναι η αλληλοκάλυψη.

Κεφάλαιο 6. Εφαρμογή και Δοκιμές των Αλγορίθμων

6.1 Κανονικοποιημένη Κοινή Πληροφορία

Η αξιολόγηση ενός αλγορίθμου σε οποιονδήποτε γράφο με ενσωματωμένη κοινοτική δομή συνεπάγεται και τον καθορισμό ενός ποσοτικού κριτηρίου το οποίο εκτιμά το πόσο καλή είναι η απάντηση που δίνει ο αλγόριθμος σε σύγκριση με την πραγματική μορφή κοινοτικής δομής που δίνεται απο τις πληροφορίες του γράφου. Η σύγκριση αυτή μπορεί να γίνει με τη χρήση κατάλληλων μέτρων ομοιότητας. Ανά τη βιβλιογραφία μέσα στα χρόνια, έχουν προταθεί διάφορα τέτοια μέτρα, εμείς στην παρούσα ανάλυση και σύγκριση αλγορίθμων θα χρησιμοποιήσουμε ένα απο τα πιο σύγχρονα και πιο αξιόπιστα, το οποίο προέρχεται απο τη θεωρία πληροφορίας. Αυτό ονομάζεται κανονικοποιημένη κοινή πληροφορία (Normalized Mutual Information-NMI), την οποία θα περιγράψουμε παρακάτω.

Για να αξιολογηθεί το περιεχόμενο της πληροφορίας Shannon (Shannon Information Content) (Mackay D. J., 2003), θεωρούμε τις κοινοτικές αναθέσεις (community assignments) $\{x_i\}$, $\{y_i\}$, όπου x_i και y_i υποδηλώνουν τις ετικέτες των συστάδων (cluster labels) της κορυφής i στις διαμερίσεις $\mathcal{X}$ και Υ, αντίστοιχα. Υποθέτουμε οτι οι ετικέτες x και y είναι οι τιμές δύο τυχαίων μεταβλητών X και Y με απο κοινού συνάρτηση πιθανότητας $P(x,y) = P(X = x, Y = y) = n_{xy}/n$, το οποίο σημαίνει οτι $P(x) = P(X = x) = n_x^X/n$ και $P(y) = P(Y = y) = n_y^Y/n$, όπου n_x^X, n_y^Y και n_{xy} είναι τα μεγέθη των συστάδων με ετικέτα x, y και της επικάλυψής τους, αντίστοιχα. Η Κοινή Πληροφορία (Mutual Information) $I(X,Y)$ δύο τυχαίων μεταβλητών ορίζεται ως

$$I(X,Y) = \sum_x \sum_y P(x,y) log \frac{P(x,y)}{P(x)P(y)}.$$

Το μέτρο $I(X,Y)$ δηλώνει πόσα μπορούμε να μάθουμε για την ποσότητα X γνωρίζοντας την ποσότητα Y και αντίστροφα. Στην πραγματικότητα $I(X,Y) = H(X) - H(X|Y)$, 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 κοινή πληροφορία $I(X,Y)$ ωστόσο δεν είναι ιδανική ως μέτρο ομοιότητας. Αυτό συμβαίνει διότι δοθείσας μιας διαμέρισης $\mathcal{X}$, όλες οι διαμερίσεις που παράγονται απο αυτή, μέσω περαιτέρω διαμερίσεων κάποιων απο τις συστάδες της, θα είχαν όλες την ίδια κοινή πληροφορία με την $\mathcal{X}$, ενώ ενδέχεται να είναι πολυ διαφορετικές μεταξύ τους. Στην περίπτωση αυτή η Κοινή Πληροφορία απλά θα ισούταν με την εντροπία $H(X)$ διότι η υπο συνθήκη εντροπία θα ήταν συστηματικά μηδέν. Για να αποφευχθεί αυτό, οι Danon et al. υιοθέτησαν την Κανονικοποιημένη Κοινή Πληροφορία *(*Normalized Mutual Information, NMI*)*, η οποία ορίζεται ως:

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

η οποία ισούται με 1 εαν οι διαμερίσεις είναι πανομοιότυπες, ενώ έχει αναμενόμενη τιμή 0 έαν οι διαμερίσεις είναι ανεξάρτητες. Η Κανονικοποιημένη Κοινή Πληροφορία χρησιμοποιείται πολύ συχνά σε ελέγχους και συγκρίσεις αλγορίθμων εντοπισμού κοινωνιών σε δίκτυα. (Lancichinetti & Fortunato, 2009)

6.2 Δίκτυα Αναφοράς και LFR Δίκτυο Αναφοράς

Όπως έχουμε έχουμε αναφέρει και σε προηγούμενα κεφάλαια, ο εντοπισμός κοινωνιών σε δίκτυα είναι ένα θεμελιώδες πρόβλημα κατά τη μελέτη τους, η επίλυση του οποίου μπορεί να δώσει σημαντικές πληροφορίες για το πως αυτά οργανώνονται αλλά και λειτουργούν. Οι μέθοδοι που έχουν αναπτυχθεί ανα τα χρόνια χρησιμοποιούν εργαλεία και τεχνικές απο ετερογενή επιστημονικά πεδία, όπως φυσική, βιολογία, εφαρμοσμένα μαθηματικά και επιστήμη των υπολογιστών. Ωστόσο δεν είναι σαφές ποιοι αλγόριθμοι απο όλα αυτά τα πεδία είναι αξιόπιστοι και πρέπει να χρησιμοποιούνται στις εκάστοτε εφαρμογές. Η ερώτηση για την αξιοπιστία των αλγορίθμων καθεαυτή ωστόσο είναι δυσνόητη, γιατί όπως έχουμε αναφέρει, απαιτεί κοινούς ορισμούς κοινότητας αλλά και τμηματοποίησης του δικτύου. Αυτό ουσιαστικά σημαίνει οτι παρόλη την τεράστια βιβλιογραφία πάνω στο θέμα, δεν υπάρχει απο κοινού συμφωνία μεταξύ των συγγραφέων για το πως τελικά μοιάζει ένα δίκτυο με κοινότητες. Στο σημείο αυτό θα κάνουμε μια περιγραφή των πιο διάσημων δικτύων αναφοράς που έχουν χρησιμοποιηθεί για τη σύγκριση των διαθέσιμων τεχνικών και θα εξηγήσουμε γιατί επιλέξαμε το LFR δίκτυο αναφοράς (LFR

benchmark) ως το δίκτυο στο οποίο θα συγκρίνουμε τους αλγορίθμους που θα χρησιμοποιήσουμε. Η πιο απλή μορφή μοντέλου δικτύου που έχει χρησιμοποιηθεί στη βιβλιογραφία, ονομάζεται μοντέλο εμφωλευμένης l −διαμέρισης (*planted l-partition model*), το οποίο έχει χρησιμοποιηθεί σε διάφορες εκδόσεις του. Στο μοντέλο δικτύου αυτό, κάθε κόμβος συνδέεται με τους υπόλοιπους κόμβους της ομάδας του με πιθανότητα p_{in} και με πιθανότητα p_{out} με κόμβους διαφορετικών ομάδων. Όσο ισχυέι $p_{in} \geq p_{out}$ οι ομάδες είναι κοινότητες, ενώ όταν $p_{in} \leq p_{out}$ το δίκτυο είναι ουσιαστικά ένας τυχαίος γράφος, χωρίς κοινοτική δομή. Η πιο δημοφιλής έκδοση του μοντέλου εμφωλευμένης l −διαμέρισης προτάθηκε απο τους Girvan και Newman, το οποίο πήρε και το όνομα του απο τα αρχικά τους, GN δίκτυο αναφοράς (GN benchmark). Το GN δίκτυο αναφοράς αποτελείται απο 128 κόμβους, κάθε ένας απο αυτούς με αναμενόμενο βαθμό 16, οι οποίοι χωρίζονται σε 4 ομάδες των 32 κόμβων. Πολλοί αλγόριθμοι έχουν συγκριθεί βασιζόμενοι στην απόδοσή τους πάνω σε αυτό το δίκτυο αναφοράς, λόγω της απλότητας της δομής του. Ωστόσο, το GN δίκτυο αναφοράς έχει δύο βασικά μειονεκτήματα. Το πρώτο είναι οτι όλοι του οι κόμβοι έχουν τον ίδιο αναμενόμενο βαθμό. Το δεύτερο είναι οτι όλες οι κοινότητες έχουν το ίδιο μέγεθος. Αυτά τα γνωρίσματα είναι μη ρεαλιστικά, καθώς τα σύνθετα δίκτυα χαρακτηρίζονται απο ετερεγονείς κατανομές βαθμών και κοινοτικών μεγεθών. Για το λόγο αυτό, έχει προταθεί μια γενίκευση των GN δικτύων αναφοράς, τα λεγόμενα LFR δίκτυα αναφοράς (Lancichinetti-Fortunato-Radicchi, LFR benchmarks) τα οποία γενικεύουν το GN δίκτυο, επιβάλλοντας στους βαθμούς και τα κοινοτικά μεγέθη του δικτύου να ακολουθούν κατανομές τύπου δυνάμεων (power law distributions), με εκθέτες r_1 και r_2, αντίστοιχα. . Ο αριθμός των κόμβων στο δίκτυο συμβολίζεται με N. Κατα την κατασκευή του δικτύου κάθε κόμβος του έχει έναν βαθμό τον οποίο και διατηρεί σταθερό μέχρι τέλους. Ως ανεξάρτητη παράμετρος του δικτύου επιλέγεται η παράμετρος μίξης μ (mixing parameter μ), η οποία εκφράζει το λόγο ανάμεσα στον εξωτερικό βαθμό ενός κόμβου και το συνολικό του βαθμό.

Ενώ οι περισσότεροι αλγόριθμοι εντοπισμού κοινωνιών σε δίκτυα έχουν πολύ καλές αποδόσεις στο GN δίκτυο, λόγω της δομικής του απλότητας, η νέα αυτή κατηγορία γράφων θέτει στους αλγορίθμους ένα δυσκολότερο έργο καθώς είναι ευκολότερο πλέον να αποκαλυφθούν τα όρια των προς σύγκριση αλγορίθμων. Επιπλέον τα LFR δίκτυα αναφοράς κατασκευάζονται γρήγορα. Η πολυπλοκότητα των αλγορίθμων κατασκευής τους είναι γραμμική όσον αφορά τον αριθμό των συνδέσμων του δικτύου και έτσι είναι δυνατή η εφαρμογή αλγορίθμων σε πολύ μεγάλα δίκτυα, δεδομένου οτι ο εκάστοτε αλγόριθμος είναι γρήγορος αρκετά για να τα αναλύσει.

Για το λόγο αυτό, χρησιμοποιήσαμε τα LFR δίκτυα αναφοράς στην ανάλυσή μας και συγκεκριμένα δύο απο τις διάφορες εκδοχές τους, τα μη κατευθυνόμενα και μη σταθμισμένα (undirected and unweighted) και και μη κατευθυνόμενα και σταθμισμένα (undirected and weighted) και πάνω σε αυτά εφαρμόσαμε κατάλληλους αλγορίθμους από αυτούς που έχουμε περιγράψει και ταιριάζουν στην εκάστοτε μορφή.

6.3 Αλγόριθμοι

Στην ενότητα αυτή θα κάνουμε μια σύντομη περιγραφή των αλγορίθμων που χρησιμοποιήσαμε στην εφαρμογή μας. Προσπαθήσαμε να κάνουμε μια επιλογή η οποία να καλύπτει όσο το δυνατόν μεγαλύτερο εύρος τεχνικών απο αυτές που αναλύθηκαν σε προηγούμενα κεφάλαια, χρησιμοποιώντας τους πιο διάσημους απο την εκάστοτε κατηγορία, δεδομένου οτι ο κώδικάς τους ήταν διαθέσιμος στις βιβλιοθήκες της R, που είναι ως επι το πλείστον η γλώσσα που χρησιμοποίησαμε στην εφαρμογή μας. Η λίστα των 9 αλγορίθμων που χρησιμοποιήθηκαν, η οποία στηρίζεται και στην ανάλυση που έχουμε ήδη κάνει, αλλά και στην επιλογή αλγορίθμων που έκαναν οι Lancichinetti και Fortunato (Lancichinetti & Fortunato, Community detection algorithms: a comparative analysis, 2009) για να κάνουν συγκρίσεις, είναι η εξής:

1. *Ο αλγόριθμος των Girvan και Newman (Algorithm of Girvan and Newman)*: Είναι ο πρώτος μοντέρνος αλγόριθμος εντοπισμού κοινωνιών σε γράφους. Είναι ενας ιεραρχικός διαιρετικός αλγόριθμος στον οποίο οι ακμές αφαιρούνται με μια επαναληπτική διαδικασία, βασιζόμενη στην μεταξύ τους τιμή (betweenness), η οποία εκφράζει τον αριθμό των συντομότερων μονοπατιών μεταξύ ζευγών κόμβων που περνάνε απο την κάθε ακμή. Στην πιο διάσημη εφαρμογή του, η διαδικασία αφαίρεσης ακμών σταματά, όταν η modularity της διαμέρισης που προκύπτει, φτάσει τη μέγιστη τιμή της. Όπως έχουμε αναφέρει, η modularity των Girvan και Newman, είναι μια συνάρτηση ποιότητας που εκτιμά το πόσο καλή είναι μια διαμέριση, συγκρίνοντας τον εν λόγω γράφο με ένα μηδενικό μοντέλο, δηλαδή μια κατηγορία τυχαίων γράφων με την ίδια αναμενόμενη ακολουθία βαθμών με τον γράφο που εξετάζουμε. Ο αλγόριθμος έχει πολυπλοκότητα $O(N^3)$ σε αραιούς γράφους.

2. *Γρήγορη Λαίμαργη βελτιστοποίηση της Modularity απο τους Clauset et al*. (*Fast Greedy modularity optimization by Clauset et al*.) : Ο αλγόριθμος αυτός είναι ουσιαστικά μια γρήγορη εφαρμογή της τεχνικής που αναφέρθηκε πιο πάνω. Ξεκινώντας απο ένα σύνολο απομονωμένων κόμβων, οι ακμές του δικτύου προσθέτονται με μια επαναληπτική διαδικασία ώστε να παράγουν τη μεγαλύτερη δυνατή αύξηση της modularity των Newman και Girvan σε κάθε βήμα. Η πολυπλοκότητα του είναι $O(log^2 N)$ σε αραιούς γράφους.
3. *Γρήγορη βελτιστοποίηση της modularity των Blondel et al.(Fast modularity optimization by Blondel et al):* Αυτή είναι μια τεχνική πολλών βημάτων που βασίζεται σε μια τοπική βελτιστοποίηση της modularity, στη γειτονιά του κάθε κόμβου. Αφού παραχθεί κατά αυτόν τον τρόπο μια διαμέριση του δικτύου, οι κοινότητες αντικαθίστανται απο υπερκόμβους (supernodes), διαδικασία που ουσιαστικά αντικαθιστά το δίκτυο με ένα μικρότερο και σταθμισμένο. Η διαδικασία επαναλαμβάνεται μέχρι η modularity, η οποία υπολογίζεται πάντα σε σχέση με το αρχικό δίκτυο, να μη βελτιστοποιείται περαιτέρω. Η μέθοδος αυτή προσφέρει μια δίκαιη ισορροπία ανάμεσα στην ακρίβεια της εκτίμησης του μεγίστου της modularity, η οποία είναι ανώτερη της λαίμαργης τεχνικής των Clauset et al, που αναφέραμε πιο πάνω και υπολογιστικής πολυπλοκότητας, η οποία είναι γραμμική. Όπως θα δούμε και στα αποτελέσματα πιο μετά, είναι ανώτερη των δύο προαναφερθέντων τεχνικών για τους λόγους που αναφέραμε.
4. *Εξαντλητική βελτιστοποίηση της modularity μέσω προσομειωμένης ανόπτησης (Exhaustive modularity optimization via simulated annealing by Guimerà and Amaral):* Εδώ ο στόχος είναι ο ίδιος με τον λαίμαργο αλγόριθμο των Clauset et al., με τη διαφορά οτι η τελική εκτίμηση του μεγίστου είναι πολύ μεγαλύτερη, εξαιτίας της εξαντλητικής βελτιστοποίησης, με το κόστος όμως, της υπολογιστικής πολυπλοκότητας. Η βελτιστοποίηση αυτή δεν μπορεί να εκφραστεί με κλειστό τύπο και εξαρτάται απο τις παραμέτρους βελτιστοποίησης που θέτει ο χρήστης.
5. *Εύρεση Κοινοτήτων στηριζόμενη στις πολλαπλαστιαστικές ετικέτες-Αλγόριθμος των Raghkavan, Albert, Kumara. (Finding communities based on propagating labels-Algorithm of Ragkhkavan et al.):* Αυτός είναι ένας γρήγορος, σχεδόν γραμμικού χρόνου αλγόριθμος για εντοπισμό κοινωνιών σε δίκτυα. Λειτουργεί τοποθετώντας σε κάθε

κόμβο του δικτύου μια μοναδική «ετικέτα» και σε κάθε βήμα του, υιοθετεί την ετικέτα που εκείνη τη στιγμή, έχουν οι περισσότεροι γείτονες του κόμβου. Σε αυτή την επαναληπτική διαδικασία, πυκνά συνδεδεμένες ομάδες κόμβων αποτελούν και μια συναίνεση για μια μοναδική ετικέτα ωστε να κατασκευαστούν κοινότητες. Στην ανάλυση που θα ακολουθήσει θα αναφερόμαστε στον αλγόριθμο αυτό ως LP (Label Propagation)

6. *Αλγόριθμος που στηρίζεται στο κύριο ιδιοδιάνυσμα του πίνακα κοινοτήτων. (Community Structure Detection Algorithm based on the leading eigenvector of the community matrix by Mark Newman):* Η καρδιά του συγκεκριμένου αλγορίθμου είναι ο ορισμός του πίνακα της modularity, B, ο οποίος ορίζεται ως $B = A - P$, όπου A ο πίνακας γειτνίασης του μη κατευθυνόμενου δικτύου και P ο πίνακας που περιέχει τις πιθανότητες συγκεκριμένες ακμές είναι παρούσες σύμφωνα με το «μοντέλο διαμόρφωσης». Δηλαδή το $P(i, j)$ στοιχείο του πίνακα είναι η πιθανότητα οτι υπάρχει μια ακμή μεταξύ των κορυφών i και j σε ένα τυχαίο δίκτυο στο οποίο οι βαθμοί όλων των κορυφών είναι ίδιοι με αυτούς του προς εξέταση γράφου. Η μέθοδος αυτή λειτουργεί υπολογίζοντας το ιδιοδιάνυσμα του πίνακα της modularity για τη μεγαλύτερη θετική ιδιοτιμή και έπειτα διαχωρίζει τις κορυφές σε δύο κοινότητες σύμφωνα με το πρόσημο της αντίστοιχης συνιστώσας του ιδιοδιανύσματος. Αν όλες οι συνιστώσες του ιδιοδιανύσματος έχουν το ίδιο πρόσημο, σημαίνει οτι το δίκτυο δεν έχει κοινοτική δομή. Στην ανάλυση που θα ακολουθήσει θα αναφερόμαστε στον αλγόριθμο αυτό ως LE (Leading Eigenvector).
7. *Ο αλγόριθμος Walktrap:* Ο αλγόριθμος αυτός προσπαθεί να βρεί πυκνά συνεκτικούς υπογράφους-κοινότητες στο δίκτυο σύμφωνα με τυχαίους περιπάτους σε αυτό. Η ιδέα είναι οτι σύντομοι τυχαίοι περίπατοι τείνουν να παραμένουν στην ίδια κοινότητα.
8. *Ο αλγόριθμος MCL:* ο Markov Cluster αλγόριθμος είναι μια μέθοδος εντοπισμού κοινοτήτων σε δίκτυα που προσομειώνει τυχαίους περιπάτους στον $\mathrm{n} \times \mathrm{n}$ πίνακα γειτνίασης του δικτύου. Εναλλάσσει ένα βήμα επέκτασης και ένα βήμα εμφύσησης εως ότου φτάσει σε μια κατάσταση ισορροπίας.
9. *Ο δυναμικός αλγόριθμος των Rosvall and Bergstrom-Infomap (Dynamic algorithm by Rosvall and Bergstrom)*: Εδώ το πρόβλημα εύρεσης της καλύτερης δομής των συστάδων στο δίκτυο, μετατρέπεται σε ένα πρόβλημα βέλτιστης συμπίεσης μιας

δυναμικής διαδικασίας που λαμβάνει χώρο στο γράφο, δηλαδή έναν τυχαίο περίπατο. Στην περίπτωση αυτή η βέλτιστη συμπίεση γίνεται μέσω βελτιστοποίησης μιας συνάρτησης ποιότητας, που ονομάζεται ελάχιστο μήκος περιγραφής (minimum description length) του τυχαίου περιπάτου. Τέτοια βελτιστοποίηση μπορεί να γίνει πολύ γρήγορα με έναν συνδυασμό λαίμαργης αναζήτησης και προσομειωμένης ανόπτησης. Στην ανάλυση που θα ακολουθήσει θα αναφερόμαστε στον αλγόριθμο αυτό ως Infomap.

6.4 Παραγωγή του LFR Δικτύου Αναφοράς.

Όπως αναφέραμε, το software για παραγωγή διαφόρων τύπων των LFR δικτύων είναι διαθέσιμο στον παγκόσμιο ιστό *(https://sites.google.com/site/santofortunato/inthepress2*) ωστόσο αυτά έχουν παραχθεί απο τον συγγραφέα, Santo Fortunato, γραμμένα στη γλώσσα C++. Στην εφαρμογή μας, χρησιμοποιήσαμε το Ubuntu, ενα λειτουργικό σύστημα βασισμένο στο GNU/Linux λογισμικό, δηλαδή μια διανομή του Linux. Εγκαταστήσαμε μια εφαρμογή του στα Windows 10, η οποία βασίζεται στο λειτουργικό σύστημα Ubuntu 16.04 και μέσω αυτού παρηγάγαμε LFR δίκτυα αναφοράς, δύο κατηγοριών:

1. Μη κατευθυνόμενα και μη σταθμισμένα LFR δίκτυα αναφοράς.
2. Μη κατευθυνόμενα και σταθμισμένα LFR δίκτυα αναφοράς.

Στον κώδικα που χρησιμοποιήσαμε (παράρτημα 2) οι συμβολισμοί σημαίνουν τα εξής:

-N**:** number of nodes: ο αριθμός των κόμβων του δικτύου.

-k**:** average degree: ο μέσος όρος του βαθμού των κόμβων του δικτύου, ουσιαστικά δηλαδή ο μέσος όρος του αθροίσματος των εξωτερικών, k_{out} και εσωτερικών, k_{in} βαθμών των κόμβων.

-maxk (maximum degree): ο μέγιστος βαθμός των κόμβων του δικτύου.

-mu (mixing parameter μ): Κάθε κόμβος μοιράζεται ένα μέρος, $1-\mu$, των ακμών του με άλλους κόμβους της κοινότητας του και ένα μέρος μ, των ακμών του, με κόμβους των άλλων κοινοτήτων, με $0 \leq \mu \leq 1$.

-minc (minimum for the community sizes): Ο μικρότερος δυνατός αριθμός κόμβων σε κάθε κοινότητα.

-maxc (maximum for the community sizes): Ο μεγαλύτερος δυνατός αριθμός κόμβων σε κάθε κοινότητα.

Στην κατηγορία των μη κατευθυνόμενων και μη σταθμισμένων LFR δικτύων αναφοράς έχουμε παράγει, κρατώντας σταθερές τις παραμέτρους $N = 1000, k = 15,\ maxk = 50,$ LFR δίκτυα για όλες τις τιμές της παραμέτρου μίξης, δηλαδή μ $= 0.1, 0.2 \ldots .0.9$ (στον παραπάνω κώδικα $\mu = mu = 0.1$) για μικρές $(minc = 20, maxc = 50)$ αλλά και μεγάλες $(minc = 20, maxc = 100)$ κοινότητες αντίστοιχα και έχουμε εξετάσει την συμπεριφορά των αλγορίθμων που προαναφέρθηκαν για τις διάφορες τιμές της παραμέτρου μίξης.

Στη δεύτερη κατηγορία LFR δικτύων αναφοράς, παρηγάγαμε μη κατευθυνόμενα και σταθμισμένα LFR δίκτυα αναφοράς και η μορφή του κώδικα που χρησιμοποιήσαμε για συγκεκριμένες τιμές των παραμέτρων δίνεται στο παράρτημα 3.

Στην περίπτωση σταθμισμένων LFR δικτύων αναφοράς, προστίθενται οι εξής παράμετροι:

-muw (mixing parameter for the weights μ_w): η αντίστοιχη παράμετρος της $\mu = \mu_t$ παραμέτρου μίξης για σταθμισμένα δίκτυα. Στα σταθμισμένα δίκτυα έχουμε δύο παραμέτρους μίξης. Την τοπολογική παράμετρο μίξης, μ_t (mu_t) που ορίσαμε πριν και την αντίστοιχή σταθμισμένη της, μ_w (mu_w) η οποία εκφράζει το λόγο (fraction) της δύναμης ενός κόμβου που βρίσκεται στις ακμές που συνδέεουν τον κόμβο αυτό με κόμβους έξω απο την κοινότητά του, σε σχέση με τη συνολική του δύναμη. Η δύναμη ενός κόμβου είναι το άθροισμα των βαρών των ακμών του.

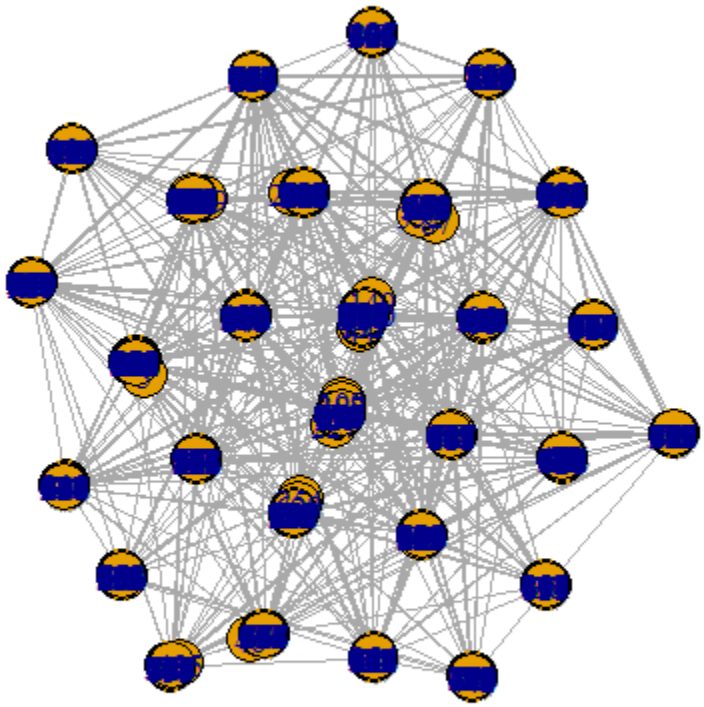
-beta (exponent for the weight distribution): ο εκθέτης της κατανομής της δύναμης των κόμβων. Τον έχουμε θέσει σταθερά στην προκαθορισμένη τιμή του, 1.5, για όλες τα σταθμισμένα LFR δίκτυα που παρηγάγαμε.

Για να μπορέσουμε να εφαρμόσουμε τους αλγορίθμους που αναφέραμε, τα network.dat αρχεία που παρήχθεισαν μέσω του Ubuntu, τα περάσαμε στην R ως data.frames για όλες τις τιμές των παραμέτρων μίξης. Ωστόσο, τα δίκτυα στην αρχική τους μορφή περιείχαν πολλαπλές ακμές. Δηλαδή κατά την ανάγνωση του δικτύου απο την R, η ακμή που συνέδεε τον κόμβο i με τον κόμβο

j, διαβαζόταν δύο φορές, μια φορά στη γραμμή που διαβάζεται ο κόμβος i και μια φορά στη γραμμή που διαβάζεται ο κόμβος j. Ο κώδικας που φαίνεται στο παράρτημα 4, είναι η τροποποίηση που εφαρμόσαμε στην αρχική μορφή του δικτύου, ωστε να καταλαβαίνει η R οτι θα χρησιμοποιήσουμε δίκτυο χωρίς πολλαπλές ακμές και να είμαστε σε θέση να εφαρμόσουμε τους αλγορίθμους χωρίς κανένα πρόβλημα. Η βιβλιοθήκη της R που χρησιμοποιήσαμε για την εν λόγω διαδικασία είναι η βιβλιοθήκη igraph.

Μια αναπαράσταση του μη κατευθυνόμενου-μη σταθμισμένου δικτύου $N = 1000$ κόμβων και για τιμές των παραμέτρων αυτές που προαναφέρθηκαν και παράμετρο μίξης $\mu = mu = 0.1$, απο τα γραφικά της R, δίνεται όπως παρακάτω.

Σχήμα 24. Μη κατευθυνόμενο και μη σταθμισμένο LFR δίκτυο αναφοράς, N=1000 κόμβων για k=15, maxk=50, minc=20, maxc=50 και mut=0.1



Κατά την κατασκευή των LFR δικτύων μέσω του Ubuntu για κάθε τιμή της παραμέτρου μίξης, εκτός απο το δίκτυο έχουμε και ένα community.dat αρχείο 1000 γραμμών και δύο στηλών,

το οποίο μας δείχνει πόσες κοινότητες υπάρχουν στο δίκτυο και σε ποια απο αυτές ανήκει καθένας απο τους 1000 κόμβους του δικτύου.

Η μορφή του στην R console δίνεται όπως παρακάτω (στην αναπαράσταση αυτή απο τους 1000 του δικτύου, έχουμε συμπεριλάβει την εμφάνιση των 12 πρώτων κόμβων και της κοινότητας στην οποία ανήκουν):

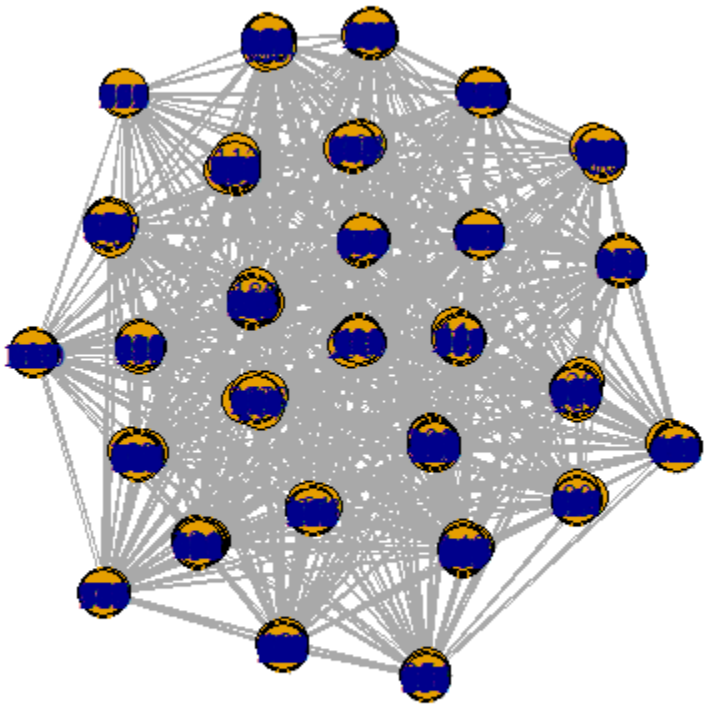
```
    V1 V2
1    1  7
2    2  8
3    3 11
4    4 21
5    5 18
6    6 20
7    7 26
8    8 19
9    9 28
10  10  2
11  11 24
12  12 16 …….
```

Το αρχείο αυτό, στην μορφή πίνακα 1000 γραμμών και 2 στηλών που βλέπουμε, θα είναι και το αρχείο με το οποίο θα συγκρίνουμε τα αντίστοιχα αποτελέσματα του εκάστοτε αλγορίθμου για να πάρουμε τις διάφορες τιμές του κριτηρίου NMI για όλες τις πιθανές τιμές της παραμέτρου μίξης, $\mu = 0.1, 0.2 \dots 0.9$. Στο σύνολο λοιπόν κατασκευάσαμε 9 μη κατευθυνόμενα και μη σταθμισμένα LFR δίκτυα αναφοράς για μεγάλες κοινότες ($minc = 20, maxc = 100$) και 9 για μικρές κοινότητες ($minc = 20, maxc = 50$), ένα για κάθε τιμή της παραμέτρου μίξης μ, κρατώντας σταθερές τις υπόλοιπες παραμέτρους $k = 15,\ maxk = 50$.

Στην περίπτωση των μη κατευθυνόμενων και σταθμισμένων LFR δικτύων αναφοράς ο κώδικας που χρησιμοποιήσαμε στην R για την ανάγνωσή τους είναι σχεδόν ο ίδιος, με τη μόνη διαφορά οτι προσθέσαμε έναν μετρητή για τα βάρη των ακμών. Κατασκευάσαμε 9 μη κατευθυνόμενα και σταθμισμένα LFR δίκτυα αναφοράς για μεγάλες κοινότες ($minc = 20, maxc = 100$) και 9 για μικρές κοινότητες ($minc = 20, maxc = 50$), ένα για κάθε τιμή της παραμέτρου μίξης των βαρών, $\mu_w = 0.1, 0.2, \dots 0.9$, κρατώντας σταθερές τις υπόλοιπες παραμέτρους, $\mu_t = mu_t = 0.5,\ k = 15,\ maxk = 50,\ beta = 1.5$. Η μορφή του κώδικα που χρησιμοποιήσαμε δίνεται στο παράρτημα 5.

Μια αναπαράσταση του μη κατευθυνόμενου και σταθμισμένου LFR δικτύου αναφοράς 1000 κόμβων, μεγάλων κοινοτήτων ($minc = 20, maxc = 100$) και για τιμές των παραμέτρων αυτές που προαναφέρθηκαν, με παράμετρο μίξης $\mu_w = mu_w = 0.1$, απο τα γραφικά της R, δίνεται όπως παρακάτω.

Σχήμα 25. Μη κατευθυνόμενο και σταθμισμένο LFR δίκτυο αναφοράς, N=1000 κόμβων για k=15, maxk=50, minc=20, maxc=100, mut=0.5, muw=0.1, beta=1.5



6.5 Εφαρμογές των αλγορίθμων.

Εφαρμόσαμε τους προαναφερθέντες αλγορίθμους για τις δύο κατηγορίες LFR δικτύων αναφοράς και τα αποτελέσματα του εκάστοτε αλγορίθμου για κάθε μια απο τις τιμές της παραμέτρου μίξης τα χρησιμοποιήσαμε για να δημιουργήσουμε τις τιμές του κριτηρίου NMI. Θεωρώντας ένα σύστημα αξόνων $x - y$, στο οποίο ο άξονας x περιέχει τις τιμές της παραμέτρου μίξης $\mu_t = mu_t$, για μη κατευθυνόμενα και μη σταθμισμένα LFR δίκτυα αναφοράς και $\mu_w = mu_w$, για μη κατευθυνόμενα και σταθμισμένα LFR δίκτυα αναφοράς και ο άξονας y τις τιμές του

κριτηρίου NMI, δημιουργήσαμε μια γραφική αναπαράσταση της συμπεριφοράς του κάθε αλγορίθμου για όλες τις δυνατές τιμές της παραμέτρου μίξης. Τα outputs του κάθε αλγορίθμου, δηλαδή ο αριθμός των κοινοτήτων, η τιμή της modularity (σε περίπτωση που αυτή υπολογίζεται), σε ποια κοινότητα ανήκει η κάθε κορυφή δίνονται σαν outputs σε διαφορετική μορφή για τον κάθε αλγόριθμο. Για το λόγο αυτό, τροποποιήσαμε τα αποτελέσματα, τοποθετώντας τα σε ένα table 1000 γραμμών και 2 στηλών σε αντίστοιχη μορφή με το community.dat αρχείο που αναφέραμε, ώστε να είναι σε θέση να γίνει η σύγκριση με αυτό και να υπολογιστούν οι τιμές του κριτηρίου NMI.

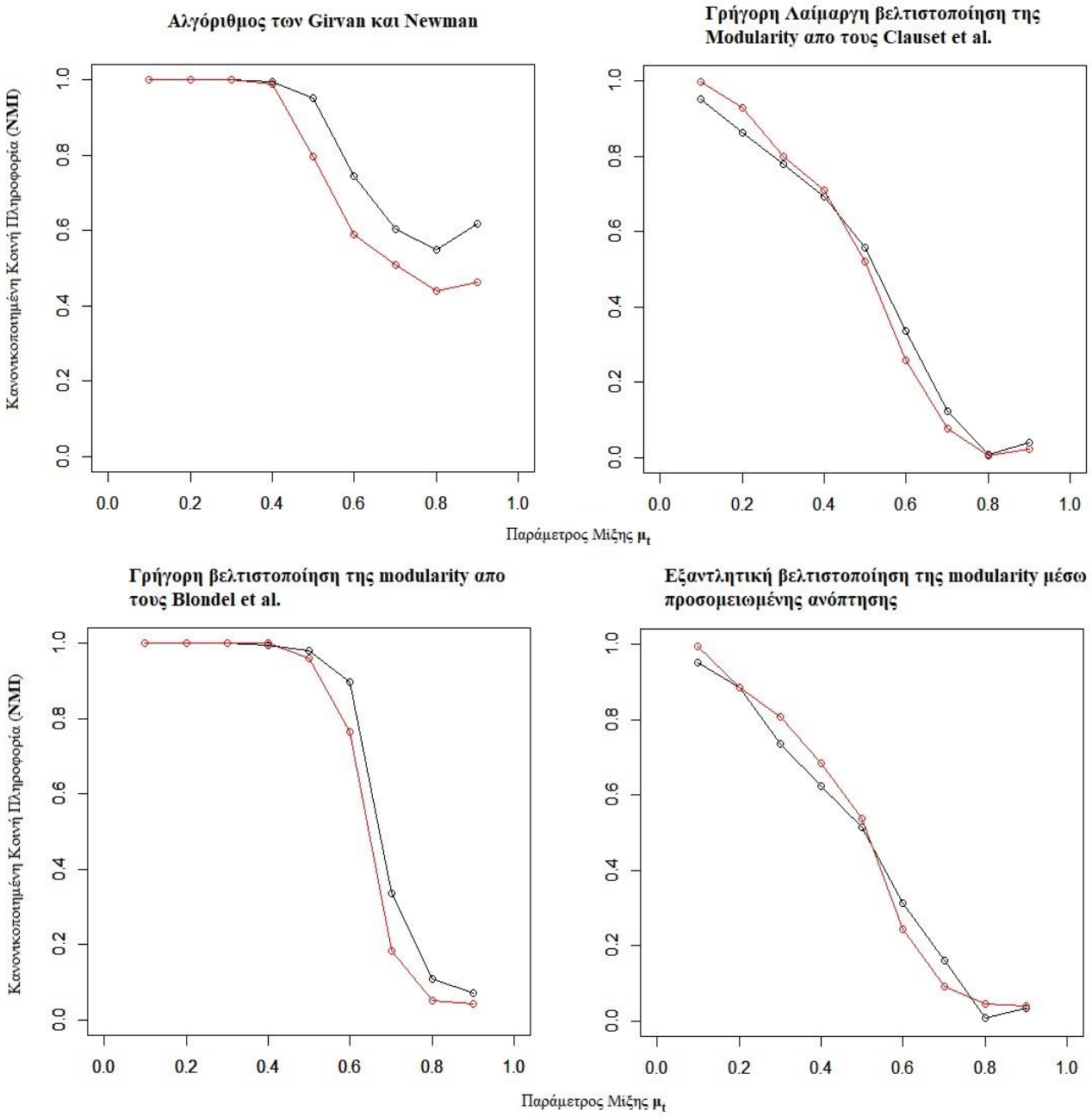
6.5.1 Εφαρμογές των αλγορίθμων για μη κατευθυνόμενα και μη σταθμισμένα LFR Δίκτυα αναφοράς

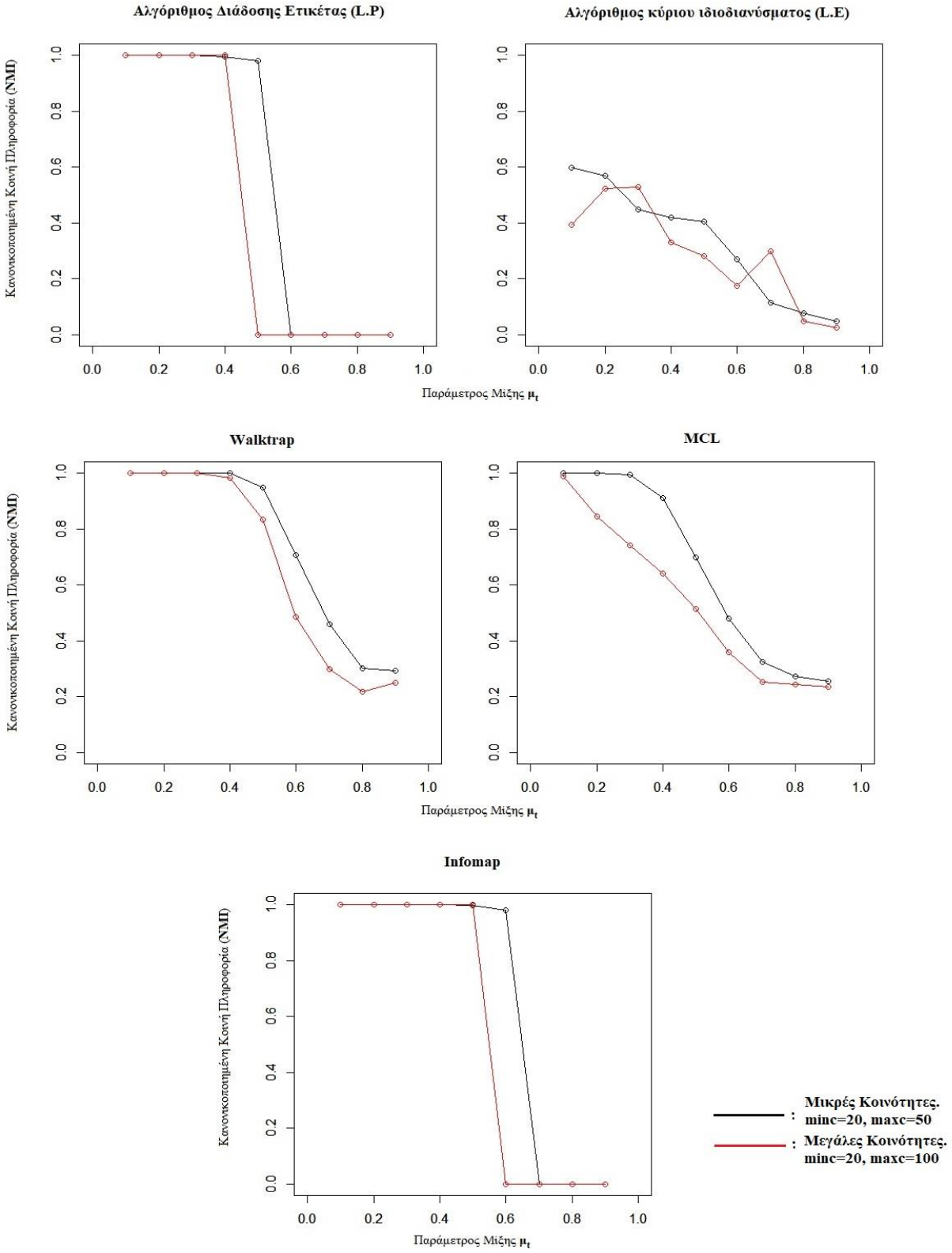
Στο παράρτημα 6 παρατίθεται ο κώδικας που χρησιμοποιήσαμε στην R για την εφαρμογή των 9 αλγορίθμων που χρησιμοποιήσαμε, ο οποίος περιέχει και την αντίστοιχη τροποποίηση των outputs σε table της μορφής του community.dat αρχείου του εκάστοτε δικτύου και τον υπολογισμό των τιμών του κριτηρίου NMI:

Στο παράρτημα 7 παρατίθεται ο κώδικας που χρησιμοποιήσαμε στην R για την κατασκευή του συστήματος αξόνων $x - y$, στο οποίο έχουμε μια γραφική απεικόνιση της συμπεριφοράς του εκάστοτε αλγορίθμου (τιμές κριτηρίου NMI) για τις διάφορες τιμές της παραμέτρου μίξης, για τα μη κατευθυνόμενα και μη σταθμισμένα LFR δίκτυα αναφοράς $N = 1000$ κόμβων, που παρηγάγαμε.

Στην επόμενη σελίδα παραθέτουμε τα γραφήματα των τιμών του κριτηρίου NMI για τον κάθε αλγόριθμο ξεχωριστά.

Σχήμα 26. Τιμές κριτηρίου NMI των 9 αλγορίθμων για μη κατευθυνόμενα και μη σταθμισμένα LFR Δίκτυα αναφοράς μικρών και μεγάλων κοινοτήτων.

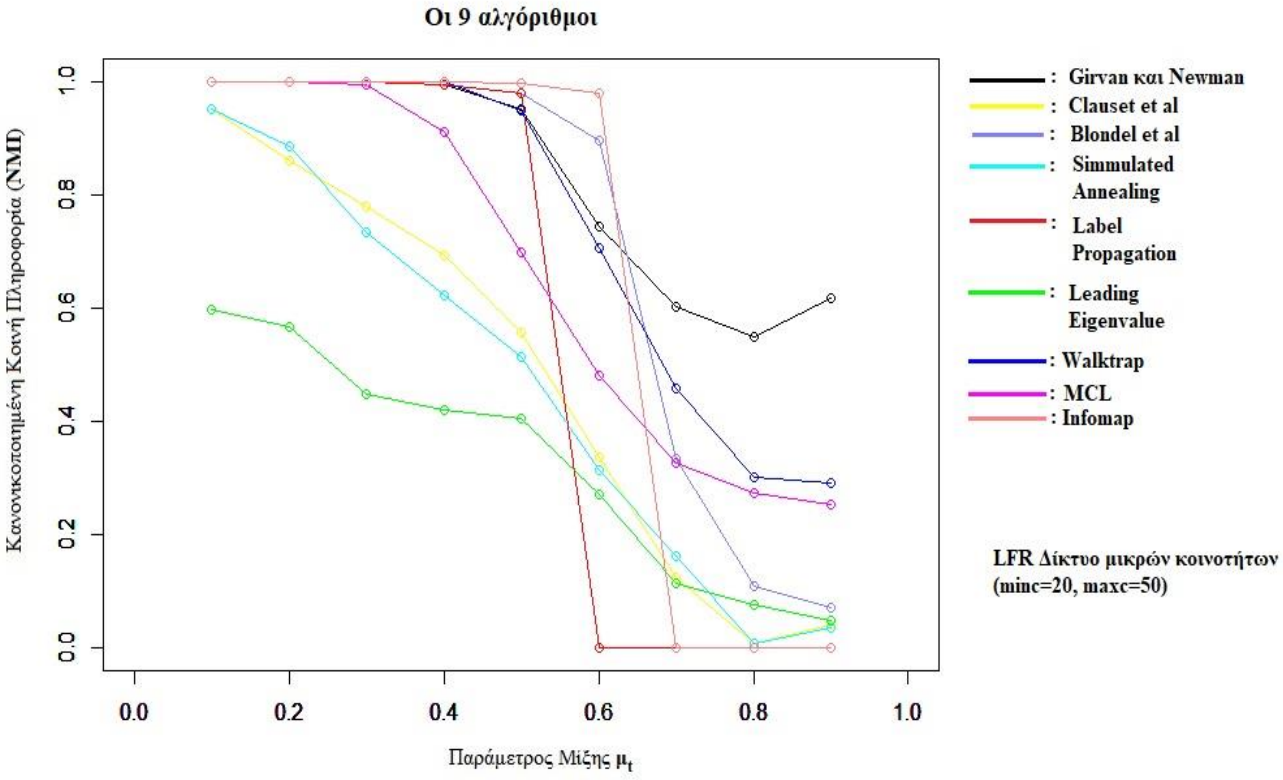



Αλγόριθμος Διάδοσης Ετικέτας (L.P)
Αλγόριθμος κύριου ιδιοδιανύσματος (L.E)
Walktrap
MCL
Infomap
Κανονικοποιημένη Κοινή Πληροφορία (NMI)
Παράμετρος Μίξης μ_t
Μικρές Κοινότητες. minc=20, maxc=50
Μεγάλες Κοινότητες. minc=20, maxc=100

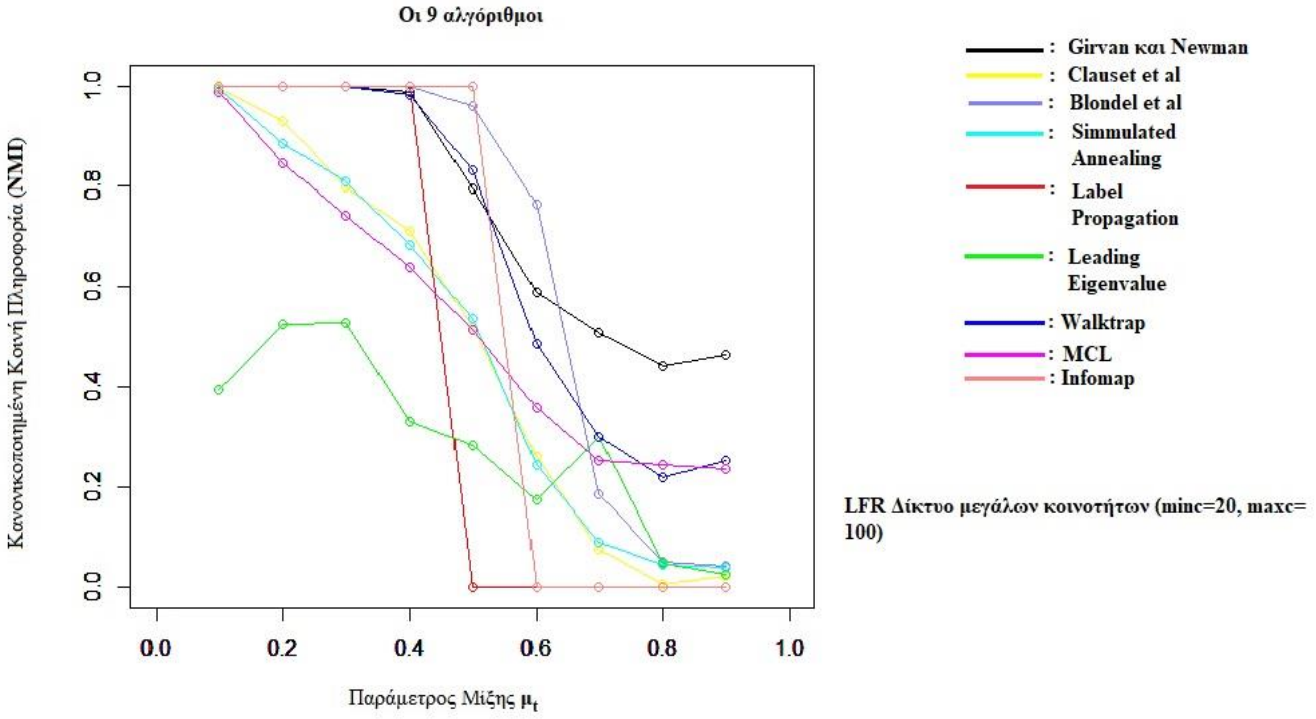
Στα εν λόγω σχήματα με μαύρο χρώμα συμβολίζουμε τις τιμές του κριτηρίου NMI για τον κάθε αλγόριθμο εφαρμοσμένο σε LFR δίκτυα αναφοράς μικρών κοινοτήτων ($minc = 20, maxc = 50$) και με κόκκινο χρώμα τις τιμές του κριτηρίου NMI για τον κάθε αλγόριθμο εφαρμοσμένο σε LFR δίκτυα αναφοράς μεγάλων κοινοτήτων ($minc = 20, maxc = 100$), δηλαδή μεταξύ 20 και 100 κόμβων. Από τις μεθόδους που στηρίζονται στη modularity, παρατηρούμε οτι τις πιο εντυπωσιακές αποδόσεις έχει ο αλγόριθμος της γρήγορης βελτιστοποίησής της, των Blondel et al. Ο αλγόριθμος των Girvan και Newman εντοπίζει και αυτός άριστα τις κοινότητες του δικτύου αλλά για μικρότερο εύρος τιμών της τοπολογικής παραμέτρου μίξης συγκριτικά με τον αλγόριθμο των Blondel et al. Ο Infomap διατηρεί άριστες τιμές του κριτηρίου NMI παραμένοντας ανεπηρέαστος απο την αύξηση της τιμής της τοπολογικής παραμέτρου μίξης μ, μέχρι $\mu = 0.5$ για LFR δίκτυα μεγάλων κοινοτήτων και $\mu = 0.6$ για LFR δίκτυα μικρών κοινοτήτων, αποτυγχάνοντας ωστόσο να εντοπίσει κοινότητες για μεγαλύτερες τιμές της τοπολογικής παραμέτρου μίξης. Απο τις μεθόδους που χρησιμοποιούν την έννοια του τυχαίου περιπάτου, και ο Walktrap και ο MCL έχουν πολύ καλές τιμές κριτηρίου NMI επίσης, αλλά για μικρότερο εύρος τιμών της τοπολογικής παραμέτρου μίξης, συγκριτικά με τον Infomap. Παρατηρούμε επίσης και οι 2 αλγόριθμοι εντοπίζουν καλύτερα μικρές κοινότητες, παρά μεγάλες καθώς υπάρχει αισθητή διαφορά στο γράφημα των τιμών NMI ανάμεσα σε μικρές και μεγάλες κοινότητες αντίστοιχα. Το διάγραμμα των τιμών του κριτηρίου NMI για τον αλγόριθμο LP (Label Propagation) θυμίζει τον Infomap, εντοπίζοντας πλήρως τις κοινότητες του δικτύου για εύρος τιμών παραμέτρου μίξης $\mu = 0.1 - 0.4$ για LFR δίκτυα αναφοράς μεγάλων και $\mu = 0.1 - 0.5$ μικρών κοινοτήτων. Οι τιμές του κριτηρίου NMI για τον αλγόριθμο εξαντλητικής βελτίωσης της modularity μέσω προσομειωμένης ανόπτησης φθίνουν με σταθερό τρόπο όσο αυξάνεται η τιμή της παραμέτρου μίξης, εντοπίζοντας ικανοποιητικά τις κοινότητες του δικτύου (τιμή NMI > 0.8) μόνο για $\mu = 0.1, 0.2$. Τη χειρότερη απόδοση απο τους 9 αλγορίθμους έχει ο Leading Eigenvalue αλγόριθμος του Newman, καθώς η μεγαλύτερη τιμή του κριτηρίου NMI που επιτυγχάνει είναι 0.597 για $\mu = 0.1$. Εν κατακλείδι, απο τους 9 αλγορίθμους που χρησιμοποιήσαμε στην ανάλυσή μας σε μη κατευθυνόμενα και μη σταθμισμένα LFR δίκτυα αναφοράς , οι δύο καλύτεροι σύμφωνα με τις NMI τιμές τους για όλες τις δυνατές τιμές της τοπολογικής παραμέτρου μίξης μ, είναι ο αλγόριθμος των Blondel et al. και ο Infomap.

Παρακάτω παραθέτουμε τους 9 αλγορίθμους που χρησιμοποιήσαμε σε κοινό γράφημα για LFR δίκτυα αναφοράς μικρών και μεγάλων κοινοτήτων αντίστοιχα.

Σχήμα 27. NMI τιμές των 9 αλγορίθμων μαζί για μη κατευθυνόμενα και μη σταθμισμένα LFR δίκτυα αναφοράς, μικρών κοινοτήτων.



Σχήμα 28. NMI τιμές των 9 αλγορίθμων για μη κατευθυνόμενα και μη σταθμισμένα LFR δίκτυα αναφοράς, μεγάλων κοινοτήτων.

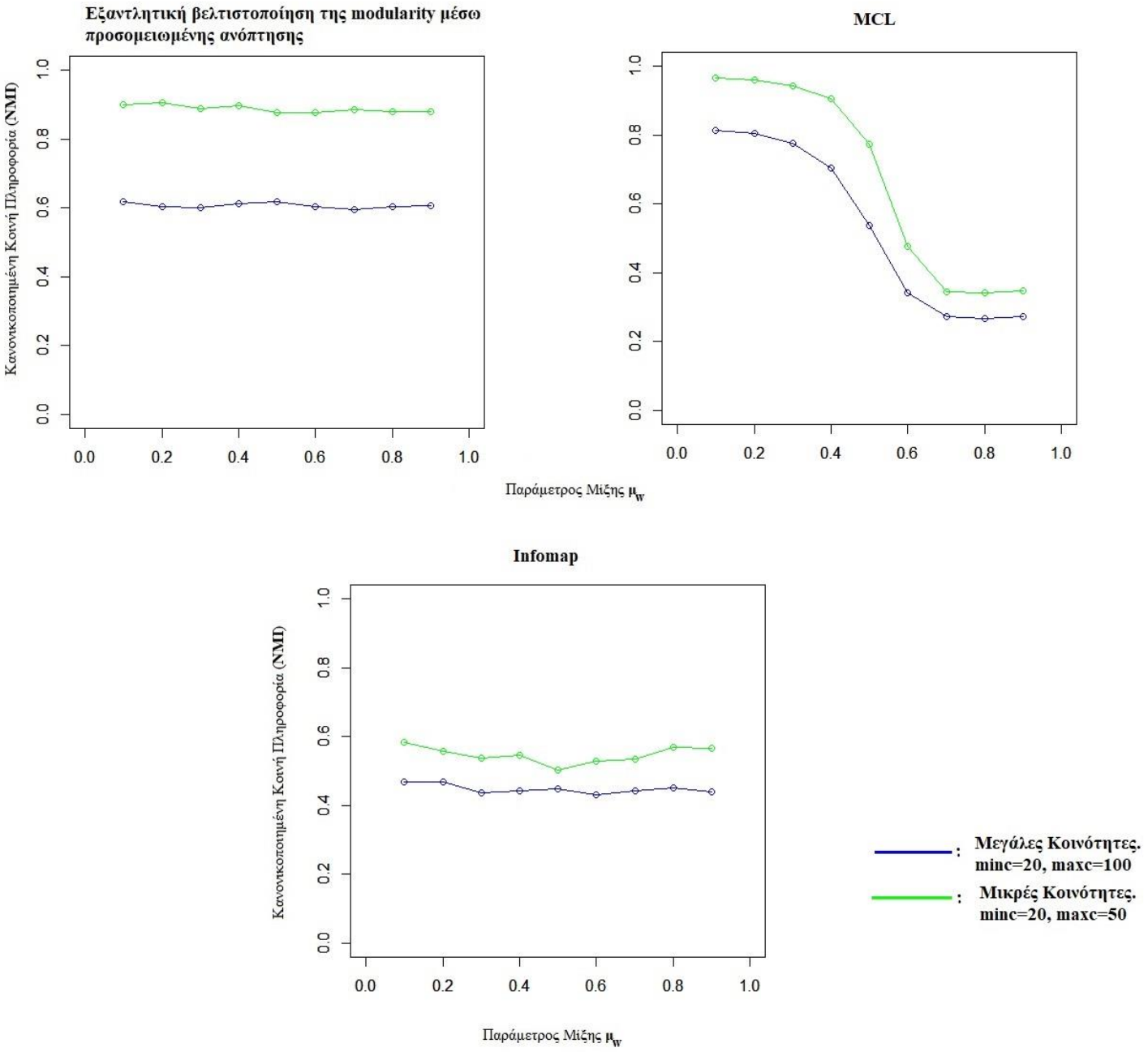


6.5.2 Εφαρμογές των αλγορίθμων για μη κατευθυνόμενα και σταθμισμένα LFR δίκτυα αναφοράς

Στο παράρτημα 8 εμφανίζεται ο κώδικας που χρησιμοποιήσαμε στην R για τους 3 αλγορίθμους που χρησιμοποιήσαμε αυτή τη φορά. Ανάμεσα στους προαναφερθέντες 9 αλγορίθμους, χρησιμοποιήσαμε αυτούς για τους οποίους υπήρχε διαθέσιμη, η αντίστοιχη τροποποίηση στον κώδικα για το σταθμισμένο τους κομμάτι. Ο κώδικας επίσης περιέχει και την αντίστοιχη τροποποίηση των outputs σε table της μορφής του community.dat αρχείου του εκάστοτε δικτύου και την κατασκευή του συστήματος αξόνων $\mathrm{x}-y$, στο οποίο έχουμε τη γραφική απεικόνιση της συμπεριφοράς του εκάστοτε αλγορίθμου (τιμές κριτηρίου NMI) για τις διάφορες τιμές της παραμέτρου μίξης, για τα μη κατευθυνόμενα και σταθμισμένα LFR δίκτυα αναφοράς $N = 1000$ κόμβων, που παρηγάγαμε, για όλες τις τιμές της παραμέτρου μίξης των βαρών, μ_w.

Στο σημείο αυτό παραθέσουμε τα γραφήματα των NMI scores για τον κάθε αλγόριθμο ξεχωριστά, για διάφορες τιμές της παραμέτρου μίξης των βαρών, μ_w αυτή τη φορά, κρατώντας σταθερή την τοπολογική παράμετρο μίξης $\mu_t = 0.5$ για όλα τα μη κατευθυνόμενα και σταθμισμένα LFR δίκτυα αναφοράς $N = 1000$ κόμβων, που παρηγάγαμε .

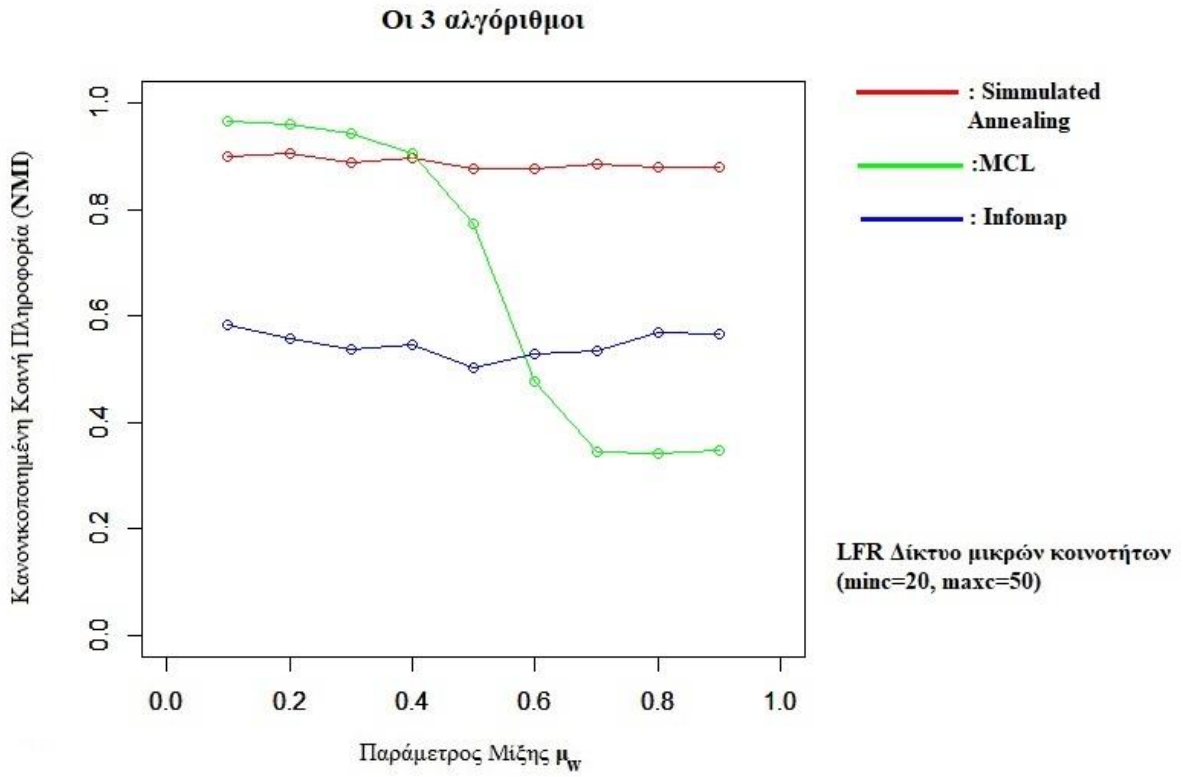
Σχήμα 29. Τιμές κριτηρίου NMI των 3 αλγορίθμων για μη κατευθυνόμενα και σταθμισμένα LFR Δίκτυα αναφοράς μικρών και μεγάλων κοινοτήτων.



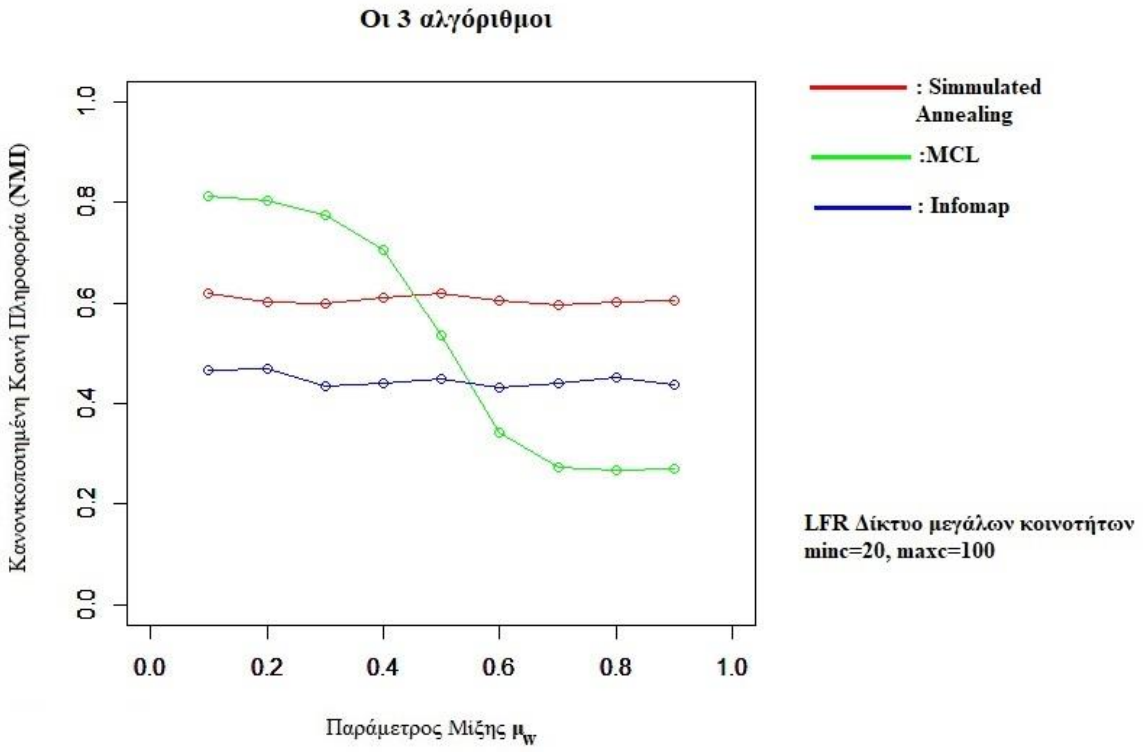
Παρατηρούμε οτι και οι 3 αλγόριθμοι που χρησιμοποιήσαμε εντοπίζουν πολύ καλύτερα μικρές παρά μεγάλες κοινότητες. Τις υψηλότερες τιμές του κριτηρίου NMI τις έχει ο MCL για εύρος παράμετρου μίξης των βαρών, $\mu_w = 0.1 - 0.4$. Ο αλγόριθμος εξαντλητικής βελτιστοποίησης της modularity μέσω προσομειωμένης ανόπτησης πετυχαίνει επίσης πολύ υψηλές τιμές του κριτηρίου NMI για όλο το εύρος των τιμών της παραμέτρου μίξης των βαρών, αλλά με πολύ υψηλότερες τιμές (NMI> 0.8) για σταθμισμένα LFR δίκτυα αναφοράς μικρών κοινοτήτων σε σχέση με τις τιμές του για τα αντίστοιχα δίκτυα μεγάλων κοινοτήτων. Απο τους 3 αλγόριθμους ο Infomap για τα συγκεκριμένα LFR δίκτυα αναφοράς, έχει τις χειρότερες τιμές του κριτηρίου NMI, με υψηλότερη τιμή 0.582 για $\mu_w = 0.1$ για σταθμισμένα LFR δίκτυα αναφοράς μικρών κοινοτήτων και υψηλότερη τιμή 0.467 για $\mu_w = 0.1$ για σταθμισμένα LFR δίκτυα αναφοράς μεγάλων κοινοτήτων.

Παρακάτω παραθέτουμε τους 3 αλγορίθμους που χρησιμοποιήσαμε σε κοινό γράφημα για LFR δίκτυα αναφοράς μικρών και μεγάλων κοινοτήτων αντίστοιχα.

Σχήμα 30. NMI τιμές των 3 αλγορίθμων μαζί για μη κατευθυνόμενα και σταθμισμένα LFR δίκτυα αναφοράς μικρών κοινοτήτων.



Σχήμα 31. NMI τιμές των 3 αλγορίθμων μαζί για μη κατευθυνόμενα και σταθμισμένα LFR δίκτυα αναφοράς μεγάλων κοινοτήτων.



Βιβλιογραφία

Lancichinetti, A., Radicchi, F., Ramasco, J., & Fortunato, S. (2011). *Finding statistically significant communities in networks.* PLoS ONE.

Adamcsek, B., Palla, G., Farkas, I., Derényi, I., & Vicsek, T. (2006). *Locating cliques and overlapping modules in biological networks.* Bioinformatics, 22:1021–1023.

Agarwal, G., & Kempe, D. (2008). *Modularity-Maximizing Graph Communities via Mathematical Programming.* Los Angeles.

Alba, R. D. (1973). *J. Math. Sociol. 3, 113.*

Arenas, A., Duch, J., Fernandez, A., & Gomez, S. (2007). *Size reduction of complex networks preserving modularity.* New Journal of Physics 9 (2007) 176.

Arora, S., Rao, S., & Vazirani, U. (2008). *Geometry, Flows, and Graph Partitioning Algorithms.* Communications of ACM.

Arora, S., Rao, S., & Vazirani, U. (2008). *Geometry, Flows, and. Graph-Partitioning Algorithms.* New York: Communications of the ACM.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). *Fast unfolding of communities in large networks.* J. Stat. Mech. (2008) P10008.

Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z., & Wagner, D. (2006). *Maximizing Modularity is hard.*

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A practical Information-Theoretic approach.* New York, USA: Springer.

Chakrabarti, D. (2004). *AutoPart: Parameter-Free Graph Partitioning and Outlier Detection.* PKDD, volume 3202 of Lecture Notes in Computer Science, page 112-124. Springer.

Clauset, A., Newman, M. E., & Moore, C. (2004). *Finding community structure in very large networks.* Phys. Rev. E 70, 066111.

Coscia, M., Giannotti, F., & Pedreschi, D. (2011). *A Classification for Community Discovery Methods in Complex Networks.* Stat. Anal. Data Min. 4 (5) 512-546.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. pp.

Dongen, S. v. (2000). *Graph Clustering by Flow Simulation, Ph.D. thesis.* Utrecht, Netherlands: Dutch National Research Institute for Mathematics and Computer Science, University of Utrecht.

Doreian, P., Batagelj, V., & Ferligoj, A. (2005). *Generalized Blockmodeling.* New York, USA: Cambridge University Press.

Farkas, I. J., Abel, D., Palla, G., & Vicsek, T. (2007). *Weighted network modules.* New J. Phys. 9, 180.

Fortunato, S. (2010). *Community detection in graphs.* Torino: Physics Reports 486, 75-174 .

Fortunato, S., & Barthélemy, M. (2007). *Resolution limit in community detection.* Proc. Natl. Acad. Sci. USA 104, 36.

Fortunato, S., & Hric, D. (2016). *Community detection in networks: A user guide.* Physics Reports 659, 1-44 .

Fortunato, S., & Hric, D. (2016). *Community detection in networks: A user guide.* Indiana University, Bloomington, USA.

Freeman, L. C. (1977). *A Set of Measures of Centrality Based on Betweenness.* Sociometry 40(1):35-41.

Golub, G. H., & Van Loan, C. F. (1989). *Matrix Computations.* Baltimore, USA: John Hopkins University Press.

Good, B. H., Montjoye, Y.-A. d., & Clauset, A. (2009). *The performance of modularity maximization in practical contexts.* Phys. Rev. E 81, 046106 (2010).

Guimerà, R., & Amaral, N. L. (2005). *Functional cartography of complex metabolic networks.* Nature 433, 895.

Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., . . . Samatova, N. (2014). *Community detection in large-scale networks: a survey and empirical evaluation.* Wiley Interdisciplinary Reviews: Computational Statistics, 6 (6), 426-439.

Hastings, M. B. (2006). *Community detection as an inference problem.* Phys. Rev. E 74, 035102.

Havemann, F., Heinz, M., & Struck, A. (2011). *Identification of Overlapping Communities by Locally Calculating Community-Changing Resolution Levels.* Berlin, Germany.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). *Stochastic Blockmodels: First Steps.* Social Networks 5 (1983) lO9- 137 .

Hu, Chen, Zhang, Li, Di, & Fan. (2008). *Comparative definition of community and corresponding identifying algorithm.* Phys. Rev. E 78(2), 026121.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). *Optimization by Simulated Annealing.* Science, New Series, Vol. 220, No. 4598., pp. 671-680.

Krishnamurthy, & Wang. (2000). *On Network-Aware Clustering of Web Clients.*

Kumpula, J. M., Kivela, M., Kaski, K., & Saramaki, J. (2008). *A sequential algorithm for fast clique percolation.* Phys. Rev. E 78(2), 026109.

Lancichinetti, A., & Fortunato, S. (2009). *Community detection algorithms: a comparative analysis.* Torino, Italy: Physical Review E 80, 056117.

Lancichinetti, A., Fortunato, S., & Kertesz, J. (2009). *Detecting the overlapping and hierarchical community structure of complex networks.* New Journal of Physics 11, 033015.

Lehmann, S., Schwartz, M., & Hansen, L. (2008). *Biclique communities.* Phys. Rev. E 78(1), 016108.

Lorrain, F., & White, H. C. (1971). *Structural equivalence of individuals in social network.* The Journal of Mathematical Sociology,1:1, 49-80.

Luccio, F., & Sami. (1969). *IEEE Trans. Circuit Th. CT 16, 184.*

Luce, R. D. (1950). *Psychometrika 15(2), 169.*

Lusseau, D. (2003). *The emergent properties of a dolphin social network.* Proc. Royal Soc. London B 270, S186.

Luxburg, U. v. (2006). *A tutorial on spectral clustering.* Technical Report 149, Max Planck Institute for Biological Cybernetics.

Mackay, D. J. (2003). *Information Theory, Inference and Learning Algorithms.* Cambridge, UK: Cambridge University Press.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations.* Berkeley, California.: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297.

Malliaros, F., & Vazirgiannis, M. (2013). *Clustering and Community Detection in Directed Networks.* Physics Reports 533, 95-142.

Mancoridis, S., Mitchell, Rorres, C., Chen, Y., & Gansner. (1998). *IWPC '98: Proceedings of the 6th International Workshop on Program Comprehension.* Washington, DC, USA.

Mokken, R. J. (1979). *Qual. Quant. 13(2), 161.*

Newman, M. E. (2006). *Finding community structure in networks using the eigenvectors of matrices.* Phys. Rev. E 74, 036104.

Newman, M. E., & Girvan, M. (2004). *Finding and evaluating community structure in networks.* Michigan, USA: Phys. Rev. E 69(2), 026113.

Newman, M. E., & Leicht, E. A. (2007). *Mixture Models and Exploratory analysis in Networks.* Proc. Natl. Acad. Sci. USA 104, 9564.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). *On Spectral Clustering: Analysis and an algorithm.* Cambridge, USA: Advances in Neural Information Processing Systems 14, 2002.

Palla, Derényi, Farkas, & Vicsek. (2005). *Uncovering the overlapping community structure of complex networks in nature and society.* Nature 435, 814.

Pons, P., & Latapy, M. (2005). *Computing Communities in Large Networks Using Random Walks.* Journal of Graph Algorithms and Applications vol. 10, no. 2, pp. 191–218 .

Radicchi, Castellano, Cecconi, Loreto, & Parisi. (2004). *Defining and identifying communities in networks.* Rome, Italy.

Ramasco, J., & Mungan, M. (2008). *Inversion method for content-based networks.* Phys. Rev. E 77(3), 036122.

Reichardt, J., & Bornholdt, S. (2006). *Statistical mechanics of community detection.* Physica D 224, 20.

Rosvall, M., & Bergstrom, C. T. (2007). *An information-theoretic framework for resolving community structure in complex networks.* Proc. Natl. Acad.Sci. USA 104, 7327.

Rosvall, M., & Bergstrom, C. T. (2008). *Maps of random walks on complex networks reveal community structure.* Proc. Natl. Acad. USA 105, 1118.

Shannon, C. E. (1948). *A Mathematical Theory of Communication.* The Bell System Technical Journal (Volume: 27 , Issue: 3).

Shen, H., Cheng, X., Cai, K., & Hu, M.-B. (2008). *Detect overlapping and hierarchical community structure in networks.*

Shi, J., & Malik, J. (2000). *Normalized Cuts and Image Segmentation.* IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888.

Winkler, R. (2003). *Introduction to Bayesian Inference and Decision.* Probabilistic Publishing, Gainesville, USA.

Παραρτήματα

Παράρτημα 1. Γεωμετρική τεχνική και τεχνική που δρομολογεί Ροές στο γράφο

Στη γεωμετρική τεχνική, ο γράφος αναπαρίσταται σε κάποιο γεωμετρικό χώρο, για διαισθητική διευκόλυνση ας θεωρήσουμε το δισδιάστατο Ευκλείδειο χώρο $\mathcal{R}^2$, με τέτοιο τρόπο ώστε η μέση απόσταση μεταξύ δυο απομακρυσμένων κορυφών να είναι μια δεδομένη σταθερά, έστω 1, ενώ η απόσταση μεταξύ δυο γειτονικών κορυφών να είναι η μικρότερη δυνατή. Αυτή η διαδικάσια αναφέρεται στη βιβλιογραφία ως ενσωμάτωση (embedding) του γράφου στο γεωμετρικό χώρο.

Η ενσωμάτωση αυτή υπονοεί οτι η εγγύτητα των κορυφών στο γεωμετρικό αυτό χώρο αντιπροσωπεύει τη συνδεσιμότητα του γράφου. Μια καλή διαμέριση του γράφου είναι ο διαχωρισμός μιας μεγάλης του περιοχής απο τον υπόλοιπο, κατά μήκος μιας γεωμετρικής καμπύλης. Αυτό που επιτυγχάνεται με την αποκοπή αυτή, δεδομένων των προϋποθέσεων της ενσωμάτωσης, είναι οτι ο αναμενόμενος αριθμός κορυφών εκατέρωθεν της αποκοπής θα είναι μεγάλος, ενώ οι σύνδεσμοι μεταξύ των δύο μερών του γράφου που προέκυψαν θα είναι λίγοι.

Η πραγματική ενσωμάτωση του γράφου, ωστόσο, απο τους αλγορίθμους που υπάγονται στην μέθοδο αυτή, δε γίνεται στον Ευκλείδειο χώρο $\mathbb{R}^2$, αλλά στην επιφάνεια της μοναδιαίας σφαίρας σε έναν $n-$διάστατο Ευκλείδειο χώρο, όπου n, το πλήθος των κορυφών του γράφου. Η απόσταση μεταξύ κορυφών στο χώρο αυτό, ορίζεται ως το τετράγωνο της Ευκλείδειας απόστασης, $d_{AB} = \sum_{k=1}^{n}(a_k - \beta_k)^2$. Αυτό σημαίνει οτι ο γράφος αναπαρίσταται με τέτοιο τρόπο ώστε το άθροισμα των τετραγώνων των μηκών των ακμών να είναι το μικρότερο δυνατό, ενώ απαιτούμε το τετράγωνο της απόστασης απομακρυσμένων ζευγών κορυφών να είναι μια σταθερά, έστω 1.

Η τεχνική που δρομολογεί ροές στο γράφο, αναπαρίσταται μέσω κυκλοφοριακών ροών σε οδικά δίκτυα. Ένα δίκτυο ροών (*flow network*) είναι ένας κατευθυνόμενος γράφος $G = (V, E)$, με μια κορυφή «πηγής» (*source vertex*) $s \in V$ και μια κορυφή «βυθίσματος» (*sink vertex*) $t \in V$. Κάθε ακμή $e = (v, w)$ από το v στο w, έχει μια προκαθορισμένη χωρητικότητα (*capacity*) η οποία περιγράφεται απο μια συνάρτηση $c : E \to \mathbb{R}$, όπου $c(e)$ η χωρητικότητα της ακμής e. Η ποσότητα της ροής ανάμεσα σε δύο κορυφές, περιγράφεται απο μια συνάρτηση $f : V \times V \to \mathbb{R}$, με τις ακόλουθες ιδιότητες:

- Χωρητικότητα (*Capacity*) : $f(v,w) \leq c(v,w) \quad \forall\, v,w \in V$
- Αντισυμμετρικότητα: (*Antisymmetry*) : $f(v,w) = -f(w,v) \quad \forall\, v,w \in V$
- Διατηρησιμότητα (*Conservation*) :$\sum_{w \in V} f(v,w) = 0, \forall\, v \in V - \{s,t\}$

Η αναπαράσταση της τεχνικής έχει ως εξής: δρομολογείται μια ροή μεταξύ κάθε ζεύγους κορυφών στο γράφο έτσι ώστε η «συμφόρηση» του συνδέσμου με τη μεγαλύτερη «συμφόρηση» να είναι η μικρότερη δυνατή. Δηλαδή δρομολογείται η κίνηση με τέτοιο τρόπο ωστε η χειρότερη δυνατή κυκλοφοριακή συμφόρηση να μην είναι πολύ κακή. Αυτό επιτυγχάνεται μέσω διοδίων, που πρέπει να «πληρωθούν» για να μεταβεί κάποιος απο έναν κόμβο σε έναν άλλο. Η απόσταση μεταξύ ενός ζεύγους κορυφών προσδιορίζεται απο το ελάχιστο διόδιο, μεταξύ όλων των μονοπατιών, που πρέπει να «πληρώσει» κάποιος για να ταξιδέψει ανάμεσα τους. Με το να θεωρήσουμε αυθαίρετα έναν κόμβο, έστω v, και όλους τους κόμβους σε απόσταση R απο αυτόν, ορίζεται και μια αποκοπή στο γράφο. Εναλλακτικά, θα μπορούσαμε να προσσεγίσουμε διαισθητικά την τεχνική και ως εξής: Τα διόδια ορίζουν μια αφηρημένη «Γεωμετρία» στο γράφο. Ένας κόμβος είναι κοντά σε κόμβους που συνδέονται με χαμηλά διόδια και απομακρυσμένος απο κόμβους που συνδέονται με υψηλά διόδια. Έχει αποδειχθεί (Leighton and Rao, 1999) οτι μια καλή αποκοπή του χώρου αυτού είναι μια τυχαία αποκοπή του.

Παράρτημα 2. Περιγραφή Γενετικου Αλγορίθμου.

Οι γενετικοί αλγόριθμοι, είναι αλγόριθμοι που χρησιμοποιούνται ευρέως στον κλάδο τεχνικών βελτιστοποίησης και η χρήση τους κρίνεται αναγκαία, όταν ο χώρος των λύσεων του προβλήματος είναι πολύ μεγάλος και η εύρεση της βέλτιστης λύσης απο αλγορίθμους ωμής βίας (αλγόριθμοι που ελέγχουν όλο το σύνολο των δυνατών λύσεων) είναι πρακτικά αδύνατη. Στηρίζονται στην προσομοίωση της δημιουργίας και εξέλιξης ενός πληθυσμού ατόμων, που χαρακτηρίζονται απο ένα χρωμόσωμα (chromosome), όπου κάθε άτομο αναπαριστά μια πιθανή λύση του προβλήματος βελτιστοποίησης. Σε κάθε βήμα της προσομοίωσης, επιλέγονται τα «καλύτερα» άτομα , τα οποία συμπεριλαμβάνονται σε μια νέα γενιά και χρησιμοποιούνται για να δημιουργηθούν νέα στοιχεία τα οποία αντικαθιστούν τα «χειρότερα» άτομα της προηγούμενης. Η σύγκλιση σε μια ικανοποιητική λύση επιτυγχάνεται εφαρμόζοντας κατάλληλες διεργασίες σε κάθε άτομο, οι οποίες εξαρτώνται κάθε φορά απο τη φύση του προβλήματος. Σε κάθε επανάληψη

αξιολογείται η προσαρμογή των μελών του πληθυσμού και τα άτομα ταξινομούνται σε αναλογία με το επίπεδο προσαρμογής τους (fitness level). Τα άτομα με τα μεγαλύτερες τιμές προσαρμογής (fitness values) αναπαράγονται για την επόμενη γενιά. Επίσης δημιουργούνται νέα άτομα μέσω συνδυασμού των «καλύτερων» ατόμων της προηγούμενης γενίας, χρησιμοποιώντας μια διαδικασία διασταύρωσης (crossover operation) και μέσω τυχαίων μεταλλάξεων (random mutations).

Σημαντικό χαρακτηριστικό των αλγορίθμων αυτών είναι η χαμηλή υπολογιστική τους πολυπλοκότητά, ενώ το μειονέκτημα τους έγκειται στον κίνδυνο του να μη βρεθεί η βέλτιστη λύση, καθώς η διαδικασία βελτιστοποίησης μπορεί να παγιδευτεί σε τοπικά μέγιστα.

Στην περίπτωση μας τώρα, όπως περιγράφουν οι συγγραφείς, η συνάρτηση προσαρμογής (fitness function) είναι η Q_{ov} και οι λύσεις του προβλήματος αναπαριστώνται από όλες τις δυνατές αλληλοκαλυπτόμενες κοινότητες στο γράφο. Για αρχή πρέπει να βρούμε μια κατάλληλη χρωμοσωμική αναπαράσταση, η οποία να ταιριάζει στο δεδομένο πρόβλημα. Στην εφαρμογή μας, το χρωμόσωμα αναπαρίσταται απο έναν πίνακα $M = (\alpha_{i,c})$, όπου $i = 1, \dots, |V|$ και όπου $c = 1, \dots, |C|$. Κάθε στοιχείο $\alpha_{i,c}$ είναι η ισχύς με την οποία ένας κόμβος i του γράφου, ανήκει σε μια κοινότητα c. Το σύνολο τιμών των $\alpha_{i,c}$ είναι το $[0.0, 1.0]$. Για κάθε κόμβο στο γράφο ισχύει ο εξής περιορισμός:

$$\sum_{c=1}^{|C|} a_{ic} = 1.0$$

Ο περιορισμός αυτός χρησιμοποιείται ωστε να αποφευχθεί η σύγκλιση του γενετικού αλγορίθμου σε λύσεις κατά τις οποίες κόμβοι ανήκουν σε διαφορετικές κοινότητες ταυτόχρονα, με ισχύ 1.0. Με άλλα λόγια μας ενδιαφέρουν λύσεις στις οποίες ένας κόμβος με ισχύ 1.0 μπορεί να ανήκει μόνο σε μια κοινότητα. Τα βήματα του αλγορίθμου είναι τα εξής:

Βασική λειτουργία του γενετικού αλγορίθμου.

Απαιτεί: *n_epochs, n_individuals,* n_{comm}

---

1. *Population=CreatePopulation (n_individuals,* n_{comm}*)*
2. **for** (*i=0 to n_epochs)* **do**
3. *population.Fitness* ( )
4. *population.SortingAndSelection* ( )
5. *population.CrossOver* ( )
6. *population.Mutation* ( )
7. *population.CleanUp* ( )
8. **end for**

Ο παραπάνω αλγόριθμος δέχεται 3 παραμέτρους ως εισόδους: *n_epochs,* δηλαδή το συνολικό αριθμό γενιών (δηλαδή τον αριθμό των κύκλων της προσομοίωσης), *n_individuals*, δηλαδή τον αριθμό των ατόμων απο τα οποία αποτελείται ο πληθυσμός και n_{comm}, δηλαδή το συνολικό αριθμό αλληλοκαλυπτόμενων κοινοτήτων που ο αλγόριθμος προσπαθεί να βρει.

Το πρώτο βήμα του γενετικού αλγορίθμου είναι η δημιουργία του αρχικού πληθυσμού των *n_individuals.* Στο βήμα αυτό, το χρωμόσωμα κάθε ατόμου αρχικοποιείται με τους συντελεστές συμμετοχής (belonging factors) να επιλέγονται τυχαία στο διάστημα $[0.0, 1.0]$ και κανονικοποείται έτσι ώστε να πληρείται ο περιορισμός που αναφέρθηκε πιο πάνω. Μετά την αρχικοποίηση ο αλγόριθμος τρέχει για *n_epochs* βήματα, εφαρμόζοντας σε κάθε επανάληψη ένα σύνολο απο γενετικούς χειρισμούς (genetic operators) στα άτομα του πληθυσμού.

Πιο συγκεκριμένα, η εκτίμηση της προσαρμογής και οι χειρισμοί επιλογής, διασταύρωσης και μετάλλαξης εφαρμόζονται στα άτομα, ως εξής:

Εκτίμηση προσαρμογής, διάταξη και επιλογή (*Fitness evaluation, sorting and selection*): Η εκτίμηση προσαρμογής έχει να κάνει με τον υπολογισμό της modularity για το χρωμόσωμα του κάθε ατόμου. Ένας δεδομένος αριθμός ατόμων με υψηλή προσαρμογή, συμπεριλαμβάνεται στην επόμενη γενιά. Με σκοπό να διατηρηθεί ο αριθμός των ατόμων, προστίθεται νέα μέλη εκ των οποίων ένας αριθμός απο αυτά να επιτυγχάνεται μέσω διασταύρωσης ανάμεσα στα καλύτερα άτομα της προηγούμενης γενιάς, ενώ τα εναπομείναντα μέλη δημιουργούνται απο την αρχή. Εν

ολίγοις η νέα γενιά εμπεριέχει τους καλύτερους αλλά και έναν συνδυασμό των καλύτερων της προηγούμενης.

Διασταύρωση και Μετάλλαξη (*Crossover and Mutation*): Η διασταύρωση είναι μια γενετική διαδικασία που αποτελείται απο ανταλλαγή μέρους των χρωμοσωμάτων δύο διαφορετικών ατόμων. Ο σκοπός της διαδικασίας είναι η δημιουργία ενός νέου ατόμου που κληρονομεί τη γενετική δομή (χρωμόσωμα) από αυτή των δύο άλλων μελών του πληθυσμού, ελπίζοντας οτι το νέο άτομο θα είναι καλύτερο απο αυτά απο τα οποία προέρχεται.

Επιπλέον, σε κάθε επανάληψη, μεταλλάξεις εφαρμόζονται σε ένα σύνολο ατόμων με σκοπό να βελτιωθούν οι πιθανότητες εύρεσης της βέλτιστης λύσης. Κατά τη διαδικασία μετάλλαξης εκτελούνται αυθαίρετες τροποποιήσεις σε σύνολα χρωμοσωμάτων των ατόμων. Στόχος της μετάλλαξης είναι να αυξηθεί η ποικιλομορφία του πληθυσμού ωστε να επιταχυνθεί η σύγκλιση.

Διαδικασία Εκκαθάρισης (*CleanUp)*: Εκτός των προαναφερθέντων χειρισμών στον προτεινόμενο γενετικό αλγόριθμο, έχει ενσωματωθεί σε αυτόν και μια συνάρτηση εκκαθάρισης (CleanUp function), με σκοπό την βελτιστοποίηση της ποιότητας της διαίρεσης του γράφου σε αλληλοκαλυπτόμενες κοινότητες. Η συνάρτηση αυτή εφαρμόζεται στο χρωμόσωμα του κάθε ατόμου και είναι στενά συνδεδεμένη με το πρόβλημα της βελτιστοποίησης. Για ένα δεδομένο χρωμόσωμα, επιλέγονται τυχαία ένας κόμβος i και μια κοινότητα c και υπολογίζεται ο μέσος συντελεστής συμμετοχής $avgNeigh(i,c)$ για την κοινότητα c και γειτονικούς στον i κόμβους, όπως επίσης και ο συντελεστής μη συμμετοχής $avgNotNeigh(i,c)$ για την κοινότητα c και μη γειτονικούς στον i κόμβους.

$$avgNeigh(i,c) = \frac{\sum_{v \in Neighborhood(i)} a_{v,c}}{|Neighborhood(i)|}$$

και

$$avgNotNeigh(i,c) = \frac{\sum_{v \notin Neighborhood(i)} a_{v,c}}{n - |Neighborhood(i)|}$$

Οι τιμές των $avgNeigh(i,c)$ και $avgNotNeigh(i,c)$ κατόπιν συγκρίνονται: Εάν $avgNeigh(i,c) > avgNotNeigh(i,c)$, τότε ο παράγοντας συμμετοχής του κόμβου i στην κοινότητα c αυξάνεται κατά ένα μικρό ποσοστό, διαφορετικά μειώνεται. Ο λόγος για τον οποίο το βήμα του αλγορίθμου αυτό είναι σημαντικό, εξηγείται απο το γεγονός οτι ένας γειτονικός στον

i, κόμβος, θα ανήκει στην κοινότητα c με κατά μέσο όρο μεγαλύτερη πιθανότητα απο τους κόμβους που δεν είναι σε γειτονιές του i. Μέσω αύξησης του συντελεστή συμμετοχής του κόμβου i στην κοινότητα c, ο αλγόριθμος συγκλίνει στη βέλτιστη λύση ταχύτερα. Μειώνοντας τον παράγοντα συμμετοχής του κόμβου i στην κοινότητα c, αντίστοιχα, λαμβάνουμε στην ουσία υπόψιν οτι ο κόμβος i ανήκει στην κοινότητα c με λιγότερη ισχύ.

Έχει αποδειχτεί μέσω πειραμάτων οτι η συνάρτηση εκκαθάρισης αυτή, βελτιώνει και την ποιότητα των διαιρέσεων αλλά και την ταχύτητα σύγκλισης του γενετικού αλγορίθμου.

Η πολυπλοκότητα του αλγορίθμου αυτού προκύπτει λαμβάνοντας υπόψιν τους χειρισμούς που αφορούν τον αριθμό n των κόμβων του δικτύου και τον αριθμό $|C|$ των κοινοτήτων. Έχει υπολογιστεί οτι η πολυπλοκότητα αυτή είναι $O(|C| * n^2)$ στη χειρότερη περίπτωση.

Παραρτήμα 3. Κώδικας Ubuntu 16.04

Η γενική μορφή του κώδικα που χρησιμοποιήσαμε για την παραγωγή μη κατευθυνόμενων και μη σταθμισμένων LFR δικτύων αναφοράς με συγκεκριμένες παραμέτρους δίνεται παρακάτω:

```
napo2d@DESKTOP-UIHQJHF:~$ cd /
napo2d@DESKTOP-UIHQJHF:/$ cd mnt/c/Users/napol/Desktop/Community\ Detection/bi-
nary_networks
napo2d@DESKTOP-UIHQJHF:/mnt/c/Users/napol/Desktop/Community\Detection/bi-
nary_networks$ ./benchmark -N 1000 -k 15 -maxk 50 -mu 0.1 -minc 20 -maxc 100
```

Γενική μορφή κώδικα για την παραγωγή μη κατευθυνόμενων και σταθμισμένων LFR δικτύων αναφοράς, με συγκεκριμένες τιμές των παραμέτρων:

```
napo2d@DESKTOP-UIHQJHF:~$ cd /
napo2d@DESKTOP-UIHQJHF:/$ cd mnt/c/Users/napol/Desktop/Community\ Detec-
tion/weighted_networks
napo2d@DESKTOP-UIHQJHF:/mnt/c/Users/napol/Desktop/Community\Detec-
tion/weighted_networks$ ./benchmark -N 1000 -k 15 -maxk 50 -mut 0.5 -muw 0.1 -beta 1.5 -
minc 20 -maxc 50
```

Παράρτημα 4. Κώδικας R για ανάγνωση μη κατευθυνόμενων και μη σταθμισμένων LFR δικτύων αναφοράς.

```
install.packages("igraph")
library(igraph)
setwd("C:/Users/napol/Desktop/Community Detection/binary_networks/mu01")
getwd()
table_graph<-read.table("network.dat")
g <- make_empty_graph(n=1000,directed = FALSE)
d=dim(table_graph)[1]
for (i in 1:d){
        a<-as.numeric(table_graph[i,1])
        b<-as.numeric(table_graph[i,2])
        if (g[a,b]==0){             #αν δεν υπάρχει σύνδεση πρόσθεσέ τη, αν δεν υπάρχει μην
την ξαναπροσθέτεις
        g<-g+edges(c(a,b))
            }
    }
plot(g)
table_community<-read.table("community.dat")
table_community
```

Παράρτημα 5. Κώδικας R για ανάγνωση μη κατευθυνόμενων και σταθμισμένων LFR δικτύων αναφοράς.

```
install.packages("igraph")
library(igraph)
setwd("C:/Users/napol/Desktop/Community Detection/weighted_networks/1k-big-
ger/muw01")
getwd()
table_graph<-read.table("network.dat")
g <- make_empty_graph(n=1000,directed = FALSE)
d=dim(table_graph)[1]
d
for (i in 1:d){
        a<-as.numeric(table_graph[i,1])
        b<-as.numeric(table_graph[i,2])
        w<-as.numeric(table_graph[i,3])
        if (g[a,b]==0){             # αν δεν υπάρχει σύνδεση πρόσθεσέ τη, αν δεν υπάρχει μην
την ξαναπροσθέτεις
                g<-g+edges(c(a,b),weight=w)
        }
```

```
}
plot(g)
table_community<-read.table("community.dat")
table_community
```

Παράρτημα 6. Κώδικας για την εφαρμογή των 9 αλγορίθμων και υπολογισμό τιμών NMI σε μη κατευθυνόμενα και μη σταθμισμένα LFR Δίκτυα Αναφοράς.

```
#algorithm of G-N(1)
communities1<-cluster_edge_betweenness(g)
communities1
d=dim(communities1[])
d
t=matrix(, nrow = 1000, ncol = 2)
count<-1
for (i in 1:d){
        j<-1
        c<-communities1[[i]][j]
        while(!is.na(c)) {
                t[count,1]<-c
                t[count,2]<-i
                count<-count+1
                j<-j+1
                c<-communities1[[i]][j]
        }
}
tt1<-as.table(t)
tt1

#Fast Greedy Clauset(2)
communities2<-cluster_fast_greedy(g)
communities2
d=dim(communities2[])
t=matrix(, nrow = 1000, ncol = 2)
count<-1
for (i in 1:d){
        j<-1
        c<-communities2[[i]][j]
        while(!is.na(c)) {
                t[count,1]<-c
                t[count,2]<-i
                count<-count+1
                j<-j+1
                c<-communities2[[i]][j]
```

```
        }
}
tt2<-as.table(t)
tt2

#Blondel et al.(3)
communities3<-cluster_louvain(g, weights = NULL)
communities3
communities3[[1]][2]
d=dim(communities3[])
d
t=matrix(, nrow = 1000, ncol = 2)
count<-1
for (i in 1:d){
        j<-1
        c<-communities3[[i]][j]
        while(!is.na(c)) {
                t[count,1]<-c
                t[count,2]<-i
                count<-count+1
                j<-j+1
                c<-communities3[[i]][j]
        }
}
tt3<-as.table(t)
tt3
#Exhaustive modularity optimization via simulated annealing(4)
install.packages("modMax")
library(modMax)
adj <- get.adjacency(g)
adj
communities4 <- simulatedAnnealing(adj,initial =c('greedy'),alpha = 1.005, fixed=10)
d<-length(communities4[[3]])
d
t=matrix(, nrow = 1000, ncol = 2)
t
for (i in 1:d){
        t[i,1]<-i
        t[i,2]<-communities4[[3]][i]
}
tt4<-as.table(t)
tt4

#Finding communities based on propagating labels(5)
communities5<-cluster_label_prop(g, weights = NULL, initial = NULL, fixed = NULL)
communities5
```

```
d=dim(communities5[])
d
t=matrix(, nrow = 1000, ncol = 2)

count<-1
for (i in 1:d){
        j<-1
        c<-communities3[[i]][j]
        while(!is.na(c)) {
                t[count,1]<-c
                t[count,2]<-i
                count<-count+1
                j<-j+1
                c<-communities3[[i]][j]
        }
}
tt5<-as.table(t)
tt5

#Leading Eigevalue of Newman(6)
communities6<-cluster_leading_eigen(g, steps = -1, weights = NULL)
communities6
d=dim(communities6[])
d
t=matrix(, nrow = 1000, ncol = 2)
count<-1
for (i in 1:d){
        j<-1
        c<-communities6[[i]][j]
        while(!is.na(c)) {
                t[count,1]<-c
                t[count,2]<-i
                count<-count+1
                j<-j+1
                c<-communities6[[i]][j]
        }
}
tt6<-as.table(t)
tt6

#Walktrap(7)
communities7<-cluster_walktrap(g, weights = NULL)
communities7
d=dim(communities7[])
d
t=matrix(, nrow = 1000, ncol = 2)
```

```
count<-1
for (i in 1:d){
        j<-1
        c<-communities7[[i]][j]
        while(!is.na(c)) {
                t[count,1]<-c
                t[count,2]<-i
                count<-count+1
                j<-j+1
                c<-communities7[[i]][j]
        }
}
tt7<-as.table(t)
tt7
```

#MCL (8)

```
install.packages("MCL")
library(MCL)
adjacency<-as_adjacency_matrix(g, type = c("both"), attr = NULL)
communities8<-mcl(adjacency, addLoops = TRUE, expansion = 2, inflation = 2, allow1 = FALSE, max.iter = 100, ESM = FALSE)
communities8
communities8$Cluster[159]
d<-length(communities8[[3]])
d
t=matrix(, nrow = 1000, ncol = 2)
for (i in 1:d){
        t[i,1]<-i
        t[i,2]<-communities8[[3]][i]
}
tt8<-as.table(t)
tt8
```

#Infomap (9)

```
communities9<-cluster_infomap(g, e.weights = NULL, v.weights = NULL, nb.trials = 10, modularity = TRUE)
communities9
d=dim(communities9[])
d
t=matrix(, nrow = 1000, ncol = 2)

count<-1
for (i in 1:d){
        j<-1
        c<-communities9[[i]][j]
```

```
        while(!is.na(c)) {
                t[count,1]<-c
                t[count,2]<-i
                count<-count+1
                j<-j+1
                c<-communities9[[i]][j]
        }
}
tt9<-as.table(t)
tt9
```

```
#Comparisons
install.packages("NMI")
library(NMI)

N1<-NMI(table_community,tt1) #Benchmark-G-N
N1

N2<-NMI(table_community,tt2)#Benchmark-Fast Greedy
N2

N3<-NMI(table_community,tt3) #Benchmark-Louvain
N3

N4<-NMI(table_community,tt4) #Benchmark-Simulated Annealing
N4

N5<-NMI(table_community,tt5) #Benchmark-Label Propagation
N5

N6<-NMI(table_community, tt6) #Benchmark-Leading Eigenvalue
N6

N7<-NMI(table_community, tt7) #Benchmark- Walktrap
N7

N8<-NMI(table_community, tt8) #Benchmark- MCL
N8

N9<-NMI(table_community, tt9) #Benchmark- Infomap
N9
```

Παράρτημα 7. Κώδικας για κατασκευή των αντίστοιχων γραφημάτων για τους 9 αλγορίθμους.

```
#NMI Scores for Small Communities
data=matrix(, nrow = 9, ncol = 9)
data[1,1:9]<-c(1,1,1,0.994,0.951,0.744,0.602,0.549,0.617) #NMI Scores of G-N
data[2,1:9]<-c(0.952, 0.861, 0.779,0.693, 0.556, 0.336, 0.123, 0.0073, 0.04) #NMI Scores of Fast Greedy
data[3,1:9]<-c(1,1,1,0.994,0.98,0.896,0.335,0.108,0.07)  #NMI Scores of Blondel et al.
data[4,1:9]<-c(0.952,0.885,0.735,0.623,0.513,0.313,0.161,0.008,0.035) #NMI Scores of Sim. Annealing
data[5,1:9]<-c(0.597,0.568,0.447,0.419,0.406,0.270,0.114,0.076,0.0482) #NMI Scores of Lead. Eigenvalue
data[6,1:9]<-c(1,1,1,0.994,0.98,0,0,0,0)  #NMI Scores of Label Propagation
data[7,1:9]<-c(1,1,1,1,0.949,0.706,0.459,0.301,0.292)  #NMI Scores of Walktrap
data[8,1:9]<-c(1,1,0.995,0.911,0.699,0.481,0.325,0.273,0.254)  #NMI Scores of MCL
data[9,1:9]<-c(1,1,1,1,0.998,0.980,0,0,0)   #NMI Scores of Infomap
data
#NMI Scores for Big Communities
data2<-matrix(, nrow = 9, ncol = 9)
data2[1,1:9]<-c(1,1,1,0.9889,0.7954,0.5899,0.5082,0.4402,0.4629) #NMI Scores of G-N
data2[2,1:9]<-c(0.9961, 0.9291,0.7984,0.7111, 0.5188, 0.2587, 0.0760, 0.0050, 0.02256) #NMI Scores of Fast Greedy
data2[3,1:9]<-c(1,1,1,1,0.9605,0.7647,0.1844,0.051,0.043)  #NMI Scores of Blondel et al.
data2[4,1:9]<-c(0.9943,0.8854,0.8090,0.6838,0.5372,0.2446,0.090,0.0449,0.040) #NMI Scores of Sim. Annealing
data2[5,1:9]<-c(1,1,1,1,0,0,0,0,0)  #NMI Scores of Label Propagation
data2[6,1:9]<-c(0.39410,0.52372,0.5289,0.3304,0.2812,0.174,0.2990,0.0482,0.02624) #NMI Scores of Lead. Eigenvalue
data2[7,1:9]<-c(1,1,1,0.9831,0.8321,0.4854,0.2990,0.2173,0.25053)  #NMI Scores of Walktrap
data2[8,1:9]<-c(0.9884,0.8454,0.7421,0.6398,0.5141,0.3575,0.2514,0.2432,0.2337)  #NMI Scores of MCL
data2[9,1:9]<-c(1,1,1,1,1,0,0,0,0)   #NMI Scores of Infomap
data2
mu<-c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)
mu
name<-matrix(, nrow = 9, ncol = 1)
name[1]<-'Algorithm of Girvan-Newman'
name[2]<-'Fast Greedy Algorithm'
name[3]<-'Algorithm of Blondel et al'
name[4]<-'Simulated Annealing'
name[5]<-'Label Propagation'
name[6]<-'Leading Eigevalue'
name[7]<-'Walktrap'
name[8]<-'MCL'
```

```
name[9]<-'Infomap'

s=plot(mu,data[i,1:9], main=name[i], type='o',col = 'black', xlim = c(0,1), ylim = c(0, 1),
ylab='Normalized Mutual Information', xlab='Mixing Parameter')
par(new=TRUE)
s=plot(mu,data2[2,1:9], main=name[i], type='o',col = 'red',
        xlim = c(0,1), ylim = c(0, 1),
        ylab='Normalized Mutual Information', xlab='Mixing Parameter')
# Plot του i αλγορίθμου για μικρές και μεγάλες κοινότητες.
color<-matrix(, nrow = 9, ncol = 3)      #Πίνακας Χρωμάτων
color[1,1:3] <- c(0,0,0)
color[2,1:3] <- c(1,1,0)
color[3,1:3] <- c(0.5,0.5,1)
color[4,1:3] <- c(0,1,1)
color[5,1:3] <- c(1,0,0)
color[6,1:3] <- c(0,1,0)
color[7,1:3] <- c(0,0,1)
color[8,1:3] <- c(1,0,1)
color[9,1:3] <- c(1,0.5,0.5)
color
i=1
r = color[i,1]                  #RGB
g = color[i,2]
b = color[i,3]

for (i in 1:9){
        r = color[i,1]                          #Plot όλων των αλγορίθμων μαζί.
        g = color[i,2]
        b = color[i,3]
        s=plot(mu,data[i,1:9],main="All 9 algorithms",type='o',col = rgb(r, g, b),
        xlim = c(0,1), ylim = c(0, 1),
        ylab='Normalized Mutual Information', xlab='Mixing Parameter')
        par(new=TRUE)
}
```

Παράρτημα 8. Κώδικας για εφαρμογή αλγορίθμων σε μη κατευθυνόμενα και σταθμισμένα LFR Δίκτυα αναφοράς, υπολογισμό τιμών κριτηρίου NMI και κατασκευή αντίστοιχων γραφημάτων.

```
#Exhaustive modularity optimization via simulated annealing (1)
install.packages("modMax")
library(modMax)
adj <- get.adjacency(g,attr="weight")
adj
communities1 <- simulatedAnnealing(adj,initial =c('greedy'),alpha = 1.005, fixed=10)
d<-length(communities1[[3]])
t=matrix(, nrow = 1000, ncol = 2)
t

for (i in 1:d){
        t[i,1]<-i
        t[i,2]<-communities1[[3]][i]
}
tt1<-as.table(t)
tt1

#MCL (2)
install.packages("MCL")
library(MCL)
adjacency<-get.adjacency(g,attr="weight")
adjacency
communities2<-mcl(adjacency, addLoops = TRUE, expansion = 2, inflation = 2, allow1 = FALSE, max.iter = 100, ESM = FALSE)
communities2
communities2$Cluster[159]
d<-length(communities2[[3]])
d
t=matrix(, nrow = 1000, ncol = 2)

for (i in 1:d){
        t[i,1]<-i
        t[i,2]<-communities2[[3]][i]
}
tt2<-as.table(t)
tt2

#Infomap (3)
communities3<-cluster_infomap(g,e.weights = w, v.weights = NULL, nb.trials = 10, modularity = FALSE)
communities3
```

```
d<-dim(communities3[])
d
t<-matrix(, nrow = 1000, ncol = 2)

count<-1
for (i in 1:d){
        j<-1
        c<-communities3[[i]][j]
        while(!is.na(c)) {
                t[count,1]<-c
                t[count,2]<-i
                count<-count+1
                j<-j+1
                c<-communities3[[i]][j]
        }
}
tt3<-as.table(t)
tt3
#Comparisons
install.packages("NMI")
library(NMI)

N1<-NMI(table_community,tt1) #Benchmark-Simulated Annealing
N1
N2<-NMI(table_community, tt2) #Benchmark- MCL
N2
N3<-NMI(table_community, tt3) #Benchmark- Infomap
N3
```

```
#NMI Scores for Big Communities
data<-matrix(, nrow = 3, ncol = 9)
data[1,1:9]<-c(0.618,0.603,0.599,0.612,0.618,0.604,0.595,0.602,0.605) #NMI Scores of Sim. Annealing
data[2,1:9]<-c(0.812, 0.804, 0.776,0.705, 0.537, 0.342, 0.272, 0.267, 0.271) #NMI Scores of MCL
data[3,1:9]<-c(0.467,0.469,0.436,0.441,0.448,0.432,0.442,0.451,0.438)  #NMI Scores of Info-map.
data
muw<-c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)
muw
data
#NMI Scores for Small Communities
data2<-matrix(, nrow = 3, ncol = 9)
data2[1,1:9]<-c(0.899,0.906,0.888,0.896,0.878,0.878,0.885,0.879,0.879) #NMI Scores of Sim. Annealing
data2[2,1:9]<-c(0.966, 0.961, 0.944,0.905, 0.773, 0.477, 0.344, 0.341, 0.346) #NMI Scores of MCL
```

```
data2[3,1:9]<-c(0.582,0.556,0.538,0.546,0.503,0.528,0.533,0.568,0.567)  #NMI Scores of Info-
map.
data2
muw=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)
muw
name<-matrix(, nrow = 3, ncol = 1)
name[1]<-'Simulated Annealing'
name[2]<-'MCL'
name[3]<-'Infomap'
name

s<-plot(muw,data[1,1:9], main=name[1], type='o',col = 'blue',xlim = c(0,1), ylim = c(0,
1),ylab='Normalized Mutual Information', xlab='MUw')
par(new=TRUE)
s<-plot(muw,data2[1,1:9], main=name[1], type='o',col = 'green',          #Plot του i αλγορίθμου
        xlim = c(0,1), ylim = c(0, 1),
        ylab='Normalized Mutual Information', xlab='MUw')
i<-1
r <- color[i,1]                  #RGB
g <- color[i,2]
b <- color[i,3]

for (i in 1:3){
        r = color[i,1]                    #Plot πολλών αλγορίθμων μαζί για μεγάλες κοινότητες
        g = color[i,2]
        b = color[i,3]
        plot(muw,data[i,1:9], main='Big Communities', type='o',col = rgb(r, g, b),
        xlim = c(0,1), ylim = c(0, 1),
        ylab='Normalized Mutual Information', xlab='MUw')
        par(new=TRUE)
}

for (i in 1:3){
        r = color[i,1]                  #Plot πολλών αλγορίθμων μαζί για μικρές κοινότητες
        g = color[i,2]
        b = color[i,3]
        plot(muw,data2[i,1:9], main='Small Communities', type='o',col = rgb(r, g, b),
        xlim = c(0,1), ylim = c(0, 1),
        ylab='Normalized Mutual Information', xlab='MUw')
        par(new=TRUE)
}
```