ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. «ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ»

ΕΙΔΙΚΕΥΣΗ : ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

<<Μελέτη και ανάπτυξη προσεγγίσεων αξιολόγησης αποτελεσμάτων clustering σε γράφους>>

Ιωάννα Γαγλία

Α.Μ. ΜΕ2003

Επιβλέπων Καθηγήτρια :

Μαρία Χαλκίδη

ΠΕΙΡΑΙΑΣ

06 / 2022

Περιεχόμενα

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια μου Μαρία Χαλκίδη, για την ανάθεση της συγκεκριμένης εργασίας, τις σημαντικές συμβουλές και την καθοδήγηση της. Επίσης ένα ευχαριστώ στους συναδέλφους και τους διδάσκοντες του μεταπτυχιακού για την συνεργασία και την γνώση που μου προσφέρθηκε.
Θα ήθελα να ευχαριστήσω την εταιρεία στην οποία εργάζομαι ERMA FIRST ESK Engineering Solutions S.A. και τους συναδέλφους μου για την πολύτιμη υποστήριξη τους.
Κλείνοντας θα ήθελα να ευχαριστήσω την οικογένεια και τους δικούς μου ανθρώπους που είναι σε κάθε μου βήμα η δύναμη μου.

Περίληψη

Στην επιστήμη των μαθηματικών οι γράφοι είναι ένας τρόπος αναπαράστασης ενός δικτύου, μιας συλλογής στοιχείων που είναι συνδεδεμένα μεταξύ τους. Οι γράφοι δομούνται από κόμβους οι οποίοι αντιπροσωπεύουν τα δεδομένα και ακμές οι οποίες χαρακτηρίζουν τις σχέσεις μεταξύ των κόμβων. Οι γράφοι οργανώνονται σε συστάδες, δηλαδή ομάδες κόμβων με περισσότερα κοινά χαρακτηριστικά εντός της ίδιας συστάδας και λιγότερο κοινά χαρακτηριστικά μεταξύ διαφορετικών συστάδων.

Σε αυτή την διπλωματική εργασία στο περιβάλλον της python θα μελετηθούν τα αποτελέσματα αλγορίθμων σε σύνολα δεδομένων γράφων τα οποία θα αξιολογηθούν από δείκτες αξιολόγησης οι οποίοι έχουν προταθεί από σχετικές μελέτες. Πιο αναλυτικά θα χρησιμοποιηθούν τρία σύνολα δεδομένων γράφων και θα τμηματοποιηθούν με την χρήση δύο διαδεδομένων αλγορίθμων συσταδοποίησης, Spectral και Louvain. Κάθε αλγόριθμος θα δώσει αποτελέσματα για διαφορετικό αριθμό συστάδων, κάθε αποτέλεσμα θα αξιολογηθεί με την χρήση δύο γνωστών δεικτών, οι οποίοι είναι ο modularity και ο conductance. Ο δείκτης modularity αξιολογεί την πυκνότητα των εσωτερικών συνδέσεων της συστάδας σε σχέση με τις εξωτερικές συνδέσεις με άλλες συστάδες. Ο δείκτης Conductance μιας τομής (conductance of a cut) συγκρίνει το μέγεθος μιας τομής, δηλαδή τον αριθμό των κομμένων άκρων και το βάρος (weight) των ακμών σε οποιονδήποτε από τους δύο υπογράφους που δημιουργούνται από την τομή. Επίσης θα χρησιμοποιηθούν τρείς δείκτες αξιολόγησης οι οποίοι έχουν αναπτυχθεί και προταθεί σε μελέτες σχετικής μελέτης. Ο δείκτης Q-graph χρησιμοποιεί το degeneracy και το density τα οποία αφορούν στην πυκνότητα των κόμβων και των ακμών, για να αξιολογήσει την συνδεσιμότητα μεταξύ των γράφων εντός και εκτός της συστάδας. Ο δείκτης CDS που χρησιμοποιείται επίσης για αξιολόγηση λαμβάνει υπόψη του την πυκνότητα του γράφου και το cohesion το οποίο αξιολογεί το πόσο κοντά είναι οι κόμβοι σε ένα δίκτυο και κατά πόσο μπορούν να διαχωριστούν από άλλους. Ένας ακόμα δείκτης που υπολογίζεται είναι ο GS* ο οποίος βασίζεται στον silhouette index και αξιολογεί την ποιότητα της συσταδοποίησης με βάση την απόσταση των κόμβων.

Οπτικοποιώντας τις συσταδοποιήσεις ενός πολύ γνωστού μικρού συνόλου δεδομένων και θα γίνει προσπάθεια ερμηνείας των αποτελεσμάτων των δεικτών αξιολόγησης και θα παρατηρηθεί πως η δομή των συστάδων επηρεάζει τα αποτελέσματα του. Οι δείκτες αξιολόγησης είναι πολύ σημαντική για την επικύρωση των αποτελεσμάτων, παρόλα αυτά δεν εντοπίζεται ένας ο οποίος μπορεί να είναι αξιόπιστος υπό όλες τις συνθήκες.

Abstract

Graphs is the way to represent networks in mathematics, networks that their structural elements are interconnected. Graph are consisted from nodes that represents the data and edges that represents the relationships between nodes. The structure of a graph gives information about the nodes, the nodes that belongs to the same cluster have more similar characteristics than with nodes that belongs to other clusters.

In this thesis will be studied the results of clustering algorithms in graph data sets by evaluating them with evaluating indices that have proposed in relevant papers. More specifically three graph data sets will be clustered by using two popular clustering algorithms, Spectral and Louvain. Each cluster algorithm will give results for several number of clusters, in each result will be evaluated the quality of clusters with the use of two popular indices, modularity and conductance. The modularity index compares the intra-linkage of a cluster that need to be denser from the inter-linkage of the cluster with the neighbor clusters. The index conductance compares the number of edges cut and their weights that induced from a cut in graph that split the graph to subgraphs. Also there will be used three evaluation indices that have developed and proposed from papers with common field of interest. The index Q-graph that is also calculated uses the degeneracy and graph density that referred to density of nodes and edges, to evaluate the connectivity of nodes in and between clusters. The index CDS that is also used for evaluation of clustering results, contains calculation regarding the structural cohesion and the graph density, cohesion evaluates the distance between the nodes and the possibility of separation. One more index that is calculated is GS* that is based to silhouette index, this index is evaluating the similarity of nodes comparing the distance between nodes belonging in the same cluster and their distance with nodes of neighboring clusters. Python environment includes tools and libraries that will be useful for the experimental study.

To interpret the results, a small but well-known graph data set will be used to visualize the clusters and to observe how the cluster structure affects the results of the cluster validation indices. Cluster validation indices have significant role to ensure the reliability of the results, the current experiment study comes to the conclusion that there is no the best index but a combination that need to be used according to graph structure.

ΚΕΦΑΛΑΙΟ 1 : Εισαγωγή

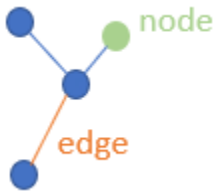
Σε διάφορες επιστημονικές περιοχές όπως μαθηματικά, φυσική, κοινωνιολογία γίνεται αναπαράσταση των δεδομένων σε μορφή γράφου ή δικτύου, για να αναπαρασταθεί και εν συνεχεία αντληθεί πληροφορία. Για πάνω από μια δεκαετία έχουν δημοσιευτεί μελέτες πάνω σε αλγόριθμους ομαδοποίησης [7], ενώ υπάρχουν και μελέτες που αφορούν την αξιολόγηση των αποτελεσμάτων που προκύπτουν. Η ομαδοποίηση είναι ένα πολύ σημαντικό θέμα για την ανάλυση και την εξερεύνηση των δεδομένων [9]. Με τη μέθοδος ομαδοποίηση (clustering) σε γράφους χωρίζονται τα δεδομένα σε κοινότητες/κλάσεις έτσι ώστε τα δεδομένα στο ίδιο σύμπλεγμα να έχουν περισσότερα όμοια χαρακτηριστικά μεταξύ τους παρά με τα δεδομένα ενός άλλου συμπλέγματος [6], [9]. Το αποτέλεσμα της ομαδοποίησης επηρεάζεται από δύο βασικούς παράγοντες ο ένας είναι ο τύπος των δεδομένων (για παράδειγμα κατηγορικά ή αριθμητικά) που επηρεάζουν την δομή του γράφου και ο δεύτερος παράγοντας είναι οι διαφορετικές παραδοχές που χρησιμοποιούνται ως είσοδοι, δηλαδή οι παράμετροι και οι συναρτήσεις κόστους που εισάγονται σε έναν αλγόριθμο και μπορεί να οδηγήσουν σε διαφορετικά αποτελέσματα συστάδων, συμπερασματικά δεν υπάρχει η δυνατότητα επιλογής του καλύτερου αλγορίθμου [3]. Για να αποκτηθεί μια αξιόπιστη πληροφόρηση σχετικά με τις ομάδες/κοινότητες (clusters) ενός γράφου άρα για τα αποτελέσματα των αλγορίθμων ομαδοποίησης σε σύνολα δεδομένων γράφων, χρησιμοποιούνται κάποιοι δείκτες αξιολόγησης που βασίζονται στη βασική παραδοχή για το τί χαρακτηρίζεται κοινότητα σε έναν γράφο και την οποία θα αναφέρουμε και παρακάτω. Αυτοί οι δείκτες αξιολόγησης έχουν σας βασική διεργασία την μέτρηση της πυκνότητας και του διαχωρισμού (compactness και separability) των συστάδων [3]. Στην συνέχεια αυτής της μελέτης στο δεύτερο κεφάλαιο θα γίνει εισαγωγή στην έννοια του γράφου και των δομικών χαρακτηριστικών του, θα γίνει αναφορά στον τρόπο αναπαράστασης των δεδομένων σε πίνακες και λίστες, θα εκτελεστεί ένα παράδειγμα οπτικοποίησης γράφου. Στο τρίτο κεφάλαιο θα αναφερθούν αρκετοί αλγόριθμοι ομαδοποίησης, ενώ στο τρίτο κεφάλαιο θα αναφερθούν μετρικές καθώς και δείκτες αξιολόγησης αποτελεσμάτων που έχουν προκύψει από μελέτες στον τομέα της αξιολόγησης αποτελεσμάτων σε αλγόριθμους ανίχνευσης κοινοτήτων σε γράφους. Στο πέμπτό κεφάλαιο θα γίνει πειραματική μελέτη επιλεγμένων αλγόριθμων ομαδοποίησης και θα αξιολογηθούν με βάση μετρικών που παρουσιάστηκαν σε αντίστοιχης θεματολογίας μελέτες. Στο έκτο κεφάλαιο θα γίνει αξιολόγηση και ερμηνεία των αποτελεσμάτων που θα οδηγήσει σε συμπεράσματα σχετικά με τις μεθόδους αξιολόγησης.

ΚΕΦΑΛΑΙΟ 2 : Η έννοια του γράφου και τα χαρακτηριστικά του

Σκοπός αυτού του κεφαλαίου είναι να παρουσιάσει την μορφή του γράφου και τα χαρακτηριστικά της, καθώς παρουσιάζονται και βασικές έννοιες για την κατανόηση και αξιολόγηση της δομής του. Η θεωρία των γράφων χρονολογείται από τη λύση του Euler του προβλήματος των γεφυρών Konigsberg το 1736 [17]. Έκτοτε έχουν μελετηθεί αρκετά οι γράφοι και οι μαθηματικές τους ιδιότητες. Πλέον είναι εξαιρετικά χρήσιμοι για την αναπαράσταση δεδομένων σε διαφορετικά επιστημονικά πεδία [17]. Ο γράφος είναι μία μη γραμμική δομή δεδομένων που αποτελείται από κόμβους(nodes/vertices) και συνδέσεις(edges/links). Οι γράφοι αντιπροσωπεύουν δίκτυα και χρησιμοποιούνται για την επίλυση πολλών προβλημάτων σε διάφορα επιστημονικά πεδία [18]. Παρακάτω εισάγονται βασικά γνωρίσματα των γράφων.

2.1 Εισαγωγή στην έννοια του γράφου

Ένα δίκτυο αποτελείται από κόμβους (nodes) που αντιπροσωπεύουν τα υπό εξέταση δεδομένα, δηλαδή οντότητες όπως πρόσωπα, οργανισμούς ή απλά αντικείμενα τα οποία συνδέονται με δυαδικές σχέσεις όπως κοινωνικές σχέσεις, εξαρτήσεις ή ανταλλαγές αυτές οι συνδέσεις (edge/links) απεικονίζουν την σχέση του κάθε σημείου με τα υπόλοιπα [8].

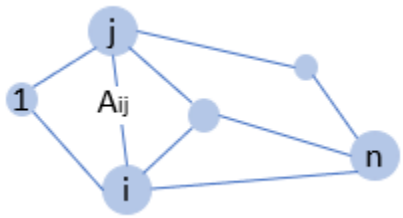


Εικόνα 1. Δομικά στοιχεία γράφου

Στα δίκτυα οι συνδέσεις (edges) μπορεί να είναι κατευθυνόμενες, μη κατευθυνόμενες ή μικτές [22]. Υπάρχουν περιπτώσεις όπου μας ενδιαφέρει η αλληλουχία των κόμβων ώστε να παρατηρηθεί η ροή μιας διεργασίας και περιπτώσεις όπου δεν αντλείται κάποια πληροφορία από την κατεύθυνση αλλά η ανάγκη είναι να διαπιστωθεί η σχέση και όποια ομοιότητα μεταξύ των κόμβων, η μορφή των δεδομένων και η πληροφορία που ενδιαφέρει. Τα χαρακτηριστικά μπορούν να είναι οποιουδήποτε τύπου και τα χαρακτηριστικά αριθμητικής σύνδεσης μπορεί να ενισχύσουν ή να αποδυναμώσουν τη σχέση μεταξύ δύο κόμβων[8].

Για την εξέταση και περαιτέρω μελέτη των ιδιοτήτων και των χαρακτηριστικών ενός γράφου θα χρησιμοποιηθεί η πιο διαδεδομένη περιγραφή που ορίζει την δομή ενός γράφου.

Έστω ένας γράφος όπου G=(V , E), όπου V ={ 1,…n } ο αριθμός των κόμβων και E ο αριθμός των συνδέσεων[1],[3]. Έστω { i , j } ∈ E, οι κόμβοι i,j είναι συνδεδεμένοι και ονομάζονται γείτονες. Ο βαθμός (degree) του κόμβου i είναι ίσος με τον αριθμό των γειτόνων του[15].



Εικόνα 2. Ενδεικτική μορφή γράφου

2.2 Χαρακτηριστικά των ακμών

Αναλόγως με το σύνολο δεδομένων οι συνδέσεις (edges) μπορεί να μεταφέρουν και άλλες πληροφορίες όπως [14]:

- το βάρος (weight) που σε έναν γράφο μπορεί για παράδειγμα να υποδεικνύει την συχνότητα επικοινωνίας μεταξύ κόμβων, δηλώνει την ενέργεια που απαιτείται για την μετακίνηση από κόμβο σε κόμβο.
- την κατάταξη (ranking) που θα μπορούσε να υποδεικνύει την σημαντικότητα που έχει τους κόμβος ως τους έναν άλλο
- το τύπο (type) σύνδεσης των κόμβων, δηλαδή την σχέση μεταξύ των κόμβων
- κάποια ιδιότητα που έχει ο κόμβος μέσα στο δίκτυο σε συνδυασμό με άλλη μετρική

Αναλόγως των πληροφοριών που φέρουν οι συνδέσεις (edges), τους γράφος μπορεί να χαρακτηριστεί κατευθυνόμενος και μη κατευθυνόμενος. Τους κατευθυνόμενους γράφους η αναπαράσταση γίνεται με την χρήση είτε με την βοήθεια τους πίνακα γειτνίασης (adjacency matrix), είτε με λίστα ακμών (edgelist), είτε με λίστα γειτνίασης (adjacency list) [14] .

2.3 Αναπαράσταση δεδομένων

Οι συνδέσεις μπορούν να αναπαρασταθούν με[14],[23]:

Πίνακας γειτνίασης (adjacency matrix), έχει την μορφή πίνακα με 1 και 0, οι δείκτες είναι οι κόμβοι και τα στοιχεία αναπαριστούν την σύνδεση μεταξύ των κόμβων :

- τους πίνακας A_{ij} όπου η διαγώνιος παίρνει την τιμή 0 εάν δεν υπάρχει σύνδεση μεταξύ i και j που υποδηλώνουν τους κόμβους και την τιμή 1 αν υπάρχει σύνδεση μεταξύ των κόμβων [Εικόνα 3,α].

- τους πίνακας A_{ii} όπου η διαγώνιος έχει την τιμή 0 εκτός και αν υπάρχει σύνδεση που ξεκινάει και καταλήγει στον ίδιο κόμβο (self-loops) [Εικόνα 3,β].

- Σε περίπτωση τους μη κατευθυνόμενου γράφου $A_{ij} = A_{ji}$, δηλαδή δεν υπάρχει μία κατεύθυνση που να χαρακτηρίζει των συνδέσεων (edges) μεταξύ των κόμβων (nodes) [Εικόνα 3,γ].

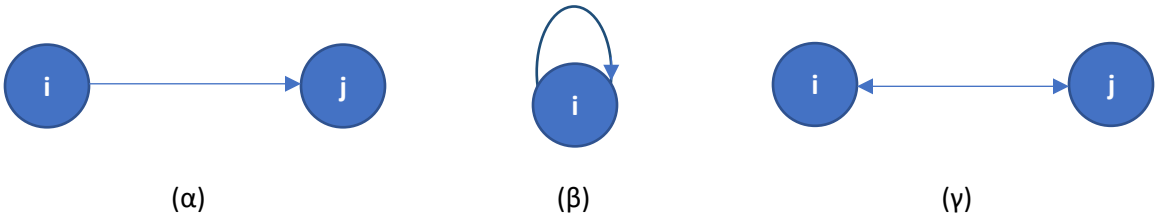
Λίστα ακμών (edgelist),πρόκειται για μία λίστα που καταγράφει κάθε ένα κόμβο σε συνάρτηση με έναν άλλο σε ζευγάρια που υποδεικνύουν την κατεύθυνση τους σύνδεσης[22]:

- i , j σε αυτή την περίπτωση θα έχουμε σύνδεση με κατεύθυνση από τι i στο j [Εικόνα 3,α].

- i , j και παρακάτω στην λίστα j , i σε αυτή την περίπτωση θα έχουμε αμφίδρομης κατεύθυνσης σύνδεση από το i στο j και από το j τους το i [Εικόνα 3,γ].

Λίστα γειτνίασης (adjacency list), είναι χρήσιμη σε μεγάλους γράφους χωρίς πολλές συνδέσεις ανάμεσα στους κόμβους, έτσι δεν δεσμεύεται χώρος για κόμβους οι οποίοι δεν συνδέονται δηλαδή δεν καταγράφεται η τιμή 0. Η λίστα καταγράφει κάθε κόμβο με τους συνδέσεις που έχει ως τους σε τους κόμβους :

- i : j,m,n ο κόμβος i έχει σύνδεση με κατεύθυνση τους τους κόμβους j,m,n και όλη αυτή η πληροφορία καταγράφεται σε μια γραμμή σε σχέση με τους δύο προηγούμενους τρόπους που κάθε σύνδεση του κόμβου i αποτελούσε ξεχωριστή καταγραφή για κάθε άλλο κόμβο που συνδέονταν.



Εικόνα 3. Κατευθύνσεις συνδέσεων

Τα δίκτυα είναι τους φυσικός τρόπος για να αναπαραστήσουν κοινωνικά, βιολογικά, τεχνολογικά και πληροφοριακά συστήματα. Σε αυτά τα δίκτυα ή αλλιώς γραφήματα οι κόμβοι οργανώνονται σε ομάδες με πυκνές συνδέσεις που αναφέρονται ως κοινότητες (communities) ή τμήματα (clusters). Υπάρχουν πολλοί λόγοι που τα δίκτυα οργανώνονται σε πυκνά συνδεδεμένα συμπλέγματα. Για παράδειγμα, η κοινωνία οργανώνεται σε κοινωνικές ομάδες, οικογένειες, χωριά και ενώσεις. Στο World Wide Web, οι σχετικές σελίδες συνδέονται πιο πυκνά μεταξύ τους [5]. Μπορεί τώρα να γίνει κατανοητή η έννοια των παρακάτω ορισμών.

2.4 Χαρακτηρισμός τμημάτων του γράφου

Ισχυρά συνδεδεμένα μέρη (Strongly Connected Components):
Είναι ένα υποσύνολο του γράφου που αποτελείται απο ένα σύνολο κόμβων όπου κάθε κόμβος είναι συνδεδεμένος με τους υπόλοιπους, μέσω των απευθείας συνδέσεων ή ακόμα και μέσω άλλων κόμβων αλλά λαμβάνοντας υπόψιν την κατεύθυνση των συνδέσεων[14],[23].

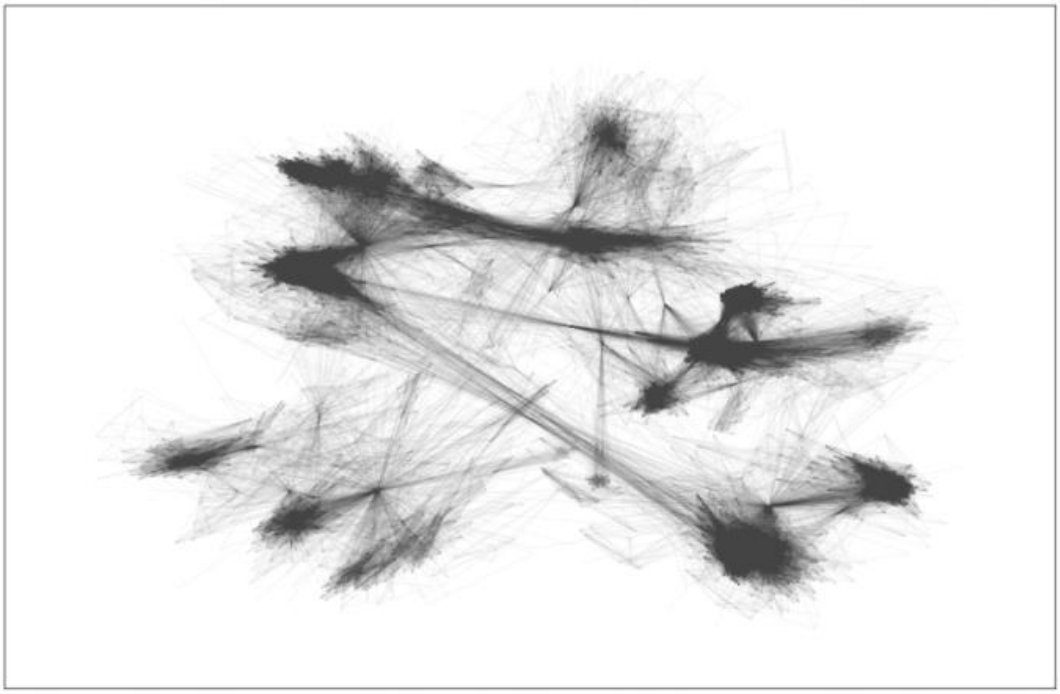
Αδύναμα συνδεδεμένα μέρη (Weakly Connected Components):
Είναι ένα σύνολο κόμβων που ενώνονται μεταξύ τους με συνδέσεις χωρίς να λαμβάνονται υπόψιν οι κατευθύνσεις των συνδέσεων[14],[23].

Παρακάτω παρατίθεται παράδειγμα με την οπτικοποίηση (visualization) ενός μεγάλου δικτύου και η ανίχνευση κοινοτήτων με σκοπό να αποκτηθεί μία εικόνα χρήσιμη για την εισαγωγή στους δείκτες αξιολόγησης και τους αλγόριθμους ομαδοποίησης.

2.5 Οπτικοποίηση δικτύου και ανίχνευση κοινοτήτων

Θα πραγματοποιηθεί η οπτικοποίηση ενός μεγάλου συνόλου δεδομένων [Εικόνα 4] ώστε να γίνει αντιληπτή η μορφή που μπορεί να έχει άλλα και οι φυσικές κοινότητες που παρουσιάζει η δομή του[11].

Με την χρήση ενός εργαλείου της python για δημιουργία, διαχείριση και μελέτη της δομής και των χαρακτηριστικών πολύπλοκων δικτύων του NenworkX , θα γίνει οπτικοποίηση του συνόλου δεδομένων Social circles: Facebook [11],[12] και η ανίχνευση φυσικών κοινοτήτων. Το συγκεκριμένο σύνολο δεδομένων είναι διαθέσιμο στην πλατφόρμα ανάλυσης δικτύων του Στάνφορντ (Stanford) [13]. Το σύνολο δεδομένων του Facebook συλλέχθηκαν από συμμετέχοντες στην έρευνα χρησιμοποιώντας αυτήν την εφαρμογή. Στο σύνολο δεδομένων οι κόμβοι αντιπροσωπεύουν προφίλ χρηστών. Περιέχονται διανύσματα χαρακτηριστικών τα οποία όμως έχουν έτσι ώστε το αρχικό σύνολο δεδομένων μπορεί να περιείχε ένα χαρακτηριστικό "political=Δημοκρατικό Κόμμα", τα νέα δεδομένα θα περιέχουν απλώς "political=anonymized χαρακτηριστικό 1". Έτσι, χρησιμοποιώντας τα ανώνυμα δεδομένα είναι δυνατό να προσδιοριστεί εάν δύο χρήστες έχουν τις ίδιες πολιτικές πεποιθήσεις, αλλά όχι τι αντιπροσωπεύουν οι μεμονωμένες πολιτικές πεποιθήσεις τους [11], [12]. Το δίκτυο αποτελείται από 4039 κόμβους και 88234 συνδέσεις, επίσης στα χαρακτηρίστηκα που είναι διαθέσιμα υπάρχουν ιδιότητες όπως ο αριθμός των κόμβων και των συνδέσεων που απαρτίζουν τα Strongly connected components και τα Weakly connected components στα οποία έγινε αναφορά παραπάνω.



Εικόνα 4. Οπτικοποίηση συνόλου δεδομένων του γράφου Facebook

Παρατηρείται ότι με μία προσεκτική παρατήρηση στο γράφο μπορεί να γίνει μια εκτίμηση σχετικά με τον αριθμό των κοινοτήτων άρα και των ομάδων όπου οι κόμβοι παρουσιάζουν ίδιες ιδιότητες.

Στο επόμενο βήμα θα χρησιμοποιηθεί ο αλγόριθμος greedy_modularity_communities [11], αυτός ο αλγόριθμος ξεκινά με κάθε κόμβο στην δική του κοινότητα και ομαδοποιεί τα ζευγάρια των κοινοτήτων που έχουν αυξημένο modularity μέχρι να μην υπάρχει αντίστοιχο ζευγάρι ή έως ότου επιτευχθεί ο αριθμός των κοινοτήτων (n_communities). Το αποτέλεσμα που προέκυψε είναι ότι το δίκτυο απαρτίζεται από 16 κοινότητες.

Το παραπάνω αποτελεί ένα παράδειγμα της μελέτης της δομής ενός γράφου, παρόλα αυτά το πρόβλημα του εντοπισμού κοινοτήτων (community detection) είναι πολύ πιο σύνθετο, με πολλές μεταβλητές που επηρεάζουν το αποτέλεσμα όπως τα χαρακτηρίστηκα του συνόλου δεδομένων που θα χρησιμοποιηθεί και οι παράμετροι που θα χρησιμοποιηθούν στον εκάστοτε αλγόριθμό. Τέλος η αξιολόγηση της αποτελεσματικότητας της διαδικασίας που ακολουθήθηκε, εξετάζοντας τα αποτέλεσμα με τον υπολογισμό δεικτών αξιολόγησης. Για παράδειγμα όσο πιο συμπαγής είναι μια συστάδα τόσο πιο σκοτεινό είναι το φόντο της. Τα μέτρα αξιολόγησης είναι απαραίτητα για να κατανοήσουμε και να διαχειριστούμε την κατάτμηση του γράφου[2].

ΚΕΦΑΛΑΙΟ 3 : Αλγόριθμοι συσταδοποίησης

Οι γράφοι είναι δυνατόν να αποτελούνται από χιλιάδες ή και εκατομμύρια κόμβους και συνδέσεις [17]. Η επιστήμη της ανάλυσης δεδομένων πλέον προσφέρει μεγάλο αριθμό μεθόδων υπολογισμού και αξιοποίησης των δεδομένων των γράφων. Ιδιαίτερα στα μεγάλα σύνολα δεδομένων η ανίχνευση κοινοτήτων είναι χρήσιμη στον εντοπισμό μοτίβων και ομοιοτήτων μεταξύ κόμβων που αντιπροσωπεύουν τα δεδομένα. Η διαδικασία της συσταδοποίησης είναι από τις πιο σημαντικές εργασίες στον τομέα της εξόρυξης δεδομένων, για τον εντοπισμό ενδιαφέροντών κατανομών στα υποκείμενα δεδομένα. Το αντικείμενο της συσταδοποίησης αφορά στην κατάτμηση του γράφου σε κοινότητες (clusters) με τέτοιο τρόπο που τα δεδομένα που ανήκουν στην ίδια ομάδα να παρουσιάζουν περισσότερες ομοιότητες μεταξύ τους παρά με τα δεδομένα άλλων ομάδων [19]. Για παράδειγμα σε μια βάση δεδομένων όπου καταγράφονται ασθενείς μιας υποκείμενης νόσου, θα μπορούσαν να αναγνωριστούν ομάδες ασθενών με κοινές συνήθεις. Η διεργασία της συσταδοποίησης έχει σαν σκοπό την οργάνωση προτύπων σε «λογικές» ομάδες για την ανάδειξη ομοιοτήτων ή διαφορών και τελικά την εξαγωγή χρήσιμων συμπερασμάτων[19]. Υπάρχουν αρκετές προσεγγίσεις όσον αφορά στη κατηγοριοποίηση των αλγορίθμων συσταδοποίησης, οι περισσότερες από αυτές ταξινομούν τους αλγόριθμους σε δύο κύριες κατηγορίες εκείνους που χρησιμοποιούν «ιεραρχική» (hierarchical) και εκείνους που χρησιμοποιούν επιμεριστική (partitional) διαδικασία [20]. Υπάρχει όμως μία μεγάλη γκάμα αλγορίθμων οι οποίοι ταξινομούνται σε κατηγορίες ανάλογα με τον τρόπο που επεξεργάζονται τα δεδομένα και φέρουν το αποτέλεσμα των συστάδων. Προκειμένου να γίνει αντιληπτός ο τρόπος επεξεργασίας των συνόλων δεδομένων με σκοπό την ομαδοποίηση, οι αλγόριθμοι ταξινομούνται στις παρακάτω κατηγορίες.

3.1 Hierarchical algorithms

Εξ ου και η ονομασία οι ιεραρχικοί αλγόριθμοι διαιρούν το σύνολο των περιπτώσεων σε ένα έναν αριθμό συστάδων που έχουν ιεραρχική δομή. Η ιεραρχική δομή σημαίνει ότι κάθε σύμπλεγμα είναι ένα υποσύνολο ενός άλλου συμπλέγματος με εξαίρεση ένα το οποίο περιέχει όλα τα άλλα. Αυτή η δομή είναι εύκολο να αναπαρασταθεί οπτικά με την βοήθεια ενός δενδρογράμματος για να παρατηρηθούν τα αποτελέσματα της ιεραρχικής ομαδοποίησης [20].

Οι ιεραρχικοί αλγόριθμοι εμπεριέχουν άλλες δύο υποκατηγορίες με βάση την δημιουργία των συστάδων, οι οποίες είναι οι αθροιστικοί (agglomerative) και διαιρετικοί (divisive) αλγόριθμοι [19],[20].

Agglomerative algorithms

Η διαδικασία εκκινείτε αντιστοιχώντας κάθε ακμή σε ένα σύμπλεγμα και στην συνέχεια συγχωνεύουν τα πλησιέστερα συμπλέγματα σε μεγαλύτερα.

Divisive algorithms

Σε αντίθεση με τους αθροιστικούς αλγόριθμους οι διαιρετικοί εκχωρούν όλες τις ακμές σε ένα σύμπλεγμα και στην συνέχεια το διαιρούν σε υποσύνολα.

3.2 Louvain algorithm

Ο Louvain είναι ένας πολύ δημοφιλής αλγόριθμος ο οποίος βασίζεται στο modularity. Ο αλγόριθμος δρα σε 2 φάσεις. Ξεκινώντας κάθε κόμβος εντάσσεται σε μία κοινότητα. Με τυχαίο ή όχι τρόπο οι κόμβοι εξετάζονται κυκλικά. Υπολογίζεται η αύξηση του δείκτη modularity κατά την μετακίνηση ενός κόμβου σε μια γειτονική κοινότητα. Ο κόμβος μετακινείται στην κοινότητα που έχει ως αποτέλεσμα την μεγαλύτερη αύξηση του modularity [15]. Ο δείκτης modularity για έναν μη κατευθυνόμενο γράφο μπορεί να υπολογιστεί από τον παρακάτω τύπο [25]:

$$\Delta Q = \frac{k_{i,in}}{2m} - \gamma \frac{\Sigma_{tot} \cdot k_i}{2m^2} \quad (1)$$

Όπου m είναι το μέγεθος του γράφου, $k_{i,in}$ είναι το άθροισμα των βαρών των ακμών για κάθε κόμβο i προς τους υπόλοιπους κόμβους της κοινότητας C, το άθροισμα των βαρών των ακμών που είναι προσκείμενες στον κάθε κόμβο i αντιπροσωπεύεται από το k_i, και Σ_{tot} είναι το άθροισμα των ακμών που είναι προσκείμενες στους κόμβους της κοινότητας C και γ είναι η παράμετρος που επηρεάζει τον αριθμό των συστάδων γνωστή ως resolution [25].

Όταν δεν έχει απομείνει κόμβος προς μετακίνηση, ξεκινά η δεύτερη φάση. Η δεύτερη φάση είναι γνωστή ως φάση συνάθροισης (aggregation phase), οι κόμβοι κάθε κοινότητας συγχωνεύονται ώστε κάθε κοινότητα να έχει ένα σταθερό συνολικό βάρος (weight). Όποια σύνδεση υπήρχε μεταξύ κοινοτήτων αντιπροσωπεύεται με σύνδεση που ξεκινάει και καταλήγει στον ίδιο κόμβο (self-loops) [15].

3.3 Spectral algorithm

Ο Spectral είναι ένας επίσης γνωστός αλγόριθμος ο οποίος χρησιμοποιεί τα ιδιοδιανύσματα και τις ιδιοτιμές του πίνακα γειτνίασης για τον υπολογισμό των συστάδων[4]. Δέχεται μία είσοδο k που αντιπροσωπεύει τον αναμενόμενο αριθμό συστάδων και η διεργασία που ακολουθεί μπορεί να αναλυθεί σε δύο φάσεις [15].
Κατά την πρώτη φάση υπολογίζονται ο αριθμός k ιδιοδιανυσμάτων u_i , … , u_k του Λαπλασιανού πίνακα (Laplacian matrix) L.

$$L = D - A \quad (2)$$

Όπου A ο πίνακας γειτνίασης και D ο πίνακας με τους βαθμούς των ακμών στην διαγώνιο. Έστω πίνακας U που περιέχει τις ακμές u_1 , … , u_k ως στήλες και $U \in \mathbb{R}^{n\times k}$.

Κατά την δεύτερη φάση κάθε κόμβος απεικονίζεται στις γραμμές του πίνακα U όπου

$$y_i, i \in [1, n]. \quad (3)$$

Έτσι η ομαδοποίηση κάθε κόμβου πραγματοποιείται μέσο μιας συτσταδοποίησης K-means που εκτελείται στο yi [15]. Ο αλγόριθμος Spectral έχει διαφορετικά αποτελέσματα αναλόγως τον Λαπλασιανό πίνακα (Laplacian matrix) που χρησιμοποιείται αναλόγως τις παραμέτρους (Symmetric, random walk) [15].

3.4 Normalized Cut algorithm

Είναι μια μέθοδος που έχει προταθεί από τους Shi and Malik [26], βρίσκει την καταλληλότερη συσταδοποίηση τελειοποιώντας την συνάρτηση της τομής (cut). Ορίζοντας το κόστος της τομής (A,B) που χωρίζει τους κόμβους V ενός γράφου G = (V,E) σε δύο τμήματα A,B | A ∪ B = V, A ∩ B = Ø, το άθροισμα των βαρών των ακμών ενώνει τους κόμβους των τμημάτων A και B. Το επιθυμητό είναι να βρεθεί η τομή που ελαχιστοποιεί το κόστος :

$$cut(A, B) = \left(\frac{1}{Vol(A)} - \frac{1}{Vol(B)}\right) \quad (4)$$

Όπου ο όγκος (Vol) του κάθε σετ που είναι το άθροισμα των βαρών των ακμών με μία τουλάχιστον ένα τελικό σημείο μέσα του [4].

Αυτή η συνάρτηση κόστους έχει σχεδιαστεί για να τιμωρεί τις περικοπές που δημιουργούν υποσύνολα με πολύ διαφορετικά μεγέθη [4]. Έτσι ελαχιστοποιώντας το κόστος δημιουργούνται υποσύνολα με χαμηλότερη ομοιότητα μεταξύ τους και με ιδανικά περισσότερη ομοιότητα των κόμβων με τους υπόλοιπους κόμβους της συστάδας όπου ανήκουν [4]. Για την εκτέλεση του αλγόριθμου απαιτείται ως παράμετρος εισόδου ο αριθμός των επιθυμητών συστάδων.

3.5 Markov algorithm

Ο αλγόριθμος Markov βασίζεται στην προσομοίωση των στοχαστικών ροών σε ένα γράφο [4], [27], [28]. Η κύρια ιδέα του αλγόριθμου είναι ότι το βασικό χαρακτηριστικό ενός γράφου είναι η απόσταση των κόμβων του. Οι μικρές αποστάσεις μεταξύ των κόμβων του κατατάσσει στην ίδια συστάδα, ενώ οι μεγαλύτερες αποστάσεις τους διαχωρίζει σε συστάδες. Για να εντοπίσει συστάδες ο αλγόριθμος βασίζεται στην πεποίθηση ότι μια τυχαία περιήγηση ξεκινώντας από

έναν κόμβο είναι πιο πιθανό να περιοριστεί μέσα στην ίδια συστάδα παρά να οδηγήσει σε κάποια γειτονική [4].

Η διαδικασία που ακολουθεί ο αλγόριθμος χωρίζεται σε δύο βήματα, επέκταση (expansion) και πληθωρισμός (inflation). Το βήμα της επέκτασης (expansion) ο αλγόριθμος χρησιμοποιεί πολλαπλασιασμό για την κανονικοποίηση του πίνακα γειτνίασης που αντιπροσωπεύει τον γράφο. Στο βήμα του πληθωρισμού (inflation) χρησιμοποιείται η έννοια Hadamard του διευρυμένου πίνακα, που ακολουθείται από ένα βήμα κλιμάκωσης για να γίνει ξανά ο πίνακας στοχαστικός, με τα στοιχεία κάθε στήλης να αντιστοιχούν σε μια τιμή πιθανότητας [4]. Η παράμετρος που απαιτείται για την εκτέλεση του αλγόριθμου ονομάζεται τιμή πληθωρισμού (inflation) όσο χαμηλότερη είναι η τιμή τόσο πιο μεγάλα τα τμήματα που θα μελετηθεί ο γράφος [4].

3.6 Bisecting K-means algorithm

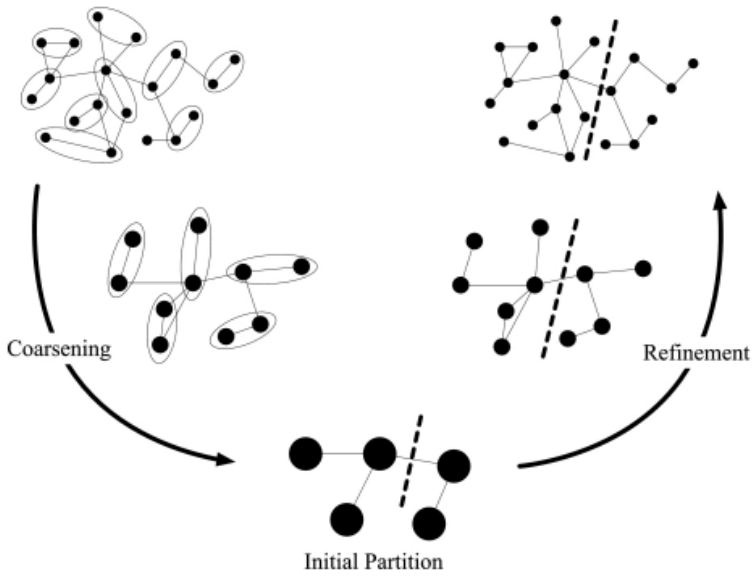
Στον παραδοσιακό αλγόριθμο K-means ο εισαγόμενος αριθμός k δηλώνει τον αριθμό των κέντρων (centroids) των συστάδων, οι κόμβοι ομαδοποιούνται αναλόγως με την κοντινότερη απόσταση που έχουν από τα κέντρα. Έτσι δημιουργούνται οι αρχικές συστάδες, υπολογίζονται τα νέα κέντρα τα οποία επηρεάζουν την συσταδοποίηση. Η διαδικασία συνεχίζεται έως ότου τα κέντρα δεν μετατοπίζονται άλλο πια [4].

Ο αλγόριθμος Bisecting K-means διαφέρει από τον παραδοσιακό K-means αλγόριθμο αντιμετωπίζοντας το γράφο ως μία συστάδα την οποία διχοτομεί χρησιμοποιώντας K-means και η αποστάσεις μεταξύ των κόμβων χρησιμοποιούνται ο συνάρτηση ομοιότητας μεταξύ των κόμβων. Η διαδικασία συνεχίζεται με την επιλογή ενός νέου τμήματος προς διχοτόμηση έως ότου συμπληρωθεί ο αριθμός συστάδων [4].

3.7 METIS algorithm

Ο αλγόριθμος METIS είναι ένα σχήμα κατάτμησης πολλαπλών επιπέδων γράφου, βασίζεται στις τομές (cut-based) k-way δηλαδή οι ακμές εκείνες που εάν θα μετακινηθούν θα τμηματοποιήσουν τον γράφο σε τουλάχιστον k συνδεδεμένα τμήματα. Σε αντίθεση με άλλους αλγόριθμους πολυεπίπεδης αναδρομικής διχοτόμησης (MultiLevelRecursive Bisection / MLRB) αρχικά αδρανοποιεί το γράφημα εισόδου δημιουργώντας υπερκόμβους από τους κόμβους που συνδέονται και επαναλαμβάνει την διαδικασία έως ότου μειωθεί σημαντικά το μέγεθος του γράφου. Η συμπύκνωση γίνεται με τέτοιο τρόπο ώστε να διατηρηθούν οι κομμένες ακμές (edge-cut). Το συμπυκνωμένο γράφημα τμηματοποιήται με την χρήση MLRB και προβάλλεται πάνω στον αρχικό γράφο εισόδου. Αποσυμπιέζεται σταδιακά και οι κόμβοι ανταλλάσσονται μεταξύ των συστάδων με την χρήση του αλγόριθμου Kernighan-Lin και με γνώμονα την μείωση της κοπής των άκρων. Από την διαδικασία παράγεται ο αριθμός k διαμερισμάτων [30]. Παρακάτω

παρατίθεται σχέδιο που βοηθάει στην χαρτογράφηση της διαδικασίας της σύμπτυξης και αποσυμπιεσής του γράφου.



Εικόνα 5. Ακολουθία Αλγόριθμου συσταδοποίησης Metis

Πηγή εικόνας [5] : Timothy A. Davis, William W. Hager, Scott P. Kolodziej, and S. Nuri Yeralan. 2020. Algorithm 1003: Mongoose, a Graph Coarsening and Partitioning Library.

Παρατηρείται ποικιλία στους αλγόριθμους συσταδοποίησης όσον αφορά στον τρόπο με τον οποίο κατηγοριοποιούνται οι κόμβοι σε συστάδες. Δύο αλγόριθμοι οι οποίοι συναντώνται σε αντίστοιχες μελέτες, έχουν διαφορετικό τρόπο συσταδοποίησης μεταξύ τους και είναι ενδιαφέρον να παρατηρηθούν τα αποτελέσματα τους είναι ο αλγόριθμος Spectral και Louvain. Ο αλγόριθμος Spectral δέχεται στην είσοδο τον επιθυμητό αριθμό συστάδων, χρησιμοποιεί πληροφορίες από τις ιδιοτιμές (φάσμα) ειδικών πινάκων που έχουν δημιουργηθεί από το γράφημα, κατατάσσει τους κόμβους στις συτστάδες. Ο αλγόριθμος Louvain όπως αναφέρθηκε χωρίζεται σε δύο φάσεις, μία κατά την οποία κατατάσσει τους κόμβους με βάση την αύξηση του δείκτη Modularity και μία όπου εκτελείται συγχώνευση κόμβων κάθε κοινότητας. Στο πειραματικό μέρος θα παρουσιαστούν τα αποτελέσματα των δεικτών αξιολόγησης με βάση τις συσταδοποιήσεις των παραπάνω αλγορίθμων.

ΚΕΦΑΛΑΙΟ 4 : Δείκτες αξιολόγησης αποτελεσμάτων συσταδοποίησης

Σε αυτό το κεφάλαιο θα παρουσιαστούν δείκτες αξιολόγησης των αποτελεσμάτων των αλγορίθμων συσταδοποίησης. Ένα πλήθος από δείκτες αξιολόγησης έχουν αναπτυχθεί στην ανάλυση δεδομένων. Η αρχή είναι η σύγκριση μεταξύ του intra-cluster και inter-cluster με την χρήση του γνωστού θεωρήματος του Huygens : Η καθολική μεταβλητότητα είναι ο διαχωρισμός των μεταβολών μεταξύ intra και inter-cluster (εντός και μεταξύ συστάδων) [2]. Αναλόγως την προσέγγιση που ακολουθείται στους παρακάτω δείκτες μελετάται η συμπαγή(compactness) σύνδεση ή η αραιή(separability) σύνδεση των κόμβων με την χρήση μεθόδων ανάλυσης βασισμένες στην διακύμανση(variance) ή την πυκνότητα(density). Η πλειοψηφία των δεικτών αξιολόγησης των συστάδων εφαρμόζεται στον Ευκλείδειο χώρο [1].

4.1 Δείκτες αξιολόγησης

Η πιο αποδεκτή έννοια της συστάδας περιγράφει τους κόμβους της ίδιας συστάδας να έχουν παρόμοια χαρακτηριστικά μεταξύ τους, παρά με κόμβους που δεν ανήκουν στην ίδια συστάδα με τους οποίους θα έχουν και αραιότερες συνδέσεις [4]. Ο στόχος της συσταδοποίησης γράφων είναι να οριστούν συμπαγείς συστάδες και καλώς διαχωρισμένες μεταξύ τους [2], οι δείκτες αξιολόγησης που εφαρμόζονται στα αποτελέσματα των αλγορίθμων συσταδοποίησης, είναι το εργαλείο που χρησιμοποιείται ώστε να αποκτηθεί η γνώση της δομής του γράφου άρα και η γνώση των πραγματικών χαρακτηριστικών των δεδομένων. Σε αυτό το κεφάλαιο θα γίνει αναφορά στους πιο δημοφιλείς δείκτες αξιολόγησης ομαδοποίησης, αλλα και σε δείκτες αξιολόγησης που έχουν προταθεί σε μελέτες αξιολόγησης αλγορίθμων εντοπισμού κοινοτήτων σε γράφους.

4.2 Ορισμός γράφου

Όπως προαναφέρθηκε έστω ένας γράφος G=(V,E), όπου V οι κόμβοι και E = (u, v) | u, v ∈ V οι συνδέσεις. Σε έναν μη κατευθυνόμενο γράφο ή όταν δεν υπάρχει σχετική πληροφόρηση θεωρείται (u, v) = (v, u). Ο αριθμός των συνδέσεων του γράφου G είναι |E(G)|=m, και ο αριθμός των συνδέσεων σε έναν συγκεκριμένο κόμβο v αντιπροσωπεύει όπως προαναφέρθηκε τον βαθμό deg(v) [4]. Οι συνδέσεις ενδέχεται να φέρουν τιμές weight w(u, v). Σε περιπτώσεις που δεν αναφέρονται βάρη (weight) θεωρείται ότι w(u, v) =1 για κάθε (u, v) ∈ E. Θεωρείται $E(C_i, C_j)|i\neq j$ οι συνδέσεις μεταξύ των συστάδων C_i και C_j και $E(C_i)$ ως το σύνολο των συνδέσεων (u, v)|u,v ∈ Ci. Τότε E(C) είναι το σύνολο όλων των εσωτερικών συνδέσεων όλων των συστάδων στο C και $\bar{E}(C)$ είναι το σύνολο όλων των συνδέσεων μεταξύ των συστάδων σε ένα γράφο$((u,v)|u\in C_i, u\in C_j, i\neq j)$ [4].

4.2.1 Modularity

Ο πιο δημοφιλής δείκτης αξιολόγησης αποτελεί τον φυσικό ορισμό μιας καλής ομαδοποίησης, δηλαδή μια ομαδοποίηση όπου οι εσωτερικές συνδέσεις(intra_linkage) των κόμβων είναι πιο πυκνές σε σχέση με τις εξωτερικές συνέσεις(inter_linkage) με άλλες ομάδες[3],[4],[15]. Το modularity Q δίνεται από την παρακάτω ισότητα όπου e είναι ένας συμμετρικός πίνακας του οποίου τα στοιχεία e_{ij} όλων των ακμών στο δίκτυο στις κοινότητες i και j, και Tr(e) είναι το ίχνος του πίνακα e, δηλαδή το άθροισμα των στοιχείων από την κύρια διαγώνιο του[4].

$$Q = Tr(e) - \|e^2\| \tag{5}$$

Ο δείκτης modularity Q μας επιστρέφει τιμές από 0 έως 1, όπου το 1 αντιπροσωπεύει μια ομαδοποίηση με πολύ ισχυρά χαρακτηριστικά κοινοτήτων. Παρόλα αυτά σε κάποιες σπάνιες περιπτώσεις το modularity αποδίδει αρνητικές τιμές, σε αυτές τις περιπτώσεις οι συστάδες έχουν 0 εσωτερικές συνδέσεις [4].

4.2.2 Conductance

Ο δείκτης Conductance μιας τομής (conductance of a cut) συγκρίνει το μέγεθος μιας τομής, δηλαδή τον αριθμό των κομμένων άκρων και το βάρος (weight) των ακμών σε οποιονδήποτε από τους δύο υπογράφους που δημιουργούνται από την τομή. Ο δείκτης conductance ϕ(G) συμβολίζει την τιμή της ελάχιστης συνδεσιμότητας όλων των ομάδων[4].

Θεωρείται μια τομή που χωρίζει τον γράφο G σε k μη αλληλοκαλυπτόμενες (non-overlapping) συστάδες C_1, C_2… C_k. Ο δείκτης Conductance για κάθε συστάδα ϕ(C_i) μπορεί να δοθεί από την παρακάτω ισότητα, όπου $a(C_i) = \sum_{u \in c_i} \sum_{u \in V} w\,(u, v)$ είναι το άθροισμα των βαρών όλων των ακμών από τις οποίες έστω η μία άκρη τους ανήκει στο C_i [4].

$$a(C_i) = \frac{\sum_{u \in c_i} \sum_{u \notin c_i} w(\{u,v\})}{min(\alpha(Ci),\alpha(\bar{C_i}))} \tag{6}$$

Εφόσον το ζητούμενο είναι να βρεθεί ο αριθμός k συστάδων, θα χρειαστούν k-1 τομές για να επιτευχθεί ο εκάστοτε αριθμός k. Ο δείκτης Conductance για την συνολική συσταδοποίηση μπορεί να θεωρηθεί το μέσο όρο των k-1 τομών όπως φαίνεται και στην παρακάτω ισότητα[4]:

$$\varphi(G) = \ avg(\varphi(\mathrm{C_i})),\ \ C_i \subseteq \ V \tag{7}$$

Ο δείκτης Conductance έχει την εξής ιδιότητα $0 \leq \phi(C_i) \leq 1$, όπου το αποτέλεσμα 0 υποδεικνύει ότι όλες οι ακμές των κόμβων είναι εντός της συστάδας , άρα ένα αποτέλεσμα όσο πιο κοντά στο 0 θεωρείται ιδανικό [24].

4.2.3 Silhouette Index

Η μετρική Silhouette Index αξιολογεί την ομοιότητα των κόμβων με βάση την απόσταση και χρησιμοποιεί έννοιες όπως της συνοχής (cohesion) και του διαχωρισμού (separation) [4].Ο συγκεκριμένος δείκτης βασίζεται στην γειτνίαση του κόμβου, μπορεί να προσδιοριστεί δηλαδή εάν ένας κόμβος είναι σωστά ομαδοποιημένος σε μια συστάδα ή όχι και να γίνει ανακατάταξη ή απόρριψη [2]. Ένας κόμβος v_i ο οποίος ανήκει στη συστάδα C_i, θεωρείται ότι με βάση το μέσο όρο απόστασης κοντά στη συστάδα C_h. Ο δείκτης Silhouette Index ορίζεται ως [2]:

$$s(v_i) = \frac{d(v_i,C_h) - d(v_i,C_i)}{max(d(v_i,C_i),d(v_i,C_h)} \quad (8)$$

Η μετρική Silhouette Index $s(v_i)$ έχει ως αποτέλεσμα τιμές από το -1 έως το 1. Οι τιμές κοντά στο 1 συμβολίζουν καλώς ομαδοποιημένους κόμβους, για τους κόμβους που η τιμή του δείκτη είναι κοντά στο 0 οι κόμβοι θα πρέπει να ομαδοποιηθούν στις πιο κοντινές τους συστάδες [2]. Ο υπολογισμός του δείκτη για μια συστάδα C_i δίνεται από τον παρακάτω τύπο [2] :

$$C_i = \frac{\sum_{i=1}^{n_i} s(v_i)}{n_i} \quad (9)$$

Όπου v οι συνδέσεις και n οι κόμβοι.

Η συνολική μετρική Silhouette Index ενός γράφου GS υπολογίζεται από τον παρακάτω τύπο [2]:

$$SG = \frac{\sum_{i=1}^{K} s_i}{K} \quad (10)$$

Στην συγκεκριμένη μελέτη [2], προτείνεται μια τροποποιημένη εκδοχή του συγκεκριμένου δείκτη η οποία λαμβάνει υπόψη το μέγεθος της συστάδας :

$$GS^* = \frac{\sum_{j=1}^{k} N_j S_j}{\sum_{j=1}^{k} N_j} = \frac{\sum_{i=1}^{N} S(v_i)}{N} \quad (11)$$

Μία μεγάλη συστάδα έχει πιο σημαντική συνεισφορά από μία μικρότερη σε αυτόν τον δείκτη.

4.2.4 Q-graph Index

Στην μελέτη [3] παρουσιάζεται ένας δείκτης αξιολόγησης ο οποίος βασίζεται χρησιμοποιείται για την αξιολόγηση των αποτελεσμάτων ομαδοποίησης σε γράφους.

Μελετιούνται οι παρακάτω δείκτες που αφορούν την συνδεσιμότητα του γράφου :

Για την μελέτη της συνδεσιμότητας των κοινοτήτων στον εκάστοτε γράφο, χρησιμοποιείται ο δείκτης degeneracy deg(G) ενός γράφου G που αντιπροσωπεύει την ύπαρξη υπογράφων όπου κάθε κόμβος έχει τουλάχιστον deg(G) γείτονες. Ο δείκτης degeneracy πιο συγκεκριμένα αφορά στον μέγιστο αριθμό των κόμβων v που ανήκουν σε ένα σύνολο κόμβων V.

$$\text{Degeneracy : } deg(G) = k_{max} - core = max_{v \in V} core(v) \quad (12)$$

Τότε ο δείκτης degeneracy μπορεί να θεωρηθεί ως μέτρο της αραιότητας του γράφου[3]. Όπως αναφέρεται και στην έρευνα [3] υπάρχουν περιπτώσεις όπου ενώ ο δείκτης degeneracy έχει όμοιο αποτέλεσμα για 2 υπογράφους, το αποτέλεσμα της κάλυψης του δείκτη degeneracy-core είναι διαφορετικό, ο δείκτης degeneracy-core αποτελεί έναν δείκτη για τον μέγιστο αριθμό κόμβων που απαρτίζουν τον πυρήνα του γράφου.

$$\text{Degeneracy-core : } dG = (dV, dE), dV \subset V \text{ και } dE \subset E \quad (13)$$

Έτσι ορίζεται η κάλυψη του δείκτη degeneracy-core : $deg\_coverage(G) = \frac{|dV|}{|V|}$ (14)

Για να αξιολογηθεί η συνδεσιμότητα μίας συστάδας c_i χρησιμοποιώντας τους παραπάνω δείκτες ορίζεται το intra_linkage που αφορά την συνδεσιμότητα των κόμβων εντός της συστάδας και αποτυπώνεται στην παρακάτω ισότητα :

$$intra\_linkage(c_i) = deg(c_i) * deg\_coverage(c_i) \quad (15)$$

Έστω ένας γράφος G, τμηματοποιημένος σε m κοινότητες C = {c_1,…c_m}, η συνδεσιμότητα μέσα στο τμήμα C ορίζεται ως ο μέσος όρος intra_linkage όλων των κοινοτήτων στο τμήμα C.

$$intraLink(C) = \frac{1}{m} \sum_{c_i \in C} intra\_linkage(c_i) \quad (16)$$

Στην συνέχεια τις δημοσίευσης [3] μελετάται μια μετρική που αξιολογεί την συνδεσιμότητα μεταξύ δύο κοινοτήτων c_i και c_j ορίζεται με βάση τις παραπάνω μετρικές που ορίστηκαν degeneracy και την κάλυψη του deg_coverage. Η μετρική ονομάζεται inter_linkage και αποδίδεται από την παρακάτω ισότητα:

$$inter\_linkage(c_i,c_j) = deg(G(c_i,c_j)) * deg\_coverage(G(c_i,c_j)) \quad (17)$$

Συμπερασματικά ο δείκτης inter_linkage ενός τμήματος σε έναν γράφο G που αποτελείται από m κοινότητες αποτυπώνεται στην παρακάτω ισότητα:

$$inter\_linkage(C) = \frac{2}{m*(m-1)} \sum_i \sum_{j,j<i} inter\_linkage(G(c_i,c_j)) \quad (18)$$

Ένα τμήμα του γράφου του οποίου η δομή αντιπροσωπεύει με τον πιο ακριβή τρόπο την δομή του γράφου G, αναμένεται να περιέχει συστάδες με υψηλό δείκτη intra linkage και χαμηλό inter linkage.

Ο δείκτης inter-density μεταξύ δύο συστάδων c_i και c_j σε έναν γράφο G, υποδεικνύει το ποσοστό των ακμών μεταξύ τους :

$$interDens(c_{i,}c_j) = \frac{|E(c_{i,}c_j)|}{|V_i||V_j|} \quad (19)$$

Η συνδεσιμότητα(connectivity) μεταξύ δύο συστάδων c_i και c_j σε έναν γράφο G, αποτυπώνεται στην παρακάτω ισότητα και αναμένεται μια χαμηλή τιμή του δείκτη inter-Connectivity για να θεωρηθούν καλώς διαχωρισμένες οι συστάδες :

$$interCon(c_{i,}c_j) = \frac{InterDens(c_{i,}c_j)}{min\{dens(c_i),dens(c_j)\}} \quad (20)$$

Έτσι το μέτρο διαχωρισμού δύο συστάδων διαμορφώνεται ως εξής :

$$Separation(C) = \frac{2}{m*(m-1)} \sum_i \sum_j \frac{1}{InterCon(c_{i,}c_j)} \quad (21)$$

Χρησιμοποιώντας τα παραπάνω δημιουργείται ο δείκτης αξιολόγησης Q-graph , όσο πιο μεγάλη η τιμή του τόσο πιο καλή η ποιότητα της ομαδοποίησης. Λαμβάνεται υπόψη η εσωτερική πυκνότητα των συστάδων σε σχέση με την πυκνότητα που έχουν οι ακμές μεταξύ των συστάδων καθώς και ο βαθμός διαχωρισμού των συστάδων [3] :

$$QGraph(C) = (intraLink(C) - interLink(C)) + Separation(C) \quad (22)$$

4.2.5 Performance

Η μετρική Performance υπολογίζει τον αριθμό των εσωτερικών ακμών σε μία κοινότητα σε συνδυασμό με τις ακμές που δεν υπάρχουν μεταξύ των κόμβων της ίδιας κοινότητας αλλά και με άλλους κόμβους που ανήκουν σε άλλες κοινότητες [4]. Αποδίδεται με τις παρακάτω ισότητες:

$$perf(C) = \frac{f(C) + g(C)}{\frac{1}{2}n(n-1)} \quad \text{όπου} \tag{23}$$

$$f(C) = \sum_{i=1}^{k} |E(C_i)| \tag{24}$$

$$g(C) = \sum_{i=1} \sum_{j>i} |\{\{u, v\} \notin E | u \in C_{i,} \; u \in C_j\}| \tag{25}$$

Αυτή η προσέγγιση αφορά έναν γράφο όπου οι ακμές του δεν περιέχουν την πληροφορία βάρους(weight). Οι τιμές των αποτελεσμάτων είναι από το 0 μέχρι το 1, με το 1 να σημαίνει ότι η συστάδα έχει εσωτερικά πυκνές συνδέσεις και εξωτερικά αραιές συνδέσεις σε σχέση με άλλες συστάδες [4].

4.2.6 CDS Index

Στην δημοσίευση [1] παρουσιάζεται ο δείκτης CDS. Σε ένα συμπαγή γράφο ο αριθμός των ακμών είναι κοντά στον μέγιστο αριθμό ακμών που θα μπορούσαν να υπάρχουν στον γράφο. Ο δείκτη density αντιπροσωπεύει την αναλογία του αριθμού των υπαρχόντων ακμών σε έναν γράφο με το ποσοστό των πιθανόν ακμών, οποίος δίνεται από την παρακάτω ισότητα :

$$dens(C) = \frac{2*|E|}{|V|*(|V|-1)} \tag{26}$$

Ωστόσο στα συνεκτικά δίκτυα οι κόμβοι είναι κοντά ο ένας στον άλλο και οι κοινότητες είναι δύσκολο να διαχωριστούν , έτσι για να αξιολογηθεί η συνδεσιμότητα των συστάδων σε ένα δίκτυο θα πρέπει να ληφθεί υπόψη πέρα από τον δείκτη density και ο δείκτης cohesion. Ο δείκτης cohesion αξιολογεί το πόσο κοντά είναι οι κόμβοι σε ένα δίκτυο και κατά πόσο μπορούν να διαχωριστούν από άλλους.

Ο δείκτης intra-connectivity σε μια συστάδα $C_i \in SC$ που αποτελείται από m συστάδες, υπολογίζεται από την παρακάτω ισότητα, όπου το w_c και w_d είναι οι τιμές που προσδιορίζουν το βάρος για το cohesion και το density αντίστοιχα [1].

$$Intra\_connectivity(C_i) = \mathrm{w_c} * \frac{1}{m} \sum_{i=1}^{m} Nconnect(c_i) + w_d * \sum_{i=1}^{m} \frac{1}{m} \quad dens(c_i) \quad (27)$$

Η συνδεσιμότητα του κόμβου (Node Connectivity) είναι το μέτρο της δομικής συνοχής των συστάδων. Η συνδεσιμότητα του κόμβου (Nconnect) της συστάδας c_j ορίζεται ως ο ελάχιστος αριθμός κόμβων που πρέπει να αφαιρεθεί και να απομείνουν δύο ή περισσότερες μη συνδεδεμένες συστάδες. Σε έναν γράφο G, ομαδοποιημένο σε m συστάδες C = {c_1…c_n}, ορίζεται VC_i το σετ των κόμβων που ανήκουν στο c_i και E(c_i) το σετ των ακμών που ενώνουν τους κόμβους στο c_i. Ο αλγόριθμος που υπολογίζει τη συνδεσιμότητα του κόμβου (Node Connectivity) δίνεται από την παρακάτω σχέση [1], [32] :

$$O\left((|VC_i| - \delta_i - 1 + \delta i(\delta_i - 1) \,/\, 2)|E(c_i)||VC_i|^{\frac{2}{3}}\right) \quad (28)$$

όπου δ_i ο μέγιστος βαθμός στο c_i. Η πολυπλοκότητα του δείκτη της εσωτερικής συνδεσιμότητας intra-connectivity(C) θα διαμορφωθεί ως εξής :

$$O\left\{(|VC_i| - \delta_i - 1 + \delta i(\delta_i - 1) \,/\, 2)|E(c_i)||VC_i|^{\frac{2}{3}}\right\}_{i=1}^{m} \quad (29)$$

επίσης η πολυπλοκότητα του δείκτη διαχωρισμού Separation (C) ορίζεται :

$$O(m^2) \quad (30)$$

Η συνδεσιμότητα μεταξύ δύο συστάδων θα πρέπει να αξιολογηθεί σε συνδυασμό με τη συνδεσιμότητα των κόμβων εντός των συστάδων. Η πυκνότητα των συνδέσεων ανάμεσα σε δύο κοινότητες αναμένεται να είναι μικρότερη της πυκνότητας εντός των συστάδων [1]. Με τον ίδιο τρόπο ο οποίος αναλύθηκε και στην υποενότητα 4.2.4 για τον δείκτη G-graph, υπολογίζονται οι δείκτες inter_connectivity και separation :

$$interCon(c_i, c_j) = \frac{InterDens(c_i, c_j)}{min\{dens(c_i), dens(c_j)\}} \quad (31)$$

$$Separation(C) = \frac{2}{m*(m-1)} \sum_{i=1} \sum_{j=1, i \neq j} \frac{1}{InterCon(c_i, c_j)} \quad (32)$$

Το τμήμα εκείνο του γράφου που περιέχει τις πιο συνεκτικές και πυκνές συστάδες αλλά είναι επίσης και καλός διαχωρισμένες μεταξύ τους, είναι το τμήμα εκείνο το οποίο χαρακτηρίζει καλύτερα τον γράφο. Επομένως το τμήμα εκείνο που είναι αντιπροσωπευτικότερο του γράφου περιέχει συστάδες με υψηλή τιμή όσον αφορά στους δείκτες Intra connectivity και Separation. Αυτοί οι δύο δείκτες έχουν συνδυαστεί για την δημιουργία του δείκτη CDS :

$$CDS(C) = Intra\_connectivity(C) + w_s * Separation(C) \quad (33)$$

Όπου το w_s υποδεικνύει την βαρύτητα του Separation συνυπολογίζοντας και το Intra connectivity. Υπάρχουν περιπτώσεις που απαιτείται ο υπολογισμός καλώς διαχωρισμένων συστάδων έχοντας λιγότερη βαρύτητα η συνοχή, έτσι το βάρος w_s επιτρέπει στον χρήστη να προσαρμόσει τον δείκτη αξιολόγησης, αναλόγως των απαιτήσεων [1]. Όσο υψηλότερη είναι η τιμή του δείκτη CDS τόσο πιο επιτυχείς θεωρούνται τα αποτελέσματα της ομαδοποίησης.

4.2.7 Dunn's Index

Ο συγκεκριμένος δείκτης αναφέρεται στη δημοσίευση [2] και βασίζεται στην διάμετρο των κοινοτήτων και την απόσταση των κόμβων. Πιο συγκεκριμένα, έστω ότι οι συστάδες C_i και C_j είναι οι κοντινότερες με βάση την απόσταση d, ενώ η συστάδα C_h είναι εκείνη με την μεγαλύτερη διάμετρο, ο υπολογισμός του δείκτη φαίνεται στην παρακάτω ισότητα :

$$D(C) = \frac{d(C_i, C_j)}{diam(C_h)} \quad (34)$$

Όπου $d(C_i,C_j)$ η απόσταση μεταξύ των συστάδων και $diam(C_h)$ η εσωτερική διάμετρος της συστάδας C_h. Η λειτουργία αυτού του δείκτη είναι να μεγιστοποιήσει την απόσταση μεταξύ των συστάδων, δηλαδή να μειώσει την συνδεσιμότητα των συστάδων και ελαχιστοποιήσει την απόσταση να εντός των συστάδων, δηλαδή να ενισχύσει την εσωτερική συνεκτικότητα τους. Επομένως, ο αριθμός των συστάδων που μεγιστοποιεί το D λαμβάνεται ως ο βέλτιστος αριθμός συστάδων. Οι υψηλές τιμές για αυτόν τον δείκτη αντιπροσωπεύουν καλές ομαδοποιήσεις [16].

4.2.8 Davies Bouldin index

Στις δημοσιεύσεις [2],[16] ένας ακόμα δείκτης που μελετά την απόσταση μεταξύ των συστάδων και την διάμετρο τους, ορίζεται ως εξής :

$$DB(C) = \frac{1}{K}\sum_{i=1}^{K} \max_{j\neq 1} \left[\frac{diam(C_i) + diam(C_j)}{d(C_i, C_j)}\right] \quad (35)$$

Όπου το K είναι ο αριθμός των συστάδων στο τμήμα C, d η απόσταση μεταξύ των συστάδων C_i και C_j . Οι μικρές τιμές στο αποτέλεσμα του συγκεκριμένου αλγόριθμου αντιστοιχούν σε συμπαγή συστάδες των οποίων τα κέντρα είναι μακριά το ένα από το άλλο [16].

4.2.9 MinMaxCut

Η συνεκτικότητα(cohesiveness) μιας συστάδας C_i μπορεί να υπολογιστεί λαμβάνοντας υπόψη μόνο τις συνδέσεις, ως ακολούθως [2]:

$$\frac{E_i'}{E_i} \quad (36)$$

Συμπερασματικά ο δείκτης MinMaxCut υπολογίζεται από την παρακάτω ισότητα :

$$MinMaxCut = \sum_{i=1}^{K} \frac{E_i'}{E_i} \quad (37)$$

Η μικρότερη τιμή του δείκτης ως αποτέλεσμα είναι εκείνη που θα υποδείξει την βέλτιστη τμηματοποίηση [2].

4.2.10 Coverage of a graph clustering

Ο δείκτης Coverage of a graph clustering αξιολογεί τις εσωτερικές συνδέσεις της συστάδας και υπολογίζεται από την παρακάτω ισότητα [2]:

$$Cov(C) = \frac{\sum_{i=1}^{K} E_i}{E} \quad (38)$$

Λαμβάνοντας υπόψη και την θεωρία της βέλτιστης συσταδοποίησης ενός γράφου γίνεται αντιληπτό ότι μια υψηλή τιμή ως αποτέλεσμα του συγκεκριμένου δείκτη υποδεικνύει καλή ποιότητα ομαδοποίησης. Είναι ένας δείκτης εύκολα υπολογίσιμος, παρόλα αυτά δεν λαμβάνει υπόψη του των αριθμό των κόμβων.

Στην παρούσα ενότητα παρουσιάστηκαν δείκτες οι οποίοι εξετάζουν διαφορετικά χαρακτηριστηκά των γράφων με σκοπό τη αξιολόγηση της ποιότητας της συσταδοποίησης. Κάποιες από τις έννοιες που εξετάζονται αφορούν στην σύγκριση των εσωτερικών συνδέσεων των κόμβων μεταξύ της ίδια συστάδας δηλαδή των ενδοκοινοτικών συνδέσεων, με το αριθμό των διακοινοτικών συνδέσεων, δηλαδή των συνδέσεων των κόμβων με γειτονικές συστάδες, εκτός των συστάδων των οποίων ανήκουν. Άλλα χαρακτηριστικά που αξιολογούνται είναι ο διαχωρισμός (Separation) και η συνοχή (Cohesion), χρησιμοποιούνται για να αξιολογήσουν την ομοιότητα ενός κόμβου με τους γειτονικούς του. Επίσης χρησιμοποιείται η αγωγιμότητα της τομής (Conductance of a cut), σε μία υποτιθέμενη τομή που χωρίζει δύο συστάδες από τις ακμές που τέμνονται υπολογίζεται η συνδεσιμότητα των συστάδων. Παρουσιάστηκε δείκτης που μελετά τον μέγιστο αριθμό κόμβων που απαρτίζει τον πυρήνα μιας ομάδας (degeneracy-core). Άλλο χαρακτηριστικό των γράφων που εξετάζεται είναι η πυκνότητα (Density) που αντιπροσωπεύει την αναλογία του αριθμού των υπαρχόντων ακμών σε έναν γράφο με το ποσοστό των πιθανόν ακμών. Αναφέρθηκαν δείκτες που βασίζονται στην διάμετρο των κοινοτήτων και την απόσταση των κόμβων. Υπάρχουν δείκτες που εξετάζουν περισσότερες από μια έννοιες προκειμένου να αξιολογήσουν την ποιότητα των κοινοτήτων. Στις επόμενες ενότητες θα παρουσιαστούν τα αποτελέσματα τις πειραματικής διαδικασίας κατά την οποία θα επιλεχθούν τρία σύνολα δεδομένων διαφορετικής δομής, με την χρήση δύο αλγορίθμων με διαφορετικό τρόπο δράσης θα πραγματοποιηθούν τμηματοποιήσεις με διαφορετικό αριθμό συστάδων. Τα αποτελέσματα των αλγορίθμων θα αξιολογηθούν από πέντε δείκτες αξιολόγησης, οι οποίοι εξετάζουν διαφορετικά χαρακτηριστικά. Σκοπός είναι να μελετηθεί η συμπεριφορά των δεικτών, δηλαδή σε ποιες περιπτώσεις σημειώνουν καλό αποτέλεσμα, όπως και αν έχουμε συμφωνία στα αποτελέσματα, διαφορετικά να παρατηρηθεί και πειραματικά ποια είναι εκείνα τα χαρακτηριστικά που επηρεάζουν την απόδοση των δεικτών.

ΚΕΦΑΛΑΙΟ 5 : Πειραματική διαδικασία και Αποτελέσματα

Στόχος της εργασίας είναι να εξεταστούν τα αποτελέσματα επιλεγμένων δεικτών αξιολόγησης αποτελεσμάτων τμηματοποίησης γράφων. Έχουν επιλεχθεί τρία σύνολα δεδομένων γράφων τα οποία παρουσιάζονται παρακάτω, τα σύνολα δεδομένων έχουν γνωστό αριθμό τμημάτων (clusters) από σχετικές μελέτες, ο οποίος αναφέρεται με τον όρο ground truth. Ο λόγος που χρειάζεται το ground truth είναι για να υπάρχει μία ένδειξη για τον αριθμό των τμημάτων που σε κάθε σύνολο δεδομένων θα πρέπει να παρατηρηθεί η μεγαλύτερη τιμή στους δείκτες αξιολόγησης, ώστε να διασφαλιστεί ότι η καλύτερη απόδοση τους είναι όσο το δυνατό πιο κοντά στον πραγματικό αριθμό των τμημάτων και όχι σε κάποιο τυχαίο φαινομενικά αποτέλεσμα. Έχουν επιλεγεί δύο αλγόριθμοι τμηματοποίησης Spectral και Louvain, η επιλογή των δύο αλγορίθμων έγινε λόγω του ότι το υπολογιστικό τους κόστος είναι ανάλογο της κοινότητας που ανιχνεύεται και όχι του μεγέθους των δικτύων, μπορώντας να διαχειριστούν δίκτυα με εκατοντάδες κόμβους, επίσης δεν απαιτείται εισαγωγή παραμέτρων, είναι δύο πολύ δημοφιλής αλγόριθμοι σε σχετικές μελέτες [5]. Χρησιμοποιώντας διαφορετικές τιμές εισόδου θα γίνει τμηματοποίηση των γράφων σε συστάδες ενώ παράλληλα θα παρουσιαστούν τα αποτελέσματα των δεικτών αξιολόγησης και θα σημειωθεί σε ποιο αριθμό τμημάτων θα παρουσιάσουν την μέγιστη απόδοση, οι δείκτες αξιολόγησης που έχουν επιλεχθεί είναι από τους προαναφερόμενους και είναι οι εξής : Qgraph [3], CDS[1], GS*[2].

Παράλληλα για κάθε τμηματοποίηση έχουν υπολογιστεί και ο δείκτης Modularity και Conductance που αναφέρθηκαν παραπάνω ώστε να μελετηθεί η συμπεριφορά τους σε σχέση με τα αποτελέσματα των δεικτών αξιολόγησης που έχουν επιλεχθεί προς μελέτη.

5.1 Σύνολα δεδομένων

Παρακάτω παρουσιάζονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την διεξαγωγή των πειραμάτων. Για κάθε σύνολο δεδομένων παρατίθεται περιγραφή όσον αφορά στα δεδομένα που αντιπροσωπεύουν καθώς και των κύριων χαρακτηριστικών που ενδιαφέρουν την συγκεκριμένη μελέτη, τα οποία είναι ο αριθμός των κόμβων, των ακμών και των συστάδων στις οποίες οργανώνονται τα δεδομένα.

Α/Α	Όνομα	Κόμβοι	Ακμές
1	email-Eu-core	1005	25571
2	karate	34	78
3	facebook_combined	4039	88234

Πίνακας 1. Σύνολα δεδομένων

Περιγραφή συνόλων δεδομένων :

1) Το σύνολο δεδομένων email-Eu-core έχει παραχθεί από δεδομένα ηλεκτρονικού ταχυδρομείου από ένα μεγάλο ευρωπαϊκό ερευνητικό ίδρυμα. Οι πληροφορίες έχουν γίνει ανώνυμες για όλα τα εισερχόμενα και εξερχόμενα email μεταξύ των μελών του ερευνητικού ιδρύματος. Μια ακμή (u, v) δημιουργείται στο δίκτυο εάν το άτομο u έστειλε στο άτομο v τουλάχιστον ένα email. Τα μηνύματα ηλεκτρονικού ταχυδρομείου αντιπροσωπεύουν μόνο την επικοινωνία μεταξύ των μελών του ιδρύματος (ο πυρήνας) και το σύνολο δεδομένων δεν περιέχει εισερχόμενα μηνύματα ή εξερχόμενα μηνύματα προς τον υπόλοιπο κόσμο. Κάθε άτομο ανήκει ακριβώς σε ένα από τα 42 τμήματα του ερευνητικού ινστιτούτου [1],[3].

   Πηγή : https://snap.stanford.edu/data/email-Eu-core.html

2) Το σύνολο δεδομένων karate περιέχει κοινωνικές συνδέσεις μεταξύ μελών μιας πανεπιστημιακής ομάδας καράτε, έχουν συλλεγεί από τον Wayne Zachary το 1977. Πρόκειται για ένα μικρό σύνολο δεδομένων, το οποίο έχει χρησιμοποιηθεί σε πολλές μελέτες και αποτελείται από 2 συστάδες [1], [3]. Τα δύο τμήματα του συλλόγου καράτε δημιουργήθηκαν από την διαφωνία των δυο καθηγητών, John.A και Mr. Hi [3].

   Πηγή : https://networkrepository.com/soc-karate.php

3) Το σύνολο δεδομένων Facebook αποτελείται από λίστες φίλων. Τα δεδομένα έχουν συλλεχθεί από απο συμμετέχοντες στην έρευνα που χρησιμοποιούσαν την εφαρμογή Facebook. Οι κόμβοι αντιπροσωπεύουν τους λογαριασμούς των χρηστών που έχουν μετατραπεί σε ανώνυμα αντικαθιστώντας κάθε χρήστη με μία τιμή. Επίσης έχουν μετατραπεί σε ανώνυμα και τα διανύσματα χαρακτηριστικών, για παράδειγμα , όπου το αρχικό σύνολο δεδομένων μπορεί να περιείχε ένα χαρακτηριστικό "political = Δημοκρατικό Κόμμα", τα νέα δεδομένα θα περιέχουν απλώς "political = anonymized χαρακτηριστικό 1". Έτσι, χρησιμοποιώντας τα ανώνυμα δεδομένα είναι δυνατό να προσδιοριστεί εάν δύο χρήστες έχουν τις ίδιες πολιτικές πεποιθήσεις, αλλά όχι τι αντιπροσωπεύουν οι μεμονωμένες πολιτικές πεποιθήσεις τους. [11], [12]. Ο ιδανικός αριθμός συστάδων με βάση σχετικές μελέτες είναι 16.

   Πηγή : http://snap.stanford.edu/data/ego-Facebook.html

5.2 Πειραματικά αποτελέσματα

Στους παρακάτω πίνακες οργανώνονται ανά σύνολο δεδομένων και με βάση τον αριθμό τμηματοποιήσεων τα αποτελέσματα που προέκυψαν για κάθε αλγόριθμο και για κάθε δείκτη αξιολόγησης. Ανά τιμή εισόδου καταγράφεται η αντίστοιχη απόδοση των δεικτών.

Σε κάθε ένα από τους πίνακες έχει σημειωθεί με γκρίζο χρώμα η σειρά εκείνη η οποία αφορά στο ground truth του συγκεκριμένου γράφου και θα αποτελέσει σημείο σύγκρισης για τα αποτελέσματα που έχουν παραχθεί με διαφορετικό αριθμό συστάδων.

Για τα σύνολα δεδομένων email-Eu-core και karate έχει βρεθεί το σύνολο δεδομένων ground truth, δηλαδή μία λίστα με τις ετικέτες των συστάδων για κάθε κόμβο με την σειρά. Θα γίνει έλεγχος για κάθε συσταδοποίηση που θα προκύψει παρακάτω κατά πόσο είναι όμοια με την συσταδοποίηση του ground truth, για την σύγκριση θα χρησιμοποιηθεί ο δείκτης Rand Index.

Ο δείκτης Rand Index υπολογίζει την ομοιότητα μεταξύ δύο συσταδοποιήσεων, μετρώντας τους κόμβους που είναι συσταδοποιημένοι στην ίδια κοινότητα ή όχι, μεταξύ των πραγματικών (labels_true) ετικετών και των υπολογισμένων (labels_pred). Οι τιμή 0 αξιολογεί την συσταδοποίηση που έχει προκύψει από τον εκάστοτε αλγόριθμο περισσότερο ως τυχαία καταχώρηση ετικετών (labels) στους κόμβους, ενώ η μέγιστη τιμή είναι η μονάδα όπου υπάρχει πλήρη ταύτιση στην ομαδοποίηση που έχει προκύψει από τον αλγόριθμο που χρησιμοποιείται και της ιδανικής συσταδοποίησης (ground truth) [33].

Spectral Algorithm						
email-Eu-core						
Clusters	Modularity	Conductance	Q-graph	CDS	GS*	Rand Index
40	0.192	0.757	0.583	1.041	-0.007	0.789
42	**0.265**	**0.755**	**0.836**	**1.122**	**-0.006**	**0.807**
51	0.185	0.803	0.538	0.969	-0.029	0.673
60	0.143	0.834	0.433	1.043	-0.053	0.639
80	0.153	0.864	0.351	1.033	-0.060	0.668

Πίνακας 2. Αποτελέσματα γράφου email-Eu-core για τον αλγόριθμο Spectral

Στον παραπάνω πίνακα (2) παραθέτονται τα αποτελέσματα που προέκυψαν από κάθε μετρική για πέντε διαφορετικές τμηματοποιήσεις με τον αλγόριθμο Spectral για το σύνολο δεδομένων του γράφου email-Eu-core. Τονισμένες με εντονότερο χρώμα φαίνονται οι τιμές όπου οι δείκτες αξιολόγησης έχουν παρουσιάσει την καλύτερη απόδοση. Παρατηρείται ότι ο Modularity αποκτά την μεγαλύτερη τιμή του στα 42 τμήματα επιβεβαιώνοντας ότι σε αυτή την δοκιμή υπάρχουν οι πιο δυνατές εσωτερικές συνδέσεις εντός των συστάδων. Με την σειρά του και ο δείκτης Conductance αποκτά την χαμηλότερη τιμή του στις 42 συστάδες . Οι δείκτες Q-graph, CDS και GS* με την σειρά τους επιβεβαιώνουν το ground truth του συγκεκριμένου συνόλου δεδομένων. Ο ιδανικός αριθμός συστάδων στο συγκεκριμένο γράφο είναι ο αριθμός 42, σε αυτό το σημείο

όλοι οι δείκτες έχουν αξιολογήσει σε σχέση με τις υπόλοιπες συσταδοποιήσεις την συγκεκριμένη ως την ποιοτικότερη. Ο δείκτης Rand Index επίσης αποκτά στις 42 κοινότητες την μέγιστη τιμή του επιβεβαιώνοντας πως πλησιάζει η συγκεκριμένη συσταδοποίηση την ιδανική αλλά δεν είναι πανομοιότυπες καθώς δεν αποκτά την τιμή 1.

Louvain Algorithm							
email-Eu-core							
Resolution	Clusters	Modularity	Conductance	Q-graph	CDS	GS*	Rand Index
3	41	**0.353**	**0.778**	**2.200**	1.930	-0.075	**0.938**
3.1	42	0.352	0.780	2.087	**2.006**	**0.002**	0.932
3.6	51	0.337	0.803	1.813	1.956	-0.080	0864
4	55	0.327	0.810	1.482	1.928	-0.100	0.687
6	78	0.293	0.853	1.154	2.002	-0.134	0.599

Πίνακας 3. Αποτελέσματα γράφου email-Eu-core για τον αλγόριθμο Louvain

Με εντονότερα ψηφία παρουσιάζονται οι τιμές όπου οι δείκτες είχαν καλύτερη απόδοση και σε αυτόν τον πίνακα (3) που παρουσιάζει της τμηματοποιήσεις του αλγόριθμου Louvain για το ίδιο σύνολο δεδομένων. Ο Modularity, ο Conductance και Q-graph, παρουσιάζουν τα καλύτερα αποτελέσματα για τις 41 τμηματοποιήσεις και έτσι υπάρχει απόκλιση από την επιθυμητή αξιολόγηση η οποία θα ήταν να επιβεβαίωναν ότι οι 42 τμηματοποίσεις είναι εκείνες με τα ιδανικότερα χαρακτηριστικά. Παρόλα αυτά οι δείκτες CDS και GS* στις 42 τμηματοποιήσεις παρουσιάζουν το καλύτερο αποτέλεσμα τους, έτσι από αυτούς τους δείκτες συμπεραίνεται η ύπαρξη καλώς διαχωρισμένων συστάδων για τις 42 συστάδες. Κατά τις τμηματοποιήσεις που εκτελέστηκαν από τον αλγόριθμο Louvain, ο δείκτης Rand Index έχει σημειώσει μεγαλύτερη τιμή για τις 41 συστάδες σε σχέση με τις υπόλοιπες τμηματοποιήσεις και με τον αλγόριθμο Spectral, συνεπώς γίνεται αντιληπτό ότι οι 41 τμηματοποιήσεις του αλγόριθμου Louvain είναι ποιο κοντά στις ιδανικές. Τις 41 τμηματοποιήσεις παρατηρείται όπως αναφέρθηκε παραπάνω οι δείκτες Modularity, ο Conductance και Q-graph να έχουν αξιολογήσει ως ποιοτικότερες, ενώ οι δείκτες CDS και GS* τις 42 συστάδες οι οποίες είναι με μικρή διαφορά του δείκτη Rand Index η επόμενη ομαδοποίηση ποιο κοντά στο ground truth.

Spectral Algorithm						
karate						
Clusters	Modularity	Conductance	Q-graph	CDS	GS*	Rand Index
2	**0.360**	**0.152**	**4.605**	2.605	**0.346**	**0.629**
3	0.339	0.205	2.433	2.371	0.262	0.626
5	0.339	0.380	1.392	2.706	0.208	0.585
7	0.334	0.512	1.494	**2.935**	0.189	0.538
8	0.316	0.561	1.670	2.886	0.190	0.597

Πίνακας 4. Αποτελέσματα γράφου karate για τον αλγόριθμο Spectral

Στην συνέχεια αναλύεται το σύνολο δεδομένων του γράφου karate, στον πίνακα 4 παρουσιάζονται τα αποτελέσματα των δεικτών αξιολόγησης, για δοκιμές τμηματοποίησεων με τον αλγόριθμο Spectral, με εντονότερη γραφή είναι επίσης οι τιμές με την καλύτερη απόδοση των δεικτών. Έχει σημειωθεί με γκρίζο χρώμα η αντίστοιχη γραμμή του πίνακα, που αναφέρεται στην τιμή των ιδανικών τμηματοποίησεων (ground truth) η οποία είναι 2 συστάδες . Ο Modularity, ο Conductance , ο Q-graph και ο GS* παρουσιάζουν τα καλύτερα αποτελέσματα υποστηρίζοντας την δομή του γράφου. Αντίθετα ο δείκτης CDS φαίνεται να μην παρουσιάζει κάποια συνοχή στα αποτελέσματα του. Ο δείκτης Rand Index για τον γράφο karate δεν έχει σημειώσει υψηλά ποσοστά ομοιότητας με τις ετικέτες που έχουν οι κόμβοι στην ιδανική συσταδοποίηση (ground truth), πρόκειται για 34 κόμβους, σε περίπτωση που έστω και ενός η ετικέτα διαφέρει θα επηρεαστεί το αποτέλεσμα. Παρόλα αυτά το μεγαλύτερο αποτέλεσμα σημειώνεται στις 2 συστάδες, που είναι ο ιδανικός αριθμός συσταδοποιήσεων (ground truth) και οι περισσότεροι δείκτες εκτός του CDS αξιολογούν ως τις ποιοτικότερες.

Louvain Algorithm							
karate							
Resolution	Clusters	Modularity	Conductance	Q-graph	CDS	GS*	Rand Index
0.3	2	0.372	**0.128**	**2.691**	2.676	**0.329**	**0.629**
0.7	3	**0.402**	0.201	2.462	2.712	0.248	0.597
1.8	6	0.360	0.475	2.419	2.731	0.167	0.576
2	7	0.335	0.514	2.066	2.794	0.180	0.554
2.3	9	0.288	0.609	2.096	**2.852**	0.149	0.535

Πίνακας 5. Αποτελέσματα γράφου karate για τον αλγόριθμο Louvain

Αντίστοιχα στον πίνακα 5 στις τμηματοποιήσεις του ίδιου συνόλου δεδομένων με τον αλγόριθμο Louvain οι δείκτες Conductance , Q-graph και GS* περιγράφουν δύο καλώς διαχωρισμένα τμήματα. Από τον δείκτη Modularity θα είχαν προκύψει καλά αποτελέσματα αλλά στις 3 τμηματοποιήσεις αυξήθηκε αρκετά η τιμή του αποτελέσματος καταστρέφοντας την συνοχή των αποτελεσμάτων. Ο δείκτης CDS όπως και για τα αποτελέσματα του πίνακα 4 δεν εμφανίζει τα επιθυμητά αποτελέσματα, αντίθετα έχει αυξανόμενη κατά κύριο λόγο τιμή όσο απομακρύνεται η συσταδοποίηση από το ground truth. Όπως για τον αλγόριθμο Spectral έτσι και για τον αλγόριθμο Louvain ο τρόπος που έχουν ομαδοποιηθεί οι κόμβοι για τον ιδανικό αριθμό συστάδων δύο, έχει διαφορές με αποτέλεσμα ο δείκτης Rand Index να σημειώνει το καλύτερο αποτέλεσμα του αλλά όχι το βέλτιστο. Αντίστοιχα δεν υπάρχει ομοιογένεια στα αποτελέσματα των δεικτών αξιολόγησης.

Spectral Algorithm					
1. facebook_combined					
Clusters	Modularity	Conductance	Q-graph	CDS	GS*
5	0.684	**0.003**	5.604	1.653	0.305
6	0.686	0.004	4.004	1.543	**0.314**
16	**0.733**	0.059	**8.551**	**3.781**	0.054
25	0.643	0.108	5.330	3.234	0.177
30	0.625	0.123	5.945	3.481	0.000

Πίνακας 6. Αποτελέσματα γράφου facebook_combined για τον αλγόριθμο Spectral

Στον παραπάνω πίνακα (6) καταγράφονται τα αποτελέσματα του αλγόριθμου συσταδοποίησης Spectral για τον γράφο facebook_combined. Ο δείκτης Modularity, CDS και Q-graph επιβεβαιώνουν με την απόδοση που έχουν την ιδανική συσταδοποίηση για τις 16 συστάδες. Αντιθέτως οι δείκτες Conductance και GS* παρουσιάζουν την καλύτερη τους απόδοση στις 5 και στις 6 ομαδοποιήσεις αντίστοιχα.

Louvain Algorithm						
1. facebook_combined						
Resolution	Clusters	Modularity	Conductance	Q-graph	CDS	GS*
0.3	11	0.824	**0.025**	**11.357**	2.731	0.160
0.5	12	0.831	0.031	10.945	2.672	0.177
1	16	**0.835**	0.054	10.249	**3.333**	**0.178**
2	22	0.830	0.095	9.636	3.493	0.139
2.5	23	0.829	0.098	8.930	3.548	0.125

Πίνακας 7. Αποτελέσματα γράφου facebook_combined για τον αλγόριθμο Louvain

Στον πίνακα 7 φαίνονται τα αποτελέσματα των δεικτών για το σύνολο δεδομένων του γράφου facebook_combined, για συσταδοποιήσεις με την χρήση του αλγόριθμου Louvain. Παρατηρείται ότι στις 16 τμηματοποιήσεις οι δείκτες Modularity, CDS, και GS* σημειώνουν τις καλύτερες τιμές τους ενώ οι δείκτες Conductance και Q-graph παρουσιάζουν τα καλύτερα αποτελέσματα στις 11 συστάδες.
Είναι επίσης ενδιαφέρον να παρατηρηθεί πως οι δείκτες αξιολόγησης, αξιολογούν την ποιότητα των συστάδων του ιδανικού συνόλου δεδομένων (ground truth).

email-Eu-core (ground truth)					
Clusters	Modularity	Conductance	Q-graph	CDS	GS*
42	0.226	0.789	1.498	1.166	0.001

Πίνακας 8. Αποτελέσματα δεικτών αξιολόγησης για τις ετικέτες ground truth του το σύνολο email-Eu-core

karate (ground truth)					
Clusters	Modularity	Conductance	Q-graph	CDS	GS*
2	0.301	0.292	0.812	1.532	0.127

Πίνακας 9. Αποτελέσματα δεικτών αξιολόγησης για τις ετικέτες ground truth του το σύνολο karate

Από τα αποτελέσματα του πίνακα 8, παρατηρείται ότι οι δείκτες Modularity, Q-graph, CDS και GS* για τις 42 συστάδες που έχουν οριστεί ως οι ιδανικές έχει αποτέλεσμα που συγκριτικά με τους πίνακες 2 και 3 σημειώνουν καλύτερο αποτελέσματα από τα αποτελέσματα του αλγόριθμου Louvain αλλά οι συστάδες που έχει ανιχνεύσει ο αλγόριθμος Spectral φαίνονται ποιοτικότερες με βάση τους ίδιους δείκτες από αυτές που έχουν οριστεί ως ιδανικές. Ο δείκτης Conductance για τις ιδανικές τμηματοποιήσεις δεν έχει σημειώσει καλό αποτέλεσμα, φαίνεται πως οι 42 συστάδες που έχουν δημιουργηθεί από τους δύο αλγόριθμους να παρουσιάζουν καλύτερο μέσο όρο αγωγιμότητας του γράφου.

Απο τα αποτελέσματα του πίνακα 9, παρατηρείται ότι τα χαρακτηριστικά των δύο τμηματοποιήσεων των 2 συστάδων που προέκυψαν από τους αλγόριθμους αξιολογούνται ποιοτικότερα από τους δείκτες αξιολόγησης σε σχέση με τα αποτελέσματα που παρουσίασαν κατά την αξιολόγηση των ιδανικών συστάδων τμηματοποιήσεων (ground truth).

Αυτό που θα ήταν αναμενόμενο είναι η αξιολογήσεις των δεικτών για τον ιδανικό αριθμό τμηματοποιήσεων (ground truth) είναι να σημειώνουν τα καλύτερα αποτελέσματα. Ωστόσο στο επιστημονικό άρθρο [31] υποστηρίζεται ότι η οργάνωση των πραγματικών δικτύων συνήθως συσχετίζεται με πολλαπλά σύνολα μεταδεδομένων. Έτσι η επιλογή ενός συγκεκριμένου συνόλου ως του ιδανικού (ground truth) για την καταχώριση των κόμβων σε συστάδες και την χρήση του ως μέτρο σύγκρισης με τις ομαδοποιήσεις που έχουν προκύψει από κάποιον αλγόριθμο να μην είναι μια ασφαλής πρακτική. Αυτό που θα μπορούσε να ειπωθεί με σιγουριά για τα παραπάνω αποτελέσματα με την χρήση του συνόλου δεδομένων ground truth, κάτι που επιβεβαιώνεται και από τον δείκτη Rand Index είναι ότι οι τμηματοποιήσεις που προέκυψαν μέσω των αλγορίθμων διαφέρουν από τις τμηματοποιήσεις που περιγράφονται από τα σύνολα δεδομένων ground truth. Όπως παρουσιάζεται στο σχετικό επιστημονικό άρθρο [31], τρία πράγματα θα μπορούσαν να συμβαίνουν i) τα μεταδεδομένα δεν έχουν σχέση με τα δεδομένα του γράφου ii) το σύνολο δεδομένων των ιδανικών τμηματοποιήσεων περιγράφουν διαφορετικά χαρακτηριστικά της δομής του γράφου iii) το δίκτυο δεν έχει την δομή ομάδων. Στην περίπτωση των παραπάνω δοκιμών με τα σύνολα δεδομένων ground truth, μπορεί να θεωρηθεί πως οι ετικέτες ορίστηκαν με την χρήση διαφορετικών κριτηρίων από αυτά που εξετάζονται από τους δύο αλγόριθμους συσταδοποίησης με αποτέλεσμα διαφορετικές ομαδοποιήσεις των κόμβων. Για την καλύτερη κατανόηση την ερμηνείας που προαναφέρθηκε, στο άρθρο παρουσιάζεται ένα παράδειγμα για το σύνολο δεδομένων του γράφου karate, που αφορά έναν κόμβο του οποίου η ετικέτα δεν ανακτάται από τους περισσότερους αλγόριθμους. Ο εν λόγω κόμβος ενώ έχει περισσότερες συνδέσεις με την ομάδα του προέδρου, επέλεξε να ενταχθεί στην ομάδα του εκπαιδευτή για να μην χάση την πρόοδο του για την μαύρη ζώνη [31]. Ομοίως σε ομαδοποιήσεις συνόλων δεδομένων γράφου μπορούν να παρατηρηθούν διαφοροποιήσεις στις ετικέτες που ανακτώνται από τους αλγόριθμους συσταδοποίησης και τις ετικέτες των συνόλων δεδομένων ground truth. Στην συνέχει θα μελετηθούν οι συστάδες που έχουν προκύψει από τους αλγόριθμους για των αριθμό των ιδανικών τμηματοποίησεων αλλά το βάρος της παρατήρησης θα επικεντρωθεί στο ποια χαρακτηριστικά είναι εκείνα που επηρεάζουν τους δείκτες αξιολόγησης.

Όσον αφορά στους δείκτες αξιολόγησης παρατηρείται ότι δεν έχουν όλοι οι δείκτες καλή απόδοση για κάθε σύνολο δεδομένων. Θεωρητικά αυτό είναι αναμενόμενο διότι η

συσταδοποίηση που εκτελείται διαφέρει για κάθε αλγόριθμο που χρησιμοποιείται, για αυτό το λόγο χρησιμοποιήθηκαν δύο αλγόριθμοι συσταδοποίησησης. Επίσης τα κριτήρια με βάση των οποίων αξιολογούν τα αποτελέσματα των συσταδοποιήσεων οι δείκτες αξιολόγησης διαφέρουν. Από τα αποτελέσματα του δείκτη Rand Index γίνεται αντιληπτό πως ακόμα και για τον ιδανικό αριθμό συστάδων οι δύο αλγόριθμοι δε έχουν επιτύχει πλήρη ομοιότητα ομαδοποίησης των κόμβων με το σύνολο δεδομένων ground truth. Στη επόμενη ενότητα θα χρησιμοποιηθεί ένα ευρέως χρησιμοποιημένο σύνολο δεδομένων και λόγο της μικρής δομής του θα μπορέσουν να γίνουν ορατοί μέσω τις οπτικοποίησης, οι λόγοι για τους οποίους κάθε δείκτης αξιολόγησης είχε τα συγκεκριμένα αποτελέσματα.

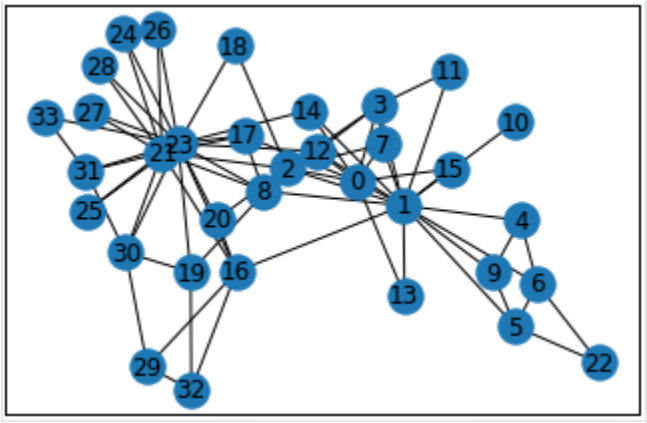
ΚΕΦΑΛΑΙΟ 6 : Ερμηνεία Αποτελεσμάτων

Από τα αποτελέσματα των αλγορίθμων συσταδοποίησης στα τρία διαφορετικά σύνολα δεδομένων παρατηρείται μια ποικιλία αποτελεσμάτων και συμπεριφοράς των δεικτών αξιολόγησης. Το σύνολο δεδομένων karate είναι ένα ιδιαίτερα γνωστό σύνολο δεδομένων και έχει χρησιμοποιηθεί για δοκιμές σε αντίστοιχες μελέτες, πρόκειται για έναν μικρό γράφο που αποτελείται από 34 κόμβους και 78 ακμές. Για να μπορέσει να γίνει κάποια ερμηνεία και αιτιολόγηση επί των αποτελεσμάτων θα γίνει οπτικοποίηση διαφορετικών ομαδοποιήσεων του συνόλου δεδομένων karate και από τους δύο αλγόριθμους που χρησιμοποιήθηκαν. Παρατηρώντας την δομή των συστάδων θα συγκριθούν τα αποτελέσματα ώστε να τα αιτιολογήσουμε σε σχέση με τα στοιχεία ενός γράφου που επηρεάζουν την απόδοση των δεικτών.

6.1 Οπτικοποίηση γράφου και σύγκριση αποτελεσμάτων μεταξύ αλγορίθμων

Αρχικά από και τα παραπάνω αποτελέσματα και από τις οπτικοποιήσεις που ακολουθούν γίνεται αντιληπτό ότι τόσο τα δομικά χαρακτηριστικά του γράφου αλλά και ο τρόπος που ορίζονται οι συστάδες αναλόγως τον αλγόριθμο που χρησιμοποιείται επηρεάζουν το αποτέλεσμα της δομής των συστάδων και συνεπώς και τα χαρακτηριστικά που αξιολογούνται από του δείκτες.

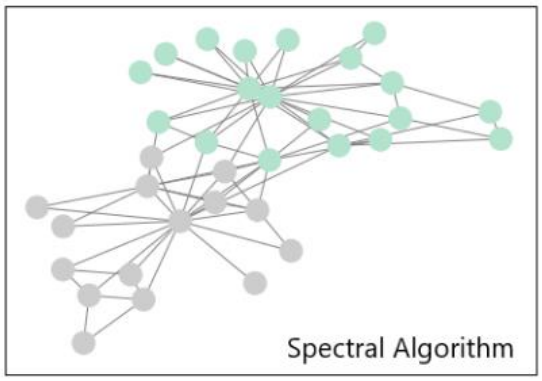
Ο γράφος δικτύου Zach's karate club είναι της παρακάτω μορφής :



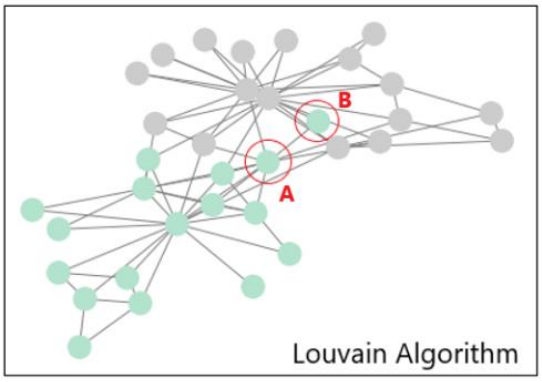
Εικόνα 6. Οπτικοποίηση του γράφου karate

Πρόκειται για έναν μη κατευθυνόμενο γράφο και αποτελεί ένα μοτίβο σχέσεων μαθητών, όπου κάθε κόμβος έχει πάνω από μία ακμή να τον συνδέουν με τους κόμβους του υπόλοιπου δικτύου. Επιλέχθηκε ο συγκεκριμένος γράφος διότι το μέγεθος του, μας επιτρέπει να παρατηρήσουμε την δομή του, καθώς και ότι μας είναι γνωστό ότι αποτελείται από 2 συστάδες που αντιπροσωπεύουν την οργάνωση μαθητών ως προς δύο καθηγητές με τις ετικέτες (labels) John A. (κόμβος 23) και Mr. Hi (κόμβος 1).

Παρακάτω παρατίθενται οπτικοποιημένα τα αποτελέσματα για 2 και 3 συστάδες από τους αλγόριθμους Spectral και Louvain. Επιλέχθηκαν τα συγκεκριμένα γιατί από ότι παρατηρείται από τους παραπάνω πίνακες αποτελεσμάτων 4 και 5 υπάρχει διαφορά στην απόδοση των δεικτών αξιολόγησης.



Εικόνα 7. Ομαδοποίηση δύο συστάδων (Spectral)



Εικόνα 8. Ομαδοποίηση δύο συστάδων (Louvain)

Μεταξύ των δύο παραπάνω σχεδίων (plots) είναι ορατές σε ένα τόσο μικρό γράφο οι διαφορές που υπάρχουν ως προς την ομαδοποίηση των κόμβων. Στο αποτέλεσμα του αλγορίθμου Louvain έχουν σημειωθεί επί του γράφου οι δύο κόμβοι που έχουν ενταχθεί στην ομάδα του καθηγητή John A. ενώ στα αποτελέσματα του Spectral είχαν κατηγοριοποιηθεί στην ομάδα με κεντρικό κόμβο τον καθηγητή Mr. Hi.

Παρατηρείται λοιπόν διαφοροποίηση τον αποτελεσμάτων συσταδοποίησης μεταξύ των δύο αλγορίθμων, αυτό συμβαίνει διότι ο τρόπος που λειτουργούν οι δύο αλγόριθμοι είναι διαφορετικός. Όπως αναφέρθηκε παραπάνω στον Louvain εξετάζεται κάθε κόμβος σε κυκλικό μοτίβο και κατατάσσεται στις κοντινές κοινότητες, οι κόμβοι θα μετακινηθούν εκεί όπου παρουσιάζεται η μεγαλύτερη αύξηση του modularity, μόλις ολοκληρωθεί η διαδικασία οι κοινότητες συγχωνεύονται ώστε να υπάρχει ένα σταθερό συνολικό βάρος (weight). Ενώ στον Spectral χρησιμοποιούνται τα ιδιοδιανύσματα και οι ιδιοτιμές του πίνακα γειτνίασης για τον υπολογισμό των συστάδων. Με αυτό το τρόπο σε μικρότερη κλίμακα από ότι σε ένα μεγάλο σύνολο δεδομένων παρατηρείται διαφορετική συσταδοποίηση των κόμβων στα παραπάνω γραφήματα.

Με βάση τις διαφορές που έχουν παρατηρηθεί στην συσταδοποίηση από τους δύο αλγορίθμους θα γίνει προσπάθεια να ερμηνευθεί η συμπεριφορά των δεικτών και να

συγκριθούν τα αποτελέσματα. Παρατηρώντας την μετρική αξιολόγησης Modularity (ισότητα 5) στους πίνακες 4 και 5, για δύο συστάδες το σύνολο δεδομένων karate παρουσιάζει την υψηλότερη του τιμή για τις τμηματοποιήσεις που προέκυψαν από τον αλγόριθμο Louvain. Εξετάζοντας τους δύο κόμβους που διαφοροποιούν το αποτέλεσμα, στο διάγραμμα του αλγόριθμου Louvain, ο κόμβος Α φαίνεται να έχει πέντε ακμές να τον συνδέουν με την συστάδα στην οποία έχει κατηγοριοποιηθεί ενώ με την άλλη συστάδα 4 ακμές. Ο κόμβος Β συνδέεται με την συστάδα που έχει κατηγοριοποιηθεί με μία ακμή αλλά και με την άλλη συστάδα με μία ακμή επίσης. Αν εξεταστεί μεμονωμένα δεν γίνεται αντιληπτό που παρουσιάζεται το υψηλότερο Modularity, αλλά εάν καταταχθεί στην συστάδα που ανήκει και ο κόμβος Α, αυξάνει τις συνδέσεις του κόμβου Α εντός της συστάδας της οποίας ανήκει, επομένως και το αποτέλεσμα του Modularity. Εικάζεται ότι ο κόμβος Α έχοντας ποιο πολλές συνδέσεις με την ¨πράσινη¨ συστάδα να επηρεάζει το συνολικό αποτέλεσμα της μετρικής Modularity και να έχουμε υψηλότερη τιμή για τον αλγόριθμο Louvain.

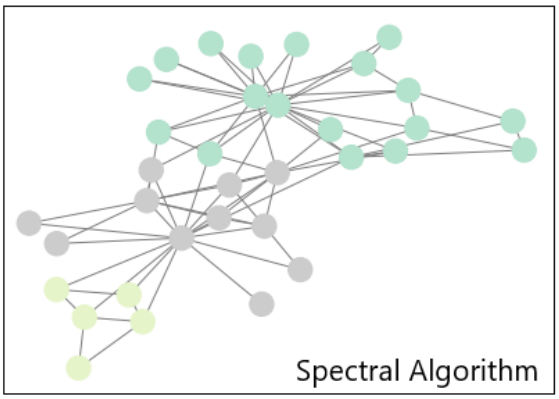
Ο Conductance (ισότητα 7) για τις δύο τμηματοποιήσεις και των δύο αλγορίθμων παρουσιάζει το μικρότερο αποτέλεσμα από όλες τις δοκιμές όταν οι συστάδες είναι δύο. Το καλύτερο αποτέλεσμα από τους δύο αλγορίθμους παρουσιάζεται στην συσταδοποίηση που εκτελείται από τον αλγόριθμο Louvain. Εικάζεται ότι η συγκεκριμένες συστάδες εξυπηρετούν το καλύτερο μέσο όρο αγωγιμότητας του γράφου, με τους κόμβους Α και Β να ανήκουν στην ¨πράσινη¨ ομάδα για τον λόγο που αναφέρθηκε και στην ερμηνεία των αποτελεσμάτων του Modularity. Ομαδοποιόντας τους κόμβους Α και Β με τον τρόπο που έχει εκτελεστεί από τον αλγόριθμο Louvain ελαχιστοποιείται η εξωτερική αγωγιμότητα δηλαδή οι συνδέσεις μεταξύ των συστάδων.

Ο δείκτης αξιολόγησης Q-graph (ισότητα 22) λαμβάνει υπόψη του τις εσωτερικές συνδέσεις (intra_linkage), τις εξωτερικές συνδέσεις (inter_linkage) και το μέτρο διαχωρισμού των συστάδων (Separation). Για τον υπολογισμό των εξωτερικών συνδέσεων και των εξωτερικών χρησιμοποιείται ο δείκτης degeneracy-core που αφορά στον μέγιστο αριθμό των κόμβων που απαρτίζουν τον πυρήνα του γράφου. Από όλους τους αριθμούς συστάδων που έγιναν δοκιμές και για τους δύο αλγόριθμους στις 2 συστάδες ο Q-Graph παρουσίασε το καλύτερο αποτέλεσμα, επιβεβαιώνοντας ότι για το σύνολο δεδομένων karate τα δύο τμήματα είναι ο ιδανικός αριθμός συστάδων. Μεταξύ των αλγορίθμων Spectral και Louvain, ο αλγόριθμος Spectral παρουσίασε την μεγαλύτερη τιμή, οι κόμβοι Α και Β όπως προαναφέρθηκε είναι η διαφορά που παρατηρείται στην συσταδοποίηση που προέκυψε από τους δύο αλγορίθμους. Παρατηρώντας την οπτικοποίηση της συσταδοποίησης του αλγόριθμου Spectral σε σύγκριση με του Louvain συμπεραίνεται ότι όταν οι κόμβοι Α και Β είναι ομαδοποιημένοι κατά την συσταδοποίηση που εκτελείται από τον αλγόριθμο Spectral ενισχύεται το degeneracy-core της άνω ομάδας με αποτέλεσμα καλύτερη ποιότητα συσταδοποίησης κατά τον συγκεκριμένο δείκτη αξιολόγησης.

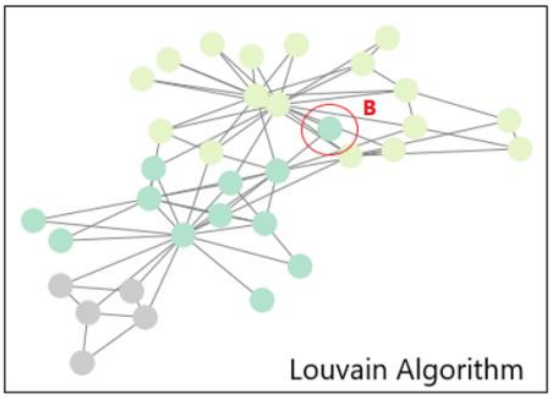
Η τιμή του δείκτη CDS (ισότητα 33) προκύπτει από την αξιολόγηση της πυκνότητας (Density) των ακμών σε σχέση με τον αριθμό τον κόμβων, για τον υπολογισμό της εσωτερικής συνδεσιμότητας (intra-connectivity) των κόμβων εντός της συστάδας όπου ανήκουν, της εξωτερικής συνδεσιμότητας των κόμβων (inter_connectivity) με κόμβους γειτονικών συστάδων και την ποιότητα διαχωρισμού των συστάδων (Separation). Όπως αναφέρεται και στην υποενότητα 4.2.6 επειδή οι κόμβοι στα συνεκτικά δίκτυα είναι κοντά ο ένας στον άλλο και οι κοινότητες είναι δύσκολο να διαχωριστούν , για να αξιολογηθεί η συνδεσιμότητα των συστάδων

σε ένα δίκτυο θα πρέπει να ληφθεί υπόψη και ο δείκτης cohesion, με τον οποίο αξιολογείται το πόσο κοντά είναι οι κόμβοι σε ένα δίκτυο και κατά πόσο μπορούν να διαχωριστούν από άλλους. Στον γράφο karate για τον υπολογισμό του δείκτη CDS, τα βάρη των μετρικών Cohesion, Density και Separation (wc, wd, και ws) έχουν ληφθεί υπόψη με τιμή ένα. Εφόσον η τιμή του βάρους του Separation είναι ένα, ο δείκτης CDS υπολογίζεται από το άθροισμα της εσωτερικής συνδεσιμότητας (intra-connectivity) των κόμβων και το μέτρο διαχωρισμού (Separation). Ο δείκτης αξιολόγησης CDS για το σύνολο δεδομένων karate, παρουσιάζει αύξηση όσο απομακρύνεται η τμηματοποίηση από το τον ιδανικό αριθμό. Παρόλα αυτά με μικρή διαφορά για τις δύο τμηματοποιήσεις ο αλγόριθμος Louvain παρουσιάζει το καλύτερο αποτέλεσμα. Από τα δεδομένα που έχουμε από το γράφο αλλά και από την οπτικοποίηση του, παρατηρείται ότι είναι χαμηλής πυκνότητας γράφος, δηλαδή η αναλογία των ακμών σε σχέση με τους κόμβους δεν είναι αρκετά μεγάλη, για να χαρακτηρίζεται από υψηλό δείκτη πυκνότητας (Density). Στην οπτικοποίηση των συστάδων του αλγόριθμου Louvain, οι κόμβοι A και B εντάσσονται στην ¨πράσινη¨ ομάδα, απομονώνοντας τις συνδέσεις αυτών των δύο κόμβων στην περίπτωση του Spectral έχουμε 4 διακοινοτικές ακμές λόγω του κόμβου A και τις συνδέσεις που έχει με την γειτονική ομάδα και 5 συνδέσεις εντός την ομάδας όπου ανήκει. Στην περίπτωση του αλγόριθμου Louvain έχουμε 4 διακοινοτικές συνδέσεις από τον κόμβο A και 5 ενδοκοινοτικές συνδέσεις λόγω του ίδιου κόμβου. Με βάση τα αποτελέσματα του δείκτη, μπορεί να θεωρηθεί ότι η ομαδοποίηση κατά τον Louvain αυξάνει το βαθμό των εσωτερικών συνδέσεων για την πράσινη ομάδα (intra-connectivity) δίνοντας καλύτερο αποτέλεσμα στον δείκτη αξιολόγησης.

Ο δείκτης GS* (ισότητα 11) αξιολογεί την ομοιότητα των κόμβων με βάση την απόσταση και χρησιμοποιεί έννοιες όπως της συνοχής (Cohesion) και του διαχωρισμού (Separation). Από τα αποτελέσματα που προέκυψαν και για τους δύο αλγόριθμους ο δείκτης εκδηλώνει την καλύτερη τιμή του για αριθμό κοινοτήτων δύο, με τον αλγόριθμο Spectral να αποδίδει καλύτερο αποτέλεσμα. Ερμηνεύοντας τα αποτελέσματα και παρατηρώντας την οπτικοποίηση οι κόμβοι A και B μπορούν να θεωρηθούν πιο κοντά στο κέντρο της άνω συστάδας άρα η συσταδοποίηση που έχει εκτελεστεί από τον Spectral μπορεί και οπτικά να εξηγηθεί για το καλύτερο αποτέλεσμα.



Εικόνα 9. Ομαδοποίηση τριών συστάδων (Spectral)



Εικόνα 10. Ομαδοποίηση τριών συστάδων (Louvain)

Το Modularity (ισότητα 5) για τον αλγόριθμο Louvain σημειώνει καλύτερο αποτέλεσμα από ότι για τον αλγόριθμο Spectral για αριθμό συστάδων τρία. Η διαφορά στην τμηματοποίηση που κάνει ο αλγόριθμος Louvain είναι ο κόμβος Β ο οποίος αποτελούσε μια από τους διαφορετικά τμηματοποιημένους κόμβους και στην προηγούμενη συσταδοποίηση. Το Modularity αυξάνεται όταν οι κόμβοι έχουν περισσότερες συνδέσεις με τους κόμβους της συστάδας στην οποία ανήκουν και λιγότερες συνδέσεις με κόμβους διπλανών συστάδων, στην περίπτωση του κόμβου Β υπάρχει μία ακμή που τον συνδέει με την ομάδα στην οποία ανήκει και μία ακμή με την γειτονική ομάδα. Η μοναδική ακμή που φέρει και τον συνδέει με τον κόμβο της συστάδα στην οποία ανήκει αυξάνει τις συνδέσεις του κόμβου αυτού και έχει ως αποτέλεσμα υψηλό βαθμό Modularity, επηρεάζοντας το τελικό αποτέλεσμα.

Για τις τρείς τμηματοιήσης το Conductance (ισότητα 7) δεν έχει τόσο καλό αποτέλεσμα από ότι στις δύο συσταδοποιήσεις και των δύο αλγορίθμων. Ομοίως με το Modularity και στην περίπτωση του Conductance υπάρχει η ίδια εικασία ότι η ομαδοποίηση του κόμβου Β στην ¨πράσινη¨ συστάδα εξυπηρετεί το καλύτερο αποτέλεσμα του συγκεκριμένου δείκτη για τις συστάδες που έχουν προκύψει από τον Louvain. Όπως και στον παραπάνω δείκτη έτσι και στο Conductance, μπορεί να παρατηρηθεί ότι αν μελετηθεί ατομικά ο κόμβος Β δεν υπάρχει κάποια ομάδα στην οποία μπορεί να καταταχθεί με κριτήριο τον αριθμό τον ακμών του. Άρα και σε αυτές της τμηματοποιήσεις μπορεί να γίνει η υπόθεση πως ενδυναμώνεται η ενδοκοινοτική συνδεσιμότητα και αποδυναμώνεται η εξωτερική συνδεσιμότητα του γειτονικού κόμβου ομαδοποιώντας τους στην ίδια συστάδα.

Ο δείκτης αξιολόγησης Q-graph (ισότητα 22) όπως προαναφέρθηκε χρησιμοποιώντας το δείκτη degeneracy-core που αφορά στον μέγιστο αριθμό των κόμβων που απαρτίζουν τον πυρήνα του γράφου, υπολογίζει την συνδεσιμότητα των κόμβων εντός συστάδας (intra_linkage), και την σύνεση των συστάδων μεταξύ τους (inter_linkage) καθώς αξιολογεί και τον διαχωρισμό των συστάδων με βάση τις διακοινοτικές συνδέσεις (Separation). Για τα αποτελέσματα των τριών τμηματοποιήσεων οι διαφορά των αλγορίθμων εντοπίζεται στον κόμβο Β, το αποτέλεσμα του Q-graph είναι με μικρή διαφορά καλύτερο για τον αλγόριθμο Louvain. Εικάζεται ότι όταν ο κόμβος Β κατηγοριοποιηθεί με βάση την ομαδοποίηση που εκτελείται από τον αλγόριθμο Louvain, να αυξάνει έμμεσα τον αριθμό των κόμβων που συμμετέχουν στον πυρήνα της συστάδας στην οποία ανήκει και να μειώνει τον βαθμό σύνδεσης των δύο συστάδων, με αποτέλεσμα καλύτερη τιμή για τον συγκεκριμένο δείκτη αξιολόγησης.

Ο δείκτης αξιολόγησης CDS (ισότητα 33) για την τμηματοποίηση σε τρείς ομάδες του συνόλου δεδομένων karate, με μικρή διαφορά ο αλγόριθμος Louvain παρουσιάζει το καλύτερο αποτέλεσμα σε σχέση με τον Spectral. Όπως προαναφέρθηκε τα βάρη των μετρικών Cohesion, Density και Separation (wc, wd, και ws) έχουν ληφθεί υπόψη με τιμή ένα. Στην οπτικοποίηση των συστάδων του αλγόριθμου Louvain, ο κόμβος Β εντάσσονται στην ¨πράσινη¨ ομάδα έχοντας μία ακμή να τον ενώνει με την ¨πράσινη¨ ομάδα και μία ακμή να τον ενώνει με την ¨κίτρινη¨ ομάδα. Αυτό πού μπορεί να εκτιμηθεί από την οπτικοποίηση είναι ότι η σύνδεση του με τον κόμβο της ¨πράσινης¨ ομάδας αυξάνει το βαθμό των εσωτερικών συνδέσεων για την πράσινη ομάδα (intra-connectivity) δίνοντας καλύτερο αποτέλεσμα στον δείκτη αξιολόγησης.

Ο δείκτης GS* όπως προαναφέρθηκε χρησιμοποιεί τους δείκτες της συνοχής (Cohesion) και του διαχωρισμού (Separation) για να αξιολογήσει την απόσταση των κόμβων και να

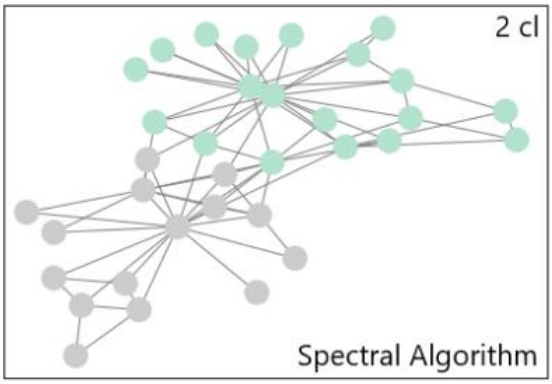
καταλήξει σε πόρισμα σχετικά με την ομοιότητα των κόμβων εντός των συστάδων άρα και την ποιότητα της συσταδοποίησης. Για τον αριθμό των τριών συστάδων στον αλγόριθμος Spectral, έχει σημειωθεί το καλύτερο αποτέλεσμα για τον δείκτη GS*. Παρατηρώντας την οπτικοποίηση των συστάδων μπορεί να δικαιολογηθεί λόγω του ότι ο κόμβος Β είναι πιο κοντά με τους κόμβους της άνω ομάδας από ότι με τους κόμβους της κεντρικής.

Παρατηρήθηκε πως οι δομή του γράφου επηρεάζει τα αποτελέσματα της συσταδοποίησης αλλά και την αξιολόγηση των δεικτών αναλόγως τον μέτρων που χρησιμοποιείται στην κάθε μία.

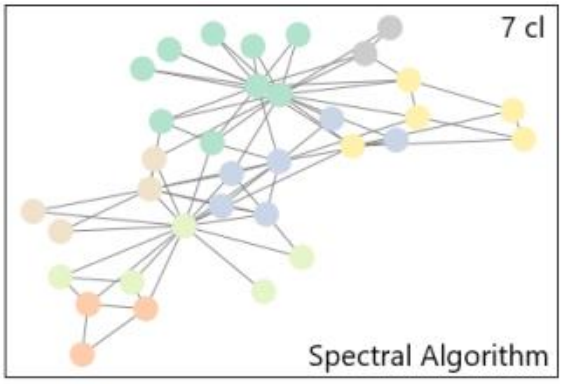
6.2 Οπτικοποίηση γράφου και σύγκριση αποτελεσμάτων μεταξύ των δεικτών αξιολόγησης για το σύνολο δεδομένων karate

Οπτικοποιόντας την δομή του γράφου και τον τρόπο που έχουν συσταδοποιηθεί οι κόμβοι, μπορούν να προκύψουν χρήσιμα συμπεράσματα για τον λόγο που κάθε δείκτης αξιολόγησης έχει υποδείξει τον αντίστοιχο αριθμό συστάδων ως τον καταλληλότερο για την ομαδοποίηση των κόμβων.

Όσον αφορά στον αλγόριθμο Spectral, θα ακολουθήσει οπτικοποίηση του γράφου karate για τις ομαδοποιήσεις των κόμβων σε δύο συστάδες όπου οι δείκτες Modularilty, Conductance, Q-Graph και GS* παρουσίασαν μεταξύ των δοκιμών που έγιναν το καλύτερο τους αποτέλεσμα και οπτικοποίηση του γράφου όπου οι κόμβοι έχουν ομαδοποιηθεί σε επτά συστάδες όπου ο δείκτης CDS παρουσίασε την καλύτερη τιμή του.



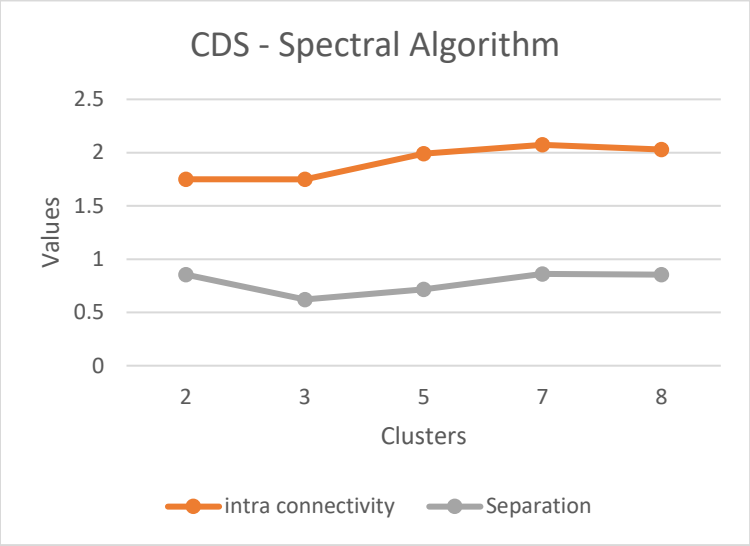
Εικόνα 11. Ομαδοποίηση δύο συστάδων (Spectral)



Εικόνα 12. Ομαδοποίηση εφτά συστάδων (Spectral)

Όπως αναφέρθηκε και στην προηγούμενη υποενότητα ο CDS χρησιμοποιώντας την πυκνότητα (Density) των ακμών σε σχέση με τον αριθμό τον κόμβων, υπολογίζει την εσωτερική συνδεσιμότητα (intra-connectivity) των κόμβων εντός της συστάδας όπου ανήκουν, και με την χρήση της εξωτερικής συνδεσιμότητας των κόμβων (inter_connectivity) με κόμβους γειτονικών συστάδων υπολογίζεται η ποιότητα διαχωρισμού των συστάδων (Separation). Εφόσον Τα βάρη

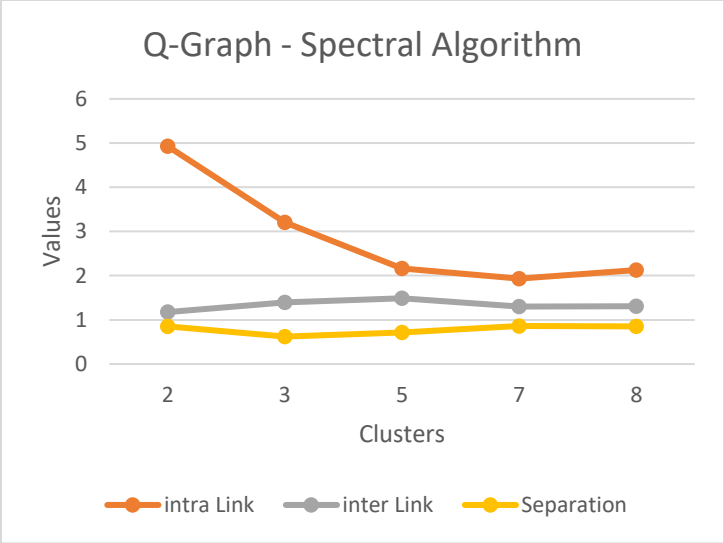
του Separation έχει ληφθεί υπόψη ως μονάδα το αποτέλεσμα του δείκτη CDS (ισότητα 33) προκύπτει από το άθροισμα της τιμής της εσωτερικής συνδεσιμότητας (intra-connectivity) των κόμβων και της ποιότητας διαχωρισμού των συστάδων (Separation). Από τις οπτικοποιήσεις των δύο και των επτά συστάδων δεν είναι εύκολο να γίνει κάποια παρατήρηση δεδομένου ότι πρόκειται για ένα σύνολο δεδομένων με λίγες συνδέσεις μεταξύ των κόμβων, μπορεί όμως να γίνει κάποια ερμηνεία αν παρατηρηθεί ταυτόχρονα η μεταβολή των τιμών της εσωτερικής συνδεσιμότητας (intra-connectivity) των κόμβων και της ποιότητας διαχωρισμού των συστάδων (Separation).



Εικόνα 13. Διαμόρφωση τιμών για τον CDS κατά τον Spectral

Παρατηρείται ότι υπάρχει μία κοινή τάση στην αυξομείωση της τιμής των δύο μετρικών. Για τις εφτά συστάδες παρατηρείται ότι και οι δύο δείκτες παρουσιάζουν την μέγιστη τους τιμή, υποδεικνύοντας ότι για αυτό τον αριθμό συστάδων, οι συνδέσεις μεταξύ των κόμβων τις ίδιας συστάδας είναι περισσότερες από τις συνδέσεις με κόμβους εξωτερικών συστάδων.

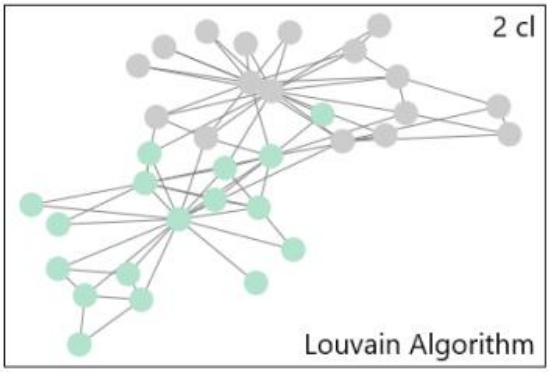
Η μετρική Q-Graph (ισότητα 22) επίσης αξιολογεί τις συσταδοποιήσεις με βάση την ποιότητα διαχωρισμού των συστάδων (Separation), τον δείκτη συνδεσιμότητας των κόμβων εντός των ομάδων όπου ανήκουν (intra linkage) και των δείκτη συνδεσιμότητας των κόμβων με γειτονικές συστάδες (inter_linkage). Η διαφορά του Q-Graph από τον δείκτη CDS είναι ότι υπολογίζει την εξωτερική και την εσωτερική συνδεσιμότητα χρησιμοποιώντας τον δείκτης degeneracy που αφορά στον μέγιστο αριθμό των κόμβων v που ανήκουν σε ένα σύνολο κόμβων V, ενώ ο CDS για τον υπολογισμό αυτών των δεικτών εξετάζει την πυκνότητα των συνδέσεων. Ο δείκτης Q-Graph υπολογίζεται από την διαφορά του δείκτη intra linkage από τον δείκτη inter_linkage συν την τιμή της ποιότητας διαχωρισμού (Separation). Η διακύμανση των τιμών των δεικτών για τον Q-Graph αποτυπώνεται στο παρακάτω γράφημα:



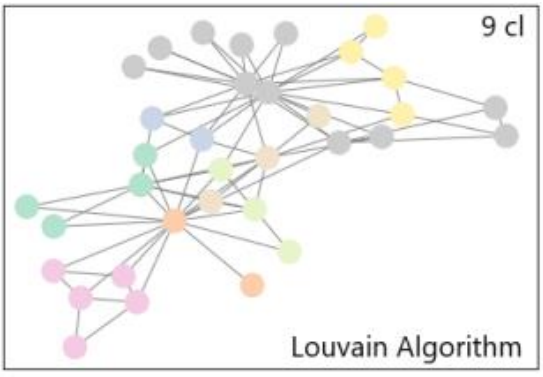
Εικόνα 14. Διαμόρφωση τιμών για τον Q-Graph κατά τον Spectral

Αυτό που παρατηρείται είναι ότι το Separation το οποίο υπολογίζεται με αντίστοιχο τρόπο και για τον Q-Graph έχει την ίδια συμπεριφορά παρουσιάζοντας την μέγιστη τιμή του για τις εφτά συστάδες. Ο δείκτης συνδέσεων των κόμβων με εξωτερικές συστάδες (inter_linkage) παρουσιάζει την ελάχιστη τιμή του για την ομαδοποίηση σε 2 συστάδες, υποδεικνύοντας ότι δεν υπάρχουν πολλές συνδέσεις μεταξύ κόμβων διαφορετικών συστάδων. Ο δείκτης των εσωτερικών συνδέσεων των κόμβων με κόμβους που ανήκουν στην ίδια συστάδα (intra linkage) επίσης παρουσιάζει την μέγιστη του τιμή για τις δυο συστάδες , αξιολογώντας την ομαδοποίηση σε δύο συστάδες ως την καλύτερη.

Αντίστοιχη είναι η συμπεριφορά των δύο δεικτών για τον αλγόριθμο Louvain, στις ομαδοποιήσεις των κόμβων σε δύο συστάδες, οι δείκτες Conductance, Q-Graph και GS* παρουσίασαν μεταξύ των δοκιμών που έγιναν το καλύτερο τους αποτέλεσμα, ο δείκτης Modularilty είχε καλύτερο αποτέλεσμα για τις τρείς ομάδες και ο δείκτης CDS παρουσίασε την καλύτερη τιμή του στην ομαδοποίηση των εννιά συστάδων. Παρατηρείται άλλη μια φορά για το ίδιο σύνολο δεδομένων ότι ο δείκτης CDS αυξάνει όσο απομακρύνεται η ομαδοποίηση από τον ιδανικό αριθμό συστάδων.

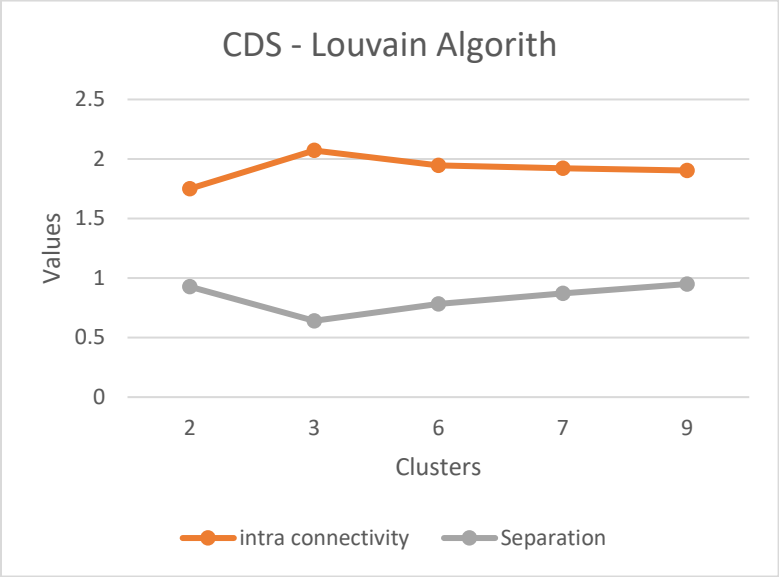


Εικόνα 15. Ομαδοποίηση δύο συστάδων Louvain



Εικόνα 16. Ομαδοποίηση εννιά συστάδων Louvain

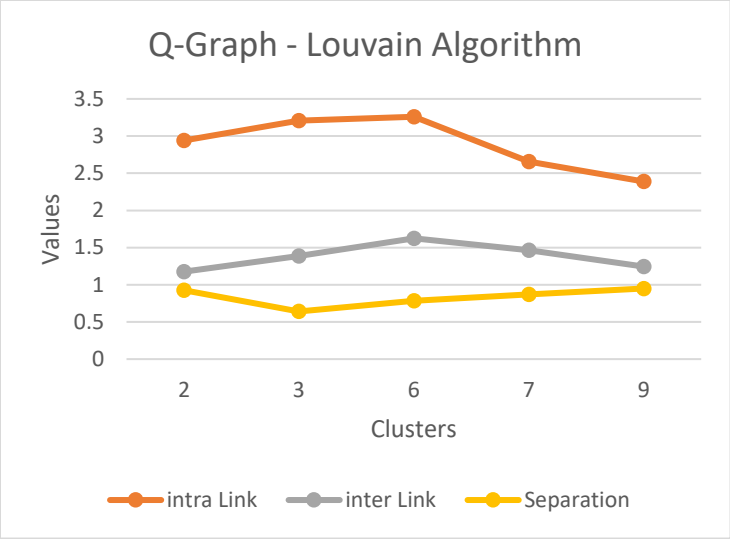
Είναι γνωστό για το σύνολο δεδομένων karate ότι ο ιδανικός αριθμός συστάδων είναι δύο, εκτός του δείκτη CDS οι υπόλοιποι δείκτες παρουσίασαν την καλύτερη τους τιμή στις δύο συστάδες ή στις τρείς συστάδες. Ο δείκτης CDS (ισότητα 33) όχι μόνο δεν επιβεβαίωσε τον ιδανικό αριθμό συστάδων αλλά είχε και αυξητική συμπεριφορά όσο μεγάλωνε ο αριθμός των συστάδων.



Εικόνα 17. Διαμόρφωση τιμών για τον CDS κατά τον Louvain

Στις δύο συστάδες ο δείκτης Separation έχει καλή τιμή σε σχέση με την τιμή που έχει για τις υπόλοιπες συσταδοποιήσεις αλλά ο δείκτης intra-connectivity παρουσιάζει την ελάχιστη τιμή του μεταβάλλοντας το αποτέλεσμα. Αντίστοιχα ο δείκτης intra-connectivity αποκτά την μέγιστη τιμή του για αριθμό συστάδων τρία αλλά ο δείκτης Separation στις τρείς συστάδες σημειώνει την μικρότερη του τιμή. Ο δείκτης CDS αξιολογεί την ομαδοποίηση των κόμβων σε εννιά συστάδες ως την καλύτερη όπου και το intra-connectivity και το Separation παρουσιάζουν αύξηση στην τιμή.

Συγκριτικά με τον δείκτη Q-Graph που επίσης χρησιμοποιεί την εσωτερική συνδεσιμότητα (intra linkage) σε σχέση με την εξωτερική συνδεσιμότητα (inter_linkage) των κόμβων, λαμβάνοντας υπόψη και το Separation παρατηρείται ότι ο συγκεκριμένος δείκτης επιβεβαιώνει τον ιδανικό αριθμό συστάδων. Στο γράφημα [Εικόνα 18] παρατηρούνται οι διακυμάνσεις των τιμών για τα κριτήρια αξιολόγησης που χρησιμοποιεί ο δείκτης Q-Graph.

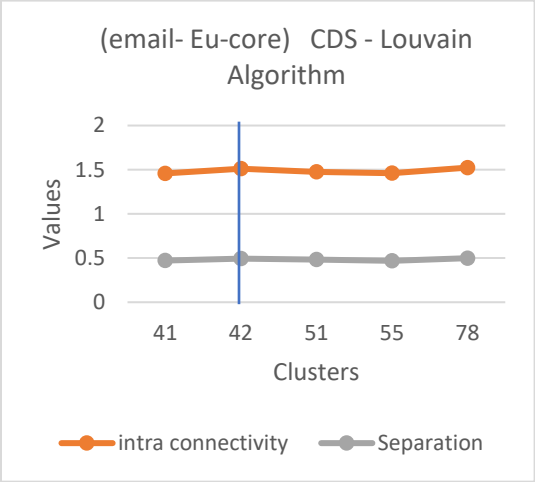


Εικόνα 18. Διαμόρφωση τιμών για τον Q-Graph κατά τον Louvain

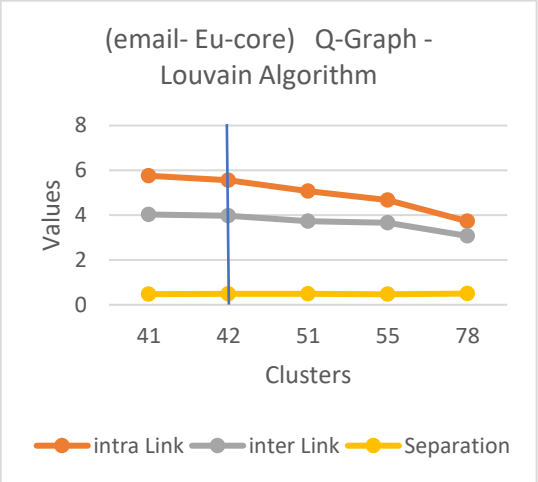
Ο δείκτης Separation παρουσιάζει αύξηση στις δύο και στις εννιά συστάδες με το αποτέλεσμα στις εννιά συστάδες να είναι το υψηλότερο. Ο δείκτης intra Linkage παρουσιάζει την μέγιστη τιμή για τις έξι συστάδες, όπου ο δείκτης inter linkage παρουσιάζει επίσης την μεγαλύτερη τιμή του και έτσι η διαφορά τους δεν είναι αρκετά μεγάλη για να αποτελέσει η συσταδοποίηση των έξι ομάδων την ποιοτικότερη. Ο ποιοτικότερος αριθμός συστάδων με βάση των συγκεκριμένο δείκτη είναι οι δύο συστάδες όπου η διαφορά μεταξύ εσωτερικής και εξωτερικής συνδεσιμότητας αποκτά την μέγιστη τους τιμή.

Παρατηρείται πως αν και ο δείκτης της ποιότητας διαχωρισμού των συστάδων(Separation) που λαμβάνει υπόψη του το ποσοστό πυκνότητας των συνδέσεων, έχει υψηλές τιμές για τις ομαδοποιήσεις κοντά στον ιδανικό αριθμό συστάδων(ground truth), οι συγκεκριμένοι δείκτες που μελετώνται λαμβάνουν υπόψη τους τουλάχιστον ένα ακόμα χαρακτηριστικό όπως αναφέρθηκε, την συνδεσιμότητα των κόμβων εντός των συστάδων αλλά και τη συνδεσιμότητα των κόμβων με κόμβους άλλων συστάδων. Κάθε δείκτης προσεγγίζει διαφορετικά τον υπολογισμό της συνδεσιμότητας για αυτό και δεν υπάρχει συμφωνία στα αποτελέσματα. Συμπεραίνεται πως λόγω τον αραιών συνδέσεων του συγκεκριμένου γράφου ο δείκτης CDS που υπολογίζει την εσωτερική και εξωτερική συνδεσιμότητα βασιζόμενος στην πυκνότητα των ακμών δεν απέδωσε σε σχέση με τον δείκτη Q-Graph ο οποίος υπολογίζει αντίστοιχα την εξωτερική και εσωτερική συνδεσιμότητα των κόμβων αλλά χρησιμοποιώντας τον δείκτη degeneracy-core που βασίζεται τον μέγιστο αριθμό κόμβων που απαρτίζουν τον πυρήνα του γράφου.

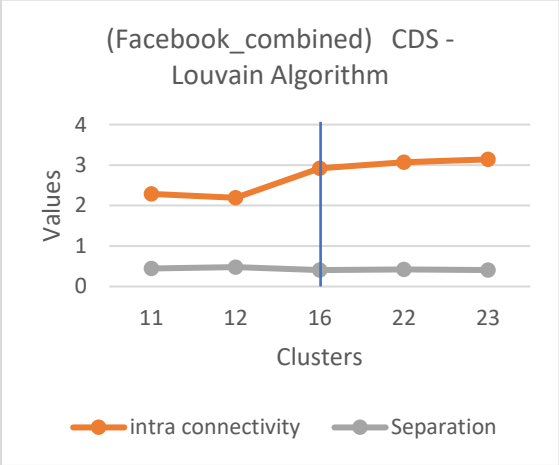
Αντιστοίχως για τα υπόλοιπα σύνολα δεδομένων τα παρακάτω γραφήματα παρουσιάζουν την διακύμανση των τιμών των δεικτών.



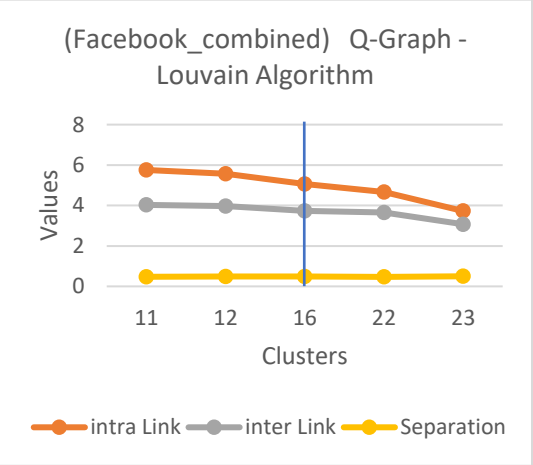
Εικόνα 19. Email-Eu-core: CDS κατά τον Louvain



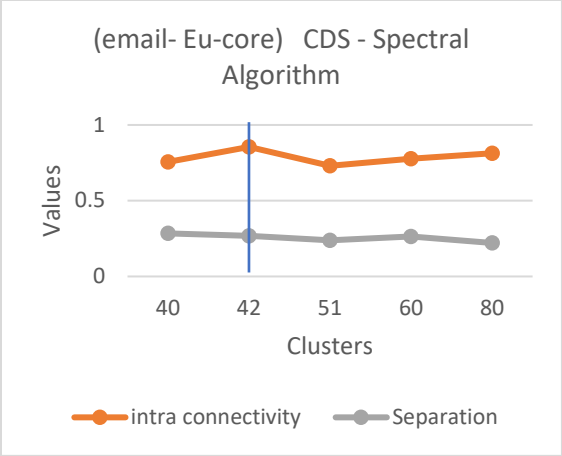
Εικόνα 20. Email-Eu-core: Q-Graph κατά τον Louvain



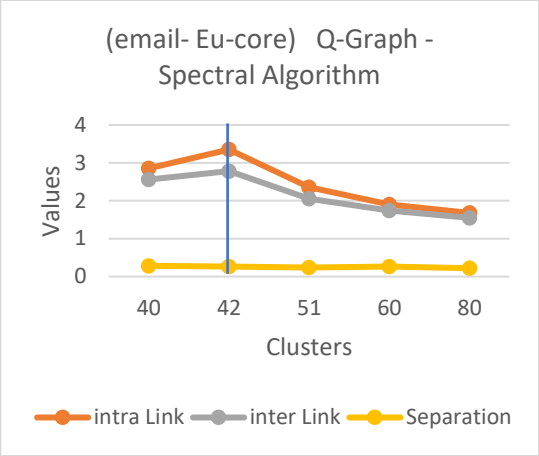
Εικόνα 21. Facebook_combined: CDS κατά τον Louvain



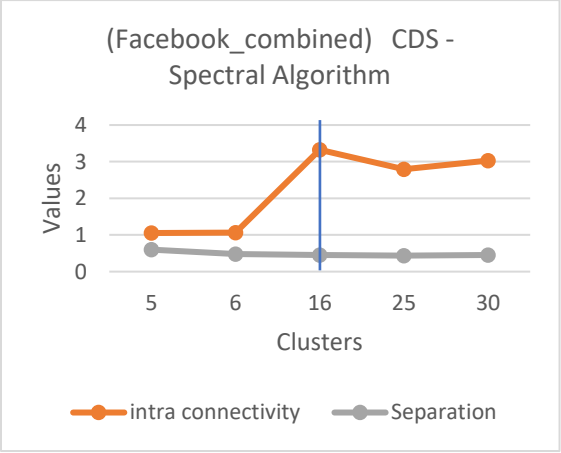
Εικόνα 22. Facebook_combined: Q-Graph κατά τον Louvain



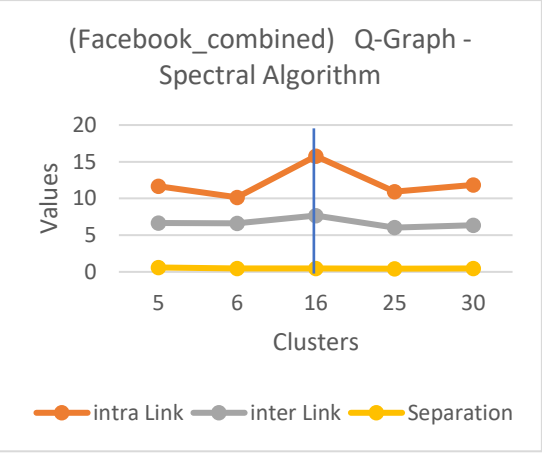
Εικόνα 23. Email-Eu-core: CDS κατά τον Spectral



Εικόνα 24. Email-Eu-core: Q-Graph κατά τον Spectral



Εικόνα 25. Facebook_combined: CDS κατά τον Spectral



Εικόνα 26. Facebook_combined: Q-Graph κατά τον Spectral

Στα γραφήματα έχει σημειωθεί με μπλε γραμμή το σημείο όπου οι τιμές των δεικτών αξιολογούν τον αριθμό των ιδανικών συσταδοποιήσεων. Ο δείκτης Separation αναφέρεται σε πολλές έρευνες και μελέτες ως ένας σημαντικός δείκτης αξιολόγησης για την ποιότητα διαχωρισμού των συστάδων. Αυτό που μπορεί να παρατηρηθεί από τα γραφήματα είναι πως ο δείκτης Separtation για τα συγκεκριμένα σύνολα δεδομένων έχει πολύ μικρές διακυμάνσεις τιμής, τις τάξεως υποδιαιρέσεων της μονάδας από δοκιμή σε δοκιμή, όπως επίσης ότι δεν έχει αξιολογήσει ως καλώς διαχωρισμένες τις συστάδες σε κάθε περίπτωση τον ιδανικό αριθμό συσταδοποίησης για κάθε σύνολο δεδομένων με βάση την βιβλιογραφία. Οι δείκτες αξιολόγησης Q-Graph και CDS λαμβάνουν υπόψη τους το Separation για το αποτέλεσμα της αξιολόγησης, παρόλα αυτά λόγω του χαμηλού εύρους τιμών και μικρής διακύμανσης αποτελεσμάτων που παρουσιάζει ο συγκεκριμένος δείκτης, σε σχέση με τις υπόλοιπες μετρικές που χρησιμοποιούνται από τους δείκτες αξιολόγησης, ο δείκτης Separation θα μπορούσε να καθορίσει το αποτέλεσμα με βάση τις πειραματικές δοκιμές μόνο στην περίπτωση που τα αποτελέσματα των υπόλοιπων δεικτών παρουσίαζαν διαφορά στην τιμή των δεκαδικών τους ψηφίων. Έτσι παρατηρείται ότι οι δείκτες intra connectivity, intra Link και inter Link λόγω του υψηλότερου εύρους τιμών και της

μεγαλύτερης διακύμανσης που παρουσιάζουν στα αποτελέσματα τους έχουν την κύρια επίδραση στο αποτέλεσμα της αξιολόγησης. Η μικρή συνεισφορά του δείκτη Separation ενδεχομένως οφείλεται στο ότι έχει χρησιμοποιηθεί στο υπολογισμό των δεικτών το βάρος διαχωρισμού (w_s) με την τιμή 1, ένας άλλος λόγος ο οποίος ίσως ευθύνεται για το χαμηλό εύρος τιμών είναι ο τρόπος που έχουν καταταχθεί σε συστάδες οι κόμβοι και έχουν οριστεί οι κοινότητες, καθώς το inter connectivity δηλαδή η διακοινοτική συνδεσιμότητα συμβάλει στον υπολογισμό της ποιότητας διαχωρισμού.

6.3 Παρατήρηση και ερμηνεία αποτελεσμάτων των συνόλων δεδομένων

Βάση αποτελεσμάτων της παρούσας ενότητας στο σύνολο δεδομένων karate μπορεί να γίνει ορατό πως η κατάταξη των κόμβων και οι συνδέσεις, μπορούν να επηρεάσουν το αποτέλεσμα των δεικτών αξιολόγησης. Πρόκειται για ένα σύνολο δεδομένων με μικρό αριθμό κόμβων και ακμών, το οποίο απαρτίζεται από δύο ομάδες. Παρατηρώντας τα αποτελέσματα για τα υπόλοιπα σύνολα δεδομένων θα μπορούσε να γίνει καποια ερμηνεία για την συμπεριφορά των δεικτών αξιολόγησης.

Ο παρακάτω πίνακας αποτυπώνει, για κάθε σύνολο δεδομένων που χρησιμοποιήθηκε ποιοι δείκτες πέτυχαν καλύτερο αποτέλεσμα για την ομαδοποίηση των κόμβων κοντά στον αριθμό της ιδανικής ομαδοποίησης και από ποιόν αλγόριθμο προέκυψε η ομαδοποίηση αυτή.

Dataset	Index	Spectral	Louvain
Email-Eu-core	Modularilty	✓	✓
	Conductance	✓	✓
	Q-Graph	✓	✓
	CDS	✓	-
	GS*	✓	-
karate	Modularilty	✓	-
	Conductance	✓	✓
	Q-Graph	✓	✓
	CDS	-	-
	GS*	✓	✓
Facebook combined	Modularilty	✓	✓
	Conductance	-	-
	Q-Graph	✓	-
	CDS	✓	✓
	GS*	-	✓

Πίνακας 10. Οι δείκτες που επέφεραν το καλύτερο αποτέλεσμα για κάθε συσταδοποίηση στα σύνολα δεδοένων

Παρατηρείται ότι ο δείκτης Modularilty έχει αποδώσει καλή τιμή για τον ιδανικό αριθμό συστάδων σε όλα τα σύνολα και για τους δύο αλγορίθμους, εκτός από τις ομαδοποιήσεις που έγιναν από τον αλγόριθμο Louvain για το σύνολο δεδομένων karate και email-Eu-core, παρόλο που ο αλγόριθμος δρα με την αύξηση του δείκτη Modularilty, η αστοχία θα μπορούσε να προέλθει από το γεγονός ότι στην περίπτωση του γράφου karate πρόκειται για ένα σύνολο δεδομένων με λίγες ακμές μεταξύ των κόμβων, και στην περίπτωση του γράφου email-Eu-core ίσως είναι αποτέλεσμα των συστάδων που προέκυψαν από τον αλγόριθμο Louvain μιας και η καλύτερη του τιμή σημειώνεται στις 41 συστάδες, πολύ κοντά στον ιδανικό αριθμό. Ο δείκτης Conductance αξιολογεί ως καλώς διαχωρισμένες τις κοινότητες που βρίσκονται κοντά στον ιδανικό αριθμό συστάδων για κάθε αλγόριθμο για το σύνολο δεδομένων karate, αντιθέτως για το σύνολο δεδομένων Facebook_combined ο δείκτης δεν επιβεβαίωσε τον ιδανικό αριθμό συστάδων. Με βάση τα αποτελέσματα που παρουσιάζονται στους πίνακες 6,7 παρατηρείται ότι ο δείκτης επιφέρει καλύτερα αποτελέσματα όσο ο αριθμός συστάδων μικραίνει, το σύνολο δεδομένων Facebook_combined αποτελείται από 88234 ακμές μπορούμε να έχουμε μια ιδέα για την δομή του και από την εικόνα 4, συσχετίζοντας τα δεδομένα θα μπορούσε να υποθεί ότι ο αριθμός περισσότερων ακμών μεταξύ των συστάδων σημαίνουν πιο ακριβά κοψίματα. Όπως διαπιστώθηκε και σε σχετική μελέτη [4], χωρίς εξισορρόπηση της εξωτερικής αγωγιμότητας με την εσωτερική αγωγιμότητα, τα αποτελέσματα θα μας δώσουν ελλιπή αποτελέσματα. Ο δείκτης Conductance επίσης δεν απέδωσε καλά αποτελέσματα για τις συστάδες που προέκυψαν από τον αλγόριθμο Louvain για το σύνολο δεδομένων email-Eu-core, είναι ένα σύνολο δεδομένων που όπως και στον γράφο facebook_combined υπάρχει μεγάλος αριθμός ακμών, ενδέχεται ο τρόπος που κατηγοριοποιήθηκαν οι κόμβοι να είχαν αυξημένη εξωτερική αγωγιμότητα στις 42 συστάδες από ότι στις 41 που ο δείκτης σημείωσε την μεγαλύτερη τιμή. Ο δείκτης Q-Graph επίσης αξιολογεί βασιζόμενος στην συνδεσιμότητα των κόμβων και συγκρίνει τις εσωτερικές συνδέσεις (intra_linkage) των κόμβων με τις εξωτερικές συνδέσεις (inter_linkage) των κόμβων με γειτονικές συστάδες, το μέτρο που χρησιμοποιείται είναι το degeneracy-core που αφορά στον μέγιστο αριθμό των κόμβων που απαρτίζουν τον πυρήνα του γράφου. Σε όλες τις δοκιμές που διεξήχθησαν στην ενότητα 5, τα αποτελέσματα του συγκεκριμένου δείκτη ήταν υψηλά όταν ο αριθμός των συστάδων ήταν ίδιος με τον ιδανικό αριθμό κοινοτήτων που ορίζεται από σχετικές μελέτες ιδιαίτερα εκείνων που προέκυψαν από τον αλγόριθμο Spectral. Στην περίπτωση όμως του συνόλου δεδομένων Facebook_combined και email-Eu-core και τις συσταδοποιήσεις που προέκυψαν από τον αλγόριθμο Louvain παρατηρήθηκε ότι όσο πιο μικρός ο αριθμός συστάδων τόσο καλύτερο το αποτέλεσμα του δείκτη Q-Graph. Οι κόμβοι του πυρήνα των συστάδων καθορίζονται από τις συνδέσεις που φέρουν οι κόμβοι, άρα η δομή των δύο γράφων Facebook_combined και email-Eu-core , λόγω του μεγάλου ποσοστού διακοινοτικών συνδέσεων δεν επιτρέπουν με την συγκεκριμένη τεχνική να αποκτηθεί μια ολοκληρωμένη εκτίμηση όσων αφορά στην ποιότητα της συσταδοποίησης. Ο δείκτης CDS αξιολογεί με βάση την πυκνότητας (Density) των ακμών σε σχέση με τον αριθμό τον κόμβων, για τον υπολογισμό της εσωτερικής συνδεσιμότητας (intra-connectivity) των κόμβων εντός της συστάδας όπου ανήκουν, της εξωτερικής συνδεσιμότητας των κόμβων (inter_connectivity) με κόμβους γειτονικών συστάδων και την ποιότητα διαχωρισμού των συστάδων (Separation). Για τα σύνολα δεδομένων Email-Eu-core και Facebook_combined για τις κοινότητες που προέκυψαν από τον αλγόριθμο Spectral απέδωσε υψηλή τιμή στον ιδανικό αριθμό τμηματοποιήσεων, με τον αλγόριθμο Louvain σημείωσε το επιθυμητό αποτέλεσμα επιβεβαιώνοντας τον ιδανικό αριθμό τμηματοποιήσεων

μόνο για το σύνολο δεδομένων Facebook_combined. Για το σύνολο δεδομένων karate ο αλγόριθμος CDS δεν έδωσε ικανοποιητικό αποτέλεσμα για κανέναν από τους δύο αλγορίθμους στις δύο κοινότητες, όπως είδαμε και παραπάνω πρόκειται για έναν γράφο με λίγους κόμβους και αραιές συνδέσεις, έτσι ένας δείκτης ο οποίος βασίζεται στην πυκνότητα των συνδέσεων φαίνεται πως δεν είναι ο κατάλληλος για την αξιολόγηση των αποτελέσμάτων ομαδοποίησης ενός γράφου με αυτή την δομή. Τέλος ο δείκτης GS* χρησιμοποιεί έννοιες όπως της συνοχής (Cohesion) και του διαχωρισμού (Separation), στις περιπτώσεις που ο δείκτης δεν έδωσε καλό αποτέλεσμα στον ιδανικό αριθμό κοινοτήτων κάθε συνόλου δεδομένων ήταν για τις κοινότητες που προέκυψαν από τον αλγόριθμο Louvain στο σύνολο δεδομένων Email-Eu-core και για τα αποτελέσματα του αλγορίθμου Spectral για τις ομαδοποιήσεις στο σύνολο δεδομένων του Facebook_combined. Γνωρίζοντας ότι ο συγκεκριμένος δείκτης βασίζεται στις αποστάσεις των συστάδων και παρατηρώντας ότι σε καμία δοκιμή δεν έχει επιφέρει αποτέλεσμα κοντά στο 1 που σημαίνει ότι οι συστάδες είναι καλώς διαχωρισμένες και οι περισσότερες προσεγγίζουν το 0, γίνεται αντιληπτό ότι οι αποστάσεις των ομάδων δεν είναι σημαντικές.

Με την παρατήρηση των αποτελεσμάτων, γνωρίζοντας τις έννοιες που χρησιμοποιεί κάθε αλγόριθμος συσταδοποίησης και λαμβάνοντας υπόψη τα δομικά χαρακτηριστικά του γράφου, ερμηνεύθηκαν τα αποτελέσματα του πειραματικού μέρους. Η ερμηνεία των αποτελεσμάτων δεν οδήγησε στην ανάδειξη κάποιου αλγόριθμου συστάδοποπιησης αλλά ούτε και κάποιου δείκτη αξιολόγησης που θα ήταν αποδοτικός σε κάθε περίπτωση. Μπορεί όμως να σημειωθεί ότι από τις δοκιμές που διεξάχθηκαν οι ομαδοποιήσεις που προέκυψαν από τον αλγόριθμο Spectral ήταν οι περισσότερες από τις οποίες οι δείκτες αξιολόγησης χαρακτήρισαν ως καλώς ομαδοποιημένες και από τους δείκτες αξιολόγησης ο δείκτης Q-Graph ήταν ο δείκτης που πέτυχε την καλύτερη τιμή στις περισσότερες ομαδοποιήσεις που έγιναν κατά τον ιδανικό αριθμό ομαδοποιήσεων (ground truth) με βάση σχετική βιβλιογραφία.

ΚΕΦΑΛΑΙΟ 7 : Συμπεράσματα

Η ομαδοποίηση των κόμβων των συνόλων δεδομένων γράφων έχει σκοπό την ομαδοποίηση κόμβων που μοιράζονται όμοια χαρακτηριστικά. Το ζήτημα που προκύπτει απο την διαδικασία της συσταδοποίησης είναι η ποιότητα των ομάδων που έχουν δημιουργηθεί για να επιβεβαιώσουν την αξιοπιστία των αποτελεσμάτων. Στην παρούσα μελέτη χρησιμοποιήθηκαν τρία διαφορετικά σύνολα δεδομένων. Από την μεγάλη ποικιλία αλγορίθμων συσταδοποίησης επιλέχθηκαν οι αλγόριθμοι Spectral και Louvain τα αποτελέσματα των οποίων αξιολογήθηκαν με τη χρήση των δεικτών αξιολόγησης Modularity, Conductance, Q-Graph, CDS και GS*. Με την χρήση του δείκτη Rand Index πραγματοποιήθηκε σύγκριση των συστάδων που προέκυψαν από τους αλγόριθμους συσταδοποίησης και των συστάδων που ορίζονται από το σύνολο δεδομένων ground truth. Όπως επίσης οι δείκτες αξιολόγησης εφαρμόστηκαν και στο σύνολο δεδομένων ground truth. Οι δύο αλγόριθμοι έχουν διαφορετικά κριτήρια εντοπισμού κοινοτήτων, καθώς και οι δείκτες ποικίλουν όσον αφορά στα χαρακτηριστικά που αξιολογούν, προκειμένου να προσδιοριστεί η ποιότητα της συσταδοποίησης. Βάση αποτελεσμάτων της παρούσας μελέτης έγινε αντιληπτό ότι η δομή του γράφου είναι πολύ σημαντικός παράγοντας τόσο για την επιλογή αλγορίθμου ομαδοποίησης αλλά και για την αξιοπιστία των αποτελεσμάτων των δεικτών αξιολόγησης. Με βάση τον τρόπο που επηρεάζονται οι δείκτες αξιολόγησης και τους λόγους που συντελούν στην διαφοροποίηση των ομαδοποιήσεων ανά αλγόριθμο, μπορεί να αναπτυχθεί μια προδιάθεση για τον τρόπο που θα είναι πιο αποδοτικός να επεξεργαστεί και συσταδοποιηθεί ένα σύνολο δεδομένων, καθώς και να επιλεχθούν οι αντίστοιχοι αλγόριθμοι ομαδοποίησης και δείκτες για την αξιολόγηση των αποτελεσμάτων της ομαδοποίησης. Από την συγκεκριμένη μελέτη προκύπτει ότι από τους δείκτες που δοκιμάστηκαν δεν υπάρχει ένας δείκτης που θα μπορέσει να αποδώσει τον χαρακτηρισμό της καλής συσταδοποίησης, για ομαδοποιήσεις που προκύπτουν από διάφορους αλγόριθμους, για όλα τα σύνολα δεδομένων. Επίσης οι ετικέτες που ορίζονται για κάθε κόμβο με βάση το σύνολο δεδομένων ground truth, δεν είναι πάντοτε το σημείο αναφοράς με το οποίο θα πρέπει να επιδιώκουμε να ταυτίσουμε τις συστάδες που έχουν προκύψει από τους αλγόριθμους ομαδοποίησης, σε πολλές περιπτώσεις δεν μπορούν να ανακτηθούν όλες οι ετικέτες από τους αλγορίθμους διότι δεν αντιπροσωπεύονται από την δομή του γράφου. Το σύνολο δεδομένων ground truth μπορεί όμως να χρησιμοποιείται για να αντληθούν χρήσιμες πληροφορίες για το σύνολο δεδομένων, σε κάποιες περιπτώσεις να δώσει μια εκτίμηση για τον κατά προσέγγιση αριθμό των συστάδων του γράφου όπως επίσης να συμμετέχει σε ενδιαφέρουσες παρατηρήσεις. Στην ερμηνεία των αποτελεσμάτων δόθηκε ιδιαίτερη σημασία στο τρόπο που η δομή του γράφου επηρεάζει την διαδικασία συσταδοποίησης αλλά και την αξιολόγηση των δεικτών. Παρατηρήθηκε ότι ο ίδιος αριθμός συστάδων για δύο διαφορετικούς αλγόριθμους δεν υποδηλώνει την ίδια ετικέτα για κάθε κόμβο. Η κατηγοριοποίηση των κόμβων και οι συστάδες που προκύπτουν διαφέρουν, όπως και ο τρόπος δράσης τους. Το να λαμβάνονται υπόψη τα χαρακτηριστικά και η δομή του γράφου είναι σημαντικό ώστε να επεξεργάζεται γρηγορότερα και αποτελεσματικότερα

ένα σύνολο δεδομένων. Η αξιολόγηση των αποτελεσμάτων είναι απαραίτητη για να επικυρώσει την αξιοπιστία των αποτελεσμάτων της συσταδοποίησης ιδιαίτερα για νέα σύνολα δεδομένων. Αυτό που προτείνεται είναι η επιλογή και δοκιμή αλγορίθμων ομαδοποίησης με βάση τα χαρακτηρίστηκα της δομής των συνόλων δεδομένων αλλά και η αξιολόγηση των αποτελεσμάτων από ανάλογους δείκτες.

ΚΕΦΑΛΑΙΟ 8 : Προτάσεις για μελλοντική μελέτη

Στη παρούσα μελέτη πραγματοποιήθηκε μια πειραματική έρευνα των παράγοντες που επηρεάζουν τους δείκτες αξιολόγησης με βάση τα κριτήρια που χρησιμοποιούν. Σε μελλοντική μελέτη θα ήταν ενδιαφέρων να μελετηθεί η συμπεριφορά νέων δεικτών αξιολόγησης που θα προκύπτουν από εκείνους που χρησιμοποιήθηκαν στην παρούσα μελέτη, αλλά διαφοροποιώντας τα κριτήρια τους ώστε να παρατηρηθεί η απόδοση τους σε συνδυασμό με τα χαρακτηριστικά των κοινοτήτων και να εξεταστεί η δημιουργία ενός μοτίβου επιλογής του κατάλληλου δείκτη. Για παράδειγμα παρατηρήθηκε ότι για τα συγκεκριμένα σύνολα δεδομένων ο δείκτης Separation είχε μικρή συνεισφορά στο τελικό αποτέλεσμα των δεικτών που το χρησιμοποίησαν, θα μπορούσε να ληφθεί υπόψη με διαφορετικό τρόπο στον υπολογισμό του δείκτη είτε και να αντικατασταθεί από κάποια άλλη μετρική για τα σύνολα δεδομένων με αντίστοιχη δομή. Όσον αφορά στην διασφάλιση της βέλτιστης επιλογής δείκτη αξιολόγησης για να υπάρξει αξιόπιστο αποτέλεσμα και να αποκτηθεί η γνώση για το αποτέλεσμα συσταδοποίησης, προτείνονται δοκιμές σε περισσότερα σύνολα δεδομένων, τα σύνολα δεδομένων θα μπορούσαν να κατηγοριοποιηθούν με βάση τα δομικά τους χαρακτηριστικά, οι αλγόριθμοι συσταδοποίησης θα μπορούσαν επίσης να κατηγοριοποιηθούν με βάση τα κριτήρια συσταδοποίησης των κόμβων, έτσι θα αποκτηθεί περαιτέρω γνώση για την επιλογή της βέλτιστης επεξεργασίας με βάση τα χαρακτηριστικά του γράφου, που θα επιτρέπει μετά το πέρας της στον παρατηρητή να καταλήξει σε χρήσιμα συμπεράσματα για τα χαρακτηριστικά των κοινωτήτων στις οποίες ανήκουν οι κόμβοι.

Βιβλιογραφική αναφορά

Η εφαρμογή αλγορίθμων clustering σε γράφους με σκοπό την ανίχνευση κοινοτήτων και η αξιολόγηση τους με την χρήση δεικτών είναι ένα θέμα που έχει κεντρίσει το ενδιαφέρων στην ακαδημαϊκή κοινότητα. Στην μελέτη [1] αξιολογούνται έννοιες όπως η συνδεσιμότητα και διαχωρισμός κοινοτήτων σε ένα γράφο. Μελετάται και αξιολογείται η ποιότητα του διαχωρισμού σε συστάδες γράφων, με την χρήση των εννοιών structural cohesion και graph density που αφορούν στη σύνδεση των κόμβων εντός και μεταξύ των συστάδων. Καθώς παρουσιάζεται και προτείνεται ένα νέος δείκτης αξιολόγησης των κοινοτήτων σε γράφους ο CDS ο οποίος θα αναφερθεί λεπτομερώς παρακάτω. Στην έρευνα [2] εξετάζονται δείκτες αξιολόγησης και προτείνονται ρυθμιζόμενα μέτρα. Προτείνονται διάφοροι δείκτες αξιολόγησης συσταδοποίησης γράφων, πολλοί από αυτούς βασίζονται στην μεταβολή των intra και inter-cluster, δηλαδή την μεταβολή της πυκνότητας εντός και εκτός μιας κοινότητας. Πιστεύεται ότι οι κατάλληλοι δείκτες αξιολόγησης συσταδοποίησης μπορούν να απεικονισθούν στο σχεδιάγραμμα των συστάδων. Για παράδειγμα όσο πιο συμπαγής είναι μια συστάδα τόσο πιο σκοτεινό είναι το φόντο της. Τα μέτρα αξιολόγησης είναι απαραίτητα για να κατανοηθεί και να διαχειριστεί την κατάτμηση του γράφου. Στην μελέτη [3] χρησιμοποιούνται κάποιοι δείκτες ώστε να αξιολογηθούν οι αλγόριθμοι συσταδοποίησης. Οι αλγόριθμοι αυτοί μετρούν την πυκνότητα και τον διαχωρισμό (compactness και separability) των συστάδων. Παρουσιάζονται και αναλύονται έννοιες όπως Degeneracy (εκφυλισμός) και graph density (πυκνότητα γραφήματος) για να αξιολογηθεί η σύνδεση των κόμβων μέσα και μεταξύ των clusters. Καθώς προτείνεται και αναλύεται ένας νέος δείκτης αξιολόγησης QGraph ο οποίος θα παρατεθεί αναλυτικά παρακάτω. Στην μελέτη [4] συγκρίνονται μερικές από τις πιο δημοφιλείς μετρικές ποιότητας για συσταδοποίηση γράφων. Αυτές οι μετρικές αξιολογούνται με βάση τα αποτελέσματα των συστάδων και των χαρακτηριστικών που θα πρέπει να έχει μια συστάδα. Επίσης παρατηρείται ότι σε πραγματικά δεδομένα ο τύπος του δικτύου (κοινωνικό, τεχνολογικό κ.ο.κ.) επιδρά στην δομή των κοινοτήτων και τα διαφοροποιεί από αυτό που τυπικά αναμενόταν να προκύψει σύμφωνα με τις μελέτες που έγιναν στον εκάστοτε αλγόριθμο συσταδοποίησης και τις μετρικές που χρησιμοποιήθηκαν. Στην έρευνα [5] μελετώνται 230 σύνολα μεγάλων κοινωνικών δικτύων συνεργασίας και πληροφόρησης όπου οι κόμβοι δηλώνουν την συμμετοχή σε μία ομάδα. Για παράδειγμα οι κόμβοι συνδέονται με ακμές σε διάφορες κοινωνικές ομάδες που βασίζονται στα ενδιαφέροντα. Χρησιμοποιούνται αυτές τις ομάδες για να αποκτηθεί μια σταθερή και αξιόπιστη εικόνα για τις κοινότητες κάτι που ονομάζεται ground truth. Προτείνεται μία μεθοδολογία ώστε να αξιολογηθούν διαφορετικοί ορισμοί των κοινοτήτων και να συγκριθούν με την γνώση του ground truth. Επιλέγονται 13 δημοφιλείς ορισμοί δόμησης ενός δικτύου και εξετάζεται η επιρροή και απόδοση τους στον εντοπισμό του ground truth. Επεκτείνεται ο αλγόριθμος Spectral clustering σε μια ευρετική μέθοδο ανίχνευσης κοινότητας χωρίς παραμέτρους που προσαρμόζεται εύκολα σε δίκτυα με περισσότερους από εκατό εκατομμύρια κόμβους. Η προτεινόμενη μέθοδος επιτυγχάνει 30% σχετική βελτίωση σε σχέση με τις τρέχουσες μεθόδους τοπικής ομαδοποίησης.

ΠΑΡΑΡΤΗΜΑ κώδικα Python :

Αλγόριθμος Spectral

```
K = 42 #Input Value Number of Clusters
A = nx.to_scipy_sparse_array(G)
C = SpectralClustering(n_clusters=K, affinity='precomputed',
random_state=2022).fit_predict(A)
```

Αλγόριθμος Louvain

```
resolution = 3.1 #Input Value of Parameter
#If Resolution is <1, the algorithm favors larger communities, if  >1 favors
smaller communities
C = louvain_communities(G, weight=None, resolution=resolution, seed=2022)
```

Μέθοδοι και δείκτες αξιολόγησης :

Υπολογισμός απόστασης μεταξύ κόμβων	`D = nx.algorithms.shortest_path_length(G))`
Υπολογισμός δείκτη Modularity	`m = modularity(G, C)`

Υπολογισμός δείκτη Conductance	`cond = []`
	`for x in C:`
	`  tmp = conductance(G, x)`
	`  cond.append(tmp)`
	
	`conductunce = np.mean(cond)`

Υπολογισμός μετρικής Silhouette Index, GS*	`def GS_star(D, C):`
	`  s = silhouette_samples(D, C, metric='precomputed')`
	`  cs = []`
	`  N = []`
	`  K = int(np.max(C))`
	`  for k in range(K + 1):`
	`  x = np.where(C == k)[0]`
	`  n = len(x)`
	`  N.append(n)`
	`  cs.append(np.mean(s[x]))`
	
	`  gs_star = 0`
	`  for i in range(len(N)):`
	`  gs_star += N[i] * cs[i]`
	`  gs_star = gs_star / sum(N)`

	`    return gs_star`

Υπολογισμός μετρικής Q-Graph	`def qgraph(G, C):`
	`    intra_linkage = []`
	`    dens = []`
	`    for c in C:`
	`        gc = G.subgraph(c)`
	`        kc = nx.algorithms.k_core(gc)`
	
	`        degrees = list(kc.degree)`
	`        degrees = [d[1] for d in degrees]`
	`        deg = max(degrees)`
	
	`        n = gc.number_of_nodes()`
	`        deg_cov = kc.number_of_nodes() / n`
	`        intra_linkage.append(deg * deg_cov)`
	`        if n == 1:`
	`            dens.append(0)`
	`        else:`
	`            dens.append(gc.number_of_edges() *2 / (n * (n - 1)))`
	
	`    inter_linkage = []`
	`    inter_conn_1 = []  # 1 / inter_conn`
	`    for i in range(len(C) - 1):`
	`        for j in range(i + 1, len(C)):`
	`            ci = C[i]`
	`            cj = C[j]`
	`            cij = ci.union(cj)`
	
	`            gc = G.subgraph(cij)`
	`            kc = nx.algorithms.k_core(gc)`
	
	`            degrees = list(kc.degree)`
	`            degrees = [d[1] for d in degrees]`
	`            deg = min(degrees)`
	
	`            n = gc.number_of_nodes()`
	`            deg_cov = kc.number_of_nodes() / n`
	`            inter_linkage.append(deg * deg_cov)`
	`            inter_dens = gc.number_of_edges() / (len(ci) * len(cj))`
	`            if (dens[i] == 0 or dens[j] == 0):`
	`                inter_conn_1.append(0)`
	`            else:`
	`                inter_conn_1.append(1 / (inter_dens / min(dens[i], dens[j])))`
	
	`    intra_link = np.mean(intra_linkage)`
	`    inter_link = np.mean(inter_linkage)`
	`    sep = np.mean(inter_conn_1)`
	
	`    qg = intra_link - inter_link + sep`

	`    return qg`

Υπολογισμός μετρικής CDS	`def cds(G, C, wc, wd, ws):   #wc=1, wd=1, ws=1`
	`    nconn = []`
	`    dens = []`
	`    for c in C:`
	`        gc = G.subgraph(c)`
	`        nconn.append(nx.algorithms.node_connectivity(gc))`
	`        n = gc.number_of_nodes()`
	`        if n == 1:`
	`            dens.append(0)`
	`        else:`
	`            dens.append(gc.number_of_edges() *2 / (n * (n - 1)))`
	
	`    inter_conn_1 = []  # 1 / inter_conn`
	`    for i in range(len(C) - 1):`
	`        for j in range(i + 1, len(C)):`
	`            ci = C[i]`
	`            cj = C[j]`
	`            cij = ci.union(cj)`
	`            gc = G.subgraph(cij)`
	`            n = gc.number_of_nodes()`
	`            inter_dens = gc.number_of_edges() / (len(ci) * len(cj))`
	`            if (dens[i] == 0 or dens[j] == 0):`
	`                inter_conn_1.append(0)`
	`            else:`
	`                inter_conn_1.append(1 / (inter_dens / min(dens[i], dens[j])))`
	
	`    intra_conn = wc * np.mean(nconn) + wd * np.mean(dens)`
	`    sep = np.mean(inter_conn_1)`
	
	`    cds = intra_conn + ws * sep`
	
	`    return cds`

Υπολογισμός δείκτη Rand Index	`def check_clusterings(labels_true, labels_pred):`
	`    """Check that the two clusterings matching 1D integer arrays."""`
	`    labels_true = np.asarray(labels_true)`
	`    labels_pred = np.asarray(labels_pred)`
	`    # input checks`
	`    if labels_true.ndim != 1:`
	`        raise ValueError(`
	`            "labels_true must be 1D: shape is %r" % (labels_true.shape,))`
	`    if labels_pred.ndim != 1:`
	`        raise ValueError(`
	`            "labels_pred must be 1D: shape is %r" % (labels_pred.shape,))`
	`    if labels_true.shape != labels_pred.shape:`

```
        raise ValueError(
            "labels_true and labels_pred must have same
size, got %d and %d"
            % (labels_true.shape[0], labels_pred.shape[0]))
    return labels_true, labels_pred

def rand_score (labels_true, labels_pred):
#"""given the true and predicted labels, it will return the
Rand Index."""
    check_clusterings(labels_true, labels_pred)
    my_pair = list(combinations(range(len(labels_true)),
2)) #create list of all combinations with the length of
labels.
    def is_equal(x):
        return (x[0]==x[1])
    my_a = 0
    my_b = 0
    for i in range(len(my_pair)):

if(is_equal((labels_true[my_pair[i][0]],labels_true[my_pair
[i][1]])) ==
is_equal((labels_pred[my_pair[i][0]],labels_pred[my_pair[i]
[1]]))
                and
is_equal((labels_pred[my_pair[i][0]],labels_pred[my_pair[i]
[1]])) == True):
                 my_a += 1

if(is_equal((labels_true[my_pair[i][0]],labels_true[my_pair
[i][1]])) ==
is_equal((labels_pred[my_pair[i][0]],labels_pred[my_pair[i]
[1]]))
                and
is_equal((labels_pred[my_pair[i][0]],labels_pred[my_pair[i]
[1]])) == False):
                 my_b += 1
    my_denom = comb(len(labels_true),2)
    ri = (my_a + my_b) / my_denom
    return ri
```

Σημείωση :	
Μεταβλητή “G”	Το σύνολο δεδομένων του εκάστοτε γράφου
Μεταβλητή “C”	Λίστα των κόμβων σε συστάδες, αποτέλεσμα των αλγορίθμων συσταδοποίησης.

Ο παραπάνω κώδικας έχει τμήματα από διαθέσιμο κώδικα στο github, από βιβλιοθήκες της python και απο αντιστοίχου περιεχομένου ιστοσελίδες.

Βιβλιογραφία

[1] Maria Halkidi, “Graph clustering evaluation in terms of cohesion and density”, University of Piraeus.

[2] Francois Boutin, Mountaz Hascoet, “Cluster Validity Indices for Graph Partitioning”, University Montpellier II, 2004.

[3] Maria Halkidi and Iordanis Koutsopoulos, “QGraph: A quality assessment index for graph clustering”, University of Piraeus.

[4] Helio Almeida, Dorgival Guedes, Wagner Meira Jr., Mohammed J. Zaki, “Is there a best quality metric for graph clusters? ”, Universidade Federal de Minas Gerais, MG, Brazil and Rensselaer Polytechnic Institute, NY, USA.

[5] Jaewon Yang, Jure Leskovec, “Defining and Evaluating Network Communities based on Ground-truth”.

[6] Scott Emmons, Stephen Kobourov, Mike Gallant, Katy Börner, “Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale”, Plos, July 8 2016.

[7] Lizhen Shi, Bo Chen, “Comparison and Benchmark of Graph Clustering Algorithms”, Department of Computer Science, Florida State University, Tallahassee, FL, USA.

[8] Ulrik Brandes, Dorothea Wagner, “Analysis and Visualization of Social Networks”, University of Passau, Department of Mathematics & Computer Science, University of Konstanz, Department of Computer & Information Science, 2004.

[9] Ulrik Brandes, Marco Gaertler, Dorothea Wagner “Experiments on Graph Clustering Algorithms”, University of Passau, Department of Mathematics & Computer Science, University of Karlsruhe, Faculty of Informatics, 2003.

[10] Scott Beamer III, “Understanding and Improving Graph Algorithm Performance”, University of California, Berkeley, 2016.

[11] Article “Community detection using NetworkX”, Orbifold Consulting, October 7 2019. https://orbifold.net/default/community-detection-using-networkx/

[12] Julian McAuley, Jure Leskovec, “Learning to Discover Social Circles in Ego Networks”. Stanford USA, 2012.

[13] http://snap.stanford.edu/index.html

[14] Μαρία Χαλκίδη, Διάλεξη Νº8, “Social network analysis”, Εξόρυξη και Προετοιμασία δεδομένων, Πανεπιστήμιο Πειραιώς, 2021.

[15] Bertrand Charpentier, “Multi-scale clustering in graphs using modularity”, KTH Royal Institute of Technology, 2019.

[16] N. Bolshakova, F. Azuaje “Cluster validation techniques for genome expression data”, Trinity College Dublin, 2001.

[17] Santo Fortunato, “Community detection in graphs” , Complex Networks and Systems Lagrange Laboratory, ISI Foundation.

[18] Article “What is a graph (data structure)?”, Educative.
https://www.educative.io/edpresso/what-is-a-graph-data-structure

[19] Maria Xalkidi, Yannis Batistakis, Michalis Vazirgiannis, “On Clustering Validation Techniques”, Journal of Intelligent Information Systems, 2001.

[20] Αθανάσιος Σκουρτανιώτης, “Μελέτη αλγορίθμων ομαδοποίησης σε περιβάλλον προγραμματισμού Python”, Πανεπιστήμιο Πειραιώς, 2016.

[21] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, “Fast unfolding of communities in large networks”, Fast unfolding of communities in large networks, LIP6, Universit´e Pierre et Marie Curie, Institute for Mathematical Sciences, 2008.

[22] Vaidehi Joshi, “A Gentle Intriduction to Graph Theory”, Medium, 2017.
(https://medium.com/basecs/a-gentle-introduction-to-graph-theory-77969829ead8)

[23] “Introduction and Graph Structure”, Site created with Jekyll using the Tufte theme. Inspired by Stanford CS 228 Notes,2020.
https://snap-stanford.github.io/cs224w-notes/preliminaries/introduction-graph-structure

[24] Shayan Oveis Gharan, “CSE 521: Design and Analysis of Algorithms I”,2017.
https://courses.cs.washington.edu/courses/cse521/17wi/521-lecture-11.pdf

[25] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks," J. Stat. Mech. , 2008.

[26] Jianbo Shi and Jitendra Malik. “Normalized cuts and image segmentation” IEEE Trans. Pattern Anal. Mach. Intell, August 2000.

[27] Stijn Van Dongen, “Graph clustering via a discrete uncoupling process” SIAM Journal on Matrix Analysis and Applications, 2008.

[28] S. M. van Dongen, “Graph Clustering by Flow Simulation” PhD thesis, University of Utrecht, The Netherlands, 2000.

[29] James Woma (jaywoma), Kim Ngo (kimanh), “Comparisons of Community Detection Algorithms in the YouTube Network”, Stanford University, 2015.

[30] Yike Liu, Neil Shah, Danai Koutra, “An Empirical Comparison of the Summarization Power of Graph Clustering Methods”, University of Michigan, Carnegie Mellon University, 2015.

[31] Leto Peel, Daniel B. Larremore, Aaron Clauset, “The ground truth about metadata and community detection in networks” , Science Advances, 2107.

[32] Abdol-Hossein Esfahanian, “Connectivity Algorithms”.

[33] Giulio Rossetti, Letizia Milli, Remy Cazabet, “CDlib: a Python Library to Extract, Compare and Evaluate Communities from Complex Networks”, KDD Lab. ISTI-CNR, University of Pisa, Universite de Lyon, 2020.