Του Φοιτητή:
Τιγγινάγκα Αλέξανδρου
ΑΜ: ΜΕ1750
Επιβλέπουσα:
Μαρία Χαλκίδη

Διπλωματική Εργασία: Μελέτη και ανάπτυξη προσεγγίσεων αξιολόγησης αποτελεσμάτων συσταδοποίησης σε γράφους

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ: «ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ»

Κατεύθυνση: ΠΡΟΗΓΜΕΝΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Περιεχόμενα

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως αντικείμενο μία μελέτη στο πρόβλημα της αξιολόγησης των αποτελεσμάτων τμηματοποίησης ενός γράφου. Το πρόβλημα της αξιολόγησης των αποτελεσμάτων συσταδοποίησης έχει μελετηθεί ευρέως και υπάρχει ένας μεγάλος αριθμός από δείκτες αξιολόγησης των αποτελεσμάτων συσταδοποίησης. Τα τελευταία χρόνια έχει γίνει αρκετή δουλειά σε τεχνικές συσταδοποίησης (clustering) γράφων μέσω web applications, social networks, βιοϊατρικά δεδομένα.

Ο στόχος της παρούσας εργασίας είναι η μελέτη επεκτάσεων προσσεγγίσεων αξιολόγησης αποτελεσμάτων συσταδοποίησης σε γράφων (graph clustering techniques) ώστε να δούμε αυτές οι προσεγγίσεις σε τι φάση βρίσκονται, ποιες μετρικές λαμβάνουν υπόψη και ποιες από αυτές μπορούν να δώσουν ασφαλή συμπεράσματα.

Η παρούσα εργασία είναι δομημένη σε τέσσερα μέλη, στο πρώτο θα περιγράψουμε τεχνικές που αφορούν τη συσταδοποίηση δεδομένων σε μορφή γράφων τις οποίες θα τις σχολιάσουμε. Στο δεύτερο μέρος θα περιγράψουμε το πρόβλημα της αξιολόγησης αποτελεσμάτων συσταδοποίησης σε γράφους, θα μελετήσουμε προσεγγίσεις που αξιολογούν την ποιότητα αυτών των αποτελεσμάτων και θα τις σχολιάσουμε.

Στο τρίτος μέρος θα γίνει η παρουσίαση μιας πειραματικής προσέγγισης των όσων περιγράφονται στα δύο πρώτα μέρη. Εκεί θα επιλέξουμε τέσσερα datasets διαφορετικού μεγέθους το καθένα, τα οποία θα έχουν μορφή γράφων. Θα τα υποβάλλουμε σε δύο διαφορετικές τεχνικές συσταδοποίησης δεδομένων σε γράφους και θα αξιολογήσουμε τα αποτελέσματα βάση τριών δεικτών. Τέλος στο τέταρτο και τελευταίο μέρος θα σχολιάσουμε τα αποτελέσματα που πήραμε από τη πειραματική μελέτη μας και θα προβούμε σε κάποια συμπεράσματα.

Abstract

In this work we study the problem of cluster validity for data with graph structure. The problem of cluster validity has been studied widely and there are enough indices to evaluating different clustering results. There is a number of indices that measure the compactness and the separability of clustering results. However enough projects that have also studied these indices, cannot make clear conclusions about a set of different clustering results.

So with this work we study the problem of graph clustering and the problem of cluster validity too. On the second part we focus on the clustering validation techniques and presented various methods with different validation concepts.

At the rest of work (third and fourth part) we present an experimental approach where we choose different datasets with graph structure that clustered within different algorithms and for different cost functions. Alongside we evaluate these clustering results with three different clustering validation indices. At the end we compare all the indices with each other and choose the one that worked correctly based on his definition.

Μέρος 1. Συσταδοποίηση Σε Γράφους (Community Detection - Graph Clustering)

1.1 Εισαγωγή

Οι απαιτήσεις για ανάλυση μεγάλου όγκου δεδομένων στην εποχή μας έχουν αυξηθεί. Οι ανάγκες των επιστημόνων που ασχολούνται αποκλειστικά με την ανάλυση δεδομένων έχουν ωθήσει την επιστήμη της πληροφορικής στο να δώσει λύσεις σε αυτές. Σε όλες τις επιστήμες πλέον όπως τη Βιολογία, το Marketing, τη Γεωγραφία, τη Πληροφορική κοκ κατά την εκπόνηση ερευνητικών εργασιών για την εξαγωγή συμπερασμάτων χρειάζεται η μελέτη μεγάλου όγκου δεδομένων. Τα δεδομένα αυτά πλέον αυτά χάρη στην εξέλιξη της τεχνολογίας αλλά και των επιστημών, πλέον μπορούν να ανιχνευθούν κατά ομάδες και όχι μεμονωμένα. Για παράδειγμα μία ομάδα μάρκετινγκ μπορεί να θέλει να μελετήσει τις τάσεις των online καταναλωτών στις χώρες που περιβάλλουν τη Μεσόγειο Θάλασσα κατά τους καλοκαιρινούς μήνες. Αντιλαμβανόμαστε όμως ότι η μελέτη ενός τόσο μεγάλου κοινού θα μπορέσει να γίνει πιο εύκολη και να οδηγήσει σε πιο ασφαλή συμπεράσματα αν γίνει κατά τμήματα.

Έτσι λοιπόν έρχεται η επιστήμη της Πληροφορικής η οποία έχει αναπτύξει εργαλεία τα οποία είναι σε θέση να επεξεργαστούν μεγάλο όγκο δεδομένων, να το σπάσουν σε ομάδες όταν χρειαστεί, που αυτές θα αναλυθούν ξεχωριστά. Η ομαδοποίηση/τμηματοποίηση αυτή, όμως, γίνεται βάση κάποιων ορισμών κι όχι αυθαίρετα. Οι υπο-ομάδες που δημιουργούνται από μεγάλα σύνολα δεδομένων θα πρέπει να αποτελούνται από στοιχεία τα οποία θα παρουσιάζουν μεταξύ τους κοινά χαρακτηριστικά αλλά σίγουρα και διαφορές με τα στοιχεία άλλων ομάδων.

Σήμερα μηχανικοί λογισμικού αλλά και επιστήμονες από τον χώρο της Πληροφορικής και των Μαθηματικών έχουν αναπτύξει μοντέλα, τεχνικές και θεωρίες οι οποίες μπορούν να βοηθήσουν στην ανάλυση τέτοιου είδους δεδομένων. Στην πληροφορική λοιπόν έχουν αναπτυχθεί προσεγγίσεις οι οποίες δίνουν τις λύσεις που χρειαζόμαστε στις απαιτήσεις που έχουν προκύψει. Χρησιμοποιώντας λοιπόν την επιστήμη

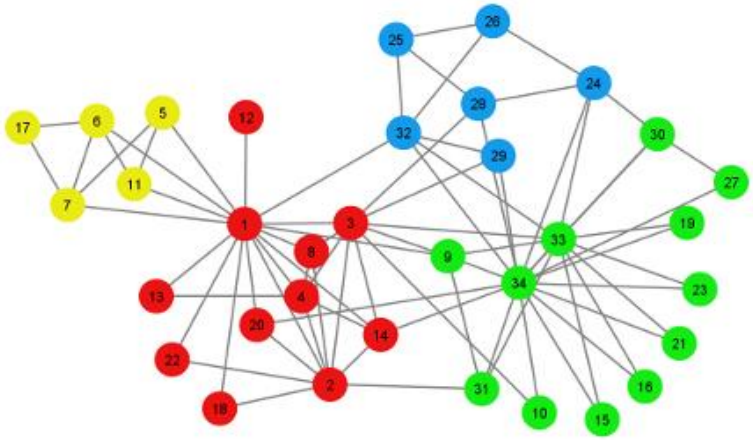
των δικτύων (network science) η οποία βρίσκει εφαρμογή σε ένα εύρος επιστημών έχουν αναπτύξει εργαλεία για την ανάλυση τέτοιου είδους δεδομένων. Η τεχνική αυτής της ανάλυσης είναι γνωστή ως Community Detection. Τα τελευταία χρόνια, πολλοί αλγόριθμοι Community Detection έχουν προταθεί για να ερμηνεύσουν τις δομικές ιδιότητες και τις δυναμικές συμπεριφορές των δικτύων (networks). Το community detection αποτελεί σημαντικό εργαλείο για την ανάλυση των σύνθετων δικτύων, επιτρέποντας τη μελέτη δομών που συχνά συνδέονται με οργανωτικά και λειτουργικά χαρακτηριστικά των υποκείμενων δικτύων.

Το community detection αποτελεί ένα από τα πιο δημοφιλή θέματα της σύγχρονης επιστήμης των δικτύων. Την παρούσα εποχή η μέθοδος αυτή, δηλαδή, ο προσδιορισμός κοινοτήτων μέσα σε μεγάλα δίκτυα είναι ένα ακαθόριστο πρόβλημα. Καθώς δεν υπάρχουν καθολικά πρότυπα για τον καθορισμό των communities με ασφάλεια. Αυτό έχει δημιουργήσει τη ανάγκη δημιουργίας κάποιων δεικτών μέσω των οποίων μπορούμε να βγάλουμε κάποια συμπεράσματα σχετικά με το κατά πόσο καλή είναι η τμηματοποίηση networks σε communities. Η εργασία αυτή παρουσιάζει μια κριτική ανάλυση του προβλήματος του clustering σε γράφους και προσεγγίζει κάποιες από τις επικρατέστερες μεθόδους αξιολόγησης αυτού.

1.2 Ανίχνευση Κοινοτήτων Σε Δίκτυα (Community Detection in Networks)

Στη βιβλιογραφία μπορεί αυτόν το όρο να τον συναντήσουμε και με την έννοια του Γράφου (Graph) ή Σύμπλεγμα Δικτύου (Network Clustering).

Τα περισσότερα networks τα οποία παρουσιάζουν ενδιαφέρον ως προς ανάλυση παρουσιάζουν δομή, η οποία είναι γνωστή στη βιβλιογραφία ως community structure. Αυτή η δομή συνδέει αντικείμενα (nodes) μεταξύ τους μέσω ακμών (edges)

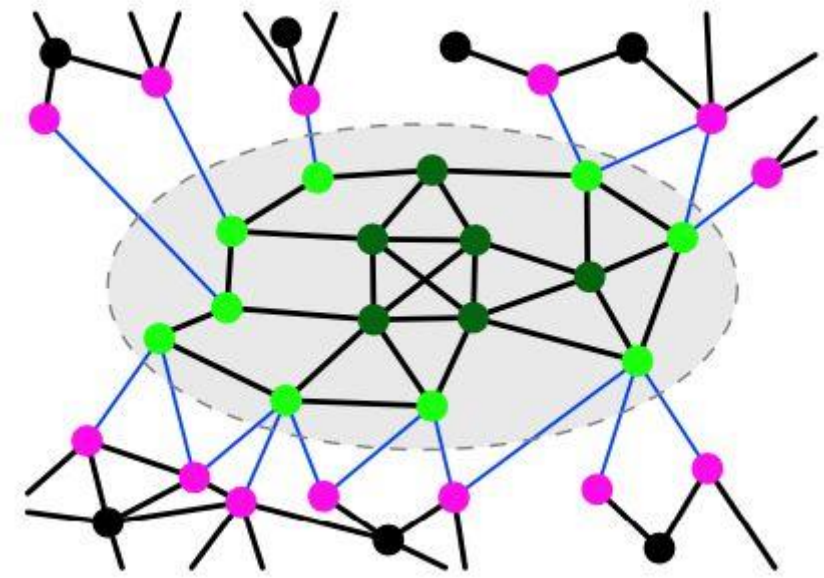


Εικόνα 1.

Στη παραπάνω εικόνα βλέπουμε ένα **Network** με **Community structure**. Τα στοιχεία (nodes) συνδέονται μεταξύ τους μέσω ακμών (edges). Οι κοινότητες που αυτά τα στοιχεία δημιουργούν (communities) φαίνονται με διαφορετικά χρώματα. Για παράδειγμα μία κοινότητα του παραπάνω δικτύου (network) σχηματίζουν τα στοιχεία (nodes) με αριθμό 17, 7, 11, 6 και 5 καθώς έχουν όλα τον ίδιο χρωματισμό.

1.3 Συσταδοποίηση Σε Γράφους: Επισκόπηση

Ξεκινώντας ας δούμε πώς είναι ένα community μέσα σε έναν γράφο, πως αυτό περιγράφεται και ποιες οι μεταβλητές του.



Εικόνα 2

Στην εικόνα 2 βλέπουμε την κοινότητα C που ανήκει στον γράφο G. Αυτή η κοινότητα C ορίζεται από τους κόμβους (nodes/vertices) [1] Πρόκειται για μία κλασσική αναπαράσταση ενός γράφου με κόμβους και ακμές.

Κόμβοι (Nodes): Είναι τα ποιο σημαντικά στοιχεία του γράφου. Αυτοί είναι οντότητες οι οποίες μέσα σε έναν γράφο παρουσιάζουν σχέσεις οι οποίες εκφράζονται μέσω των άκρων (edges). Αν ένας γράφος αποτελείται από 2 κόμβους και μία κατευθυνόμενη άκρη μεταξύ τους, τότε αυτή εκφράζει μία αμφίδρομη σχέση μεταξύ αυτών. [1] [2]

Ακμές (Edges): Οι ακμές είναι τα στοιχεία που χρησιμοποιούνται για την αντιπροσώπευση των σχέσεων μεταξύ των κόμβων μέσα στους γράφους. Μεταξύ δύο κόμβων μία ακμή εκφράζει μια αμφίδρομη σχέση. [1] [2]

Στον εικονιζόμενο γράφο G ο αριθμός των κόμβων και των ακμών θεωρούμε πως είναι n, m και για τον γράφο C θεωρούμε αντίστοιχα αριθμό κόμβων και ακμών n_c, m_c. Κάθε γράφος της μορφής G μπορεί να αναπαρασταθεί και μαθηματικά μέσω ενός πίνακα (Adjacency Matrix). Ένας τέτοιος πίνακας απεικονίζει ποια σημεία/κόμβοι παρουσιάζουν γειτνίαση μεταξύ τους – έχουν σχέση μεταξύ και ενώνονται μέσω μία

ακμής. [3] Στη περίπτωσή μας μπορούμε να ορίσουμε έναν τέτοιο πίνακα A για τον γράφο G. Τα στοιχεία αυτού του πίνακα θεωρούμε πως είναι A_{ij} και κάθε φορά που η τιμή του πίνακα για τα σημεία i,j παίρνει την τιμή 1 τότε αυτά είναι γειτονικά, όταν όμως η τιμή είναι 0 τότε αυτά δεν έχουν κάποια σχέση μεταξύ τους.

Ας θεωρήσουμε τώρα ότι τα σημεία του γράφου C παρουσιάζουν συνδέσεις μεταξύ τους καθώς αυτός αναπαριστά μια κοινότητα – community. Οι κόμβοι αυτού του C περικλείονται από μία διακεκομμένη γραφή (εικόνα 3) ώστε να ξεχωρίζει το community από τα σημεία του υπόλοιπου γράφου. Έτσι τα σημεία που αναπαρίστανται με mangenta απόχρωση είναι εκτός του εν λόγω community παρά τη σύνδεση που φαίνεται να έχουν με τους κόμβους που υπάρχουν μέσα σε αυτό. Τέλος, τα σημεία με μαύρο χρώμα είναι μέσα στον γράφο, αποτελούν μέρος του Network αλλά δεν έχουν καμία σύνδεση με τον C. Κάθε μπλε γραμμή αφορά τη σύνδεση (edges) μεταξύ των nodes του C με τα υπόλοιπα σημεία του γράφου/network.

Ο αριθμός των των ακμών των nodes i του γράφου C που είναι συνδεδεμένα μεταξύ τους είναι ένα σύνολο που μπορεί να αναπαρασταθεί ως εξής k_i^{int} . Και με k_i^{ext} μπορούμε να ορίσουμε το σύνολο αυτών των ακμών που συνδέσουν τα στοιχεία του community με κάποια από τον γράφο G. Και οι δύο αυτοί δείκτες μπορούν να εκφραστούν από τον Adjacency Matrix / Πίνακας Γειτνίασης των nodes ως εξής:

a. $$k_i^{int} = \Sigma_{jEC} A_{ij}$$
b. $$k_i^{ext} = \Sigma_{jEC} A_{ij}$$

Έτσι συμπερασματικά κατανοούμε ότι αν ο δείκτης a. είναι μεγαλύτερος από το μηδέν τα nodes i εμφανίζουν γειτνίαση μόνο με τα σημεία που βρίσκονται εντός του γράφου C. Αν όμως συμβεί το εξής, ο δείκτης b. να είναι μεγαλύτερος του μηδενός ενώ το ίδιο ισχύει και για τον a. τότε τα σημεία i εμφανίζουν γειτνίαση μόνο εκτός του γράφου C, δηλαδή θα υπάρχει σύνδεση μόνον με αυτά που συνδέονται μέσω των πράσινων edges (εικόνα 3).

Οι δύο αυτοί δείκτες των κορυφών μπορούν να μας αποκαλύψουν πόσο ισχυροί είναι οι δεσμοί μεταξύ των κορυφών μέσα στο Community. Η σχέση αυτή περιγράφεται ως εξής:

$$\Xi_i = k_i^{int} / k_i$$

Όσο μεγαλύτερος είναι ο παραπάνω δείκτης τόσο πιο ισχυροί είναι οι δεσμοί των κορυφών μέσα στο Community που μελετάμε. [4] [6]

Εκτός από τον παραπάνω δείκτη υπάρχει και ένας ακόμη ο οποίος όπως είναι εύλογο δείχνει τον βαθμό που οι εξωτερικές συνδέσεις του γράφου C, $\mathbf{k_i^{int}}$ έχουν ισχυρούς δεσμούς σε σχέση με το σύνολο των δεσμών του γράφου G. Αυτός ο δείκτης υπολογίζεται ως εξής:

$$M_i = k_i^{ext} / k_i$$

Με λογική συνέχεια να ισχύει ο ορισμός: $M_i = 1 - \Xi_i$

Στην ανάλυση των communities σε σχέση με τα networks τα οποία ανήκουν χρήσιμη είναι και η γνώση των παρακάτω μεταβλητών.

Σε πρώτη φάση παρουσιάζουμε το πως θα περιγράψουμε μετρικά το πως ένα community εμφανίζει συνεκτικότητα. Δηλαδή αν τα στοιχεία που το ορίζουν έχουν όντως σχέση μεταξύ τους.

- Internal Degree k_c^{int}: Πρόκειται για το σύνολο των εσωτερικών στοιχείων / nodes που βρίσκονται μέσα στον C. Επίσης μπορεί να οριστεί και ως εξής:

$$k_c^{int} = \Sigma_{i,j \in C} A_{i,j}$$

- Average internal degree $k_c^{avg\text{-}int}$: Ο δείκτης αυτός δείχνει τον μέσο όρο των nodes του community C, λαμβάνοντας υπόψη μόνο τις εσωτερικές συνδέσεις (edges)

$$k_c^{avg\text{-}int} = k_c^{int} / n_c$$

- Internal Edge Density δ_c^{int}: Μέσω αυτής της μετρικής ορίζουμε την αναλογία μεταξύ του αριθμού των εσωτερικών άκρων (edges) του C και του αριθμού όλων των πιθανών εσωτερικών άκρων (edges):

$$\Delta_c^{int} = k_c^{int} / n_c (n_c - 1)$$

Σε δεύτερη φάση παρουσιάζουμε μετρικές οι οποίες βασίζονται στην εξωτερική σύνδεση του community και πως αυτό ενσωματώνεται μέσω του γράφου σε ολόκληρο το network. Επίσης μέσω των παρακάτω μετρικών εμφανίζεται και το πώς έχει τμηματοποιηθεί το community από το υπόλοιπο network.

- External degree k_c^{ext}. Το συνολο των nodes που βρίσκονται εξωτερικά του C. Αυτό μας δίνει το νούμερο των εξωτερικών ακμών (edges) που συνδέουν το C με τα σημεία του network (Βλέπε τις μπλε γραμμές στην εικόνα 3)

$$k_c^{ext} = \Sigma_{i \in C, j \notin C} A_{i,j}$$

- Average external degree, $k_C^{avg\text{-}ext}$. Ο δείκτης που ορίζει τα nodes του C ως προς τις εξωτερικές συνδέσεις (Βλέπε μπλε γραμμές εικόνα 3)

$$k_c^{avg-ext} = k_c^{ext} / n_c$$

- External Egde Density, δ_c^{ext}. Η αναλογία μεταξύ του αριθμού των εξωτερικών edges του C με τον αριθμό όλων των πιθανών εξωτερικών edges:

$$\delta_c^{ext} = k_c^{ext} / n_c (n - n_c)$$

Τέλος, έχουμε υβριδικά μέτρα, που συνδυάζουν την εσωτερική και εξωτερική συνεκτικότητα. Αξιοσημείωτα παραδείγματα είναι τα παρακάτω:

Name	Symbol	Definition	Name	Symbol	Definition
Internal degree	k_i^{int}	$\sum_{j\in C} A_{ij}$	Internal strength	w_i^{int}	$\sum_{j\in C} W_{ij}$
External degree	k_i^{ext}	$\sum_{j\notin C} A_{ij}$	External strength	w_i^{ext}	$\sum_{j\notin C} W_{ij}$
Degree	k_i	$\sum_j A_{ij}$	Strength	w_i	$\sum_j W_{ij}$
Embeddedness	ξ_i	$\frac{k_i^{int}}{k_i}$	Weighted embeddedness	ξ_i^w	$\frac{w_i^{int}}{w_i}$
Mixing parameter	μ_i	$\frac{k_i^{ext}}{k_i}$	Weighted mixing parameter	μ_i^w	$\frac{w_i^{ext}}{w_i}$

TABLE I Basic vertex community variables, for unweighted and weighted networks. A and W are the adjacency and the weight matrix, respectively.

	Unweighted networks			Weighted networks		
	Name	Symbol	Definition	Name	Symbol	Definition
Internal	Internal degree	k_C^{int}	$\sum_{i,j\in C} A_{ij}$	Internal strength	w_C^{int}	$\sum_{i,j\in C} W_{ij}$
	Average internal degree	$k_C^{avg\text{-}int}$	$\frac{k_C^{int}}{n_C}$	Average internal strength	$w_C^{avg\text{-}int}$	$\frac{w_C^{int}}{n_C}$
	Internal edge density	δ_C^{int}	$\frac{k_C^{int}}{n_C(n_C-1)}$	Internal weight density	$\delta_{w,C}^{int}$	$\frac{w_C^{int}}{\bar{w}n_C(n_C-1)}$
External	External degree	k_C^{ext}	$\sum_{i\in C, j\notin C} A_{ij}$	External strength	w_C^{ext}	$\sum_{i\in C, j\notin C} W_{ij}$
	Average external degree	$k_C^{avg\text{-}ext}$	$\frac{k_C^{ext}}{n_C}$	Average external strength	$w_C^{avg\text{-}ext}$	$\frac{w_C^{ext}}{n_C}$
	External edge density	δ_C^{ext}	$\frac{k_C^{ext}}{n_C(n-n_C)}$	External weight density	$\delta_{w,C}^{ext}$	$\frac{w_C^{ext}}{\bar{w}n_C(n-n_C)}$
Total	Total degree	k_C	$\sum_{i\in C, j} A_{ij}$	Total strength	w_C	$\sum_{i\in C, j} W_{ij}$
	Average degree	k_C^{avg}	$\frac{k_C}{n_C}$	Average strength	w_C^{avg}	$\frac{w_C}{n_C}$
	Conductance	C_C	$\frac{k_C^{ext}}{k_C}$	Weighted conductance	$C_{w,C}$	$\frac{w_C^{ext}}{w_C}$

TABLE II Basic community variables, for unweighted and weighted networks. A and W are the adjacency and the weight matrix, respectively, n_C the number of vertices of the community, n the total number of vertices of the graph, $\bar{w}$ the average weight of the network edges.

Εικόνα 4.

Επίσης πρέπει λόγος να γίνει και για τις εξής μετρικές οι οποίες είναι χρήσιμες στην ανάλυση των communities που εμφανίζονται μέσα στα networks.

- Total degree, k_c: Το σύνολο του αριθμού των nodes για τον γράφο C.

$$K_c = \Sigma_{i\ E\ C,j}\ A_{ij} \ \ \acute{\eta} \ \ k_c = k_c^{int} + k_c^{ext}$$

- Average degree k_c^{avg}. Η μέση τιμή των nodes στο γράφο C.

$$k_c^{\ avg} = k_c + n_c$$

- Conductance C_c. Η αναλογία μεταξύ του external degree με του total degree για τον γράφο C.

$$C_c = k_c^{ext} / k_c$$

Μέσω αυτής της ενότητα σας παρουσίασα κάποιες βασικές μετρικές οι οποίες είναι απαραίτητες για την περιγραφή και την ανάλυση γράφων στο πεδίο του community detection.

Ένα σύνθετο δίκτυο μπορεί να αντιστοιχεί σε ένα γράφο της μορφής G (V,E) όπου V είναι το σύνολο των nodes και E το σύνολο των ακμών (edges). Στη παραπάνω περιγραφή ανιχνεύσαμε μία τμηματοποίηση (community detection) του network G, την οποία θεωρήσαμε ως C (v,e) όπου v και e είναι τα σύνολα των nodes και edges που ανήκουν σε αυτόν τον υπο-γράφο.

Καθημερινά τα δεδομένα αυξάνονται εκθετικά. Έτσι είναι λογικό να αναπτύσσονται εκθετικά και τα δίκτυα (networks) ως προς τη ποικιλία, το μέγεθος και την πολυπλοκότητα. Αυτή η ανάπτυξη φέρει με τη σειρά της αλλαγές στα δίκτυα επικοινωνίας στην τεχνολογία του Internet of Things, στα μέσα κοινωνικής δικτύωσης, στη νεφοϋπολογιστική κ.α. Οι λειτουργίες των networks πληθαίνουν και αναπτύσσονται μη γραμμικά. Η εξέλιξη αυτή μας βοηθά να ανακαλύψουν νέα χαρακτηριστικά των δικτύων τα οποία θα μας οδηγήσουν σε καλύτερη ανάλυση αυτών.

Οι κοινότητες (Communities) μέσα στα δίκτυα (networks) είναι ένα σύνολο κόμβων (nodes) τα οποία παρουσιάζουν ισχυρούς δεσμούς σύνδεσης (highly connected) μεταξύ τους σε σχέση με τους υπόλοιπους κόμβους που απαρτίζουν το δίκτυο (Yang et al. 2010). Μέσω του Community Detection είμαστε σε θέση να οδηγηθούμε σε χρήσιμες πληροφορίες σχετικές με τα υπό ανάλυση δίκτυα. [6]

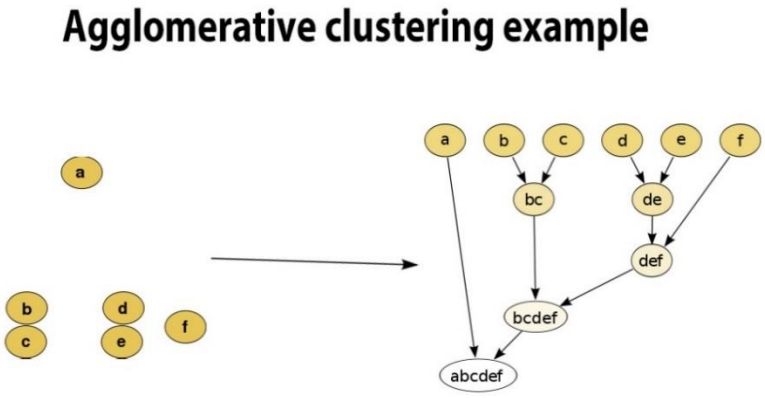
Αυτή τη στιγμή υπάρχουν διαθέσιμα αρκετά εργαλεία τα οποία μας βοηθούν να ορίσουμε communities μέσα σε μεγάλα networks. Τα περισσότερα από αυτά βασίζονται σε τεχνικές που ορίζουν κυρίως αλγόριθμοι.

1.4 Συσταδοποίηση Γράφων: Τεχνικές

Μέσω διάφορων τεχνικών Clustering μπορούμε να ομαδοποιήσουμε networks βάση των χαρακτηριστικών τους ώστε να τα μελετήσουμε καλύτερα προκειμένου να οδηγηθούμε σε συμπεράσματα από την ανάλυση αυτών. Υπάρχουν στη διάθεσή μας αλγόριθμοι οι οποίοι μας βοηθούν να κάνουμε αυτές τις τμηματοποιήσεις των data σε Clusters και

αυτοί κατηγοριοποιούνται ως εξής: Hierarchical, Partitioning, Density, Grid based algorithms and Graph Based Algorithm.

1. Hierarchical Clustering[5,6,8,9]: μία μέθοδος ανάλυσης σε clusters που επιδιώκει να οικοδομήσει μία ιεραρχία μεταξύ των τμηματοποιήσεων. Υπάρχουν δύο τύποι για αυτή τη τεχνική. (Fortunato 2010; Friedman et al. 2001)
   A. Agglomerative, Η μέθοδος αυτή ιεραρχικά συγχωνεύει στοιχεία των clusters με βάση την απόσταση που αυτά παρουσιάζουν μεταξύ τους. Συνήθως τα πιο κοντινά στοιχεία ομαδοποιούνται στο ίδιο cluster. (Εικόνα 7)



Εικόνα 5

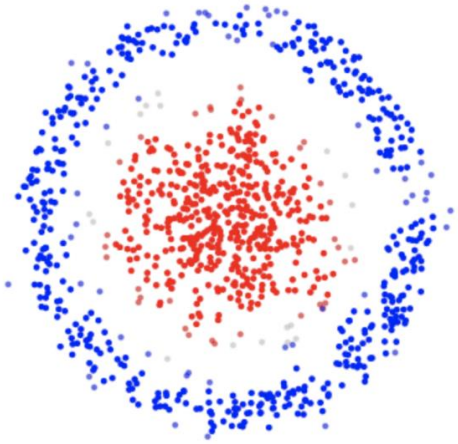
   B. Divisive, Η δεύτερη προσέγγιση της εν λόγω τεχνικής ξεκινά με το να ομαδοποιεί όλα τα στοιχεία σε ένα cluster. Στη συνέχεια ορίζει ένα βασικό αντικείμενο του cluster το βάση του οποίου μελετά τα υπόλοιπα και όσα από αυτά δεν έχουν κοινά χαρακτηριστικά με το βασικό ορίζονται εκτός του cluster. Μόλις ολοκληρωθεί η διαδικασία διαχωρισμού των στοιχείων του πρώτου έχουν έτοιμο ένα δεύτερο cluster για το οποίο ακολουθείτε η ίδια διαδικασία κοκ.

2. Partitioning[5,6,8,9]: Η μέθοδος αυτή ξεχωρίζει από την προηγούμενη. Εδώ υπάρχουν εργαλεία τα οποία μας βοηθούν να «σπάσουμε» τα δεδομένα μας σε partitions και στη συνέχεια μέσω διάφορων δεικτών αξιολογούμε κατά πόσο ο διαχωρισμός αυτός

ορίζει clusters στοιχείων με όμοια χαρακτηριστικά. (Jin and Han 2011; Fortunato 2010; Dhumal and Kamde 2015; Furht 2010; Slaninová et al. 2010)

Συγκεκριμένα παίρνουμε ένα dataset και το χωρίζουμε σε k clusters τα οποία είναι διακριτά μεταξύ τους. Στόχος αυτής της τεχνικής είναι να διαιρέσει τα σημεία των δεδομένων σε clusters ώστε να μεγιστοποιήσει ή να ελαχιστοποιήσει την ομοιότητα μεταξύ των χαρακτηριστικών που εμφανίζουν οι κόμβοι (nodes). Παραδείγματα τέτοιων τεχνικών αποτελούν το k-mean clustering (MacQueen 1967) και το fuzzy k-mean clustering (Bezdek 2013; Dunn 1973)

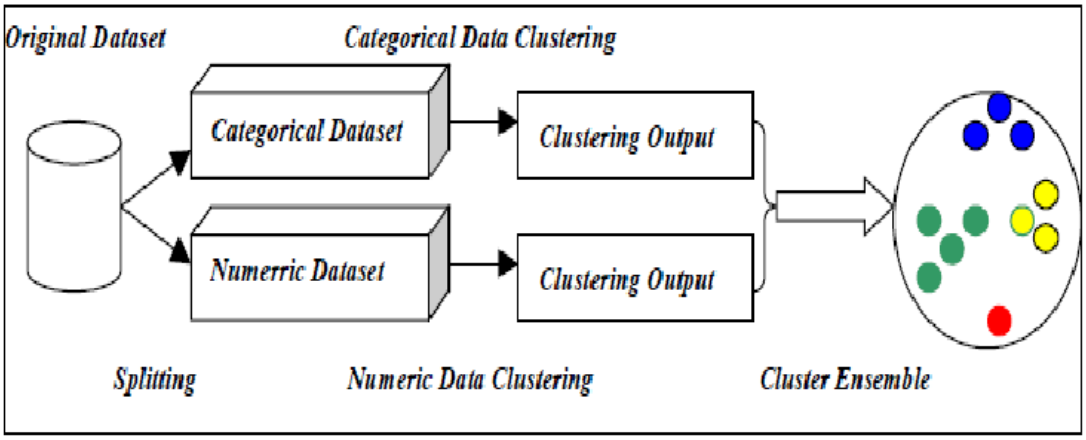
3. Spectral Clustering[5,6,8,9]: Αυτή η μέθοδος περιλαμβάνει τεχνικές οι οποίες χρησιμοποιούν μεθόδους που ορίζει ο αλγόριθμος Spectral και μέσω αυτής διαιρούμαι τα networks σε clusters με τους κόμβους που τα αποτελούν να παρουσιάζουν ομοιότητα μεταξύ τους (Fortunato 2010; Dhumal and Kamde 2015). Παραδείγματα αυτής της τεχνικής περιλαμβάνονται στη μέθοδο του Fiedler: Laplacian Spectral Partioning (1973) [9: III. MODULE IDENTIFICATION]
4. Divisive Algorithms[5,6,8,9]: Εδώ δημιουργούμε clusters μέσα από ένα network, αφαιρούμε τα edges μεταξύ αυτών που έχουν χαμηλή ομοιότητα μεταξύ τους (Murata 2010). Πρόκειται για μία πιο επιθετική τεχνική clustering η οποία διακόπτει τη σύνδεση μεταξύ των κόμβων που παρουσιάζουν μικρή ομοιότητα ως προς τα χαρακτηριστικά τους και ορίζει clusters των οποίων οι κόμβοι εξ αρχής έχουν ισχυρούς δεσμούς.
5. Modularity Optimisation Based Community Detection Techniques[5,6,8,9]: Εδώ εμφανίζονται τεχνικές οι οποίες ορίζουν clusters μέσω της συνάρτησης του Modularity. Όταν η συνάρτηση αυτή παίρνει τη μεγαλύτερη τιμή της (μέγιστη) τότε έχουμε ορίσει την καλύτερη τμηματοποίηση του network σε clusters.
6. Density Based[5,6,8,9]: Στη τεχνική αυτή ορίζονται clusters τα οποία περιέχουν στοιχεία τα οποία εμφανίζονται με ισχυρή πυκνότητα σε μία συγκεκριμένη περιοχή. (Εικόνα 8)



Εικόνα 6

7. Grid based[5,6,8,9]: Αυτή η τεχνική δημιουργεί πλέγματα τα οποία ομαδοποιούν τα στοιχεία των clusters και συνέχεια ακολουθούνται 5 βήματα τα οποία ορίζουν αν τα πλέγματα αυτά έχουν ορίσει σωστά τη τμηματοποίηση των στοιχείων. Μία τέτοια τεχνική είναι αυτή του αλγορίθμου Louvain. (Clauset et al. 2004) / Creating the grid structure, i.e., partitioning the data space into a finite number of cells / Calculating the cell density for each cell / Sorting of the cells according to their densities. / Identifying cluster centers. / Traversal of neighbor cells. Πηγή: *https://epubs.siam.org/doi/abs/10.1137/1.9780898718348.ch12?mobileUi=0&*

8. Dynamic Community Detection Algorithms[5,6,8,9]: Αυτή τη τεχνική θα τη συναντήσουμε αποκλειστικά σε δυναμικά networks. Τέτοια είναι το Twitter, Linkedin, Facebook κτλ. Όπως είναι κατανοητό αυτά παρουσιάζουν αλλαγές ως το χρόνο έτσι δημιουργούνται συνέχεια νέες σχέσεις μεταξύ των στοιχείων που ορίζουν τα networks οπότε και χρήζουν άλλης μεταχείρισης ως προς τη τεχνική της τμηματοποίησής τους (Shang et al. 2016).



Εικόνα 7

Η παραπάνω εικόνα μέσω απλών συμβολισμών μπορεί να μας περιγράψει ακριβώς πως εκτελείτε μια διαδικασία clustering, η οποία έχει ως εξής: Σε πρώτη φάση ορίζουμε το dataset το οποίο θέλουμε να εφαρμόσουν κάποια τεχνική community detection. Θεωρούμε λοιπόν ως input το dataset και στη συνέχεια ορίζουμε μια τεχνική η οποία βάση κάποιων χαρακτηριστικών θα ακολουθήσει μία διαδικασία η οποία θα μας κατηγοριοποιήσει τα στοιχεία του dataset σε ομάδες και θα μας δώσει έναν αριθμό clusters. Τα clusters αυτά θα είναι το output της διαδικασίας και σε ιδανικές συνθήκες θα μας παρουσιάζει το dataset του input τμηματικά. Όπως φαίνεται και στη εικόνα βλέπουμε ομάδες στοιχείων οι οποίες γίνονται διακρικές μέσω του διαφορετικού χρώματος που παρουσιάζουν. Ενδεικτικά τα στοιχεία που παρουσιάζονται με πράσινο χρώμα ορίζουν ένα cluster κοκ.

Συμπερασματικά το graph clustering / community detection πρόκειται για μία διαδικασία η οποία ανακαλύπτει / ανιχνεύει σύνολα αντικειμένων (objects) που τα ομαδοποιεί βάση κάποιων κοινών χαρακτηριστικών. Πολλά ερευνητικά κέντρα που ασχολούνται με τεχνικές clustering έρχονται αντιμέτωπα με ένα σοβαρό πρόβλημα για παράδειγμα την αναγνώριση groups. Υπάρχει ένα όγκος αλγορίθμων οι οποίοι είναι ικανοί τεχνικά να λύσουν το πρόβλημα του clustering, αλλά αρκετοί από αυτοί είναι πολύ ευαίσθητοι και οδηγούν σε λάθος outputs. Ως εκ τούτου, είναι πολύ σημαντικό να εκτιμηθεί το αποτέλεσμα της ομαδοποίησης των networks σε υποσύνολα. Είναι πολύ δύσκολο να προσδιοριστεί εάν ένα αποτέλεσμα clustering είναι αποδεκτό ή όχι. Έτσι

η επιστήμη κατάφερε να αποφύγει αυτό το αδιέξοδο και να ορίσει τεχνικές οι οποίες μπορούν να αξιολογήσουν αποτελέσματα clustering.

Αυτή η εργασία έχει σκοπό να μελετήσει το πρόβλημα αξιολόγησης αποτελεσμάτων συσταδοποίησης γράφων μέσω clustering, θα περιγράψει τεχνικές αξιολόγησης που είναι διαθέσιμες στη βιβλιογραφία και να υλοποιήσει κάποιους δείκτες οι οποίοι αφορούν τις τεχνικές αυτές. Οι τεχνικές αυτές είναι γνωστές στην επιστημονική κοινότητα ως εξής: cluster validity index. Οι δείκτες εγκυρότητας που χρησιμοποιούνται συχνότερα εισάγονται και εξηγούνται σε σύγκριση με πειραματικά αποτελέσματα. [5,6,8,9]

Μέρος 2. Αξιολόγηση Αποτελεσμάτων Συσταδοποίησης (Cluster Validation)

2.1 Παρουσίαση Προβλήματος

Όπως αναφέραμε ένα από τα πιο γνωστά προβλήματα στο πεδίο του data mining είναι το Clustering. Πολλές φορές όταν υλοποιούμαι τεχνικές τέτοιου ίδιους ερχόμαστε αντιμέτωποι με το εξής ερώτημα: πόσο καλά έχει γίνει το partition ενός γράφου και αν τα αντικείμενα που ορίζουν τα groups/partitions που αποτελούν το output της μεθόδου clustering που ακολουθήσαμε παρουσιάζουν όντως ομοιότητες μεταξύ τους και ισχυρές διαφορές με αντικείμενα που έχουν κατηγοριοποιηθεί ώστε να ορίζουν άλλες ομάδες του υπό εξέταση network.

Αντιλαμβανόμαστε ότι το clustering σαν μία διαδικασία χρήζει εποπτείας. Κατά συνέπεια η αξιολόγηση των αποτελεσμάτων των clustering αλγορίθμων είναι πάρα πολλή σημαντική για τους data analysts. Φανταστείτε για παράδειγμα αν μία ομάδα μάρκετινγκ έχει λάβει αποτελέσματα από το συνεργαζόμενο τμήμα data analysis αποτελέσματα clustering για ένα network που ορίζουν ως αντικείμενα οι καταναλωτές με τις συνήθειες τους για μία συγκεκριμένη γεωγραφική περιοχή. Αν όμως αυτά τα partitions που έχει ορίσει η τμηματοποίηση που έχουν χρησιμοποιήσει οι τεχνικοί του τμήματος data analysis δεν έχουν αξιολογηθεί υπάρχει σοβαρή πιθανότητα οι συνεργάτες τους από το τμήμα μάρκετινγκ να λάβουν λάθος αποφάσεις αφού δεν θα έχουν πιθανώς και τη σωστή ανάλυση δεδομένων από τους πρώτους.

Τα τελευταία χρόνια παράλληλα με τις τεχνικές clustering εξελίσσονται και οι τεχνικές που αξιολογούν τα αποτελέσματα των πρώτων. Κατά τη συνέχεια της εργασίας πέρα από ιδιότητες που αφορούν τεχνικές clustering θα σας παρουσιάσω μερικές μετρικές αξιολόγησης κάποιες από αυτές είναι αρκετά καλά ορισμένες και έχουν επικρατήσει και κάποιες από αυτές όχι.

2.2 Εισαγωγή

Στο πρώτο μέρος αναφερθήκαμε σε Community Detection Τεχνικές ή αλλιώς τεχνικές clustering. Αυτές αφορούν τις επικρατέστερες οι οποίες είναι κατηγοριοποιημένες κατά ομάδες. (βλέπε σελ 11)

Τα αποτελέσματα ενός αλγορίθμου συσταδοποίησης γράφων σε clusters μπορεί να ποικίλουν για το ίδιο σύνολο δεδομένων. Καθώς κάθε φορά που ορίζουμε ένα dataset προς clustering θα πρέπει μαζί με αυτό να θέσουμε και μία τιμή εισόδου. Αυτή η τιμή ορίζεται από τον αλγόριθμο/τεχνική που θα χρησιμοποιήσουμε. Αν για παράδειγμα επιλέξουμε να δούμε τα αποτελέσματα για τον αλγόριθμο Clustering LOUVAIN τότε η τιμή εισόδου που πρέπει να ορίσουμε αφορά το resolution. Αν όμως κάνουμε χρήση του Spectral τότε θα πρέπει ως τιμή εισόδου να ορίσουμε το αριθμό των clusters που θέλουμε ιδανικά ως output.

Τα διαφορετικά αποτελέσματα τα οποία προκύπτουν από τις τεχνικές Clustering θα πρέπει να αξιολογούνται. Τη λύση σε αυτό το πρόβλημα δίνουν οι τεχνικές **Cluster Validation**.

Ο στόχος λοιπόν του Cluster Validation είναι να βρεθούν μετρικές οι οποίες θα ορίζουν κατά πόσο καλά θα έχει γίνει η τμηματοποίηση σε clusters των υπό εξέταση datasets. Στις περισσότερες περιπτώσεις οι ερευνητές – μηχανικοί λογισμικού που ασχολούνται με Clustering Τεχνικές εκτελούν τους αλγορίθμους μια η περισσότερες φορές με διαφορετικές παραμέτρους εισόδου για τα ίδια σύνολα δεδομένων. Έτσι οι δείκτες αυτοί εγκυρότητας (Cluster Validity Indexes) μπορούν να χρησιμοποιηθούν για την ιδανική επιλογή τμηματοποίησης ενός dataset σε clusters μέσα από τα διαφορετικά αποτελέσματα που θα μας έχει εξάγει ο αλγόριθμος clustering. Τα τελευταία χρόνια έχουν αναπτυχθεί διάφοροι δείκτες ποιότητας και έχουν εισαχθεί σε διάφορες εργασίες που αφορούν το πρόβλημα αυτό.

2.3 Συσταδοποίηση Γράφων και Αξιολόγηση (Graph Clustering & Validation): Επισκόπηση

Ο στόχος του Graph Clustering είναι να ορίσει και να δημιουργήσει καλά χωρισμένα clusters χρησιμοποιώντας τη δομή γράφων. Το κατά πόσο συμπαγές είναι ένα cluster εξαρτάται από το dataset αλλά και από τη μέθοδο συσταδοποίησης που χρησιμοποιήσαμε.

Διαφορετικοί δείκτες αξιολόγησης τμηματοποίησης γράφων σε clusters έχουν προταθεί για χρήση στο πεδίο της ανάλυσης δεδομένων. Οι περισσότεροι δείκτες που είναι διαθέσιμοι αυτή τη στιγμή βασίζονται στη μεταβλητότητα του intra και inter connectivity και του cohesion. Μέσω των δεικτών αυτών πέρα από την αξιολόγηση μπορούμε και να διαχειριστούμε μεγάλους γράφους.

Η ποιότητα του αποτελέσματος που προκύπτει από μία Clustering μέθοδο εξαρτάται από την δομή του γράφου καθώς και την τεχνική clustering που θα ακολουθήσουμε. Έτσι μέσω των δεικτών ποιότητας μπορούμε να αναλύσουμε επαρκώς τη δομή των clusterred graphs ενώ παράλληλα μας δίνεται και η δυνατότητα να δούμε και την ποιότητα της μέθοδου clustering που χρησιμοποιήσαμε. Έτσι αντιλαμβανόμαστε ότι χρειαζόμαστε τα κατάλληλα εργαλεία τα οποία θα μας βοηθήσουν πέρα από το να αξιολογήσουμε τη δομή των clustered γράφων να αξιολογήσουμε και τις τεχνικές clustering που υπάρχουν διαθέσιμες.

Πριν μελετήσουμε κάποιες από τις τεχνικές που αφορούν την αξιολόγηση των αποτελεσμάτων της συσταδοποίησης σε clusters θα πρέπει να περιγράψουμε κάποιους από τους ορισμούς που αφορούν τις μετρικές αυτές. Ύστερα θα παρουσιαστούν δείκτες αξιολόγησης που δεν αφορούν μόνον τη μεταβλητότητα του intra & inter connectivity αλλά και δείκτες που αφορούν την αξιολόγηση της συνδεσιμότητα των nodes (κόμβων) με τα γειτονικά τους.

Όπως αναφέραμε οι αλγόριθμοι που έχουν δημιουργηθεί ώστε να αποδίδουν clusters από networks δίνουν διαφορετικές λύσεις ανάλογα πάντα με τη συνάρτηση που τους ορίζει. Δεν υπάρχει μεμονωμένα η καλύτερη επιλογή αλγορίθμου για κάθε πιθανό dataset που θα χρειαστεί να υποστεί cluster analysis. Για τους περισσότερους αλγόριθμους Clustering το αποτέλεσμα που θα αποδώσουν σε Clusters ποικίλει και

εξαρτάται από ένα σύνολο παραμέτρων που πρέπει να ορίσουμε πριν τους θέσουμε σε λειτουργία.

Μέσω του Cluster Validity μπορούμε να βρούμε πολύ καλές μεθόδους μέσω των οποίων θα είμαστε σε θέση ώστε να αξιολογήσουμε την ποιότητα των αποτελεσμάτων που πήραμε από τους αλγορίθμους που αναφέραμε. Παράλληλα μπορούμε να συγκρίνουμε τέτοιους αλγορίθμους μεταξύ τους για το ποιος μπορεί να μας δώσει καλύτερα αποτελέσματα. Επίσης μπορούμε να ορίσουμε ποιος είναι ο καλύτερος αριθμός τμηματοποίησης ενός dataset σε Clusters. Συμπερασματικά, κατανοούμε τη σχέση που μπορεί να έχει μία διαδικασία Clustering με μία διαδικασία που αφορά το validation. Πολλές αλλά και διαφορετικές προσεγγίσεις Cluster Validation υπάρχουν διαθέσιμες στη βιβλιογραφία (Βλέπε Σελίδες 21 - 36)

Γενικά οι τεχνικές Validity κατηγοριοποιούνται ως εξής: Σε αυτές που αφορούν το Inter Connectivity και σε αυτές που αφορούν το Intra Connectivity των αποτελεσμάτων σε Clusters. Το πρόβλημα του προσδιορισμού του αριθμού των συστάδων – clusters επιλύεται με την ανίχνευση ενός σημείου που ορίζεται μεταξύ των τιμών validity index για τις διαφορετικές τιμές Clusters που έχουν βρεθεί σε ένα εύρος M = [Mmin, Mmax]. Το σημείο αυτό αφορά τον αριθμό των clusters που παρατηρείται έντονη αλλαγή των τιμών του δείκτη. Οι δείκτες που προτιμάτε να χρησιμοποιούνται για αξιολόγηση αποτελεσμάτων Clustering είναι αυτοί που παίρνουν μέγιστη ή ελάχιστη τιμή στο σημείο αυτό. Είναι όμως αποδεκτό αυτοί οι δείκτες να έχουν τοπικά ελάχιστα και τοπικά μέγιστα στο εύρος των τιμών των Clusters που εξετάζουμε οπότε δε μπορούμε να κρίνουμε με ασφάλεια πιο αποτελέσματα φέρει την ιδανική τμηματοποίηση.

Στην επόμενη ενότητα θα δούμε βασικές τεχνικές για Cluster Validity.

2.4 Χρήσιμοι Ορισμοί

Πριν προχωρήσουμε στη περιγραφή τεχνικών αξιολόγησης αποτελεσμάτων συσταδοποίησης σε γράφους πρέπει να ξεκαθαρίσουμε κάποιους ορισμούς οι οποίοι θα μας βοηθήσουν στην κατανόησή τους.

- Θεωρούμε ένα γράφο G, με N κόμβους v_i και ακμές E e_{ij} μεταξύ των κόμβων v_i και v_j
- Ο γράφος G είναι clustered σε ένα σετ clusters που ορίζεται ως εξής: {C_1, C_2……C_k}
- Θεωρούμε ότι N_i είναι ο αριθμός των κόμβων που ορίζουν το C_i, Cluster
- Θεωρούμε ότι E_i είναι ο αριθμός των ακμών που ορίζουν το C_i, Cluster
- Επιπλέον το E_{ij} ορίζει τον αριθμό των ακμών (links) που συνδέουν τον κόμβο C_i με τον C_j
- Με τον συμβολισμό E_i θεωρούμε τον αριθμό των links που συνδέουν το C_i cluster με τα υπόλοιπα clusters.
- Η απόσταση μεταξύ v_i και v_j εκφράζεται ως εξής: d(v_i, v_j) και ορίζει το μήκος της κοντινότερης απόστασης μεταξύ των δύο αυτών κόμβων στον γράφο G.
- Η απόσταση μεταξύ δύο Clusters εκφράζεται ως d(C_i, C_j)
- Κατά τον ίδιο τρόπο μπορούμε να ορίζουμε και την απόσταση του κόμβου v_i από το Cluster C_j: d(v_i, C_j)
- Στη βιβλιογραφία επίσης για την εξαγωγή συμπερασμάτων αξιολόγησης μέσω των τεχνικών αυτών χρησιμοποιούν και την διάμετρο του Cluster, που ορίζεται από την απόσταση των δύο πιο απομακρυσμένων κόμβων (remote nodes) ή από την μέση απόσταση που ορίζουν όλα τα nodes του Cluster μεταξύ τους. Έτσι ως diam(C_k) θεωρούμε τη διάμετρο του Cluster C_k.

Σε αυτή την ενότητα αρκετοί δείκτες ποιότητας (validity indices) παρουσιάζονται. Αυτοί χρησιμοποιούνται αποκλειστικά για να αξιολογίσουν την ποιότητα της τμηματοποίησης σε Clusters που δημιούργησαν διαφορετικοί αλγόριθμοι Clustering ή προέκυψαν μέσω της χρήσης των ίδιων αλλά για διαφορετικές τιμές εισόδου κάθε φορά (input parameter values). **Σημειώνω ότι οι μετρικές που παρουσιάζονται αφορούν cluster τα οποία δεν έχουν σημεία τομής μεταξύ τους (no overlapping between partitions is allowed)**

2.5 Δείκτες Εγκυρότητας Συστάδας Για Γράφους: Τεχνικές & Εργαλεία

Σε αυτή την ενότητα θα περιγράψουμε κάποιες τεχνικές οι οποίες αφορούν το Cluster Validity. Έως και σήμερα έχουν οριστεί αρκετές τεχνικές αλλά πολλές από αυτές όμως έχουν υποστεί αρκετές βελτιώσεις μέχρι να φέρουν την τελική τους μορφή ως προς τον ορισμό τους. Οι δείκτες που μπορούν να αξιολογήσουν την ποιότητα τμηματοποίησης ενός γράφου σε clusters συνήθως συγκρίνουν τη συνοχή σύνδεσης των μεταξύ των clusters (inter connectivity). Μετρήσεις αξιολόγησης όμως μπορούν να γίνουν και για το κάθε ένα cluster ξεχωριστά (intra connectivity) ώστε να ελέγξουμε πόσα καλά ορίζεται αυτό βάση της σχετικότητα που παρουσιάζουν τα στοιχεία που το αποτελούν (nodes) Τέλος υπάρχουν και μετρικές αξιολόγησης οι οποίες αξιολογούν τον γράφο συνολικά. Αυτές για να υπολογιστούν λαμβάνουν υπόψη αποκλειστικά τις συνδέσεις ακμών E καθώς και τον αριθμό των κόμβων N. Μέσω αυτών των τύπων αξιολόγησης μπορούμε να συγκρίνουμε γράφους που μπορεί να έχουν τον ίδιο αριθμό Nodes αλλά να παρουσιάζουν τελείως διαφορετική δομή.

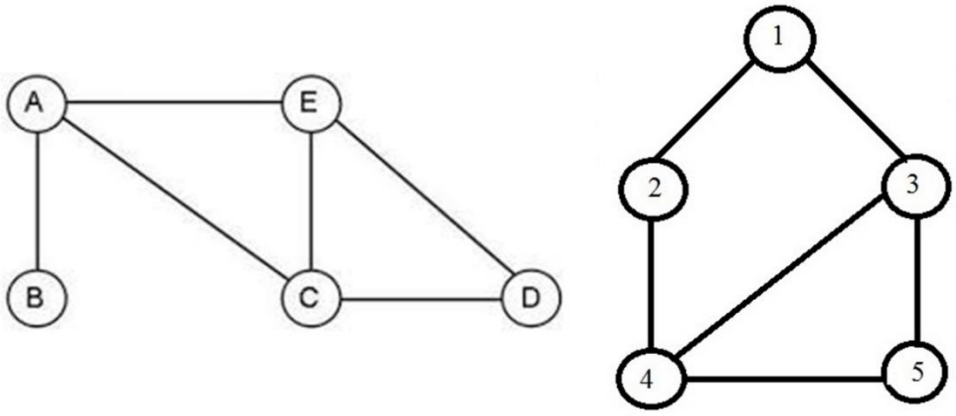
Τεχνική 1. Compactness based on edge density

Ένας απλός δείκτης που μέσα από τον υπολογισμό του μπορούμε να περιγράψουμε την συνεκτηκότητα που παρουσιάζει ο υπό εξέταση γράφος.

Ο Τύπος Υπολογισμού:

$$\frac{E}{N} \text{ or } \frac{E}{N^2}$$

Πρόκειται για έναν εύκολο υπολογισμό και από τους πρώτους που έχουν εμφανιστεί στη βιβλιογραφία και αφορά την αξιολόγηση ενός γράφου. Όπως παρατηρούμε πρόκειται για έναν τύπο που λαμβάνει υπόψη μόνο τον αριθμό των ακμών E (edges) και τον κόμβων N (Nodes) του γράφου, χωρίς όμως να συμπεριλαμβάνει στον ορισμό του κάτι για την δομή που έχει ο γράφος. Υπάρχουν γράφοι μπορούν να έχουν τον ίδιο αριθμό κόμβων και ακμών χωρίς όμως να εμφανίζουν την ίδια δομή.



Εικόνα 8

Γεγονός που επιβεβαιώνεται και με την παραπάνω εικόνα. Αν λοιπόν βάζαμε στον τύπο E/N τα nodes και τα edges των δύο αυτών γράφων θα παίρναμε το ίδιο αποτέλεσμα καθώς έχουν ίδιο αριθμό ακμών και κόμβων E/N = 6/6.

Τεχνική 2. Compactness Index Cp

Ένας πολύ καλός δείκτης που ορίζει τη συνεκτικότητα ενός γράφου λαμβάνοντας υπόψη την δομή του καθώς και τη συνδεσιμότητα των στοιχείων του (graph connectivity) είναι αυτός που θα δούμε παρακάτω:

Ας θεωρήσουμε ένα γράφο G. Ο δείκτης Cp υπολογίζεται ως εξής [13]:

$$Cp = \frac{Max - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d(v_i, v_j)}{Max - Min}$$

Διευκρινίζουμε ότι με MAX και MIN θεωρούμε τη μέγιστη και την ελάχιστη τιμή των αποστάσεων που ορίζουν οι κόμβοι $\Sigma_i \Sigma_j$ d (v_i, v_j).

Σύμφωνα με τη θεωρία υπάρχει ένα σύνολο N (N – 1) /2 ζευγαριών που ορίζουν την απόσταση $v_i - v_j$. Αν θεωρήσουμε ότι Q είναι η μέγιστη απόσταση δύο κόμβων μέσα στον υπό μελέτη γράφο τότε η τιμή των

αποστάσεων d (vi, vj) θα είναι μεταξύ των τιμών 1 και Q. Έτσι αποδεικνύεται ότι:

$$MIN = N(N-1)/2$$

και

$$MAX = Q * \{ N (N-1)/2 \}$$

Ο παραπάνω τύπος θα μας δώσει ορθά αποτελέσματα μόνο για γράφους που είναι καλά συνδεδεμένοι (connected graphs). Για γράφους όμως που δεν έχουν συνδέσεις μεταξύ τους, η τιμή του Q ορίζεται αυθαίρετα [13].

Συμπερασματικά αντιλαμβανόμαστε ότι για αυτόν τον τύπο στον υπολογισμό μας για την εξαγωγή αποτελέσματος εξαρτόμαστε αποκλειστικά από την τιμή που θα πάρει το Q.

Η τεχνική αυτή όμως με την πάροδο των χρόνων παρουσιάστηκε βελτιωμένη στη βιβλιογραφία καθώς η σταθερά Q δεν θα ληφθεί υπόψιν για τον νεό μας υπολογισμό, αυτό του Cp*. Ας δούμε λοιπόν πως ορίζεται ο δείκτης αυτός.

Θεωρούμε δύο κόμβους (nodes) v_i, v_j για τους οποίους ορίζουμε ως ομοιότητα (similarity) την παρακάτω σχέση:

$Sim(v_i, v_j) = 1 / d(v_i, v_j)$ αν οι κόμβοι αυτοί είναι συνδεδεμένοι.

Και

Sim(vi, vj) = 0 αν αυτοί δεν παρουσιάζουν σύνδεση

Οι τιμές μεταξύ των τιμών 0 και 1 δεν ορίζουν γράφο καλά ορισμένο. Έτσι ο νέο τύπος του Cp* ορίζεται ως εξής:

$$Cp^* = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} sim(v_i, v_j)}{N(N-1)/2}$$

Ο Cp* είναι ένας δείκτης δεσμευμένος, ο οποίος μπορεί να δώσει μόνο τις τιμές 0 και 1 και μπορεί να μας ορίσει αν ένας γράφος είναι καλά ορισμένος (complete graph).

Στην συνέχεια θα παρουσιαστούν τεχνικές οι οποίες μας δίνουν αποτελέσματα βασισμένες στις τιμές που ορίζονται από το inter και intra connectivity των υπό μελέτη γράφων.

Τεχνική 3. Dunn's Index

Η τεχνική αυτό είναι γνωστή ως και Dunn's Index και ο ορισμός της φαίνεται από τον ακόλουθο τύπο [10]:

$$D = \min_{i=1\ldots n_c} \left\{ \min_{j=i+1\ldots n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1\ldots n_c} (diam(c_k))} \right) \right\}$$

Όπου,

$$d(c_i, c_j) = \min_{x \in c_i, y \in c_j} \{d(x, y)\} \text{ and } diam(c_i) = \max_{x, y \in c_i} \{d(x, y)\}$$

Ο δείκτης Dunn συγκρίνει την ελάχιστη απόσταση του Cluster με την μέγιστη διάμετρό του. Έτσι αν τα υπό εξέταση dataset είναι καλά χωρισμένα μεταξύ τους (well separated clusters) η απόσταση μεταξύ των Clusters είναι σχετικά μεγάλη και η τιμή των διαμέτρων που υπολογίζει αυτός ο τύπος πρέπει να είναι μικρή. Έτσι η καλύτερη τμηματοποίηση θα ανιχνεύεται εκεί που ο παραπάνω τύπος παίρνει τη μέγιστη τιμή του.

Υπάρχουν όμως και μειονεκτήματα τα οποία έχουν εντοπιστεί σε αυτή την τεχνική Cluster Validity και είναι τα παρακάτω:

Ο υπολογισμός του δείκτη μπορεί να χρειαστεί αρκετή ώρα μέχρι να δώσει αποτελέσματα αλλά και παρουσιάζει πολλές ευαισθησίες στον υπολογισμό της μέγιστης τιμής της διαμέτρου του Cluster. [10] Ο τύπος αυτός όπως και αυτοί που παρουσιάζονται στις παραπάνω τεχνικές έχει υποστεί βελτιώσεις οι οποίες αφορούν έναν διαφορετικό τρόπο υπολογισμού των αποστάσεων των cluster και των διαμέτρων τους. [11]

Τεχνική 4. Davies-Bouldin Index (DBI)

Πρόκειται για άλλη μία τεχνική αξιολόγησης αποτελεσμάτων clustering. Αφορά ένα σχήμα μεθόδου που λαμβάνει υπόψιν την εσωτερική δομή

του cluster (inter connectivity) με το αποτέλεσμα αυτού να μας εμφανίζει το πόσο καλά έχει γίνει η ομαδοποίηση. Ο ορισμός αυτής της τεχνικής βασίζεται αποκλειστικά στην ομοιότητα των clusters (Rij) η οποία με τη σειρά της βασίζεται στο μέτρο διασποράς του cluster (s_i) καθώς και του μέτρου ανομοιότητας που παρουσιάζουν τα clusters μεταξύ τους (d_{ij}). [10, 13]

Για να κατανοήσουμε τον ορισμό αυτής της τεχνικής θα πρέπει να περιγραφούν ποιες συνθήκες πρέπει να ικανοποιεί το μέτρο ομοιότητας των clusters R_{ij} οι οποίες εμφανίζονται παρακάτω:

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- $\text{if } s_i = 0 \text{ and } s_j = 0 \text{ then } R_{ij} = 0$
- $\text{if } s_j > s_k \text{ and } d_{ij} = d_{ik} \text{ then } R_{ij} > R_{ik}$
- $\text{if } s_j = s_k \text{ and } d_{ij} < d_{ik} \text{ then } R_{ij} > R_{ik}$

Στις περισσότερες περιπτώσεις το μέτρο διασποράς των clusters είναι η μέση απόσταση από το κέντρο των συστάδων. Έτσι το $R_{i,j}$ ορίζεται βάση του ακόλουθου τύπου

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$d_{ij} = d\left(v_i, v_j\right), \; s_i = \frac{1}{\left\|c_i\right\|} \sum_{x \in c_i} d\left(x, v_i\right)$$

Συμπερασματικά ο Davies – Bouldin δείκτης μας αποκαλύπτει για κάθε cluster ποιο από τα υπόλοιπα είναι πιο όμοια με αυτό. Ύστερα αθροίζει τη μέγιστη ομοιότητα των cluster ώστε να μας δώσει το αποτέλεσμα.

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i, \text{ where}$$

$$R_i = \max_{j=1...nc, i\neq j}\left(R_{ij}\right), \; i = 1...n_c$$

Τέλος αν ο δείκτης μας δώσει τιμές πολύ μικρές σημαίνει ότι τα clusters μεταξύ τους παρουσιάζουν ισχυρές διαφορές οπότε το κάθε ένα από αυτά έχει οριστεί καλά αφού αυτά είναι συμπαγή. Αυτό επιβεβαιώνει την ιδέα ότι κανένα σύμπλεγμα/cluster δεν πρέπει να είναι παρόμοιο με ένα άλλο, και ως εκ τούτου το καλύτερο σχήμα ομαδοποίησης ελαχιστοποιεί ουσιαστικά τον δείκτη!

Τεχνική 5. MinMaxCut

Η συνεκτικότητα ενός Cluster Ci, μπορεί να υπολογιστεί ως εξής:

$$Ei' / Ei$$

Ο τύπος που ορίζει αυτή την τεχνική ορίζεται παρακάτω:

$$MinMaxCut = \sum_{i=1}^{K} \frac{E_i'}{E_i}$$

Αντιλαμβανόμαστε ότι πρόκειται για μια απλή τεχνική η οποία λαμβάνει υπόψη μόνο τις συνδέσεις μέσω των ακμών χωρίς να κάνει κάποιον υπολογισμό για τις κορυφές (nodes).

Τεχνική 6. Conductance of a cut

Η intercluster σύνδεση των τμηματοποιήσεων ονομάζεται cut. Η τεχνική αυτή συγκρίνει το μέγεθος των τμηματοποιήσεων και τον αριθμό των ακμών μέσα σε ένα μικρότερο γράφο (subgraph) που έχει δημιουργηθεί από κάποια μέθοδο Clustering. Ας θεωρήσουμε την τμηματοποίηση που έχει προκαλέσει μια μέθοδος Clustering και από αυτήν έχουν προκύψει δύο νέα Clusters τα Ci και Cj. Ως Conductance αυτής της τμηματοποίησης ορίζουμε το εξής:

$$Conductance(C_i,C_j)=\frac{E_{ij}}{\min(E_i,E_j)}$$

Τα Clusters λοιπόν που παρουσιάζουν μικρές τιμές στον παραπάνω υπολογισμό είναι καλά χωρισμένα.

Τεχνική 7. Coverage of a graph clustering

Η τεχνική αυτή ορίζεται από το κλάσμα που δημιουργούν οι τιμές των intra-cluster ακμών με το συνολικό αριθμό αυτών:

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

Σε αυτή την περίπτωση, όταν ο παραπάνω τύπος πάρει τη μεγαλύτερη τιμή μου τότε έχουν και την καλύτερη ποιοτικά τμηματοποίηση του του Clustering C. Αυτός ο δείκτης είναι πολύ εύκολο να υπολογιστεί αλλά όπως και ο προηγούμενος που παρουσιάζουμε στην τεχνική 6. δεν λαμβάνει υπόψη στον υπολογισμό του τον αριθμό των nodes μέσα στο Cluster που έχει δημιουργηθεί [10, 13].

Τεχνική 8. Performance of clustering

Από εδώ και στο εξής θα δούμε κάποιους δείκτες οι οποίοι έχουν πιο πολύπλοκους υπολογισμούς καθώς λαμβάνουν υπόψη τα nodes και τα edges που έχουν προκύψει από το clustering ενός γράφου.

Ξεκινώντας έντονη παρουσία στη βιβλιογραφία έχει ο παρακάτω τύπος:

$$Perf(G) = 1 - \frac{\sum_{i<j}^{K} E_{ij} + \sum_{i=1}^{K}\left(\frac{N_i(N_i-1)}{2} - E_i\right)}{\frac{N(N-1)}{2}}$$

ή

$$Perf(G) = 1 - \frac{\|False+\| + \|False-\|}{\frac{N(N-1)}{2}}$$

Σημειώνουμε ότι ως False- θεωρούμε τον αριθμό των ακμών μεταξύ των clusters (inter – cluster links) και ως False + τον αριθμό των ζευγαριών των nodes/κόμβων (v_i,v_j) που βρίσκονται στο ίδιο cluster αλλά δεν συνδέονται μεταξύ τους [13].

Τεχνική 9. Modularization

Πρόκειται ένα μέτρο το οποίο χαρακτηρίζει τη δομή των clusters που έχουν προκύψει από κάποια τεχνική τμηματοποίησης (portioning) ενός γράφου. Σχεδιάστηκε για να μετρήσει τη δύναμη της διαίρεσης ενός δικτύου σε ενότητες (που ονομάζονται επίσης ομάδες, ομάδες ή κοινότητες). Τα δίκτυα με υψηλή αρθρωτότητα (modularity) έχουν πυκνές συνδέσεις μεταξύ των κόμβων εντός των clusters, αλλά αραιές συνδέσεις μεταξύ κόμβων σε διαφορετικά clusters. Ο modularity ως δείκτης validity χρησιμοποιείται συχνά σε μεθόδους βελτιστοποίησης για την ανίχνευση της κοινοτικής δομής στα δίκτυα. Ωστόσο, έχει αποδειχθεί ότι αυτός υποφέρει από ένα όριο ανάλυσης και ως εκ τούτου δεν είναι σε θέση να ανιχνεύσει μικρές κοινότητες. Η modularity είναι το κλάσμα των ακμών που εμπίπτουν στις συγκεκριμένες ομάδες μείον το αναμενόμενο κλάσμα, αν οι άκρες κατανέμονται τυχαία. Η τιμή της modularity για μη σταθμισμένα και μη κατευθυνόμενα γραφήματα βρίσκεται στην περιοχή των τιμών [-1 / 2,1]. Είναι θετικό αν ο αριθμός των άκρων (edges) εντός των ομάδων (clusters) υπερβαίνει τον αναμενόμενο αριθμό με βάση την τύχη. Για μια δεδομένη διαίρεση των κορυφών του δικτύου σε ορισμένες ενότητες, η modularity αντανακλά τη συγκέντρωση των άκρων εντός των ενοτήτων σε σύγκριση με την τυχαία κατανομή των συνδέσεων μεταξύ όλων των κόμβων ανεξάρτητα από τις ενότητες.

Η modularity Q ορίζεται στη συνέχεια ως το κλάσμα των άκρων που εμπίπτουν στην ομάδα Ci ή Cj, μείον τον αναμενόμενο αριθμό ακμών εντός των ομάδων Ci και Cj για ένα τυχαίο γράφημα με την ίδια κατανομή βαθμού κόμβου με το δεδομένο δίκτυο.

Πιο συγκεκριμένα η συνάρτηση η οποία ορίζει το Modularity ενός γράφου εκφράζεται μέσω της διαφοράς του intra & inter cluster connectivity [12].

Οπότε το intra cluster connectivity ενός Cluster Ci υπολογίζεται βάση του τύπου:

$$intra(C_i) = \frac{E_i}{N_i(N_i-1)/2}$$

όπου, Ni(Ni-1)/2 αφορά τον υπολογισμό του μέγιστου αριθμού του intra-cluster edges του Ci.

Επίσης,

Το Inter – Cluster – Connectivity αφορά μια μετρική η οποία ορίζει την σύνδεση μεταξύ δύο Clusters, Ci & Cj και υπολογίζεται ως εξής:

$$inter(C_i,C_j) = \frac{E_{ij}}{N_i N_j}$$

Όμως, υπάρχει πιθανότητα ο αριθμός των clusters που έχουν προκύψει από την τμηματοποίηση ενός γράφου να είναι αρκετά μεγάλος τότε για τον υπολογισμό του Modularity Q θα πρέπει να προσδιορίσουμε τις μέσες τιμές των intra και inter cluster connectivity.

Επομένως,

$$\overline{intra} = \frac{\sum_{i=1}^{K} \frac{E_i}{N_i(N_i-1)/2}}{K} \quad and \quad \overline{inter} = \frac{\sum_{i<j}^{K} \frac{E_{ij}}{N_i N_j}}{K(K-1)/2}$$

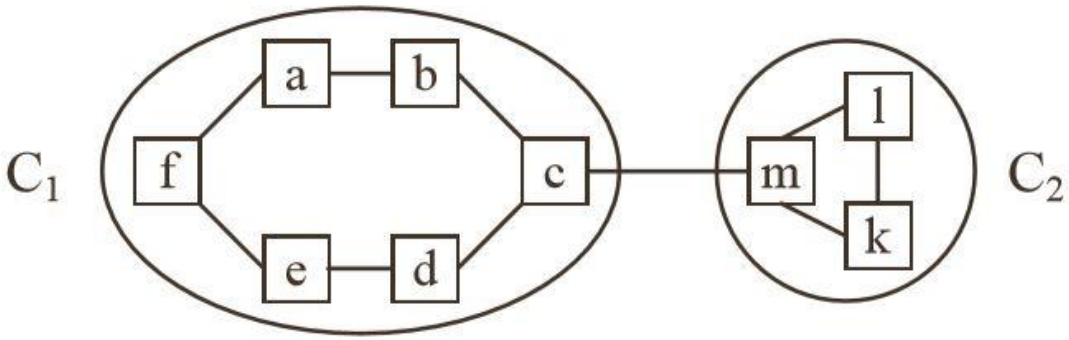
Έτσι η τελική μορφή του δείκτη Modularity Q είναι η ακόλουθη:

$$MQ = \overline{intra} - \overline{inter} = \frac{\sum_{i=1}^{K} \frac{E_i}{N_i(N_i-1)/2}}{K} - \frac{\sum_{i<j}^{K} \frac{E_{ij}}{N_i N_j}}{K(K-1)/2}$$

Για τον συγκεκριμένο δείκτη όμως προέκυψαν βελτιώσεις και αυτός τροποποιήθηκε [13] και υπάρχει στη βιβλιογραφία ως MQ* index:

$$MQ^* = \frac{\sum_i E_i}{\sum_i \frac{N_i(N_i-1)}{2}} - \frac{\sum_{i<j} E_{ij}}{\sum_{i<j} N_i N_j}$$

Ο δείκτης που ορίζεται από τον δεύτερο τύπο MQ* λαμβάνει υπόψη στον υπολογισμό του και το τρόπο σύνδεσης των clusters μεταξύ τους αλλά και το μέγεθός τους. [13]



Εικόνα 9

Στην εικόνα 9 φαίνεται ότι το Cluster, C1 είναι μεγαλύτερο από το C2 και παρουσιάζει μικρότερη τιμή στο intra-cluster-connectivity. Έτσι, η μετρική MQ* παίρνει μικρότερη τιμή από αυτή της MQ (MQ*= 0.44 και MQ = 0.64) [13] Κάτι που επιβεβαιώνει ότι ο δεύτερος τύπος λαμβάνει υπόψη του τη δομή του cluster ως προς το μέγεθος καθώς και τη συνδεσιμότητά του.

Στις προηγούμενες πέντε τεχνικές Validity για την αξιολόγηση τμηματοποίησης ενός γράφου σε communities περιγράψαμε δείκτες που αφορούν τη συνεκτικότητα των αποτελεσμάτων σε clusters

(compactness indices). Στη συνέχεια θα δούμε κάποιες τεχνικές που αφορούν την αξιολόγηση αποτελεσμάτων βάση της γειτνίασης που παρουσιάζουν οι κόμβοι (nodes) μεταξύ τους. Μέσω τέτοιου είδους τεχνικών μελετάμε αν κάθε κόμβος του cluster που έχει οριστεί παρουσιάζει συσχέτιση με τα υπόλοιπα ή όχι. Έτσι ώστε όσα παρουσιάσουν διαφορές να αφαιρεθούν από το cluster ή να γίνει εκ νέου τμηματοποίηση ώστε να οριστούν νέα clusters.

Τεχνική 10. Sihouette index.

Ξεκινώντας την περιγραφή αυτής της τεχνικής ας θεωρήσουμε ως ένα partition ενός dataset X σε clusters C = {C_1,C_2,…C_k}. Αυτός ο δείκτης ορίζεται ως εξής παρακάτω:

$$S(\mathbf{x}) = \frac{b(\mathbf{x}) - a(\mathbf{x})}{max\left(a(\mathbf{x}), b(\mathbf{x})\right)}$$

Όπου υπάρχει το α(x) θεωρούμε την μέση απόσταση μεταξύ του κόμβου x ο οποίος ανήκει στο cluster C_k, με τα υπόλοιπα στοιχεία που x_k που ορίζουν αυτό το Cluster. Αυτή η τιμή ορίζεται ως εξής:

$$a(\mathbf{x}) = \frac{1}{n_k - 1} \sum_{\mathbf{x}_k \in C_k} d\left(\mathbf{x}, \mathbf{x}_k\right)$$

όπου n_k θεωρήστε τα μέρη (nodes) που ορίζουν το cluster C_k. Επίσης με b(x) θεωρούμε την ελάχιστη απόσταση του σημείου – κόμβου x από τα υπόλοιπα του ίδιου Cluster, x_i. Αυτή η απόσταση υπολογίζεται ως εξής:

$$b(\mathbf{x}) = \min_{\substack{\iota=1 \\ \iota \neq k}}^{K} \delta(\mathbf{x}, \mathbf{x}_\iota)$$

όπου η απόσταση δ για το Cluster C_i ορίζεται με τον παρακάτω τύπο:

$$\delta(\mathbf{x}, \mathbf{x}_\iota) = \frac{1}{n_\iota} \sum_{\mathbf{x}_\iota \in C_\iota} d(\mathbf{x}, \mathbf{x}_\iota)$$

τέλος η τιμή n_i ορίζει τον αριθμό των κόμβων που ορίζουν το Cluster C_i. Έτσι η τιμή για τον δείκτη Silhouette για τα δοσμένα C_k , Clusters υπολογίζεται από τον τύπο:

$$S(C_k) = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} S(\mathbf{x})$$

Η μετρική που ορίζεται από τον δείκτη Silhouette είναι ένα μέτρο που εγκρίνει ή απορρίπτει πόσο παρόμοιο είναι ένα αντικείμενο (κόμβος) με αυτά που ορίζουν το δικό του σύμπλεγμα (cluster) σε σύγκριση με άλλα συμπλέγματα (διαχωρισμός). Το εύρος των τιμών του δείκτη αυτού κυμαίνεται από − 1 έως 1, όπου η υψηλότερη τιμή μεταξύ αυτών των τιμών δείχνει ότι το αντικείμενο είναι καλά ταυτισμένο με το δικό του σύμπλεγμα (cluster) και δε παρουσιάζει ομοιότητες με τις γειτονικές συστάδες. Εάν τα περισσότερα αντικείμενα έχουν υψηλή τιμή, τότε η ρύθμιση παραμέτρων συμπλέγματος είναι κατάλληλη. Εάν πολλά σημεία έχουν χαμηλή ή αρνητική τιμή, τότε η ρύθμιση παραμέτρων συμπλέγματος μπορεί να έχει πάρα πολλά ή πολύ λίγα συμπλέγματα.

Παρακάτω παρουσιάζομε τον σταθμικό μέσο της παρούσας μετρικής ο οποίος προσαρμόζει τον δείκτη αυτό για τον υπολογισμό περισσότερων clusters αλλά και λαμβάνει υπόψη το μέγεθός τους.

$$GS^* = \frac{\sum_{j=1}^{K} N_j S_j}{\sum_{j=1}^{K} N_j} = \frac{\sum_{i=1}^{N} s(v_i)}{N}$$

Τεχνική 11. Coverage measures

Μία ακόμη μετρική για την αξιολόγηση των αποτελεσμάτων Clustering ενός Γράφου είναι και αυτή που είναι γνωστή στη βιβλιογραφία ως Naïve coverage measure [13]

Ξεκινώντας την περιγραφή της θεωρούμε v_i έναν κόμβο ο οποίος ανήκει στο Cluster, Cj. Ως $N(v_i)$ θεωρούμε το σύνολο των γειτονικών κόμβων του v_i. Για να υπολογίσουμε όμως την τιμή αυτής της μετρικής θα πρέπει να λάβουμε υπόψη μας και τα εξής:

False Positive Set και False Negative Set (*τα οποία τα συναντήσαμε και στην τεχνική νο 6)*

Θυμίζουμε ότι το $False_i+$ αφορά το σύνολο των κόμβων που υπάρχουν στο cluster C_i τα οποία δεν ανήκουν στο σύνολο των γειτονικών κόμβων του v_i – $N(v_i)$ και το $False_i-$ αφορά το σύνολο των κόμβων που είναι γειτονικά στο v_i – $N(v_i)$ χωρίς όμως να ανήκουν στο Cluster, C_j.

Ο τύπος του Naïve Coverage Measure είναι [13] [16]:

$$Cov(v_i) = 1 - \frac{\|False_i+\| + \|False_i-\|}{N-1}$$

Με το αποτέλεσμα αυτού του τύπου να μας δίνει μία τιμή η οποία είναι ίδια με αυτή που θα πάρουμε από την μετρική που παρουσιάσαμε στην Τεχνική 8: Perfomance of clustering (όταν αξιολογούμε δύο ίδια αποτελέσματα clustering γράφων)

Τεχνική 12. Cluster Index in Small World graphs

**Ως small world graph ορίζεται ο γράφος του οποίου τα περισσότερα nodes δεν παρουσιάζουν ισχυρή γειτνίαση μεταξύ τους*

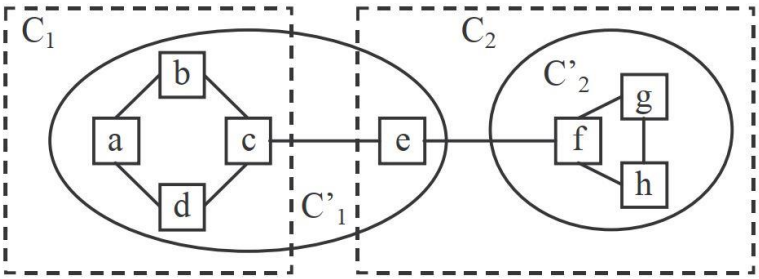
Ένας ακόμη τύπος που μπορεί να χρησιμοποιηθεί για Clustering Validation έχει παρουσιαστεί στο [17]. Με αυτή τη μετρική να υπολογίζει την πυκνότητα των ακμών των γειτονικών κόμβων του κόμβου v_i. [13] Θεωρούμε πάλι ότι ως $N(v_i)$ ορίζεται το σύνολο των γειτονικών κόμβων γύρω από τον κόμβο v_i , όπου n ο αριθμός των κόμβων και e ο αριθμός των ακμών που ορίζουν το υπό εξέταση Cluster. Έτσι, εμείς υπολογίζουμε τον τύπο $c(v_i)$ με τον εξής τρόπο:

$$c(v_i)=\frac{e}{\frac{n(n-1)}{2}}$$

Για το σύνολο όμως των Clusters σε έναν Γράφο ο παραπάνω τύπος παίρνει τη μορφή:

$$c(G)=\frac{\sum_{i=1}^{N} c(v_i)}{N}$$

Τέλος πριν προχωρήσουμε στο τρίτο μέρος της εργασίας καλό είναι για να καλύψουμε όλο το εύρος των τεχνικών validity που έχουν απασχολήσει την επιστημονική κοινότητα θα περιγράψουμε και αυτές που αφορούν τις εξωτερικές συνδέσεις μεταξύ των Clusters.



Εικόνα 10.

Στην εικόνα 10 βλέπουμε δύο τμηματοποιήσεις, τις P και P':

P = {C_1, …. C_k } και την

P' = {C'_1,…,C'_t }

Τεχνική 13. Indices based on co – clusteredness

Πρόκειται για μία τεχνική η οποία είναι βασισμένη στην κατανομή των N(N-1)/2 των γκρουπ που ορίζουν οι κόμβοι (v_i, v_j) για τις τμηματοποιήσεις P και P' [13]. Τα οποία παρουσιάζονται στο παρακάτω πινακάκι.

Partition of $\{(v_i,v_j)\}$	Same cluster in P'	Different clusters in P'
Same cluster in P	a	c
Different clusters in P	b	d

Jaccard Coefficient:

Σε πρώτη φάση για να ορίσουμε τον τύπο του δείκτη θα πρέπει να προσδιορίσουμε την σχετικότητα που παρουσιάζεται μεταξύ των P και P', η οποία βρίσκεται από το δείκτη του Jaccard Coefficient.

$$J = \frac{a}{a+b+c}$$

Ο δείκτης J υπολογίζει την πιθανότητα να ανήκουν δύο κόμβοι στο cluster σε μία partition καθώς και την πιθανότητα να βρίσκονται στο ίδιο cluster αλλά σε άλλο partition.

Folks and Mallows Index:

Οι Folks and Mallows παρουσιάζουν [13] έναν ακόμη δείκτη ο οποίος υπολογίζεται ως εξής:

$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

όπου,

το κλάσμα α/(α+b) αφορά την πιθανότητα δύο κόμβων να ανήκουν στο ίδιο Cluster μέσα στην P αν αυτά ανήκουν μέσα στο ίδιο Cluster στην P' και το αντίστροφο [13].

Rand Statistic:

Το Rand Statistic υπολογίζει την ομοιότητα μεταξύ P και P' μέσω του τύπου:

$$R = \frac{a+d}{a+b+c+d}$$

Ο δείκτης αυτός εισάγει την τιμή d ώστε να δώσει αποτέλεσμα. Υπολογίζει την πιθανότητα δύο κόμβων να ανήκουν στο ίδιο cluster ή σε διαφορετικά Clusters ενώ βρίσκονται στο ίδιο partioning P, P'.

Hubert and Arabie's Statistic:

Οι Hubert και Arabie [13] αναδιαμορφώσαν την μετρική R και προέκυψε ένας νέο τύπος:

$$Hubert = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Την μετρική του Hubert θα την βρούμε στη βιβλιογραφία και ως Phi Statistic. [13]

Τεχνική 14. Q – Graph Index

Στη δημοσίευση [18] παρουσιάζεται ένας δείκτης ποιότητας αξιολόγησης αποτελεσμάτων συσταδοποίησης δεδομένων. Μέσω αυτού εισάγεται μια νέα μετρική η οποία προχωρά στην αξιολόγηση των αποτελεσμάτων μέσω μια διαφορετικής προσέγγισης. Η προσέγγιση αυτή γίνεται μέσω δύο τμημάτων υπολογισμού. Το πρώτο αφορά τον ορισμό της πυκνότερης περιοχής μεταξύ και ενδιαμέσου των υπό μελέτη συστάδων (clusters) και το δεύτερο αφορά μία τη συνδεσιμότητα των clusters (clusters connectivity).

Πιο αναλυτικά για τον υπολογισμό του δείκτη Qgraph θα πρέπει να κάνουμε τους εξής υπολογισμούς:

a. intraLink(C)

b. interLink(C)

c. separation(C)

Ο a. υπολογισμός για να γίνει θα πρέπει να υπολογιστούν οι τιμές του υπό εξέταση γράφου οι οποίες αφορούν: α. το degeneracy* και β. το degeneracy core.

Η μετρική degeneracy για έναν γράφο αφορά τον μέγιστο αριθμό των κόμβων (vertices) σε ένα σύνολο V: deg(G) = kmax-core = maxv E V core (v).

*The degeneracy of a graph is a measure of how sparse it is, and is within a constant factor of other sparsity measures such as the arboricity of a graph. [18]

Μέσω του degeneracy μπορούμε να αξιολογήσουμε ποιοτικά πόσο ισχυρά συνεκτικές είναι οι συστάδες μέσα στους γράφους.

Η μετρική degeneracy core για έναν γράφο G ορίζεται ως εξής:

dG = (dV,dE), dV **c** V και dE **c** E (όπου V θεωρούμε τους κόμβους και E τις ακμές του γράφου G) με την τιμή της να δίνεται από τον τύπο

$$\textbf{deg-coverage(G)} = |\mathbf{dV}| / |\mathbf{V}|$$

Συμπερασματικά προκύπτει ότι για τον υπολογισμό του intraLink, θεωρώντας μια τμηματοποίηση ενός γράφου G σε συστάδες m C = $\{c_1,c_2 \ldots c_m\}$ η τιμή linkage για την τμηματοποίηση αυτή υπολογίζεται ως η μέση τιμή του intra_linkage σε σχέση με όλες τις τμηματοποιήσεις:

$$\textbf{IntraLink(C)} = \mathbf{1/m \{ \Sigma_{ci\ E\ C}\ intralinkage(c_i) \}}$$

Επίσης άλλη μία μετρική μέσω της οποίας μπορούμε να αξιολογήσουμε έναν γράφο σε σχέση με τη συνέκτικότητα του είναι αυτή του density. [19] Με τον υπολογισμό αυτής να γίνεται ως εξής για μια συστάδα - της μορφής $c_i = (V_i, E_i)$ – $dens(c_i)$:

$$dens(c_i) = \frac{2 \cdot |E_i|}{|V_i|(|V_i|-1)}$$

Προκύπτει λοιπόν ότι το density ενός cluster μας δείχνει την συνδεσιμότητα μεταξύ των κόμβων μέσα σε αυτό ενώ το intra_linkage στοχεύει με τον υπολογισμό του στα ποια πυκνά τμήματα του υπό εξέταση cluster.

Στην συνέχεια για να ολοκληρώσουμε τον υπολογισμό του Qgraph πρέπει να βρούμε τις τιμές για τα b και c.

interLink(C): Αυτή η μετρική για ένα ζευγάρι συστάδων ci,cj που ανήκουν στον ίδιο γράφο $G(c_i,c_j)$ υπολογίζεται λαμβάνοντας υπόψη τα a. degeneracy του G(ci,cj) και b. μέσω του επαναληπτικού υπολογισμού για κάθε σετ clusters του degeneracy core (c_i,c_j)

Οπότε ο τύπος για ένα σύνολο τμηματοποιήσεων είναι:

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

separation(C): Αυτή η μετρική η μετρική μας ορίζει πόσο καλά έχει γίνει η τμηματοποίηση σε συστάδες. Για να την υπολογίσουμε όμως θα πρέπει να βρούμε τις τιμές του interDensity μεταξύ των συστάδων μιας τμηματοποίησης και στη συνέχεια να βρούμε τι συμβαίνει με το InterConnectivity των ίδιων. Οι τιμές των Inter – Density και Inter Connectivity μας ορίζουν το ποσοστό των αναμενόμενων ακμών (edges) μέσα στις συστάδες που εμφανίζονται μέσα στον γράφο: **interDens(c_i,c_2) =** |E(ci,c2)| / |Vi| |Vj|

και

πόσο καλά είναι συνδεδεμένες οι συστάδες μέσα στον γράφο μέσω του τύπου:

$$InterCon(c_i, c_j) = \frac{InterDens(c_i, c_j)}{min\{dens(c_i), dens(c_j)\}}$$

αντίστοιχα.

Η τμηματοποίηση λοιπόν ενός γράφου C σε συστάδες είναι καλά ορισμένη όταν παρουσιάζει μικρή τιμή το inter – connectivity, έτσι ως separation(C) ορίζεται [18]:

$$Separation(C) = \frac{2}{m \cdot (m-1)} \sum_i \sum_j \frac{1}{InterCon(c_i, c_j)}$$

Τέλος ο τύπος του Qgraph ορίζεται [18]:

$$QGraph(C) = (intraLink(C) - interLink(C)) + Separation(C)$$

Τεχνική 14. CDS – Index

Ένας ακόμη δείκτης που έχει γνωστοποιηθεί [20] για την αξιόληγηση την τηματοποίησης ενός γράφου σε συστάδες αφορά αυτόν του CDS. Πλήθος προσεγγίσεων δεικτών αξιολόγησης χρησιμοποιούν κυρίως την

τοπολογική πληροφορία γύρω από το πόσο κοντά βρίσκονται οι κόμβοι μέσα στους γράφους (δηλαδή edge density).

Το density λοιπόν είναι μια από τις πιο συνηθισμένες μετρικές που μπορούν να μας δείξουν το κατά πόσο συνδεδεμένοι είναι κόμβοι μέσα στους γράφους [29].

$$\text{Dens}(G) = E / |V| (|V|-1)$$

Ωστόσο μελέτες έχουν δείξει ότι το density ενός γράφου δεν μπορεί να αξιολογήσει με το κατάλληλο τρόπο τη δομή ενός γράφου τμηματοποιημένου σε clusters καθώς δεν λαμβάνει πληροφορίες για το cohesion των clusters [19,20]. Υπάρχουν γράφοι που παρουσιάζουν κόμβους με υψηλή cohesive τιμή, γεγονός που κάνει δύσκολη την αξιολόγηση της τμηματοποίησης. Υπάρχουν αρκετές ενδείξεις γράφων όπου οι συστάδες τους παρουσιάζουν ίδια τιμή στο density αλλά διαφορετική δομή έτσι και διαφορετική τιμή στο cohesion [20].

Έτσι μόνο του το density σαν μετρική δεν μπορεί να δώσει με έγκυρο τρόπο το κατά πόσο καλά μπορεί να γίνει μία τμηματοποιήση σε συστάδες. Άρα, θα πρέπει να λαμβάνεται υπόψη και το cohesion.

Inter-Connectivity of clusters. Ο όρος του node connectivity σαν μετρική έχει εισαχθεί για τον υπολογισμό του cohesion των συστάδων. Με αυτόν το όρο ορίζεται ο ελάχιστος αριθμός των κόμβων που πρέπει να αφαιρέσουμε από ένα cluster, C_j ώστε να αποσυνδεθεί από αυτό. [20]

Το cohesion αποτελεί μία σημαντική μετρική για να καταλάβουμε πόσο κοντά είναι τα στοιχεία των υπό εξέταση συστάδων και κατά πόσον εύκολη είναι η τμηματοποίηση για αυτά. [21] Ωστόσο σημαντικό εξίσου είναι να αξιολογείται και η εξάρτηση που έχουν τα στοιχεία ενός γράφου μεταξύ τους και έτσι ο αριθμός των δεσμών μεταξύ των μελών θα πρέπει να είναι ο πιο ενδεικτικός ως κριτήριο για τον υπολογισμό του intra – cluster connectivity.

Ο υπολογισμό λοιπόν του Intra Connectivity (C_i) για μια τμηματοποιήση C_i σε m συστάδες ορίζεται ως εξής.

$$Intra-connectivity(C_i) = w_c \cdot \frac{1}{m} \sum_{i=1}^{m} Nconnect(c_j) + w_d \cdot \frac{1}{m} dens(c_i)$$

Όπου w_c και w_d είναι οι τιμές που προσδιορίζουν την βαρύτητα του cohesion και του density αντίστοιχα. Οι τιμές βαρύτητας αντανακλούν τη σημασία αυτών των δύο μετρήσεων.

Για τον υπολογισμό του δείκτη Cds πέρα από την τιμή του Intra Connectivity πρέπει να ορίσουμε και την τιμή του Separation για τον υπό εξέταση γράφο [19] που όπως έχουμε δει και σε προηγούμενη ενότητα με αυτό να δίνεται από τον τύπο:

$$Separation(C) = \frac{2}{m \cdot (m-1)} \sum_i \sum_j \frac{1}{InterCon(c_i, c_j)}$$

Τέλος ο τύπος του CDS ορίζεται ως εξής [19]:

$$CDS(C) = Intra - connectivity(C) + w_s \cdot Separation(C)$$

Η τιμή λοιπόν του CDS ιδανική παίρνει την μέγιστη τιμή της για μία τμηματοποίηση ενός γράφου G όταν οι δύο όροι που καθορίζουν το παραπάνω άθροισμα πάρουν τη μέγιστη τιμή τους. Συμπερασματικά βασισμένοι στο παραπάνω ορισμό την καλύτερη και ποιοτικότερη τμηματοποίηση ενός γράφου σε συστάδες θα την έχουμε όταν ο CDS για αυτή παίρνει τη μέγιστη τιμή του [29].

2.5 Σύνοψη

Σε αυτό το κεφάλαιο κάναμε μία επισκόπηση στους δείκτες που αφορούν την αξιολόγηση αποτελεσμάτων clustering σε γράφους.

Αρχικά έγινε μια περιγραφή γενικών δεικτών γράφων: Η πυκνότητα των ακμών (Edge Density) έχει έναν εύκολο υπολογισμό, χωρίς όμως να λαμβάνει υπόψη κάποια τιμή που αφορά το connectivity του γράφου. Ο δείκτης Cp υπολογίζει το πόσο συμπαγές είναι το αποτέλεσμα clustering που μελετάμε συναρτήσει της συνδεσιμότητας που παρουσιάζει αυτός αλλά και της τιμή Q για αυτό και παρουσιάστηκε μία βελτίωση του δείκτη αυτού, ο Cp*.

Σε δεύτερη φάση παρουσιάστηκαν δείκτες οι οποίοι εκφράζουν το πόσο καλά είναι συνδεδεμένο το cluster συναρτήσει του intra-cluster-connectivity. Οι δείκτες του Dunn και Davies δίνουν αποτελέσματα

αξιολόγησης λαμβάνοντας υπόψη τη διάμετρο αλλά και την απόσταση μεταξύ των Clusters.

Η μετρική Dunn έχει έναν εύκολο υπολογισμό χωρίς όμως να παρουσιάζει την ευρωστία του δείκτη Davies. [13]

Δείκτες κοπής και κάλυψης (Cut indices and Coverage index) λαμβάνουν υπόψη για την εξαγωγή τιμών τον αριθμό του inter αλλά και του intra – cluster – connectivity, χωρίς όμως να λαμβάνουν υπόψη το μέγεθος των clusters που εξετάζουν. Επιπλέον οι δείκτες Perfomance και MQ (ο οποίος αφορά την ποιότητα της μοντελοποίησης) προκύπτουν συναρτήσει του αριθμού των κόμβων και του αριθμού των inter και intra – cluster – edges. Με τον δείκτη MQ* να αποτελεί τη βελτιστοποίηση του δείκτη MQ.

Περιγράψαμε επίσης δείκτες που στηρίζονται στους κανόνες γειτνίασης των κόμβων που ορίζουν τα clusters (Neighborhood Connectivity). Έτσι μέσω του δείκτη Silhouette (GS) μπορούμε να ορίσουμε για αν ένας κόμβος έχει τοποθετηθεί σωστά σε ένα cluster ή όχι. Παράλληλα είδαμε και τον σταθμισμένο μέσο του δείκτη αυτού, GS*. Αναφορά έγινε και στον δείκτη C ο οποίος αφορά SMALL WOLRD GRAPHS.

Η περιγραφή κατάληξε με τους δείκτες Rand Statistic, Jaccard Coefficient, Folks & Mallows και Hubert οι οποίοι έχουν απλούς υπολογισμούς και αφορούν τμηματοποιήσεις οι οποίες περιέχουν κόμβους που ανήκουν σε περισσότερα από ένα Clusters.

Στη παρούσα ενότητα λοιπόν περιγράψαμε και σχολιάσαμε κάποιες από τις πιο διάσημες τεχνικές αξιολόγησης συσταδοποίησης γράφων. Όπως είδαμε το πρόβλημα του Cluster Validity έχει μελετηθεί ευρέως και υπάρχει ένας μεγάλος αριθμός προσεγγίσεων για την αξιολόγηση αποτελεσμάτων που έχουν προκύψει από κάποια μέθοδο τμηματοποίησης. Συμπεραίνουμε επίσης βάση των δεικτών που παρουσιάστηκαν παραπάνω ότι αυτοί μελετούν τα αποτελέσματα των τμηματοποιήσεων σε συστάδες (clusters) υπολογίζοντας την συνεκτικότητα και τη διαχωριστικότητα τους βασισμένοι στη διακύμανση αλλά και την πυκνότητα αυτών. Η πλειοψηφία των δεικτών που αφορούν το cluster validity βρίσκουν εφαρμογή στον Ευκλείδιο Χώρο ενώ υπάρχουν μόλις λίγοι αποκλειστικά για δεδομένα σε μορφή γράφων.

Έτσι λοιπόν με την εξέλιξη των απαιτήσεων στην ανάλυση των δεδομένων ήρθαμε αντιμέτωποι με το πρόβλημα του cluster validation καθώς οι απαιτήσεις για την εποπτεία αλλά και την ποίοτητα των αποτελεσμάτων μεθόδων clustering έχει αυξηθεί εκθετικά. Καθώς σε πολλές περιπτώσεις ανάλυσης δεδομένων σε μορφή γράφων δεν υπάρχει η δυνατότητα οπτικής παρουσίασης των απολεσμάτων τμηματοποίησης είναι δύσκολο να ορίσουμε για το πόσο καλά έχει γίνει η τμηματοποίηση των δεδομένων αυτών.

Μέρος 3. Πειραματική μελέτη δεικτών εγκυρότητας συστάδας για γράφους

3.1 Εισαγωγή

Μέσω αυτής της εργασίας ο έχουμε ως σκοπό να μελετήσουμε το πρόβλημα του cluster validity (αξιολόγησης αποτελεσμάτων) για δεδομένα που έχουν τη μορφή γράφων. Έτσι θα επιλέξουμε τέσσερα datasets σε μορφή γράφων. Στην συνέχεια αυτά θα τα τμηματοποιήσουμε σε clusters κάποια μέσω διαφορετικών αλγορίθμων για διαφορετικές τιμές εισόδων, έτσι ώστε να έχουμε ένα ικανό σύνολο αποτελεσμάτων προς εξέταση. Οι αλγόριθμοι που επιλέχθηκαν για να μας δώσουν διαφορετικά αποτελέσματα είναι ο Louvain και ο Spectral.

Σε δεύτερη φάση αφού ολοκληρώσαμε αρκετές τμηματοποιήσεις προχωρήσαμε στην υλοποίηση τριών μετρικών οι οποίες θα μας βοήθησαν να αξιολογήσουμε αυτά τα αποτελέσματα. Οι μετρικές αυτές είναι δείκτες αξιολόγησης αποτελεσμάτων τμηματοποίησης γράφων σε clusters. Στην παρούσα εργασία πέρα από την υλοποίηση τους θα δούμε πως αυτές λειτουργούν, τι αποτελέσματα δίνουν και πως συμπεριφέρονται σε σχέση με αυτά.

Τέλος για να υπάρχει κάποια σοβαρή εκτίμηση γύρω από τα αποτελέσματα που έχουμε παράξει αξιολογήσαμε τις τμηματοποιήσεις που έχουμε στα χέρια μας σε σχέση με τις ιδανικές τμηματοποιήσεις που ορίζονται για αυτά τα datasets από τη διαθέσιμη βιβλιογραφία (Ground Truths).

Το πρόβλημα του Cluster Validity έχει ευρέως μελετηθεί και υπάρχει ένας μεγάλος αριθμό δεικτών για την αξιολόγηση αποτελεσμάτων clustering. Όπως είδαμε στο μέρος δεύτερο της παρούσας εργασίας, οι δείκτες αυτοί υπολογίζουν τη συνεκτικότητα αλλά και το διαχωρισμό των clusters χρησιμοποιώντας κυρίως τη διακύμανση ή την πυκνότητα για την ανάλυση τους. Η πλειοψηφία των δεικτών αξιολόγησης βρίσκουν εφαρμογή στον Ευκλείδιο χώρο όπου και υπάρχουν λίγες εργασίες για δεδομένα γράφων.

Σε αυτό το μέρος λοιπόν θα μελετήσουμε πώς συμπεριφέρονται δύο δείκτες οι οποίοι δεν έχουν σημαντικές αναφορές στη βιβλιογραφία

ώστε να μπορέσουμε να προτείνουμε αν αυτοί μπορούν να μας αξιολογήσουν ορθά αποτελέσματα clustering σε γράφους.

3.2 Διαδικασία Α (Επιλογή Datasets σε μορφή Graphs)

Για την πραγμάτωση της πειραματικής διαδικασίας επιλέχθηκαν τέσσερα σύνολα δεδομένων τα οποία είναι τα παρακάτω:

1. AS:

Πρόκειται για ένα σύνολο δεδομένων που αποτυπώνει τις συνδέσεις μεταξύ Autonomus Systems οι οποίες όπου εμφανίζονται, αφορούν την ύπαρξη επιχειρηματικών αποφάσεων μεταξύ των AS ζευγαριών (Marián Boguná, Fragkiskos Papadopoulos, and Dmitri Krioukov. 2010. Sustaining the internet with hyperbolic mapping. Nature communications 1 (2010), 62Cora_Full

2. Cora:

Το σύνολο δεδομένων της Cora αποτελείται από 2708 επιστημονικές δημοσιεύσεις που ταξινομούνται σε μία από τις επτά τάξεις. Το δίκτυο παραπομπών αποτελείται από 5429 συνδέσμους. Κάθε δημοσίευση στο σύνολο δεδομένων περιγράφεται από ένα διάνυσμα λέξεων με τιμή 0/1 που υποδεικνύει την απουσία / παρουσία της αντίστοιχης λέξης από το λεξικό. Το λεξικό αποτελείται από 1433 μοναδικές λέξεις.

3. Email Eu:

Το δίκτυο δημιουργήθηκε χρησιμοποιώντας δεδομένα ηλεκτρονικού ταχυδρομείου από ένα μεγάλο ευρωπαϊκό ερευνητικό ίδρυμα. Για μια περίοδο από τον Οκτώβριο του 2003 έως τον Μάιο του 2005 (18 μήνες) έχουμε ανώνυμα στοιχεία σχετικά με όλα τα εισερχόμενα και εξερχόμενα μηνύματα ηλεκτρονικού ταχυδρομείου του ερευνητικού ιδρύματος. Για κάθε μήνυμα αποστολής ή λήψης e-mail γνωρίζουμε την ώρα, τον αποστολέα και τον παραλήπτη του μηνύματος ηλεκτρονικού ταχυδρομείου. Συνολικά έχουμε 3.038.531 μηνύματα ηλεκτρονικού ταχυδρομείου μεταξύ 287.755 διαφορετικών διευθύνσεων ηλεκτρονικού ταχυδρομείου. Σημειώστε ότι έχουμε ένα πλήρες γράφημα ηλεκτρονικού ταχυδρομείου μόνο για 1.258 διευθύνσεις ηλεκτρονικού ταχυδρομείου που προέρχονται από το

ίδρυμα έρευνας. Επιπλέον, υπάρχουν 34.203 διευθύνσεις ηλεκτρονικού ταχυδρομείου που στέλνουν και λαμβάνουν μηνύματα ηλεκτρονικού ταχυδρομείου εντός του εύρους του συνόλου δεδομένων μας. Όλες οι άλλες διευθύνσεις ηλεκτρονικού ταχυδρομείου είναι είτε μη υπάρχουσες, είτε με σφάλματα ή spam. Με δεδομένο ένα σύνολο μηνυμάτων ηλεκτρονικού ταχυδρομείου, κάθε κόμβος αντιστοιχεί σε μια διεύθυνση ηλεκτρονικού ταχυδρομείου. Δημιουργούμε μια κατευθυνόμενη άκρη μεταξύ των κόμβων i και j, αν έστειλαν τουλάχιστον ένα μήνυμα στο j.

4. Karate Wayne W Zachary:

Αυτό είναι το γνωστό και πολύ-χρησιμοποιημένο δίκτυο Zachary karate club. Τα στοιχεία συλλέχθηκαν από τα μέλη ενός πανεπιστημιακού συλλόγου καράτε από τον Wayne Zachary το 1977. Κάθε κόμβος αντιπροσωπεύει ένα μέλος του συλλόγου και κάθε άκρο αντιπροσωπεύει μια ισοπαλία μεταξύ δύο μελών του συλλόγου. Το δίκτυο είναι μη κατευθυνόμενο.

Τα παραπάνω σύνολα δεδομένων χρησιμοποιούνται κατά κόρων στην βιβλιογραφία και είναι γνωστά για όλα τα ground truths, τα οποία είναι αναγκαία στην μελέτη μας για την εξαγωγή συμπερασμάτων στο τέταρτο κεφάλαιο της εργασίας μας.

*links από τα οποία βρέθηκαν τα σύνολα δεδομένων είναι τα παρακάτω:

https://relational.fit.cvut.cz/dataset/CORA

http://konect.uni-koblenz.de/networks/ucidata-zachary

http://snap.stanford.edu/data/email-EuAll.html

3.3 Διαδικασία Β. (Louvain and Spectral)

Συνεχίζοντας το πειραματικό κομμάτι της παρούσας εργασίας επιλέξαμε δυο διαφορετικούς αλγορίθμους οι οποίοι μας έδωσαν διαφορετικές τμηματοποιήσεις για τα 4 παραπάνω σύνολα δεδομένων.

Αλγόριθμος Spectral

Ο πρώτος αλγόριθμος που επιλέχθηκε είναι ο Spectral. Πρόκειται για έναν γνωστό αλγόριθμο ο οποίος χρησιμοποιείται κατά κόρων στο machine learning. Η διαδικασία υλοποίησης του είναι απλή και έχει εφαρμοστεί από πολλές ερευνητικές ομάδες καθώς μπορείς αποτελεσματικά αξιοποιώντας μεθόδους γραμμικής άλγεβρας να εξάγεις αποτελέσματα συσταδοποίησης δεδομένων σε μορφή γράφων.

Για αυτή τη μέθοδο τμηματοποίησης δεδομένων σε clusters, το μέτρο που λαμβάνεται υπόψη για τη οργάνωση το γράφων σε υπο-γράφους είναι η συγγένεια (affinity) και όχι η θέση των στοιχείων που περιγράφεται από την απόλυτη θέση (absolute location. i.e. k-means). Μέσω λοιπόν της συγγένειας αυτής, ο αλγόριθμος επιλέγει ποια σημεία (nodes) θα δημιουργήσουν μια συστάδα (cluster). Αυτή η μέθοδος θεωρείται κατάλληλη για δεδομένα τα οποία έχουν τη μορφή γράφων τα οποία απεικονίζονται με περίπλοκα σχήματα [25].

Τυπικά το σύνολο δεδομένων που αυτός ο αλγόριθμος διαβάζει αποτελείται από ένα σύνολο από γραμμές όπου παρουσιάζονται τα στοιχεία με τις συνδέσεις τους (Laplacian Matrix). Στην πράξη ο Spectral είναι ένας πολύ χρήσιμος αλγόριθμος όταν η δομή των δεδομένων σε συστάδες παρουσιάζει μία κυρτότητα (πχ συμπλέγματα / συστάδες που είναι ένθετα σε κυκλικό επίπεδο) [25].

Κάθε αλγόριθμος clustering για να ξεκινήσει να μας δίνει αποτελέσματα σε συστάδες δέχεται μία τιμή εισόδου. Αυτή η τιμή – παράμετρος στον συγκεκριμένο αλγόριθμο είναι ο αριθμός των Clusters που θέλουμε να μας δώσει. [25,26]

Ένα παράδειγμα υλοποίησης του αλγόριθμου αυτού φαίνεται στην παρακάτω εικόνα:

```
>>> from sklearn.cluster import SpectralClustering
>>> import numpy as np
>>> X = np.array([[1, 1], [2, 1], [1, 0],
...               [4, 7], [3, 5], [3, 6]])
>>> clustering = SpectralClustering(n_clusters=2,
...         assign_labels="discretize",
...         random_state=0).fit(X)
>>> clustering.labels_
array([1, 1, 1, 0, 0, 0])
>>> clustering
SpectralClustering(assign_labels='discretize', n_clusters=2,
    random_state=0)
```

Εικόνα 11

Και οι μέθοδοι που χρησιμοποιούνται φαίνονται εδώ:

Methods

fit(self, X[, y])	Perform spectral clustering from features, or affinity matrix.
fit_predict(self, X[, y])	Perform spectral clustering from features, or affinity matrix, and return cluster labels.
get_params(self[, deep])	Get parameters for this estimator.
set_params(self, **params)	Set the parameters of this estimator.

Υλοποιημένα σε Python

Εικόνα 12

Η λογική του Spectral βασίζεται στα εξής βήματα:

Spectral Clustering

Three basic stages:

1. Construct a matrix representation of the dataset.

2. Compute eigenvalues and eigenvectors of the matrix
& Map each point to a lower-dimensional representation
based on one or more eigenvectors.

3. Assign points to two or more clusters,
based on the new representation. (Grouping)

Αλγόριθμος Louvain

Η μέθοδος Louvain βρίσκει εφαρμογή και αυτή στον τομέα της συσταδοποίησης δεδομένων σε γράφους και αποτελεί έναν αλγόριθμο που εξυπηρετεί γενικά τεχνικές Community Detection. Αυτή η μέθοδος για να δώσει αποτελέσματα ακολουθεί μία διαδικασία αξιολόγησης της σύνδεσης που παρουσιάζουν οι κόμβοι που αποτελούν τα υπό εξέταση σύνολα δεδομένων.

Ο Louvain λοιπόν καθώς εκτελείται για να μας δώσει αποτελέσματα θα πρέπει ως τιμή εισόδου να του δώσει ο προγραμματιστής έναν αριθμό που αφορά το resolution. Στη συνέχεια αυτός θα εκτελέσει μια συνάρτηση η οποία θα του δώσει την καλύτερη τμηματοποίηση (based on modularity) σύμφωνα με την τιμή εισόδου που του δόθηκε και στην συνέχεια θα μας δώσει σαν έξοδο ένα νέο σύνολο δεδομένων το οποίο θα έχει υποστεί συσταδοποίηση. [28] Ένας τρόπος υλοποίησης σε python φαίνεται στην παρακάτω εικόνα:

```
import community
import networkx as nx
import matplotlib.pyplot as plt

# Replace this with your networkx graph loading depending on your format !
G = nx.erdos_renyi_graph(30, 0.05)

#first compute the best partition
partition = community.best_partition(G)

#drawing
size = float(len(set(partition.values())))
pos = nx.spring_layout(G)
count = 0.
for com in set(partition.values()) :
    count = count + 1.
    list_nodes = [nodes for nodes in partition.keys()
                                if partition[nodes] == com]
    nx.draw_networkx_nodes(G, pos, list_nodes, node_size = 20,
                                node_color = str(count / size))


nx.draw_networkx_edges(G, pos, alpha=0.5)
plt.show()
```

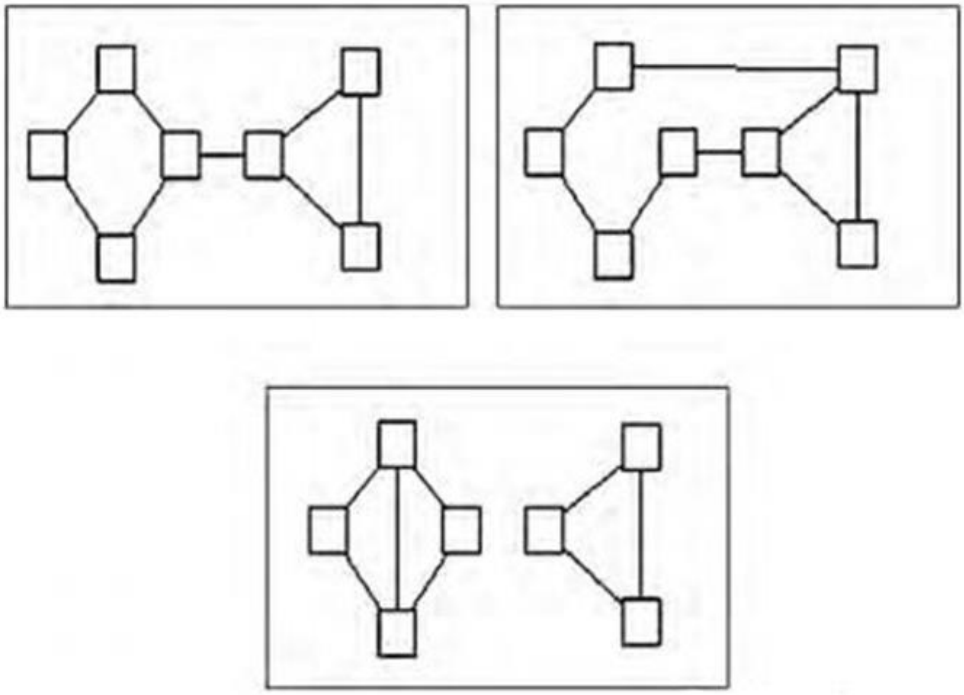
Εικόνα 13

**** Οι εικόνες 11, 12 και 13 προέρχονται από διαθέσιμα αρχεία κώδικα στο github.com*

Τα Datasets a. και b. μας χρησιμοποιήθηκαν ώστε να μας δώσουν 5 διαφορετικές τμηματοποιήσεις από τον αλγόριθμο Spectral. Για να έχουμε όμως ένα ικανό δείγμα προς μελέτη χρησιμοποιήσαμε τον αλγόριθμο Louvain ώστε να έχουμε ένα σύνολο με περισσότερες τμηματοποιήσεις για τα datasets c. και d. Ο λόγος που έγινε αυτό είναι για να δούμε πως συμπεριφέρονται οι δείκτες αξιολόγησης και για περισσότερες τμηματοποιήσεις και κατά πόσων μεταβάλλονται οι τιμές τους για ένα μεγαλύτερο σύνολο αξιολογήσεων ως προς το ίδιο dataset.

3.4 Δείκτες αξιολόγησης: Υλοποίηση Modularity Index, QGraph & Cds

Οι δείκτες που επιλέξαμε για την εκπόνηση του πειραματικού τμήματος της εργασίας είναι οι Modularity, Q Graph & Cds (αναφορά στο Μέρος 2. Δείκτες Εγκυρότητας Συστάδας Για Γράφους: Τεχνικές & Εργαλεία). Ο πρώτος δείκτης είναι ένα δείκτης που έχει χρησιμοποιηθεί πάρα πολύ στη σχετική βιβλιογραφία αλλά και σε σχετικές μελέτες. Υπάρχουν ενδείξεις ότι δεν μπορεί να δουλέψει σωστά για ένα όγκο διαφορετικών αποτελεσμάτων συσταδοποίησης γράφων ώστε να μας δώσει την καλύτερη τμηματοποίηση. Όμως οι άλλοι δύο δείκτες παρουσιάζουν μία πολυπλοκότητα στου υπολογισμούς τους καθώς πέρα από τη μετρική του density ενός γράφου για τον υπολογισμό του cohesion εισάγουν τους όρους Graph Degenenarecy και Node Connectivity. Η παρακάτω εικόνα μας δείχνει συστάδες γράφων που ενώ παρουσιάζουν το ίδιο density, εμφανίζουν διαφορετική τιμή για το cohesion.



3.5 Αποτελέσματα

Στην προηγούμενη ενότητα κάναμε μια περιγραφή για τον τρόπο που προσεγγίσαμε το πειραματικό μέρος ως προς τις τμηματοποιήσεις συνόλων δεδομένων σε συστάδες (μέσω Louvain και Spectral). Στη συνέχεια αφού δημιουργήθηκε ένα σύνολο δεδομένων σε μορφή clusters, αξιολογήσαμε τα αποτελέσματα των τμηματοποιήσεων μέσω των δεικτών που αναφέρουμε στην ενότητα 3.4 (Qgraph, Cds, Modularity)

Όμως για να εξάγουμε συμπεράσματα, για τα αποτελέσματα που έχουμε ύστερα από τις τμηματοποιήσεις καθώς και για τις τιμές που μας δίνουν οι μετρικές αξιολόγησης αυτών, χρησιμοποιήσαμε τις ιδανικές τμηματοποιήσεις των υπό εξέταση συνόλων δεδομένων ως «δείκτες» για να δούμε πως συμπεριφέρονται οι Qgraph, Modularity και CDS όταν τα υπό εξέταση σύνολα δεδομένων έχουν την καλύτερη τμηματοποιήση τους. Για να γίνει αυτό χρησιμοποιήσαμε τον δείκτη Rand Index για όλες τις τμηματοποιήσεις μας ώστε να βρούμε ποιες από αυτές είναι κοντά στα Ground Truths των συνόλων δεδομένων που μελετήσαμε [23, 24]. Θυμίζουμε ότι τα σύνολα δεδομένων που επιλέξαμε έχουν γνωστά Ground Truths.

Spectral Results

SPECTRAL RESULTS

Τμηματοποιήσεις	Dataset1 (as)				
	No. of Clusters	Modularity	Qgraph	CDS	Rand Index
Spectral Alternate 1	30	0,136153851	0,139767442	NA	0,009234
Spectral Alternate 2	20	0,122860798	0,243161582	NA	0,015613
Spectral Alternate 3	160	0,4506	0,2323	0,3434	0,67
Spectral Alternate 4	40	0,14007852	0,188190888	0,39984601	0,006841
Spectral Alternate 5	176	0,630687	0,358266	0,36264	1

Τμηματοποιήσεις	Dataset2 (cora)				
	No. of Clusters	Modularity	Qgraph	CDS	Rand Index
Spectral Alternate 1	50	0,093647815	0,5482196	1,48455965	0,0029094
Spectral Alternate 2	60	0,08667761	0,5598935	1,5197142	0,002899
Spectral Alternate 3	70	0,5165296	0,87235	1,22608	1
Spectral Alternate 4	30	0,1040861	0,5266509	1,4345247	0,00314994
Spectral Alternate 5	20	0,11346406	0,4784077	1,34478055	0,00294739

Τμηματοποιήσεις	Dataset3 (email)				
	No. of Clusters	Modularity	Qgraph	CDS	Rand Index
Spectral Alternate 1	30	0,0948415	0,89333056	NA	0,0499776
Spectral Alternate 2	40	0,08980213	0,730949	1,2639535	0,050529732
Spectral Alternate 3	42	0,31376	3,5596	2,23505	1
Spectral Alternate 4	25	0,09401263	1,014509	NA	0,04934278
Spectral Alternate 5	20	0,086471926	0,8497337	1,345841	0,0345829

Τμηματοποιήσεις	Dataset4 (karate)				
	No. of Clusters	Modularity	Qgraph	CDS	Rand Index
Spectral Alternate 1	2	0,37146614	2,0456259	4,03769	1
Spectral Alternate 2	3	0,30761	1,538617	3,04642905	0,590285
Spectral Alternate 3	4	0,26520381	1,438331	2,956363	0,464591
Spectral Alternate 4	5	0,1687	1,313309	2,5221	0,4166
Spectral Alternate 5	6	NA	1,2402938	3,0687948	0,239161

Όπως φαίνεται σύμφωνα με τον παραπάνω πίνακα ο οποίος μας έδωσε τμηματοποιήσεις για όλα τα σύνολα δεδομένων μέσω της τεχνικής συσταδοποίησης του Spectral αλγόριθμου. Υπήρξαν και στις τέσσερις περιπτώσεις αποτελέσματα σε clusters όπου ταυτίστηκαν με τις ιδανικές τμηματοποιήσεις που ορίζουν τα ground truths των συγκεκριμένων datasets. (Σημείωση: Αποτελέσματα με κόκκινο χρώμα)

Σύμφωνα με τα αποτελέσματα αυτά παρατηρούμε ότι: βασιζόμενοι λοιπόν στην υπόθεση που ορίζεται από τους ορισμούς των τριών αυτών δεικτών παρατηρούμε ότι οι δείκτες που τις ικανοποιούν είναι ο Modularity και ο Qgraph. Καθώς αυτοί παίρνουν τη μέγιστη τιμή τους εκεί που οι τμηματοποιήσεις παίρνουν τιμή στο Rand Index ίση με 1.

Αντίθετα παρατηρούμε ότι ο δείκτης CDS παίρνει τη μέγιστη τιμή του εκεί που δύο από τα τέσσερα παραπάνω dataset είναι κοντά στην ιδανική τμηματοποίηση.

Για να ισχυροποιήσουμε τα συμπεράσματά μας σχετικά με τη συμπεριφορά των δεικτών γύρω από διαφορετικές τμηματοποιήσεις προχωρήσαμε τη πειραματική μας διαδικασία με το να τμηματοποιήσουμε σε clusters δύο από τα τέσσερα σύνολα δεδομένων με τον αλγόριθμο του Louvain. Έτσι θα προκύψουν άλλες 5 τμηματοποιήσεις για τα Dataset των Email και Karate αντίστοιχα.

Πήραμε λοιπόν επιπλέον άλλες 10 τμηματοποιήσεις για τα δύο τελευταία σύνολα δεδομένων και προχωρήσαμε στην αξιολόγηση αυτών που προέκυψαν.

Louvain Results

Τμηματοποιήσεις	Dataset3 (email)					
	Τιμή Εισόδου	No. of Clusters	Modularity	Qgraph	CDS	Rand Index
Louvain Alternate 1	0,3	68	0,362496	2,462629	2,177454	0,597715
Louvain Alternate 2	0,5	44	0,405256	2,314121	1,390808	0,50999
Louvain Alternate 3	0,8	31	0,424741	2,089922	0,7976697	0,451449
Louvain Alternate 4	1	27	0,4137053	1,900004	0,5582149	0,33506
Louvain Alternate 5	1,2	12	0,2254273	1,568009	0,5363161	0,23781

Τμηματοποιήσεις	Dataset4 (karate)					
	Τιμή Εισόδου	No. of Clusters	Modularity	Qgraph	CDS	Rand Index
Louvain Alternate 1	0,3	12	0,225427	1,1673335	3,603804	0,13995
Louvain Alternate 2	0,5	7	0,344838	1,7120943	3,927618	0,249171
Louvain Alternate 3	0,8	4	0,4188034	1,9529868	3,9263814	0,392238
Louvain Alternate 4	1	4	0,419789	1,966381	3,6572192	0,464591
Louvain Alternate 5	1,2	4	0,419789	1,966381	3,6572192	0,46459

Παρατηρούμε ότι από τις τμηματοποιήσεις που προέκυψαν από τα δύο σύνολα δεδομένων δεν μπορούμε να βγάλουμε ασφαλή συμπεράσματα καθώς όλες είναι μακριά από το Ground Truth όπως μας δείχνει ο δείκτης Rand Index (Rand Index: Όταν το αποτέλεσμά του προσεγγίζει την τιμή 1 τότε είμαστε κοντά στην ιδανική τμηματοποιήση σε συστάδες).

Όμως ενδιαφέρον σε πρώτη φάση παρουσιάζεται στο εξής:

Ότι αφορά το πρώτο dataset που βλέπουμε παραπάνω, παρατηρούμε ότι όσο το Rand Index πλησιάζει στη τιμή 1, οι τιμές του Qgraph και Cds τείνουν να πάρουν τη μέγιστη τιμή τους για τις συγκεκριμένες τμηματοποιήσεις σε συστάδες.

Αυτό που πρέπει όμως να παρατηρήσουμε και έχει ενδιαφέρον, είναι ότι στον δείκτη Modularity βλέπουμε την τιμή του να μεγαλώνει στις τμηματοποιήσεις που προέκυψαν από τον Louvain οι οποίες όμως βρίσκονται μακριά από τις ιδανικές ξεπερνώντας την αντίστοιχη τιμή που παίρνει ο δείκτης όταν προκύπτει η ιδανική τμηματοποίηση μέσω του Spectral. [Πίνακας Spectral Results] Άρα ο δείκτης αυτός δεν δούλεψε σωστά σε ένα σύνολο 10 τμηματοποιήσεων για το dataset αυτό.

Σύμφωνα με την θεωρία εκεί που το σύνολο δεδομένων Email βρίσκει την ιδανική τμηματοποίηση σε συστάδες μέσω του Spectral, ο δείκτης Qgraph παίρνει τη μέγιστη τιμή του 3,55 όμοια και ο CDS 2,23. Δεν συμβαίνει όμως το ίδιο και για το Modularity καθώς σύμφωνα με τις τμηματοποιήσεις που προέκυψαν από τον Louvain βρέθηκε τιμή μέγιστη για αυτόν τον δείκτη όταν σε τμηματοποίηση μακριά από το Ground Truth του εν λόγω dataset.

Τέλος για το σύνολο δεδομένων που αφορά το Karate, παρατηρούμε ότι ο δείκτης Qgraph με τον CDS δουλεύει κανονικά και εξυπηρετούν τα αποτελέσματά τους την υπόθεση των ορισμών τους.

Μέρος 4. Συμπεράσματα / Μελλοντική Εργασία

4.1 Συμπεράσματα

Ο σκοπός της παρούσας διπλωματικής εργασίας ήταν η μελέτη του προβλήματος αξιολόγησης αποτελεσμάτων συσταδοποίησης από σύνολα δεδομένων σε μορφή γράφων. Έτσι επιλέξαμε δύο διαφορετικούς τρόπους τμηματοποίησης για τέτοιου είδους δεδομένα. Στη συνέχεια είδαμε πόσο κοντά είναι τα αποτελέσματα που προέκυψαν στις ιδανικές τμηματοποιήσεις και παρατηρήσαμε πώς συμπεριφέρονται τρεις διαφορετικοί δείκτες αξιολόγησης γύρω από αυτά και κατά πόσο αυτοί ικανοποιούν με τις τιμές που μας έδωσαν τους ορισμούς μέσω των οποίων αυτοί περιγράφονται.

Στο τρίτο και πειραματικό λοιπόν μέρος της εργασίας μελετήσαμε πως συμπεριφέρονται τρείς διαφορετικοί δείκτες αξιολόγησης σε ένα σύνολο διαφορετικών τμηματοποιήσεων. Προέκυψε ότι μόνον ένας κατάφερε να ικανοποιήσει τον ορισμό του στο σύνολο των πειραμάτων μας καθώς έπαιρνε τη μεγαλύτερη τιμή του εκεί που είτε βρισκόμασταν κοντά στην ιδανική τμηματοποίηση είτε είχαμε ακριβώς την ιδανική (καθώς υπήρξε πλήρης ταύτιση με τα Ground Truths) και αυτός ήταν ο Qgraph. Στα αποτελέσματα συσταδοποίησης που προέκυψαν από τον αλγόριθμο Louvain παρατηρήσαμε ότι όσο περισσότερο πλησίαζε ο δείκτης Rand Index τη τιμή 1, η τιμή του Qgraph είχε την τάση να μεγιστοποιείται. Το ίδιο συμβαίνει για τα συγκεκριμένα σύνολα δεδομένων και για τον CDS, όχι όμως και για τον Modularity.

Ιδιαίτερο ενδιαφέρον παρουσιάζουν τα αποτελέσματα που πήραμε από τις Spectral τμηματοποιήσεις μας. Καθώς καταφέραμε μέσα από ένα σύνολο διαφορετικών επαναλήψεων να πετύχουμε το Ground Truth και να πάρουμε Rand Index ίσο με 1. Ο δείκτης που μας δούλεψε και τις τέσσερις φορές ήταν ο Qgraph καθώς κατάφερε να ικανοποιήσει τον ορισμό του και να πάρει τη μέγιστη τιμή του εκεί όπου προέκυψε η καλύτερη τμηματοποίηση. Με τους άλλους δύο δείκτες modularity και cds να μη δουλεύουν σωστά αντίστοιχα.

Έτσι συμπεραίνουμε ότι ο δείκτης Q-graph δούλεψε σωστά, όπως και ορίζεται από τον τύπο υπολογισμού του. Ο δείκτης αυτός αξιολογεί μία τμηματοποίηση σε συστάδες ενός συνόλου δεδομένων σε μορφή γράφου βάση του separation, του density μεταξύ των clusters, με την

μέτρηση αυτή να ισχυροποιείται καθώς σε αυτόν τον τύπο εισάγεται και η μετρική του intra και inter linkeage μέσω των οποίων ορίζονται καλύτερα οι περιοχές με τους κόμβους (nodes) με τις πιο ισχυρές συνδέσεις. Αυτό δεν συμβαίνει με τους άλλους δύο δείκτες, καθώς βασίζονται σε πιο γενικούς υπολογισμούς. Έτσι ως καταλληλότερος δείκτης αξιολόγησης αποτελεσμάτων προτείνεται ο Qgraph καθώς σε μία ποικιλία αποτελεσμάτων κατάφερε να ικανοποιήσει την υπόθεση του ορισμού.

Ο δείκτης του Qgraph βασίζεται στην έννοια του Graph Degeneracy ώστε να υπολογίσει το Connectivity ανάμεσα στους κόμβους των γράφων. Η έννοια αυτή αξιολογεί και ανιχνεύει τις πιο συνεκτικές ομάδες (groups) μέσα στους γράφους.

Αντίθετα ο δείκτης CDS βασίζεται στη έννοια του cluster cohesion που μετριέται με το node connectivity και στο density του γράφου.

Συμπερασματικά είδαμε ότι σε ένα σύνολο διαφορετικών αποτελεσμάτων η έννοια του Degeneracy φαίνεται να δουλεύει καλύτερα ως μέτρο συνδεσιμότητας. Έτσι και προτείνετε από πλευράς μας ως η ικανότερη για αξιολόγηση αποτελεσμάτων συσταδοποίησης σε μορφή γράφων.

4.2 Μελλοντική εργασία

Μέσω αυτής της εργασίας μελετήσαμε το πρόβλημα του Cluster Validity. Περιγράψαμε και μελετήσαμε την πορεία των προσεγγίσεων μέσω διάφορων αναφορών στην υπάρχουσα βιβλιογραφία. Πλέον οι δείκτες αξιολόγησης τμηματοποίησης δεδομένων σε μορφή γράφων έχουν εξελιχθεί και αξιολογούν τα αποτελέσματα καλύτερα βασιζόμενοι περισσότερο στη δομή τους και όχι σε στατιστικές προσεγγίσεις.

Προτείνουμε λοιπόν σε συνέχεια της παρούσας εργασίας να γίνει μια σειρά από τμηματοποιήσεις για διαφορετικά σύνολα δεδομένων σε μορφή γράφων μέσω διαφορετικών τεχνικών (πχ Modularity Based) και να εξεταστούν τα αποτελέσματα που προκύπτουν σε συστάδες βάση των δυο μετρικών CDS και QGraph ώστε να εγκρίνουμε την καταλληλόλητά τους.

Τέλος, είναι καλό για τους ερευνητές που ασχολούνται με την πειραματική μελέτη δεικτών εγκυρότητας συστάδας για γράφους να αρχίζουν να χρησιμοποιούν μετρικές αξιολόγησης που για το

υπολογισμό τους χρησιμοποιούν σύνθετα μέτρα τα οποία βασίζουν τον υπολογισμό τους στην δομή των υπό εξέταση συστάδων όπως για παράδειγμα και αυτή του Q-Graph.

Μέρος 5. Βιβλιογραφία / Αναφορές

1. https://www.hackerearth.com/practice/algorithms/graphs/graph-representation/tutorial/
2. https://www.geeksforgeeks.org/graph-data-structure-and-algorithms/
3. Lecture #2: Directed Graphs - Transition Matrices (http://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture2/lecture2.html)
4. Community detection in networks: A user guide (Dated: November 4, 2016)
5. Community Detection in Networks with Node Attributes / Jaewon Yang, Julian McAuley, Jure Leskovec (28 Jan 2014)
6. Network Community Detection: A Review and Visual Survey
7. https://www.researchgate.net/publication/259707543_Community_Detection_and_Visualization_in_Social_Networks_Integrating_Structural_and_Semantic_Information
8. GRAPH CLUSTERING by FLOW SIMULATION [17,28 σελ]
9. Survey Graph clustering / Satu Elisa Schaeffer∗ / Laboratory for Theoretical Computer Science, Helsinki University of Technology TKK, P.O. Box 5400, FI-02015 TKK, Finland [38 – 48]
10. Cluster Validity Measurement Techniques / CSABA LEGÁNY, SÁNDOR JUHÁSZ AND ATTILA BABOS
11. S. Theodoridis and K. Koutroubas: Pattern Recognition, Academic Press, 1999
12. Dunn J. “Well separated clusters and optimal fuzzy partitions” J. Cybernetics vol.4 1974, σελίδες 95-104
13. Cluster Validity Indices For Graph Partioning, Francois Boutin, Mountaz Hascoet
14. A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters
15. Performance Evaluation of the Silhouette Index
16. Van Dongen S., Performance critera for graph clustering and Markov cluster experiments. Technical Report INS – R0012, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam 2000.
17. Chiricota Y. , Jordan F., Softoware Component Capture using Graph Clustering. IWPC 2003
18. Qgraph: A Quality Assessment index for Graph Clustering / Maria Halkidi
19. Clustering by Vertex Density in a Graph / Chapter · January 2004
20. Evaluating the quality of graph clusters: CDS
21. Graph Cohesion
22. Steve Borgatti MGT 780 / http://www.analytictech.com/
23. Rand Index function (clustering performance evaluation) (https://stackoverflow.com/questions/49586742/rand-index-function-clustering-performance-evaluation)
24. Rand index – Wikipedia

25. Spectral Clustering Algorithm Implemented From Scratch (towardsdatascience.com)
26. GRAPH PARTITIONING and SPECTRAL CLUSTERING Εργαστήριο Ηλεκτρονικής (ELLAB) Τμήμα Φυσικής, Παν. Πατρών Θεοχαράτος Χ.
27. Community detection algorithms (neo4j.com) 6.1. The Louvain algorithm
28. Community detection for NetworkX's documentation (https://python-louvain.readthedocs.io/en/latest/)
29. CDS - Evaluating the quality of graph clusters / Maria Halkidi

Ευχαριστίες

Με την περάτωση της παρούσας διπλωματικής εργασίας θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια μου, κα Χαλκίδη για την υποστήριξή της, αλλά και για τις χρήσιμες υποδείξεις της καθ' όλη τη διάρκεια της συνεργασίας μας καθώς συνέβαλε τα μέγιστα για την εκπόνηση της.

Τιγγινάγκας Αλέξανδρος.