ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ στα

ΠΟΛΥΠΛΟΚΑ ΣΥΣΤΗΜΑΤΑ και ΔΙΚΤΥΑ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΓΕΩΛΟΓΙΑΣ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τίτλος Εργασίας

Μέθοδοι Συσταδοποίησης για τον Εντοπισμό Κοινοτήτων σε Πολύπλοκα Δίκτυα

Clustering Methods for Community Detection in Complex Networks

Κάτανος Λουκάς

ΕΠΙΒΛΕΠΩΝ: Κουγιουμτζής Δημήτρης. Καθηγητής Α.Π.Θ.

Θεσσαλονίκη, Δεκέμβριος 2018

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ στα

ΠΟΛΥΠΛΟΚΑ ΣΥΣΤΗΜΑΤΑ και ΔΙΚΤΥΑ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΓΕΩΛΟΓΙΑΣ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τίτλος Εργασίας

Μέθοδοι Συσταδοποίησης για τον Εντοπισμό Κοινοτήτων σε Πολύπλοκα Δίκτυα

Clustering Methods for Community Detection in Complex Networks

Κάτανος Λουκάς

ΕΠΙΒΛΕΠΩΝ: Κουγιουμτζής Δημήτρης
Καθηγητής Α.Π.Θ.

Εγκρίθηκε από την Τριμελή Εξεταστική Επιτροπή την 21η Δεκεμβρίου 2018.

……	……	……
Δ. Κουγιουμτζής	Ι. Αντωνίου	Β. Καραγιάννης
Καθηγητής Α.Π.Θ.	Καθηγητής Α.Π.Θ.	Δρ. Μαθηματικών

Θεσσαλονίκη, Δεκέμβριος 2018

……..

Λουκάς Α. Κάτανος
Πτυχιούχος Μαθηματικός Α.Π.Θ.

Περίληψη

Τα δίκτυα χρησιμοποιούνται ευρέως για να περιγράφουν πολύπλοκες σχέσεις μεταξύ οντοτήτων και η δομή κοινοτήτων είναι ένα σημαντικό χαρακτηριστικό που εμφανίζεται κυρίως στην μελέτη των πραγματικών δικτύων.

Ο σκοπός της εργασίας είναι η ανάπτυξη μεθοδολογίας για την ανίχνευση κοινοτήτων σε πολύπλοκα δίκτυα με την μέθοδο συσταδοποίησης k-Means. Ο αλγόριθμος k-Means βασίζεται σε αποστάσεις σημείων ορισμένων σε ένα μετρικό χώρο, όπου αρχικά επιλέγονται τυχαία k από αυτά ως τα κέντρα των συστάδων. Συνεπώς, δημιουργείται το πρόβλημα του εντοπισμού των αρχικών κέντρων (initial seeds) αλλά και της επιλογής του αριθμού αυτών, ώστε η λύση του αλγορίθμου να είναι η βέλτιστη.

Στην παρούσα εργασία απαντώνται τα συγκεκριμένα προβλήματα και προτείνεται ένας τροποποιημένος αλγόριθμος K-Means ο οποίος μετατρέπει αρχικά το δίκτυο σε μετρικό χώρο και στην συνέχεια για την επιλογή των k αρχικών κέντρων κάνει την παραδοχή ότι τα αρχικά κέντρα θα πρέπει να είναι σημαντικοί κόμβοι του δικτύου οι οποίοι θα απέχουν όσο το δυνατόν περισσότερο μεταξύ τους. Ως εκ τούτου, οι κόμβοι ταξινομούνται ως προς δύο μεταβλητές, την Pagerank κεντρικότητα και την απόσταση από τους άλλους κόμβους με μεγαλύτερη Pagerank. Συνεπώς, η Pagerank ενισχύεται από την από την απόσταση καθότι οι τιμές της ενδέχεται να μην έχουν μεγάλη διακύμανση.

Από την ταξινόμηση των κόμβων προκύπτει ένα δισδιάστατο διάγραμμα απόφασης από το οποίο επιλέγεται ο αριθμός k των συστάδων αλλά και τα αρχικά κέντρα που θα χρησιμοποιήσει ο αλγόριθμος για τη συσταδοποίηση. Αφού, επιλεγούν τα κέντρα, ο αλγόριθμος εξάγει τις κοινότητες του δικτύου οι οποίες αξιολογούνται ποσοτικά από τον δείκτη Silhouette.

Για την αξιολόγηση του αλγορίθμου πραγματοποιήθηκε σύγκριση αυτού με έναν κλασικό αλγόριθμο εύρεσης κοινοτήτων σε έξι τεχνητά δίκτυα που η δομή κοινοτήτων ήταν γνωστή εκ των προτέρων και σε ένα πραγματικό δίκτυο.

Τέλος, ο προτεινόμενος αλγόριθμος εφαρμόστηκε σε ένα πραγματικό δίκτυο που δημιουργήθηκε από αναφορές μεταξύ λογαριασμών στο Twitter και εντοπίστηκε σε ποια κοινότητα ανήκουν οι πιο σημαντικοί κόμβοι του.

Λέξεις κλειδιά: Complex Networks, Community Detection, Clustering, K-means, Page Rank, Silhouette, Twitter Network.

Abstract

Community detection is a very important problem in social network analysis. Classical clustering approach, K-means, has been shown to be very efficient to detect communities in networks. However, K-means is quite sensitive to the initial centroids or seeds, especially when it is used to detect communities. Additionally, in K-means, the number of communities should be specified in advance. Till now, it is still an open problem on how to select initial seeds and how to determine the number of communities.

To address this problem in this study, a modification of the classical K-Means algorithm is proposed, which selects the top-K nodes with the highest rank centrality as the initial seeds, and updates these seeds by using an iterative technique like K-means. First, based on the fact that good initial seeds usually have high importance and are dispersedly located in a network, a modified PageRank centrality is proposed to evaluate the importance of each node, and drew a decision graph to depict the importance and the dispersion of nodes. Then, the initial seeds and the number of communities are selected from the decision graph actively and intuitively as the 'start' parameter of K-means.

The proposed algorithm is compared with a classical algorithm for community detection on artificial networks whose community structure is already known, as well as a real network. Then, the outputs are evaluated by the Silhouette index and the relevant diagram.

Finally, the K-Means algorithm is applied in a real Twitter network formed by references among Twitter accounts-ids, detecting the important nodes as well as the community that they belong to.

Key Words: Complex Networks, Community Detection, Clustering, K-means, Page Rank, Silhouette, Twitter Network.

Πίνακας περιεχομένων

1.Εισαγωγή

1.1Δίκτυα

Τα δίκτυα είναι μια μαθηματική δομή που χρησιμοποιείται για να μοντελοποιεί σχέσεις αλληλεπίδρασης μεταξύ αντικειμένων ή οντοτήτων και είναι ιδιαίτερα χρήσιμα για την μελέτη εξαρτημένων δεδομένων. Κάθε δίκτυο ορίζεται από τον τύπο (1.1).

$$G = (V,E) \qquad \text{Τύπος (1.1)}$$

Όπου $V(G) = \{V_1, V_2, \ldots, V_n\}$ είναι το πεπερασμένο σύνολο των οντοτήτων ή κόμβων και E(G) είναι το σύνολο των ακμών που αποτελούνται από ζεύγη των κόμβων V(G) καθότι ενώνουν αυτές τις οντότητες (Wilson, 1996). Οι κόμβοι και οι συνδέσεις μπορούν να περιγραφούν από τον πίνακα γειτνίασης που ορίζεται από τον τύπο (1.2) με n να είναι ο αριθμός των κόμβων.

$$A = [A_{ij}]_{nxn}, \qquad \text{Τύπος (1.2)}$$

Εάν οι κόμβοι V_i και V_j του δικτύου συνδέονται, είναι $A_{ij}=1$ αλλιώς $A_{ij}=0$. Επίσης, εάν οι ακμές του δικτύου δεν είναι κατευθυνόμενες τότε ο πίνακας A είναι συμμετρικός και σε περιπτώσεις όπου το δίκτυο είναι με βάρη ο πίνακας είναι $A_{ij}=\lambda$, όπου $\lambda \in R\backslash\{0\}$, εάν οι αντίστοιχοι κόμβοι V_i και V_j συνδέονται.

1.2Πολύπλοκα Δίκτυα

Τα πολύπλοκα δίκτυα είναι συστήματα που χρησιμοποιούνται για να περιγράφουν ποικίλα φαινόμενα της φύσης και της κοινωνίας που αποτελούνται από ένα μεγάλο αριθμό υψηλά διασυνδεδεμένων δυναμικών οντοτήτων που αλληλοεπιδρούν μεταξύ τους. Ο όρος «πολύπλοκα» αναφέρεται στην παρουσία αναδυόμενων ιδιοτήτων στο σύστημα οι οποίες εμφανίζονται ως αποτέλεσμα της αλληλεπίδρασης των μερών του συστήματος. Η άλλη έννοια του «όρου» πολύπλοκα αναφέρεται στην ποσότητα της πληροφορίας που είναι αναγκαία για να καθοριστεί το σύστημα (Biswajit et al., 2015).

Μερικά παραδείγματα πολύπλοκων δικτύων είναι το διαδίκτυο (ιστότοποι και σύνδεσμοι μεταξύ τους), τα κοινωνικά δίκτυα (άτομα και κοινωνικές σχέσεις μεταξύ αυτών), τα οικονομικά δίκτυα (μετοχές, επιχειρήσεις, οργανισμοί και αλληλεπιδράσεις αυτών), τα χημικά δίκτυα που συνδέουν χημικές αλληλεπιδράσεις, τα δίκτυα μεταφοράς και τα δίκτυα αναφορών επιστημονικών άρθρων.

Το ερώτημα λοιπόν είναι αν υπάρχει κάποια αρχή, βάση της οποίας δημιουργείται η τοπολογία των συνδέσεων και πώς τα τεράστια συστήματα που αποτελούνται από αλληλοεπιδρώμενα μέλη συμπεριφέρονται συλλογικά, δεδομένης της ατομικής δυναμικής του καθενός. Μια προσέγγιση για την ανάλυση των συλλογικών ιδιοτήτων τέτοιων πολύπλοκων συστημάτων είναι η μοντελοποίηση τους μέσω γραφημάτων. Επιπρόσθετα, έχει ευρέως αναγνωριστεί ότι η τοπολογία και η εξέλιξη των πραγματικών δικτύων διέπεται από ποικίλες οργανωτικές αρχές τοπολογίας και δυναμικότητας. Μία από τις πιο εκπληκτικές ανακάλυψες της σύγχρονης θεωρίας δικτύων είναι η καθολικότητα της τοπολογίας των δικτύων, δηλαδή ότι πολλά πραγματικά δίκτυα σε διαφορετικά επίπεδα, από το κύτταρο μέχρι το Internet, συγκλίνουν

σε παρόμοιες δομές (Barabási, 2009). Δύο τέτοιες ανακαλύψεις χαρακτηριστικών και δομών στα πραγματικά δίκτυα είναι η δυναμοκατανομή των βαθμών και των κάποιων κεντρικοτήτων αλλά και η ιδιότητα του μικρόκοσμου που διέπει τα πραγματικά δίκτυα (Watts and Strogatz, 1998). Επομένως, η μελέτη τέτοιων δομών και η ανακάλυψη προτύπων όπως είναι οι κοινότητες, αποτελεί ένα κύριο πυλώνα της διεπιστημονικής έρευνας στον εικοστό πρώτο αιώνα.

2. Κοινότητες δικτύων

2.1 Ορισμός

Τα δίκτυα κατά κανόνα είναι συστήματα τα οποία δεν είναι κανονικά, δηλαδή σπάνια συναντώνται στη φύση δίκτυα, τα οποία έχουν ίδιο αριθμό ακμών για κάθε κόμβο. Επιπρόσθετα, έχει παρατηρηθεί ότι τα δίκτυα στη φύση δεν είναι τυχαία, που συνεπάγεται πως η κατανομή των ακμών ανά κόμβο δεν είναι κανονική. Αυτό οδηγεί στο συμπέρασμα ότι υπάρχει ένα είδος τάξης και οργάνωσης στην δομή των πραγματικών δικτύων. Αυτή η τάξη δεν εμφανίζεται μόνο συνολικά στο δίκτυο αλλά και τοπικά, όπου σε ορισμένες ομάδες από κόμβους εμφανίζονται πολλές ακμές μέσα στην ομάδα και λίγες μεταξύ των ομάδων. Το χαρακτηριστικό αυτό ονομάζεται δομή κοινοτήτων (Girvan and Newman, 2002).

Η δομή κοινοτήτων είναι μια πολύ σημαντική ιδιότητα των πραγματικών δικτύων. Τα πολύπλοκα δίκτυα έχουν δομή κοινότητας εάν οι κόμβοι του δικτύου μπορούν να ομαδοποιηθούν σε σύνολα στα οποία οι κόμβοι που ανήκουν στο ίδιο σύνολο είναι πυκνότερα συνδεδεμένοι (ή περισσότερο όμοιοι) σε σύγκριση με τους υπόλοιπους κόμβους του δικτύου. Ακόμη, σε πολλά δίκτυα συναντώνται επικαλυπτόμενες κοινότητες, όπου η διακριτοποίησή τους είναι ένα πιο δύσκολο έργο.

Από τον ορισμό της κοινότητας προκύπτει, ότι για κάθε τυχαίο ζεύγος κόμβων του δικτύου, υπάρχει μεγαλύτερη πιθανότητα να συνδέονται εάν βρίσκονται στην ίδια κοινότητα από ότι εάν βρίσκονται σε διαφορετικές. Ακόμη, αυτό που κάνει τον εντοπισμό των κοινοτήτων ενός δικτύου σημαντικό, είναι ότι οι κόμβοι της ίδιας κοινότητας μοιράζονται συνήθως τις ίδιες ιδιότητες και τα ίδια χαρακτηριστικά γι' αυτό ανιχνεύοντας την δομή των κοινοτήτων ενός δικτύου μπορούν να εξαχθούν εξαιρετικά συμπεράσματα για όλο το δίκτυο (Jierui, 2013).

Ο εντοπισμός κοινοτήτων ενός δικτύου είναι καίριας σημασίας για την κατανόηση της λειτουργίας και της τοπολογίας του δικτύου και η σημαντικότητα της εύρεσής του διαφέρει ανάλογα με το πεδίο στο οποίο βρίσκεται. Οι εντοπισμένες κοινότητες αναδεικνύουν τη λειτουργία του συστήματος που μοντελοποιεί το δίκτυο καθώς οι κοινότητες αντιστοιχούν συχνά σε σημαντικές λειτουργικές μονάδες του συστήματος όπως παρατηρείται σε βιολογικά και κοινωνικά δίκτυα. Επιπρόσθετα, στη μελέτη μεγάλων και πολύπλοκων δικτύων, η διαμέριση του δικτύου σε συστάδες με όμοια χαρακτηριστικά είναι απαραίτητη για να αποκαλυφθούν κρυμμένα μοτίβα ώστε να πραγματοποιηθεί περεταίρω μελέτη και κατανόηση του δικτύου (Fortunato, 2010).

Μια άλλη σημαντική πτυχή που σχετίζεται με τη δομή της κοινότητας είναι η ιεραρχική δομή που εμφανίζεται στα περισσότερα δίκτυα στον πραγματικό κόσμο. Τα πραγματικά δίκτυα αποτελούνται συνήθως από κοινότητες, συμπεριλαμβανομένων μικρότερων κοινοτήτων, οι οποίες με τη σειρά τους περιλαμβάνουν μικρότερες κοινότητες κλπ. Το ανθρώπινο σώμα είναι ένα παράδειγμα ιεραρχικής οργάνωσης καθώς αποτελείται από όργανα, τα όργανα αποτελούνται από ιστούς, οι ιστοί από κύτταρα και ούτω καθεξής. Άλλωστε, η ιεραρχία διαδραματίζει ένα κρίσιμο ρόλο στη δομή και την εξέλιξη πολύπλοκων συστημάτων (Simon, 1962).

Τέλος, στην μελέτη των κοινωνικών δικτύων ο εντοπισμός των κοινοτήτων είναι πρωτεύουσας σημασίας καθώς στις κοινότητες ενός τέτοιου δικτύου αντανακλάται όλη η δομή και η σχέση των κόμβων και μπορούν να εξαχθούν χρήσιμα κοινωνικά συμπεράσματα.

2.2 Μέθοδοι εντοπισμού κοινοτήτων

Έχουν γίνει πολλές προσπάθειες με πληθώρα προσεγγίσεων για τον εντοπισμό της δομής κοινοτήτων σε πολύπλοκα δίκτυα. Εντούτοις, η αναζήτηση για τον βέλτιστο αλγόριθμο είναι ένα ανοικτό πρόβλημα διότι οι προσεγγίσεις διαφέρουν λόγω του διαφορετικού είδους των δικτύων για αυτό και δεν μπορεί να προταθεί ένας καθολικός αλγόριθμος ο οποίος θα είναι αποτελεσματικότερος και αποδοτικότερος από τους υπολοίπους σε όλα τα δίκτυα. Αυτό που μπορεί να μελετηθεί είναι η απόδοση των εκάστοτε αλγορίθμων σε ένα συγκεκριμένο πεδίο δικτύων (Jaewon et al., 2013).

Οι αλγόριθμοι για τον εντοπισμό κοινοτήτων σε πολύπλοκα δίκτυα μπορούν να διαιρεθούν αρχικά σε αλγορίθμους διαμέρισης και αλγορίθμους βελτιστοποίησης, όπου στόχος αυτών είναι η βελτιστοποίηση της αντικειμενικής συνάρτησης που χρησιμοποιεί ο κάθε ένας αλγόριθμος. Γενικότερα, η μελέτη για τον εντοπισμό κοινοτήτων σε δίκτυα με τη μορφή γράφων ξεκίνησε στις αρχές του 1970 με τις παραδοσιακές μεθόδους να εμφανίζονται και οι οποίες αποσκοπούσαν στην διαμέριση του δικτύου σε μικρότερα υποδίκτυα (Muhammad et al., 2018).

2.2.1 Αλγόριθμοι διαμέρισης

Οι αλγόριθμοι διαμέρισης ή αλλιώς παραδοσιακοί μέθοδοι διαμέρισης αποτελούνται από τους αλγορίθμους συσταδοποίησης, τους ιεραρχικούς αλγορίθμους, τους φασματικούς και τους διαιρετικούς.

Οι αλγόριθμοι συσταδοποίησης διαχωρίζουν τους κόμβους δικτύου σε k ομάδες, μεγιστοποιώντας ή ελαχιστοποιώντας μια συνάρτηση απώλειας με βάση την απόσταση μεταξύ τους. Ορισμένες τεχνικές ομαδοποίησης είναι οι Minimum K-clustering, Sum K-clustering, K-center και K-median. Αυτοί οι αλγόριθμοι θα μελετηθούν εκτενέστερα σε κατοπινό κεφάλαιο.

Η ιεραρχική μέθοδος βασίζεται στην ιδέα ενός δυαδικού δέντρου στο οποίο οι όμοιοι κόμβοι ενώνονται. Το πλεονέκτημα της μεθόδου είναι ότι δεν χρειάζεται εκ των προτέρων να δοθεί ο αριθμός των κλάσεων, σε αντίθεση με τους αλγορίθμους συσταδοποίησης. Ακόμη, σε πολλά δίκτυα εμφανίζεται η ιεραρχική δομή, δηλαδή μέσα στις ήδη υπάρχουσες κοινότητες είναι εμφωλευμένες άλλες μικρότερες, συνεπώς σε τέτοια δίκτυα η ιεραρχική μέθοδος είναι αρκετά αποδοτική (Zhang et al., 2014). Δυο προσεγγίσεις της ιεραρχικής μεθόδου είναι οι αθροιστική μέθοδος (Agglomerative method) και η διαιρετική (Divisive method). Στην αθροιστική κάθε κόμβος αρχικά είναι και μια κοινότητα και ακολούθως ενώνονται οι πιο όμοιοι κόμβοι, ενώ στην διαιρετική μέθοδο όλοι οι κόμβοι αποτελούν μια κοινότητα και ακολούθως αφαιρείται σε κάθε επανάληψη η ακμή με την υψηλότερη betweenness ή αυτή που ενώνει κόμβους με τη μικρότερη ομοιότητα (Newman and Girvan, 2004).

Η φασματική μέθοδος περιλαμβάνει όλους τους αλγορίθμους οι οποίοι διαμερίζουν το δίκτυο σε κλάσεις χρησιμοποιώντας ως συντεταγμένες τα ιδιοδιανύσματα του πίνακα γειτνίασης ή του Λαπλασιανού πίνακα. Η συγκεκριμένη μέθοδος μετατρέπει τους κόμβους σε σημεία ενός πολυδιάστατου χώρου (Donetti and Munoz, 2004).

Οι διαιρετικοί αλγόριθμοι χωρίζουν το δίκτυο σε k προκαθορισμένες κοινότητες με βάση τον ελάχιστο αριθμό των ακμών που είναι μεταξύ αυτών (Pothen, 1997).

Τέλος, οι αλγόριθμοι που βασίζονται στην ιδέα της διάδοσης και του τυχαίου περιπατητή όπου κάνουν την υπόθεση ότι εάν ένα δίκτυο έχει ισχυρή δομή κοινοτήτων τότε ένας τυχαίος περιπατητής θα ξοδεύει αρκετό χρόνο μέσα σε μια κοινότητα λόγω της υψηλής της πυκνότητας. Συνεπώς, μέσα από την επαναληπτική διαδικασία δημιουργείται η κατάσταση ισορροπίας από την οποία προκύπτουν οι κοινότητες του δικτύου (Masuda et al., 2017).

2.2.2 Αλγόριθμοι βελτιστοποίησης

Οι αλγόριθμοι βελτιστοποίησης έχουν ως στόχο τη μεγιστοποίηση του modularity του οποίου ο ορισμός δίνεται στον τύπο (2.1).

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

Τύπος (2.1)

Όπου $\sum_{\iota} e_{\text{ii}}$ είναι το σύνολο των ακμών που συνδέουν κόμβους της ίδιας κοινότητας ενώ $\sum_{\iota} a_i^2$ είναι το σύνολο των ακμών που εμπίπτουν σε κόμβους της i κοινότητας. Με άλλα λόγια, το modularity είναι η διαφορά των ακμών που ενώνουν κόμβους της ίδιας κοινότητας από τις ακμές που ενώνουν κόμβους διαφορετικών κοινοτήτων. Θεωρητικά, στα τυχαία δίκτυα το modurarity είναι Q≈0 και υπάρχει η παραδοχή ότι εάν το Q ενός δικτύου είναι μεγαλύτερο από 0.3, το δίκτυο παρουσιάζει δομή κοινοτήτων (Clauset et al., 2004).

Η πρώτη κατηγορία των αλγορίθμων βελτιστοποίησης του modularity είναι η ευρετική μέθοδος Extremal optimization η οποία βασίζεται στη δυναμική των συστημάτων μακράν της ισορροπίας που εξηγούν φαινόμενα αυτό-οργάνωσης (Boettcher and Percus, 2001). Ουσιαστικά, η συνάρτηση του modularity βελτιστοποιείται δηλαδή μεγιστοποιείται, καθώς προσαρμόζεται μια αντικειμενική τιμή μέσω ενός γενετικού αλγορίθμου. Άρα, ο αλγόριθμος αρχικά χωρίζει το δίκτυο σε δύο κοινότητες και στη συνέχεια σε κάθε επανάληψη χωρίζει τον κόμβο που έχει τη μικρότερη αντικειμενική τιμή με τους υπόλοιπους της ίδιας κοινότητας με σκοπό την μεγιστοποίηση του modularity.

Η φασματική βελτιστοποίηση (Spectral optimization) είναι μια τεχνική που χρησιμοποιεί τα ιδιοδιανύσματα και τις ιδιοτιμές του πίνακα γειτνίασης ή του Λαπλασιανού πίνακα για να μεγιστοποιήσει το modularity (Newman, 2006).

Η Greedy βελτιστοποίηση είναι μια μέθοδος που πρότεινε ο Newman (Newman, 2004) η οποία χρησιμοποιεί την ιδέα του αθροιστικού ιεραρχικού αλγορίθμου στον οποίο οι κόμβοι αρχικά είναι και μια κοινότητα και στη συνέχεια συνδέονται με σκοπό την αύξηση του modularity.

Η τελευταία κατηγορία αλγορίθμων βελτιστοποίησης είναι οι γενετικοί αλγόριθμοι οι οποίοι έχουν εμπνευστεί τη διαδικασία της μετάλλαξης από την βιολογία (Pizzuti, 2013). Οι αλγόριθμοι αυτοί δεν απαιτούν τον αριθμό των συστάδων από την αρχή και θεωρούν τους κόμβους ως γονίδια όπου μέσα από τις επαναλήψεις προσεγγίζουν την βέλτιστη λύση μέσω μίας ή περισσοτέρων αντικειμενικών συναρτήσεων για την εύρεση πυκνά συνδεδεμένων κόμβων.

3. Μέθοδοι Συσταδοποίησης

3.1 Συσταδοποίηση

Η συσταδοποίηση είναι μια μέθοδος μη επιβλεπόμενης μηχανικής μάθησης και είναι η διαδικασία εκείνη κατά την οποία ένα σύνολο από n «αντικείμενα», διαχωρίζεται σε k λογικές ομάδες. Η καταχώρηση αντικειμένων στην ίδια ομάδα μεταφράζεται ως ομοιότητα ή συσχέτιση των αντικειμένων αυτών και αντίστροφα, δηλαδή αντικείμενα που ανήκουν σε διαφορετικές ομάδες είναι ανόμοια. Η ομοιότητα ή μη, μεταξύ των αντικείμενων, ουσιαστικά εξαρτάται από το συγκεκριμένο πρόβλημα και τη μορφή των αντικειμένων ή οντοτήτων. Επίσης, όσο πιο όμοια είναι τα αντικείμενα μέσα στην συστάδα και ανόμοια με τα άλλα των υπολοίπων, τόσο καλύτερη είναι η επιτευχθείσα συσταδοποίηση.

Η συσταδοποίηση ουσιαστικά είναι η διαδικασία της ομαδοποίησης αντικειμένων όταν δεν υπάρχει εκ των προτέρων καμία πληροφορία για αυτά. Για παράδειγμα, στην ταξινόμηση (classification) ο αριθμός των ομάδων είναι γνωστός εκ των προτέρων και σκοπός είναι να ταξινομηθούν τα αντικείμενα σύμφωνα με αυτές. Ωστόσο, η συσταδοποίηση αποσκοπεί στο να ανακαλύψει τα κρυμμένα μοτίβα των αντικειμένων τα οποία ταξινομεί σε ομάδες που δημιουργούνται από τα ίδια τα αντικείμενα σύμφωνα με την ομοιότητα που έχει οριστεί.

Υπάρχει ένα μεγάλο πλήθος από αλγόριθμους συσταδοποίησης που έχουν προταθεί με τον κάθε ένα από αυτούς να βασίζεται σε διαφορετική φιλοσοφία. Σχεδόν όλοι οι αλγόριθμοι αρχικά δέχονται ως είσοδο ένα σύνολο παραμέτρων όπως το πλήθος των συστάδων, τα διανύσματα αρχικοποίησης που απαιτούνται καθώς επίσης και κάποιες υποθέσεις που αφορούν την πυκνότητα των διανυσμάτων στο χώρο. Διαφοροποιώντας αυτές της παραμέτρους προκύπτει ένα σύνολο από αλγορίθμους σε κάθε βασική κατηγορία. Οι δύο κύριοι αλγόριθμοι πάνω στους οποίους βασίζονται οι υπόλοιποι, είναι ο αλγόριθμος K-means (των Κ μέσων) και οι Ιεραρχικοί αλγόριθμοι (Tan et al., 2018).

3.2 K-means

Η ομαδοποίηση k-μέσων (K-Means) είναι μία δημοφιλής μέθοδος στην ανάλυση συστάδων και στόχο έχει να διαχωρίσει τις n παρατηρήσεις σε k ομάδες (k<n), έτσι ώστε κάθε παρατήρηση να ανήκει στη συστάδα με της οποίας τον μέσο έχει τη μικρότερη απόσταση σε σύγκριση με τους υπόλοιπους k-1 μέσους. Συνεπώς, αυτό οδηγεί σε μια διαμέριση του χώρου δεδομένων σε k κλάσεις (MacQueen, 1967).

Η μέθοδος K-Means χειρίζεται τις παρατηρήσεις των δεδομένων ως αντικείμενα τα οποία έχουν θέσεις και αποστάσεις μεταξύ τους σε κάποιο γεωμετρικό χώρο. Έτσι, τοποθετεί τις παρατηρήσεις σε Κ ομάδες έτσι ώστε οι παρατηρήσεις μέσα στην ίδια ομάδα να είναι όσο το δυνατόν πιο «κοντά» μεταξύ τους και όσο το δυνατόν «μακριά» από τις παρατηρήσεις των υπόλοιπων Κ-1 ομάδων. Αυτό όμως συμβαίνει χωρίς προηγούμενη γνώση για τα δεδομένα και τις σχέσεις μεταξύ τους.

Το κύριο πλεονέκτημα της μεθόδου είναι ότι ενημερώνεται σε κάθε επανάληψη από τα ίδια τα δεδομένα (unsupervised machine learning) χωρίς να πρέπει αρχικά κάποιος χρήστης να εκπαιδεύσει τον αλγόριθμο σχετικά με αυτά (supervised machine learning). Συνεπώς, ο αλγόριθμος είναι αρκετά γρήγορος στη σύγκλιση κάτι που είναι αρκετά χρήσιμο σε μεγάλα και πολύπλοκα δεδομένα (Tan et al., 2018). Επιπρόσθετα, εκεί που ο αλγόριθμος K-Means υπερτερεί ως προς τους Ιεραρχικούς αλγορίθμους είναι ότι είναι εύκολα εφαρμόσιμος, έχει μικρότερο υπολογιστικό κόστος, συγκλίνει ταχύτερα στην βέλτιστη λύση και έχει πιο διακριτές συστάδες, χαρακτηριστικά που του δίνουν προβάδισμα σε μεγάλα και πολύπλοκα δεδομένα.

3.2.1 Τα βήματα του αλγορίθμου K-Means

Ο αλγόριθμος στη συνήθη γενική μορφή του καθώς έχει πολλές παραλλαγές, λαμβάνει αρχικά από τον χρήστη τον αριθμό k των συστάδων. Η παράμετρος k δηλώνεται ως είσοδος και προσαρμόζεται ανάλογα με τα δεδομένα αλλά και τα συμπεράσματα που θέλει ο αναλυτής να βγάλει. Τα βήματα έχουν ως εξής:

1. Ο αλγόριθμος επιλέγει αυθαίρετα k σημεία ως αρχικά κέντρα των συστάδων.
2. Κάθε παρατήρηση ή σημείο του συνόλου των δεδομένων αντιστοιχίζεται στην «κοντινότερη» συστάδα, με βάση την απόσταση από τα αρχικά επιλεγμένα κέντρα της αντίστοιχης συστάδας.
3. Κάθε κέντρο των συστάδων επαναπροσδιορίζεται ως ο μέσος όρος των σημείων της συστάδας ή το σημείο που έχει την μικρότερη απόσταση από τα υπόλοιπα της ίδιας συστάδας.
4. Τα βήματα 2 και 3 επαναλαμβάνονται έως ότου επέλθει σύγκλιση που σημαίνει ότι οι συστάδες δεν αλλάζουν σημαντικά καθώς εκτελούνται τα βήματα 2 και 3 ή μέχρις ότου συμπληρωθεί ο αριθμός των επαναλήψεων που ο χρήστης έχει ορίσει.

3.2.2 Προβλήματα του K-Means

Τα μειονεκτήματα του αλγορίθμου K-Means είναι ότι είναι ευαίσθητος στην τυχαία επιλογή των αρχικών συνθηκών ή κέντρων, για αυτό και μπορεί να συγκλίνει σε τοπικά βέλτιστη λύση και όχι σε γενική βέλτιστη. Μια πρακτική λύση σε αυτό είναι να τρέξει ο αλγόριθμος αρκετές φορές με διαφορετικές αρχικές συνθήκες και να συγκριθούν τα αποτελέσματα. Ακόμη, ένα άλλο πρόβλημα είναι η επιλογή του αριθμού k των συστάδων, διότι μέχρι και σήμερα αποτελεί ανοικτό πρόβλημα η ακριβής επιλογή των αρχικών συνθηκών και ο προσδιορισμός του αριθμού k των συστάδων.

Βέβαια, η χρησιμοποίηση του αλγορίθμου K-Means απαιτεί έναν γεωμετρικό χώρο στον οποίο να ορίζεται απόσταση. Άρα, για την εφαρμογή του αλγορίθμου σε δίκτυα (που στην ουσία είναι οι κόμβοι και οι συνδέσεις μεταξύ τους) υπεισέρχεται το πρίσμα της θεωρίας των δικτύων.

Αυτό σημαίνει ότι θα πρέπει οι k αρχικοί σπόροι (seeds) να είναι κόμβοι αλλά και ότι το δίκτυο θα πρέπει να παρασταθεί σε γεωμετρικό χώρο προκειμένου να βρει εφαρμογή ο αλγόριθμος K-Means (Arthur and Vassilvitskii, 2007).

4. Ο προτεινόμενος αλγόριθμος K-Means

4.1 Μετατροπή Δικτύου σε μετρικό χώρο

Ο αλγόριθμος K-Means εφαρμόζεται σε παρατηρήσεις οι οποίες ανήκουν σε κάποιο γεωμετρικό χώρο και τις κατηγοριοποιεί με βάση την απόσταση ή κάποια ομοιότητα. Εντούτοις, οι κόμβοι του δικτύου, όπως ορίστηκε από τον τύπο (1.1), δεν ανήκουν σε κάποιον μετρικό χώρο άρα είναι αναγκαίο να παρασταθεί το δίκτυο σε γεωμετρικό χώρο ώστε να εφαρμοστεί ο προτεινόμενος αλγόριθμος για την εύρεση των κοινοτήτων.

Για τη λύση του συγκεκριμένου προβλήματος προτείνεται η μέθοδος signal similarity (Hu, 2008) η οποία μετατρέπει την τοπολογική δομή του δικτύου των n κόμβων σε διανύσματα που ανήκουν στον n-διάστατο διανυσματικό Ευκλείδειο χώρο. Με την μέθοδο signal similarity κάθε κόμβος ορίζεται ως μια πηγή η οποία αρχικά έχει ένα σήμα και το διαδίδει στο υπόλοιπο δίκτυο, στο οποίο οι υπόλοιποι κόμβοι καταγράφουν το ποσό του 'σήματος' που λαμβάνουν. Σε κάθε βήμα οι κόμβοι στέλνουν το σήμα τους στους γείτονές τους και ταυτόχρονα λαμβάνουν και το σήμα αυτών. Έτσι, μετά από κάποια βήματα η κατανομή των σημάτων των κόμβων ορίζεται ως η επιρροή της κάθε πηγής στο υπόλοιπο δίκτυο. Η κεντρική ιδέα είναι ότι κάθε κόμβος θα επηρεάζει σε μεγαλύτερο βαθμό τους κόμβους της κοινότητας που ανήκει και λιγότερο της υπόλοιπους (Jiang, 2013).

Ως είσοδος για την εφαρμογή του προτεινόμενου αλγορίθμου απαιτείται ο πίνακας γειτνίασης του δικτύου. Έτσι, η διαδικασία μετατροπής του δικτύου σε γεωμετρικό χώρο έχει ως εξής:

1. Αρχικά ορίζεται ο πίνακας:

$$S = (A + I)^{\tau} \qquad \text{Τύπος (4.1)}$$

όπου A είναι ο πίνακας γειτνίασης, I ο n-διάστατος ταυτοτικός πίνακας και τ το βήμα που επιλέγεται (τ=3 στον αλγόριθμο που θα εφαρμοστεί). Ο ταυτοτικός πίνακας προστίθεται έτσι ώστε η κύρια διαγώνιος του πίνακα να μην έχει μηδενικά στοιχεία. Σε κάθε βήμα ο πίνακας (A + I) πολλαπλασιάζεται με τον εαυτό του, με συνέπεια οι ήδη υπάρχουσες τιμές να αυξάνονται και κάποιες μηδενικές να παίρνουν τιμές γιατί όταν ο πίνακας γειτνίασης υψώνεται στην λ δύναμη τότε προκύπτει ο πίνακας συνδέσεων του δικτύου με λ βήματα.

2. Έπειτα κάθε στοιχείο του πίνακα κανονικοποιείται ως εξής:

$$\bar{S}_{ij} = \frac{sij}{\sqrt{\sum_j sij}} \qquad \text{Τύπος (4.2)}$$

Η κανονικοποίηση συμβαίνει ώστε τα διανύσματα να έχουν καλύτερο μέτρο για την εφαρμογή του αλγορίθμου.

3. Τελικά, δημιουργούνται τα διανύσματα $\bar{S}_1, \bar{S}_2, \ldots, \bar{S}_n$ που ανήκουν στον n-διάστατο Ευκλείδειο χώρο όπου το κάθε ένα αντιπροσωπεύει την επίδραση που έχει στους υπόλοιπους κόμβους του δικτύου.

Με τα τρία παραπάνω βήματα ορίστηκαν τα διανύσματα των κόμβων στον n-διάστατο χώρο στον οποίο μπορεί να οριστεί απόσταση. Επομένως, η απόσταση δύο κόμβων ορίζεται ως η αντίστροφη επιρροή μεταξύ των δύο διότι όσο ισχυρότερα ένας κόμβος επιδρά σε έναν άλλον, τόσο η απόσταση μεταξύ των δύο θα είναι μικρότερη αντίστοιχα. Στα μηδενικά στοιχεία, η απόσταση λαμβάνει τυχαία μια μεγάλη τιμή.

4.2 Επιλογή του αριθμού k και των αρχικών κέντρων

Ένα από τα κύρια μειονεκτήματα του αλγορίθμου συσταδοποίησης K-Means είναι το γεγονός ότι ο αριθμός k των συστάδων θα πρέπει να ορίζεται από την αρχή. Ως μέση τετραγωνική απόσταση ορίζεται η μέση ευκλείδεια απόσταση κάθε κόμβου μιας συστάδας από τον κόμβο που είναι το κέντρο αυτής. Η μέση τετραγωνική απόσταση όλων των σημείων μειώνεται καθώς το k αυξάνεται. Δηλαδή όταν το k, που είναι ο αριθμός των συστάδων, είναι ίσο με 1 τότε η μέση τετραγωνική απόσταση έχει την μέγιστη τιμή, και όταν το k είναι ίσο με τις παρατηρήσεις (k=n) τότε η μέση τετραγωνική απόσταση είναι 0. Συνεπώς, για να έχει νόημα και αξία η συσταδοποίηση, μια καλή επιλογή του k είναι στο σημείο όπου η μέση τετραγωνική απόσταση πέφτει απότομα, καθώς στη συνέχεια διατηρείται περίπου σταθερή μέχρι να συγκλίνει στο 0.

Επιπρόσθετα, η επιλογή των k αρχικών κέντρων (initial seeds) εξ' ορισμού είναι τυχαία, κάτι που ενδέχεται να οδηγήσει σε τοπικά βέλτιστη λύση και όχι σε ολικά βέλτιστη. Εντούτοις, στην μελέτη αυτή απαντάται το συγκεκριμένο ερωτήματα κάτω από το πρίσμα των δικτύων με βάση δύο υποθέσεις.

Οι υποθέσεις που γίνονται σχετικά με την επιλογή των αρχικών κέντρων είναι:

1. Σημαντικότητα, δηλαδή κάθε κόμβος με υψηλή κεντρικότητα είναι πιθανότερο να επιλεγεί ως αρχικό κέντρο.
2. Διασπορά, δηλαδή τα κέντρα αναμένεται να διανεμηθούν ομοιόμορφα στο δίκτυο, επειδή ένα κέντρο θα πρέπει να έχει σχετικά μεγάλη απόσταση από τα υπόλοιπα κέντρα.

4.2.1 Σημαντικότητα κόμβων μέσω της PageRank κεντρικότητας

Η κεντρικότητα PageRank αναπτύχθηκε από τους Brin and Page (Page, 1999) και υποθέτει ότι ένας τυχαίος περιπατητής ακολουθεί την τοπολογία του δικτύου μέσω του πίνακα μετάβασης P και σε κάθε επανάληψη μεταπηδάει από τον έναν κόμβο σε άλλον με πιθανότητα $\frac{1}{n}$, όπου n ο αριθμός των κόμβων του δικτύου. Η κεντρικότητα PageRank είναι μια προσαρμογή της Katz κεντρικότητας, όπου υπάρχουν τρεις παράγοντες που καθορίζουν το μέγεθος της PageRank κεντρικότητας του κόμβου:

1. Ο αριθμός των ακμών που λαμβάνει.
2. Η τάση της σύνδεσης των γειτόνων.
3. Η κεντρικότητα των γειτόνων.

Άρα, στην PageRank κεντρικότητα κάθε κόμβος είναι σημαντικός εάν συνδέεται από τους σημαντικούς και συνδέεται με τους ασήμαντους, ή σε περιπτώσεις μη κατευθυνόμενων ακμών αν είναι συνολικά συνδεδεμένος με γείτονες που κι αυτοί είναι αρκετά συνδεμένοι. Συνεπάγεται

δηλαδή ότι κάθε κόμβος με μεγάλο δείκτη PageRank είναι σημαντικός για το δίκτυο, άρα έχει μεγαλύτερη πιθανότητα να επιλεχθεί ως αρχικό κέντρο.

Το n-διάστατο διάνυσμα υ της PageRank από την επαναληπτική δυναμική μέθοδο (Franceschet, 2011) ορίζεται ως:

$$v^{t+1} = \left((1-\beta)P + e\frac{\beta}{n}\right)v^t$$

Τύπος (4.3)

Και συγκλίνει έως ότου:

$$E = |v^{t+1} - v^t| = \sum_{i=1}^{n}(v_i^{t+1} - v_i^t)^2 < \epsilon$$

Τύπος (4.4)

Όπου β>0 είναι η σταθερά απόσβεσης της μεταβολής του PageRank, e ένα μοναδιαίο διάνυσμα, t το χρονικό βήμα και P ο πίνακας μετάβασης που ορίζεται ως:

$$P_{ij} = \frac{A_{ij}}{\sum_j A_{ij}}$$

Τύπος (4.5)

Ωστόσο, εάν όλοι οι κόμβοι δεν παρουσιάζουν μεγάλο εύρος τιμών PageRank κεντρικότητας, είναι δύσκολο να βρεθούν οι κατάλληλοι αρχικοί κόμβοι. Για τη λύση αυτού του ζητήματος εισάγεται μια τροποποιημένη κεντρικότητα PageRank η οποία ενισχύεται από την απόσταση του κάθε κόμβου από τους υπόλοιπους (Li, 2015) και ορίζεται ως εξής:

$$\bar{P}_i = P_i + \sum_j exp\left(-\frac{d(i,j)^2}{P(i)}\right)$$

Τύπος (4.6)

Όπου d(i,j) χρησιμοποιείται η Ευκλείδεια απόσταση, όπου απόσταση ορίστηκε η αντίστροφη επιρροή του κάθε κόμβου στους υπόλοιπους από την μέθοδο signal similarity και τον τύπο (4.2). Στην μετασχηματισμένη αυτή μορφή της PageRank, εάν ένας κόμβος βρίσκεται κοντά σε πολλούς άλλους κόμβους, η κεντρικότητα PageRank ενισχύεται αρκετά, ενώ εάν κάποιος είναι απομακρυσμένος κόμβος, ενισχύεται ελάχιστα από τον μετασχηματισμό, καθώς $\exp(-x^2) = (0,1]$.

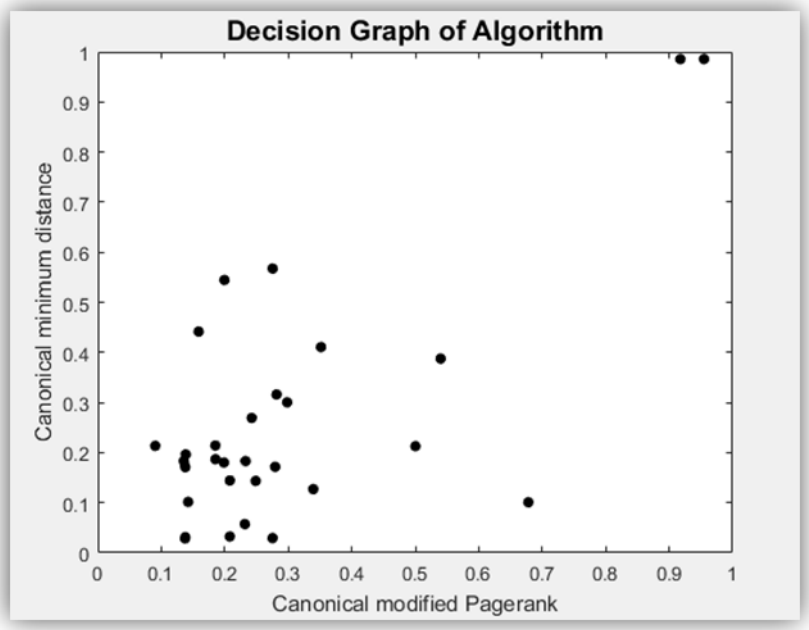
4.2.2 Απόσταση σημαντικών κόμβων

Για τον έλεγχο της διασποράς των αρχικών κέντρων σύμφωνα με την δεύτερη υπόθεση, χρησιμοποιείται η απόσταση δ_i (i= 1,2,…,n) από τον τύπο (4.7) που υπολογίζει την απόσταση κάθε κόμβου P_i από τους υπόλοιπους κόμβους οι οποίοι έχουν υψηλότερη PageRank. Έτσι, ο αλγόριθμος ταξινομεί αρχικά σε φθίνουσα σειρά τους κόμβους ως προς την PageRank κεντρικότητα και στη συνέχεια υπολογίζει την μικρότερη απόσταση. Βέβαια όπως είναι λογικό από τον υπολογισμό εξαιρείται ο κόμβος με την μεγαλύτερη Pagerank ο οποίος αποτελεί εξ' ορισμού αρχικό κέντρο.

$$\delta_i = \min_{j:\bar{P}_j > \bar{P}_i} (d_{ij})$$ Τύπος (4.7)

Για παράδειγμα, σε έναν πίνακα γειτνίασης 4x4 όπου υπάρχουν τέσσερις κόμβοι, έστω ότι η επιρροή του κόμβου 1 στους υπολοίπους είναι S_1 = [4, 2, 0.5, 0.2], τότε η κανονικοποιημένη απόσταση είναι $\bar{S}_1$ = [0.05, 0.1, 0.4, 1]. Έπειτα εξαιρούνται οι κόμβοι με την μεγαλύτερη PageRank και υπολογίζεται η απόσταση. Η απόσταση ορίζεται ως η Ευκλείδεια απόσταση των διανυσμάτων του κάθε κόμβου και χρησιμοποιείται συμβατικά το min ως κάτω φράγμα.

Στη συνέχεια κατασκευάζεται ένα γράφημα απόφασης 2 διαστάσεων το οποίο αξιολογεί τις δύο υποθέσεις στις δύο διαστάσεις (σημαντικότητα-διασπορά) και οι k κόμβοι που διανέμονται στο δεξιό άνω μέρος του γραφήματος μπορούν να επιλεγούν ως αρχικά κέντρα. Επίσης, οι δύο δείκτες Pagerank και απόσταση, κανονικοποιούνται για τον σχεδιασμό του γραφήματος. Ένα παράδειγμα τέτοιου γραφήματος με 34 κόμβους από τους οποίους ξεχωρίζουν οι δύο, βρίσκεται στην Εικόνα (4.1).



ΕΙΚΟΝΑ (4.1) *Η απεικόνιση ενός γραφήματος απόφασης*

Έπειτα, για την καλύτερη αποσαφήνιση του γραφήματος απόφασης δημιουργείται το γράφημα της Ευκλείδειας απόστασης των σημείων του γραφήματος που ορίζεται από τον τύπο (4.8).

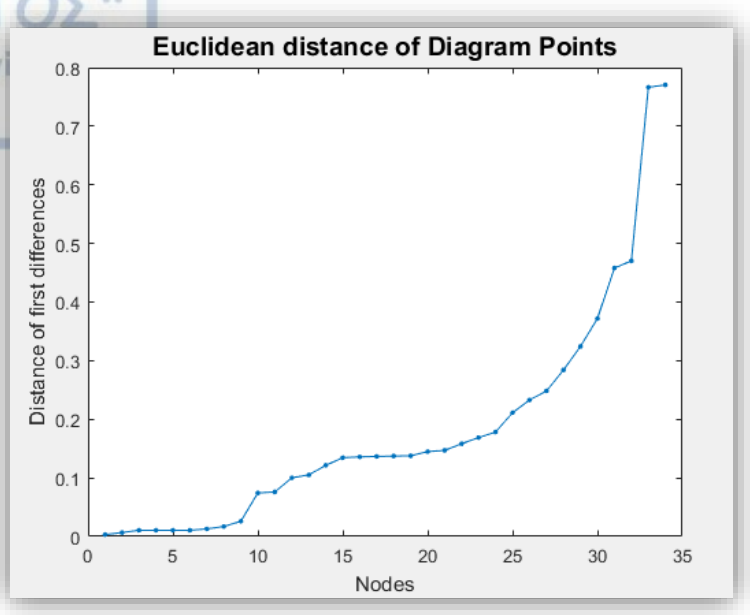
$$d(C_i) = \sqrt{(C_{xi} + C_{yi})^2}$$ Τύπος (4.8)

όπου Cx_i οι τετμημένη και Cy_i η τεταγμένη του σημείου i.

Στη συνέχεια τα σημεία ταξινομούνται σε φθίνουσα σειρά και δημιουργείται ένα νέο διάγραμμα που δίνεται στην Εικόνα (4.2) το οποίο υπολογίζει τις πρώτες διαφορές σύμφωνα με τον τύπο (4.9)

$$d(c_{i+1}) - d(c_i) \quad \text{με } i=1, \ldots, n-1$$ Τύπος (4.9)

όπου n ο αριθμός των κόμβων του δικτύου.



ΕΙΚΟΝΑ (4.2) *Η απεικόνιση της αύξουσας Ευκλείδειας απόστασης*

Στην Εικόνα (4.2) παρατηρείται η μεγαλύτερη πτώση μετά τον δεύτερο κόμβο, συνεπώς η επιλογή του k=2 είναι η βέλτιστη σύμφωνα με το διάγραμμα.

Τέλος, για τις περιπτώσεις όπου δεν υπάρχει ξεκάθαρη εικόνα για το πού συμβαίνει η απότομη πτώση, η επιλογή του αριθμού k γίνεται από το γράφημα απόφασης.

4.3 Ποσοτική αξιολόγηση της μεθόδου

Για την ποσοτική αξιολόγηση της συσταδοποίησης χρησιμοποιείται η μέθοδος Sihouette (Rousseeuw, 1987) η οποία μετρά πόσο καλά κάθε κόμβος εκχωρείται στην κοινότητα που ανήκει. Συγκεκριμένα, ορίζεται από τον τύπο (4.10).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Τύπος (4.10)

Όπου a(i) είναι η μέση απόσταση του κόμβου i με τους υπόλοιπους της ίδιας κοινότητας που δηλώνει το μέτρο της ανομοιότητας μέσα στην ίδια κοινότητα. Αντίθετα, το b(i) είναι η μέση απόσταση του κόμβου i με τους κόμβους των άλλων κοινοτήτων και δηλώνει την ανομοιότητα μεταξύ των κοινοτήτων.

Από τον ορισμό προκύπτει και ο τύπος (4.11).

$$-1 \leq s(i) \leq 1$$

Τύπος (4.11)

Συνεπώς, καθώς το s(i) τείνει στο 1 σημαίνει ότι b(i) γίνεται πολύ μεγαλύτερο από το a(i) για κάθε κόμβο i, δηλαδή υπάρχει καλή συσταδοποίηση. Αντίθετά όταν s(i) τείνει στο -1 σημαίνει ότι υπάρχει αρνητική συσταδοποίηση και θα πρέπει ο κάθε κόμβος να τοποθετηθεί σε γειτονικές κοινότητες. Τέλος, όταν s(i) τείνει στο 0, τότε υπάρχει οριακά καλή συσταδοποίηση.

Το ποσοτικό μέτρο λοιπόν που εισάγεται για να αξιολογηθεί η ποιότητα της συσταδοποίησης είναι ο μέσος όρος του s(i) όλων των κόμβων του δικτύου όπου αποκαλύπτεται η καταλληλόλητα της διαμέρισης αλλά και της επιλογής του αριθμού k των συστάδων.

4.4Περιγραφή βημάτων Αλγορίθμου

Ο προτεινόμενος αλγόριθμος είναι μια παραλλαγή του κλασικού k-Means με γνώμονα την εφαρμογή στα δίκτυα αλλά παράλληλα να ελαχιστοποιούνται οι συγκλίσεις σε τοπικά βέλτιστες λύσεις. Τα βήματα του αλγορίθμου είναι τα εξής:

Είσοδος: ο πίνακας γειτνίασης A του δικτύου.

1. Το δίκτυο μετατρέπεται σε γεωμετρική δομή διανυσμάτων του n-διάστατου χώρου με τη μέθοδο signal similarity.
2. Υπολογίζεται η μετασχηματισμένη PR_i τιμή για κάθε κόμβο i του δικτύου.
3. Υπολογίζεται η μικρότερη απόσταση (md_i) κάθε κόμβου από τους υπόλοιπους με την μεγαλύτερη κεντρικότητα από τον ίδιο.
4. Σχεδιάζεται ένα κανονικοποιημένο διάγραμμα διασποράς δύο διαστάσεων (PR_i- md_i) και υπολογίζεται η Ευκλείδεια απόσταση των σημείων του διαγράμματος ώστε να εντοπιστούν τα σημεία που έχουν ταυτόχρονα μεγάλη PageRank και μεγάλη απόσταση από τους υπόλοιπους με μεγάλη κεντρικότητα.
5. Εκτελείται ο κλασικός αλγόριθμος K-Means μέχρι η αντιστοίχηση κάθε κόμβου στην εκάστοτε κοινότητα να παραμένει αναλλοίωτη.
6. Εκτιμάται το ποσοστό βέλτιστου διαχωρισμού με τη μέθοδο Silhouette.

Έξοδος: Οι K κοινότητες των δικτύου C.

Αρχικά, ως είσοδος απαιτείται ένας πίνακας γειτνίασης κατευθυνόμενου ή μη κατευθυνόμενου δικτύου και από τον τύπο (4.1) με την μέθοδο signal similarity ο πίνακας γειτνίασης μετατρέπεται σε πίνακα επιρροής ή αντίστροφης απόστασης του κάθε κόμβου με τους υπόλοιπους του δικτύου. Επομένως από την αντιστροφή των στοιχείων του πίνακα επιρροής παράγεται ο πίνακας αποστάσεων του δικτύου.

Έπειτα υπολογίζεται η PageRank κεντρικότητα για κάθε κόμβο του δικτύου από τον τύπο (4.3) ορίζοντας β=0.75. Στη συνέχεια υπολογίζεται η PageRank κεντρικότητα η οποία

μετασχηματίζεται σύμφωνα με την απόσταση του κάθε κόμβου από τους υπόλοιπους με τον τύπο (4.6) και οι κόμβοι ταξινομούνται σε φθίνουσα σειρά.

Ακολούθως υπολογίζεται η ελάχιστη απόσταση του κάθε κόμβου από τους υπόλοιπους κόμβους του δικτύου, με τη μεγαλύτερη κεντρικότητα από τον τύπο (4.7). Να σημειωθεί ότι σε αυτό το βήμα εξαιρείται ο κόμβος με την μεγαλύτερη κεντρικότητα, ο οποίος ορίζεται από την υπόθεση που έγινε, ως αρχικό κέντρο καθώς δεν υπάρχει άλλος με μεγαλύτερη για να συγκριθεί.

Από τους δυο δείκτες που δημιουργήθηκαν (μετασχηματισμένη PageRank και ελάχιστη απόσταση), κατασκευάζεται ένα δισδιάστατο διάγραμμα για τους κόμβους του δικτύου. Έπειτα, υπολογίζεται η Ευκλείδεια απόσταση των σημείων του διαγράμματος και ταξινομούνται τα σημεία σε φθίνουσα σειρά. Έτσι, ο εντοπισμός της μεγαλύτερης πρώτης διαφοράς δίνει και τον αριθμό k των συστάδων, εκτός βέβαια των περιπτώσεων που δεν ξεχωρίζει ιδιαίτερα κάποια πρώτη διαφορά.

Όταν επιλεγούν οι αρχικοί κόμβοι, εκτελείται ο αλγόριθμος K-Means ο οποίος εφαρμόζεται στον πίνακα ομοιότητας και διαμερίζει το δίκτυο τοποθετώντας κάθε κόμβο στην αντίστοιχη κοινότητα. Τέλος, ο αλγόριθμος παρουσιάζει τις παραγόμενες κοινότητες και γίνεται η ποσοτική αξιολόγηση της διαμέρισης με τον δείκτη Silhouette.

4.5Υπολογιστική πολυπλοκότητα του αλγορίθμου

Ο προτεινόμενος αλγόριθμος περιλαμβάνει τρία κύρια στάδια για την εξαγωγή των κοινοτήτων του δικτύου από τον πίνακα γειτνίασης. Το πρώτο είναι η μετατροπή του δικτύου σε Ευκλείδειο χώρο, το δεύτερο η επιλογή των k αρχικών κέντρων και το τρίτο η εφαρμογή του αλγορίθμου K-Means. Η υπολογιστική πολυπλοκότητα του πρώτου σταδίου που είναι η χρήση της μεθόδου Signal Similarity με τον τύπο (4.1) και είναι $O(\tau(d+1)n^2)$, όπου d είναι ο μέσος βαθμός του δικτύου, τ η επιλογή των βημάτων και n ο αριθμός των κόμβων του δικτύου (Li, 2015).

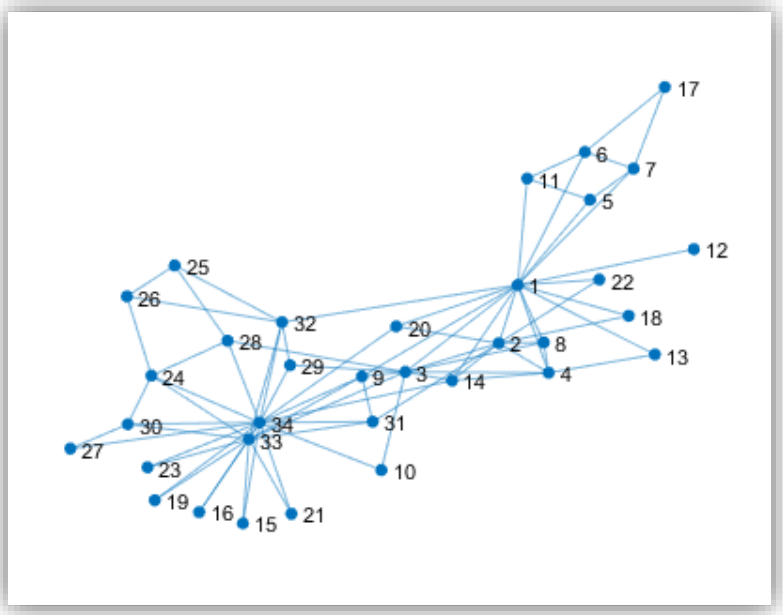
Για την επιλογή των k αρχικών κέντρων χρησιμοποιείται ο αλγόριθμος PageRank με πολυπλοκότητα O(mr) από τον τύπο (4.3) με m τον αριθμό των ακμών και r τον αριθμό των επαναλήψεων. Επίσης, το δεύτερο στάδιο περιλαμβάνει και τον υπολογισμό της ελάχιστης απόστασης του κάθε κόμβου από τους υπόλοιπους με τη μεγαλύτερη PageRank κεντρικότητα, όπου η πολυπλοκότητα του υπολογισμού της απόστασης είναι $O(n^2)$. Άρα η πολυπλοκότητα του δεύτερου σταδίου συνολικά είναι $O(mr+n^2)$.

Στο τελευταίο τμήμα, η πολυπλοκότητα του αλγορίθμου K-Means είναι $O(kn^2t)$ όπου k ο αριθμός των συστάδων, n ο αριθμός των κόμβων και t ο αριθμός των επαναλήψεων. Συνεπώς η συνολική υπολογιστική πολυπλοκότητα είναι $O(\tau(d+1)n^2+mr+n^2+kn^2t)$ ή $O((kt+1+\tau(d+1)n^2+ mr)$, όπου παρατηρείται ότι η καλή επιλογή των αρχικών κέντρων μειώνει σημαντικά την πολυπλοκότητα του αλγορίθμου.

4.6Παράδειγμα Εφαρμογής Αλγορίθμου (Δίκτυο Zachary's karate club)

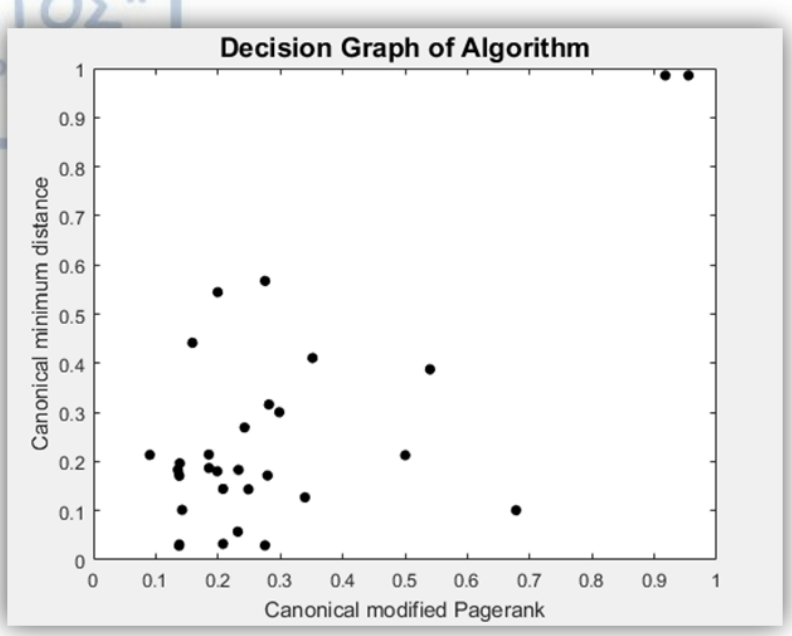
Το δίκτυο Zachary's karate club είναι ένα γνωστό κοινωνικό δίκτυο (Zachary, 1977) που μελετήθηκε από τον Wayne W. Zachary για μια περίοδο τριών ετών (1970-1972) και περιγράφει τη σχέση μεταξύ των 34 μελών μια λέσχης καράτε τα οποία έχουν φιλική σχέση εκτός λέσχης. Επιπλέον, κατά τη διάρκεια της μελέτης δημιουργήθηκε σύγκρουση μεταξύ του ιδιοκτήτη και ενός καθηγητή η οποία οδήγησε σε διαχωρισμό των μελών της λέσχης σε δύο ομάδες. Το γεγονός αυτό, όπως παρατηρήθηκε και στην ανάλυση του W. Zachary δημιούργησε μια διαμέριση του δικτύου σε δύο κοινότητες.

Η απεικόνιση του δικτύου όπου οι κόμβοι με αύξοντα αριθμό είναι τα μέλη του club και οι ακμές αντιπροσωπεύουν την φιλική σχέση μεταξύ αυτών βρίσκεται στην Εικόνα (4.3).



ΕΙΚΟΝΑ (4.3) *Η απεικόνιση του δικτύου Zachary's karate club*

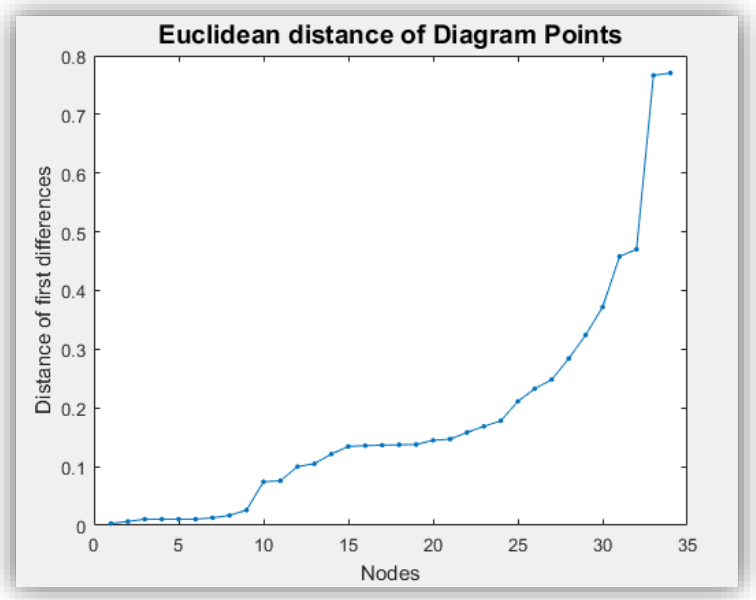
Σύμφωνα λοιπόν με τον προτεινόμενο αλγόριθμο, αρχικά υπολογίζεται η τροποποιημένη PageRank κεντρικότητα και η ελάχιστη απόσταση, από τα οποία δημιουργείται το κανονικοποιημένο διάγραμμα απόφασης των αρχικών κέντρων που απεικονίζεται στην Εικόνα (4.4).



ΕΙΚΟΝΑ (4.4) *Η απεικόνιση του γραφήματος απόφασης*

Στο διάγραμμα απόφασης ξεχωρίζουν στο πάνω και δεξιό μέρος οι κόμβοι οι οποίοι έχουν ταυτόχρονα μεγάλη ενισχυμένη PageRank κεντρικότητα αλλά και μεγάλη απόσταση από τους υπόλοιπους κόμβους μεγαλύτερης κεντρικότητας. Όπως παρατηρείται στο συγκεκριμένο διάγραμμα απόφασης, ξεχωρίζουν με μεγάλη διαφορά δύο κόμβοι, κάτι που αποτελεί ισχυρή ένδειξη ύπαρξης δύο κοινοτήτων στο δίκτυο.

Στην Εικόνα (4.5) παρουσιάζεται το διάγραμμα της Ευκλείδειας απόστασης των πρώτων διαφορών των σημείων του διαγράμματος απόφασης.



ΕΙΚΟΝΑ (4.5) *Το διάγραμμα της ευκλείδειας απόστασης*

Με βάση το διάγραμμα της Ευκλείδειας απόστασης των σημείων επιλέγεται ο αριθμός k των συστάδων ο οποίος είναι ίσος με τον αριθμό των κόμβων πριν την μεγαλύτερη «πτώση» των πρώτων διαφορών. Δηλαδή από τις πρώτες διαφορές των Ευκλείδειων αποστάσεων σύμφωνα με τον τύπο (4.9) σημειώνεται η μεγαλύτερη διαφορά, δηλαδή εκεί που το διάγραμμα έχει τη

μεγαλύτερη 'πτώση'. Και όταν εντοπίζεται η θέση αυτής γίνεται η αφαίρεση N - n (N = ο συνολικός αριθμός των κόμβων και n η θέση του κόμβου πριν την μεγαλύτερη διαφορά) και εξάγεται ο αριθμός k. Ως εκ τούτου, ο αλγόριθμος παράγει ως output τα εξής αποτελέσματα όπως φαίνονται στην Εικόνα (4.6).

```
number_k =

     2

NUMBER OF CLUSTERS: 2
THE CENTERS ARE: 1
THE CENTERS ARE: 34
```

ΕΙΚΟΝΑ (4.6) *Το εξαγόμενο αποτέλεσμα του αλγορίθμου για τον αριθμό k*

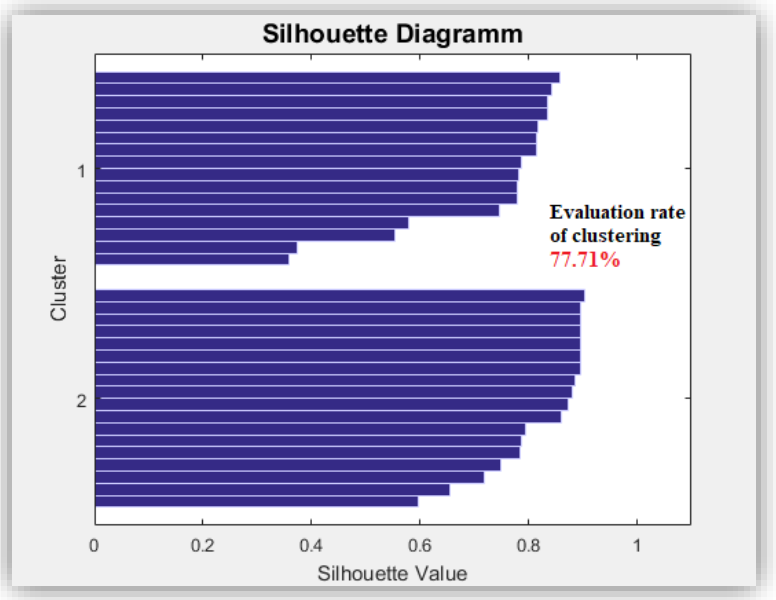
Δηλαδή, ο αριθμός των συστάδων θα είναι 2 και οι κόμβοι που ορίζονται ως αρχικά κέντρα είναι οι κόμβοι 1 και 34.

Μετά τον ορισμό των αρχικών κέντρων, ο αλγόριθμος K-Means τοποθετεί τους κόμβους του δικτύου σε δύο κοινότητες οι οποίες παρουσιάζονται στον Πίνακα (4.1), στον οποίο με έντονη γραφή είναι οι αρχικοί κόμβοι-κέντρα που επιλέχθηκαν για την εφαρμογή του αλγορίθμου.

1η Κοινότητα	**1**, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22
2η Κοινότητα	9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, **34**

ΠΙΝΑΚΑΣ (4.1) *Οι κοινότητες του karate δικτύου*

Παρατηρείται ότι η πρώτη κοινότητα αποτελείται από 16 κόμβους ενώ η δεύτερη από 18. Όσον αφορά την ποσοτική αξιολόγηση της διαμέρισης όπως έχει αναφερθεί στο Κεφάλαιο (4.3), χρησιμοποιείται το διάγραμμα Silhouette στην Εικόνα (4.7) που εμφανίζει ποσοτικά της ποιότητα της διαμέρισης.



ΕΙΚΟΝΑ (4.7) *Το διάγραμμα Silhouette της K-Means συσταδοποίησης*

Η μέση τιμή των τιμών του διαγράμματος υπολογίζεται στο 0.7771, άρα το ποσοστό αξιολόγησης της διαμέρισης του δικτύου Zachary's karate club με τον αλγόριθμο K-Means είναι 77.71%.

5. Σύγκριση με τον κλασικό Αλγόριθμο μεγιστοποίησης του modularity

Ένας από τους κλασικότερους και πολυχρησιμοποιημένους αλγορίθμους για την εύρεση κοινοτήτων είναι ο Αλγόριθμος «Greedy agglomerative modularity maximization method» που προτάθηκε από τον Newman (Newman, 2004). Ο αλγόριθμος βασίζεται στο modularity το οποίο ορίζεται από τον τύπο (2.1) και αποτελεί ένα μέτρο διαμέρισης του δικτύου. Θεωρητικά, στα τυχαία δίκτυα το modurarity είναι Q≈0 και υπάρχει η παραδοχή ότι εάν το Q ενός δικτύου είναι μεγαλύτερο από 0.3, το δίκτυο παρουσιάζει δομή κοινοτήτων.

Έτσι λοιπόν ο στόχος είναι να βελτιστοποιηθεί το modularity κάνοντας όλες τις πιθανές διαμερίσεις στο δίκτυο, κάτι όμως που απαιτεί μεγάλη υπολογιστική ισχύ. Για να ξεπεραστεί το πρόβλημα με την υπολογιστική πολυπλοκότητα χρησιμοποιείται η προσεγγιστική αθροιστική ιεραρχική μέθοδος “greedy optimization agglomerative hierarchical algorithm” με την οποία κάθε κόμβος είναι αρχικά μια κοινότητα και στη συνέχεια ενώνονται οι κόμβοι που προκαλούν την μεγαλύτερη αύξηση του modularity, όταν βέβαια υπάρχει ακμή που τους συνδέει. Η επαναληπτική μέθοδος του αλγορίθμου σταματάει όταν δεν υπάρχει πια άλλη αύξηση του modularity με όποια ακμή και αν διαλέγει. Τέλος, η υπολογιστική πολυπλοκότητα του αλγορίθμου είναι O((m+n)n) όπου n και m ο αριθμός των κόμβων και των ακμών αντίστοιχα.

5.1 Παραδείγματα Σύγκρισης

Θα πραγματοποιηθεί η σύγκριση της εύρεσης κοινοτήτων σε έξι δίκτυα που αντιπροσωπεύονται από τους αντίστοιχους πίνακες γειτνίασης στα οποία γνωρίζεται εκ των προτέρων η δομή των κοινοτήτων. Στο πρώτο παράδειγμα θα πραγματοποιηθεί αναλυτική επεξήγηση των αποτελεσμάτων του αλγορίθμου, ενώ στα επόμενα θα παρουσιαστούν με κάποια σχόλια τα αποτελέσματα.

Η σύγκριση θα είναι μεταξύ του προτεινόμενου αλγορίθμου με τον κλασικό αλγόριθμο εύρεσης κοινοτήτων που χρησιμοποιεί τη μέθοδο βελτιστοποίησης Greedy (Newman, 2004). Ως ποσοτικά μέτρα σύγκρισης με την πραγματική αλήθεια θα χρησιμοποιηθούν ο δείκτης γραμμικής συσχέτισης Pearson (Pearson, 1948) που ορίζεται από τον τύπο (5.1) και ο δείκτης ομοιότητας Jaccard (Niwattanakul et al., 2013) που ορίζεται από τον τύπο (5.2). Επίσης θα γίνει και η ποσοτική αξιολόγηση της δομής των κοινοτήτων από τον δείκτη Silhouette.

$$r = \frac{\sum_{i-1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Τύπος (5.1)

Όπου n το μέγεθος του δείγματος των παρατηρήσεων, x_i και y_i οι παρατηρήσεις και $\bar{x}$, $\bar{y}$ οι μέσες τιμές των παρατηρήσεων x και y αντίστοιχα.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

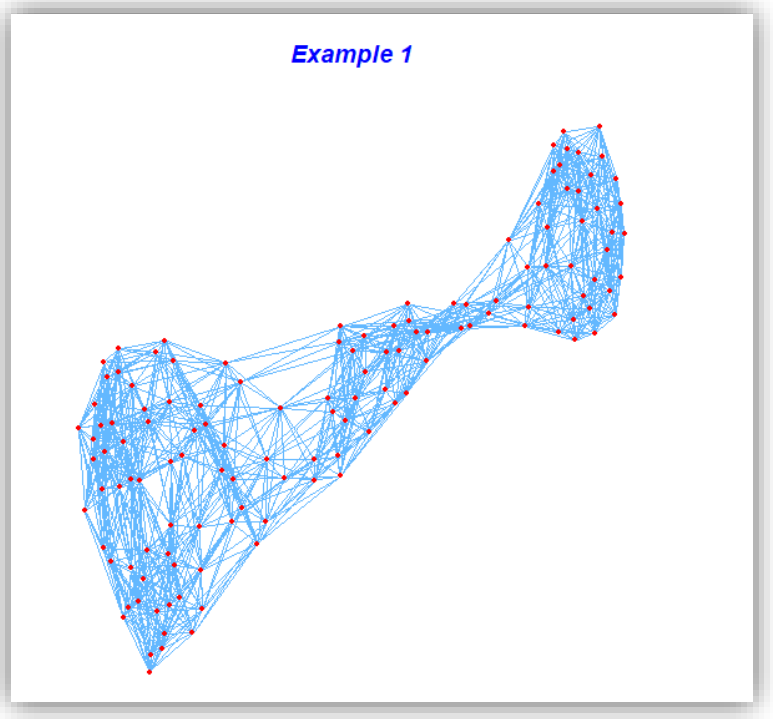
Τύπος (5.2)

Όπου Α και Β τα δύο σύνολα των παρατηρήσεων.

Τέλος, θα γίνει η σύγκριση των δύο μεθόδων στο πραγματικό δίκτυο Dolphin social network (Lusseau, 2007).

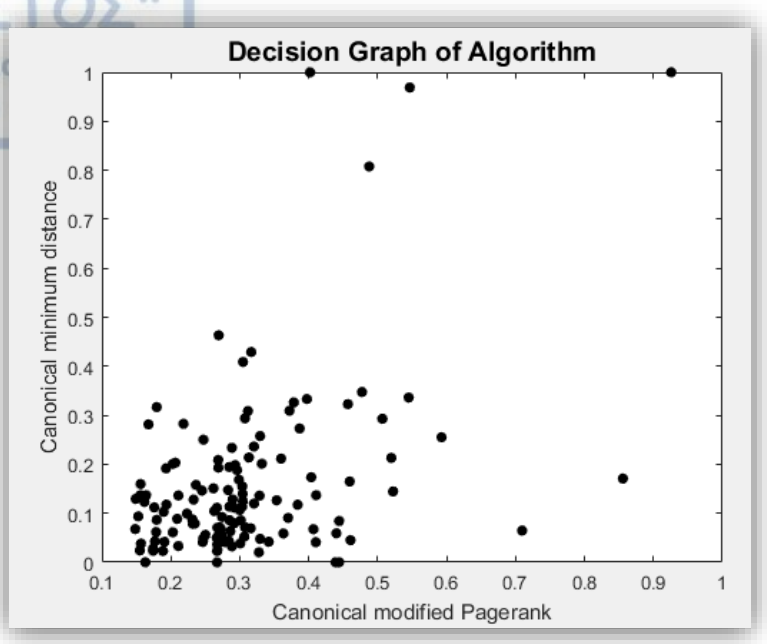
5.1.1 (Adj1: 130x130)

Το πρώτο παράδειγμα αναφέρεται σε δίκτυο με 130 κόμβους και 1199 ακμές. Το δίκτυο κατασκευάστηκε μέσω γεννήτριας σημείων p_i, με i=1,..,130, από κανονική δισδιάστατη κατανομή με γνώση εκ των προτέρων της δομής κοινοτήτων, όπου τα πρώτα 40 στοιχεία επιλέχθηκαν στην πρώτη κοινότητα ή κλάση, τα επόμενα 30 στη δεύτερη και τα υπόλοιπα 60 στην τρίτη κλάση. Τα σημεία των κλάσεων ορίζονται αρχικά στο επίπεδο όπου η πρώτη κλάση έχει κέντρο το (0,0), η δεύτερη το (2.5,2.5) και η τρίτη το (5,5). Προφανώς ισχύει ότι όσο πιο κοντά μεταξύ τους είναι τα κέντρα τόσο δυσκολότερο είναι να ξεχωρίσουν οι κλάσεις. Τα σημεία σε κάθε κλάση δίνονται από κανονική δισδιάστατη κατανομή με τα δεδομένα κέντρα και πίνακα συν διασπορών $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, με διασπορά 1 κάθε μιας από τις δύο μεταβλητές και μηδενική συσχέτιση μεταξύ των στοιχείων σε κάθε μια από τις δύο διαστάσεις. Τέλος, για κάθε σημείο εντοπίζονται οι 15 κοντινότεροι γείτονες και με βάση τα γειτονικά σημεία ορίζεται ο πίνακας γειτνίασης όπου για κάθε i γραμμή του πίνακα που αντιστοιχεί στο σημείο p_i, υπάρχουν 15 μονάδες στα κελιά που αντιστοιχούν στους δείκτες των 15 γειτονικών σημείων και τα υπόλοιπα είναι 0. Έτσι, η απεικόνιση του κανονικού δικτύου με χρήση της R είναι στην Εικόνα (5.1).



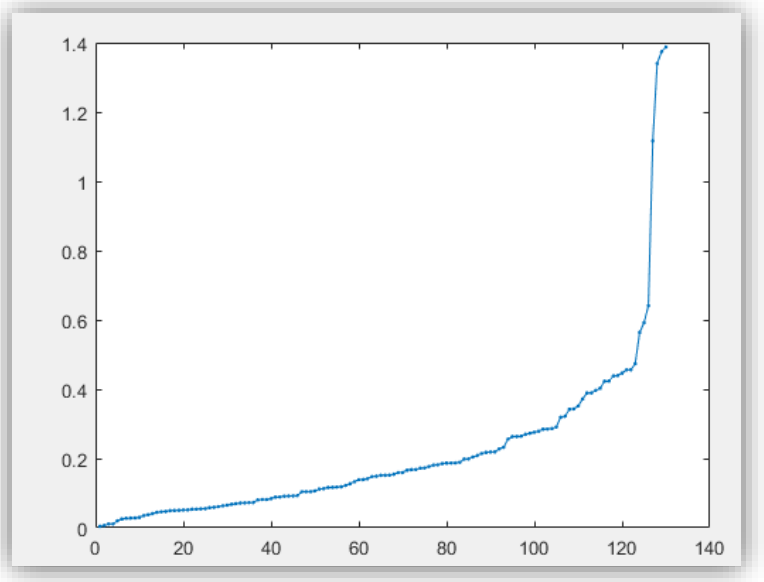
ΕΙΚΟΝΑ (5.1) *Απεικόνιση του δικτύου Adj1*

Ο αλγόριθμος υπολογίζει τη μέση απόσταση του κάθε κόμβου (τεταγμένη στο διάγραμμα απόφασης) από τους υπόλοιπους του δικτύου οι οποίοι έχουν μεγαλύτερη PageRank κεντρικότητα. Επίσης υπολογίζει την PageRank κεντρικότητα του κάθε κόμβου ενισχυμένη με την απόσταση του κόμβου από τους υπόλοιπους του δικτύου (τετμημένη στο διάγραμμα απόφασης). Ως αποτέλεσμα, παράγεται το διάγραμμα απόφασης που απεικονίζεται στην Εικόνα (5.2).



ΕΙΚΟΝΑ (5.2) *Το Διάγραμμα απόφασης*

Από το διάγραμμα διακρίνονται οι κόμβοι που ξεχωρίζουν οι οποίοι είναι κυρίως τέσσερις. Για την επιλογή των αρχικών κέντρων και κατ' επέκταση του αριθμού k χρησιμοποιείται η Ευκλείδεια απόσταση του κάθε κόμβου του διαγράμματος από την αρχή των αξόνων. Για τον λόγο αυτό, παρουσιάζονται οι πρώτες διαφορές σε αύξουσα σειρά στην Εικόνα (5.3).

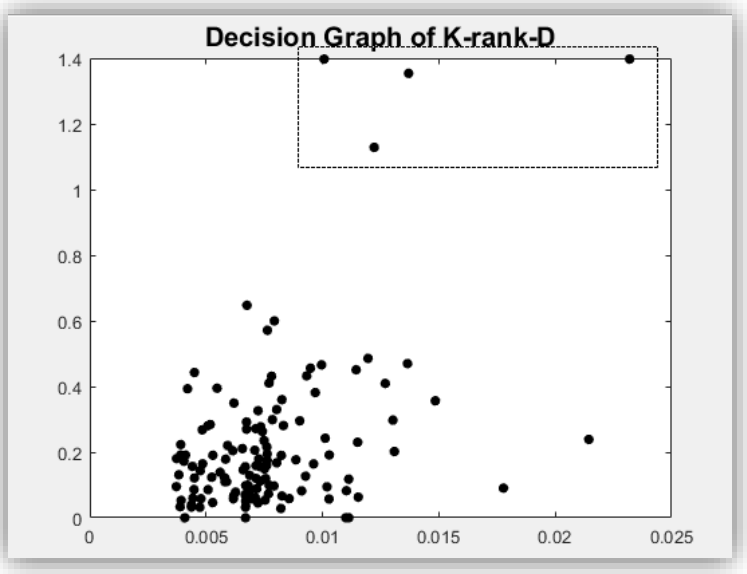


ΕΙΚΟΝΑ (5.3) *Το διάγραμμα της Ευκλείδειας απόστασης των σημείων του διαγράμματος απόφασης σε αύξουσα σειρά*

Από τις πρώτες διαφορές των Ευκλείδειων αποστάσεων σημειώνεται η μεγαλύτερη διαφορά, δηλαδή εκεί που το διάγραμμα έχει τη μεγαλύτερη 'πτώση'. Έτσι, ονομάζοντας number_k τη διαφορά N - n (N = ο συνολικός αριθμός των κόμβων και n η θέση του κόμβου πριν την μεγαλύτερη διαφορά) εξάγεται μέσω του αλγορίθμου ο αριθμός k=4 των συστάδων.

Βέβαια να σημειωθεί ότι στο number_k δεν επιλέγεται πάντα ο σωστός αριθμός και αυτή η λύση αποτελεί μια τοπικά βέλτιστη και όχι ολικά. Για το συγκεκριμένο θέμα μπορεί να υπάρξει περαιτέρω ανάλυση.

Έτσι λοιπόν στο διάγραμμα επιλέγονται οι 4 κόμβοι με την μεγαλύτερη Ευκλείδεια απόσταση, όπου ο αριθμός 4 αντιστοιχεί και στις εξαγόμενες συστάδες σύμφωνα με την Εικόνα (5.4).



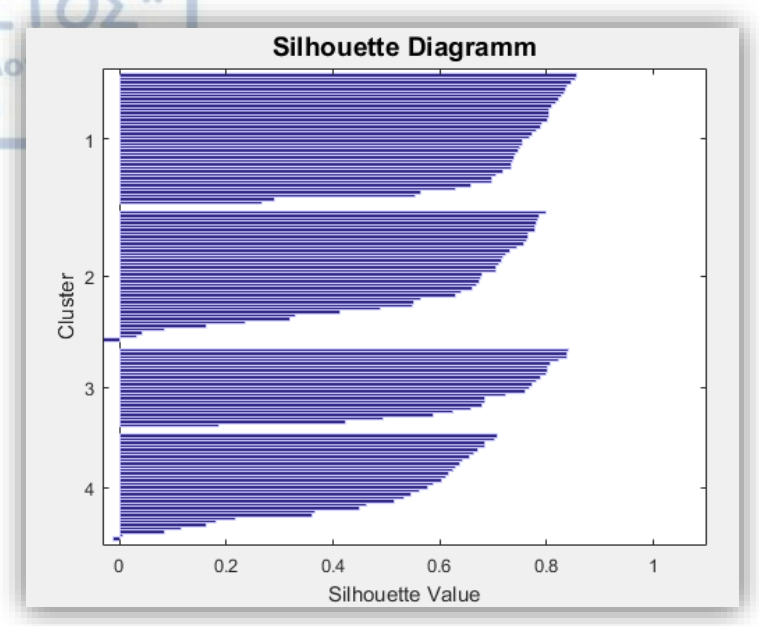
ΕΙΚΟΝΑ (5.4) *Η επιλογή των αρχικών κόμβων-κέντρων*

Και τα αποτελέσματα του K-means για τον εντοπισμό κοινοτήτων, όπου με έντονο είναι τα επιλεγμένα κέντρα κάθε συστάδας, είναι στον Πίνακα (5.1).

1η Κοινότητα	1, 2, 3, 4, **5**, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 38, 39, 40
2η Κοινότητα	36, 37, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, **53**, 54, 55, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 70, 71, 91, 98, 110, 111, 128, 130
3η Κοινότητα	**74**, 76, 78, 79, 80, 81, 84, 85, 88, 89, 90, 92, 95, 101, 108, 115, 116, 118, 121, 123, 125, 127, 129
4η Κοινότητα	63, 72, 73, 75, 77, 82, 83, 86, 87, 93, 94, 96, 97, 99, 100, 102, **103**, 104, 105, 106, 107, 109, 112, 113, 114, 117, 119, 120, 122, 124, 126

ΠΙΝΑΚΑΣ (5.1) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο K-Means*

Η γραμμική συσχέτιση Pearson των κοινοτήτων με την πραγματική αλήθεια είναι σε ποσοστό 90.91% και η Jaccard ομοιότητα είναι 55.8% καθώς 99 κόμβοι ταξινομήθηκαν σωστά. Επίσης, το διάγραμμα Silhouette του K-means, που απεικονίζει το πόσο κοντά το κάθε σημείο μιας συστάδας είναι με αυτά των γειτονικών συστάδων είναι στην Εικόνα (5.5).



ΕΙΚΟΝΑ (5.5) *Το διάγραμμα Silhouette της K-means συσταδοποίησης*

Η μέση τιμή του διαχωρισμού (mean Silhouette Value) είναι 0.6214, δηλαδή η ποσοτική αξιολόγηση του διαχωρισμού είναι 62.14%.

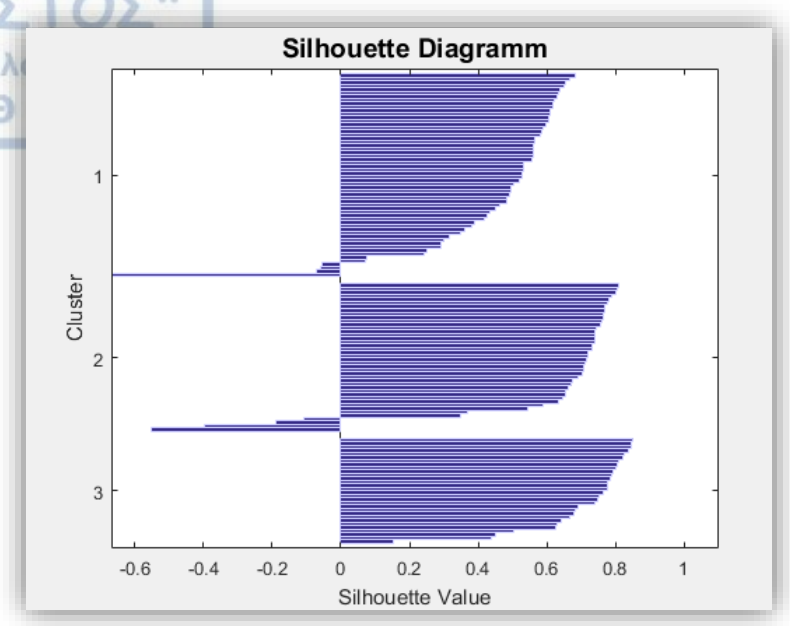
Αντίστοιχα, τα αποτελέσματα των κοινοτήτων του αλγορίθμου για modularity (ModulMax3) για το δίκτυο Adj1 είναι στον Πίνακα (5.2).

1η Κοινότητα	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 38, 39, 40,
2η Κοινότητα	21, 32, 36, 37, 41, 42, 43, 44, 45, 46, 47, 48, 51, 52, 55, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 70, 110, 111, 113
3η Κοινότητα	49, 50, 53, 54, 63, 71, 74, 76, 78, 79, 80, 81, 84, 85, 87, 89, 90, 91, 92, 95, 98, 101, 104, 105, 108, 115, 116, 117, 118, 121, 123, 125, 127, 128, 130,
4η Κοινότητα	72, 73, 75, 77, 82, 83, 86, 88, 93, 94, 96, 97, 99, 100, 102, 103, 106, 107, 109, 112, 114, 119, 120, 122, 124, 126, 129,

ΠΙΝΑΚΑΣ (5.2) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο Greedy*

Η συσχέτιση Pearson με την πραγματική αλήθεια των κοινοτήτων είναι 86.06% και η ομοιότητα Jaccard είναι 52.23%, καθώς 91 κόμβοι αντιστοιχίστηκαν σωστά. Η συσχέτιση Pearson μεταξύ των δύο μεθόδων βρέθηκε 94.61%.

Το διάγραμμα της ποσοτικής αξιολόγησης του modularity maximization αλγορίθμου για το δεύτερο δίκτυο με την μέθοδο Silhouette είναι στην Εικόνα (5.6).



ΕΙΚΟΝΑ (5.6) *Το διάγραμμα Silhouette της Greedy διαμέρισης*

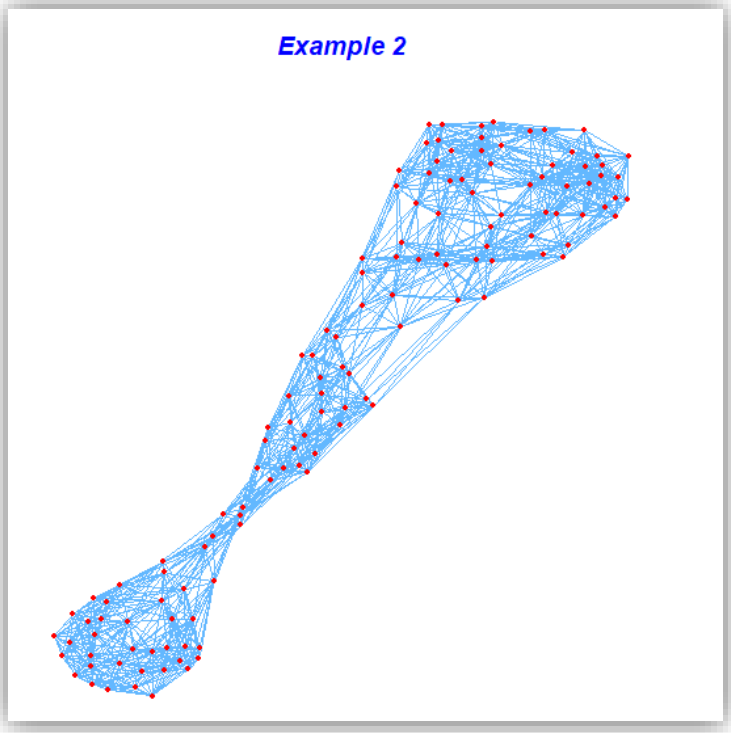
Και η μέση τιμή του διαχωρισμού (mean Silhouette Value) είναι 0.5862, δηλαδή η ποσοτική αξιολόγηση του διαχωρισμού είναι 58.82%. Αυτό σημαίνει ότι με την προτεινόμενη μέθοδο K-Means επιτεύχθηκε καλύτερη συσταδοποίηση διότι ισχύει 62.14 >58.82.

Τέλος, να τονιστεί πως παρόλο που ο αριθμός των κλάσεων είναι ο ίδιος, η ποσοτική αξιολόγηση του διαχωρισμού είναι διαφορετική στις δύο περιπτώσεις διότι οι κόμβοι που ανήκουν στην ίδια κλάση διαφοροποιούνται μεταξύ των δύο μεθόδων.

5.1.2 (Adj2: 130x130)

Το δεύτερο δίκτυο έχει 130 κόμβους και 1199 ακμές και κατασκευάστηκε από γεννήτρια σημείων στο δισδιάστατο χώρο. Τα πρώτα 40 στοιχεία είναι στην πρώτη κλάση, τα επόμενα 30 στη δεύτερη και τα υπόλοιπα 60 στην τρίτη κλάση. Τα σημεία των κλάσεων ορίζονται αρχικά στο επίπεδο όπου η πρώτη κλάση έχει κέντρο (0,0), η δεύτερη (2.5,2.5) και η τρίτη (5,5). Τα σημεία σε κάθε κλάση δίνονται από κανονική δισδιάστατη κατανομή με πίνακα συν διασπορών $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Όμως στη συνέχεια οι τιμές των συντεταγμένων των σημείων υψώνονται στο τετράγωνο με αποτέλεσμα να αλλάζει η κατανομή από κανονική σε X^2 με 1 βαθμό ελευθερίας. Αυτό αποτελεί ένα πιο δύσκολο σενάριο για τον εντοπισμό των κοινοτήτων από το προηγούμενο που τα σημεία κατασκευάστηκαν από κανονική κατανομή. Τέλος, για κάθε σημείο εντοπίζονται οι 15 κοντινότεροι γείτονες, ορίζεται ο πίνακας γειτνίασης και κατασκεύζεται το κανονικό δίκτυο το οποίο απεικονίζεται στην Εικόνα (5.7).



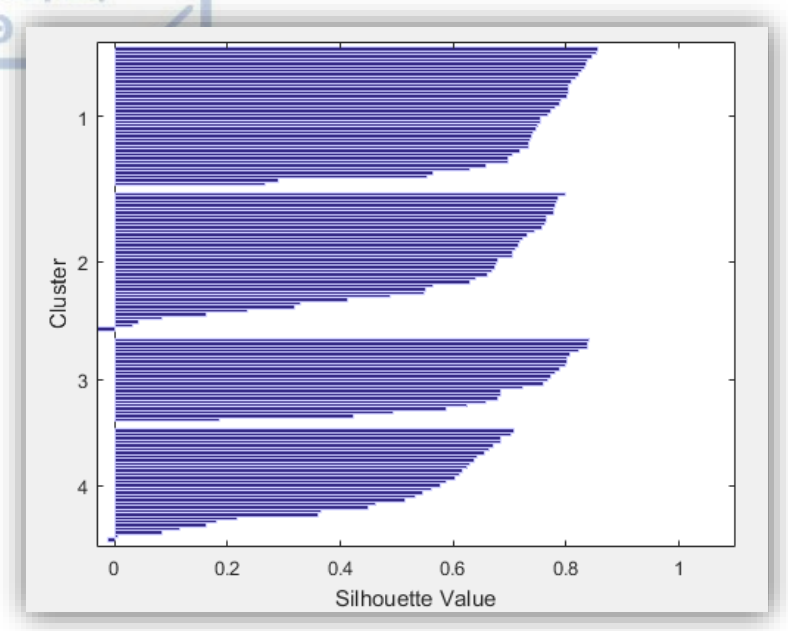
ΕΙΚΟΝΑ (5.7) *Απεικόνιση του δικτύου Adj2*

Οι εντοπισμένες κοινότητες με την K-Means είναι στον Πίνακα (5.3).

1η Κοινότητα	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 38, 39, 40
2η Κοινότητα	36, 37, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 70, 71, 91, 98, 110, 111, 128, 130
3η Κοινότητα	74, 76, 78, 79, 80, 81, 84, 85, 88, 89, 90, 92, 95, 101, 108, 115, 116, 118, 121, 123, 125, 127, 129
4η Κοινότητα	63, 72, 73, 75, 77, 82, 83, 86, 87, 93, 94, 96, 97, 99, 100, 102, 103, 104, 105, 106, 107, 109, 112, 113, 114, 117, 119, 120, 122, 124, 126

ΠΙΝΑΚΑΣ (5.3) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο K-Means*

Το διάγραμμα της ποσοτικής αξιολόγησης του K-Means για το δεύτερο δίκτυο με την μέθοδο Silhouette είναι στην Εικόνα (5.8).



ΕΙΚΟΝΑ (5.8) *Το διάγραμμα Silhouette της K-means συσταδοποίησης*

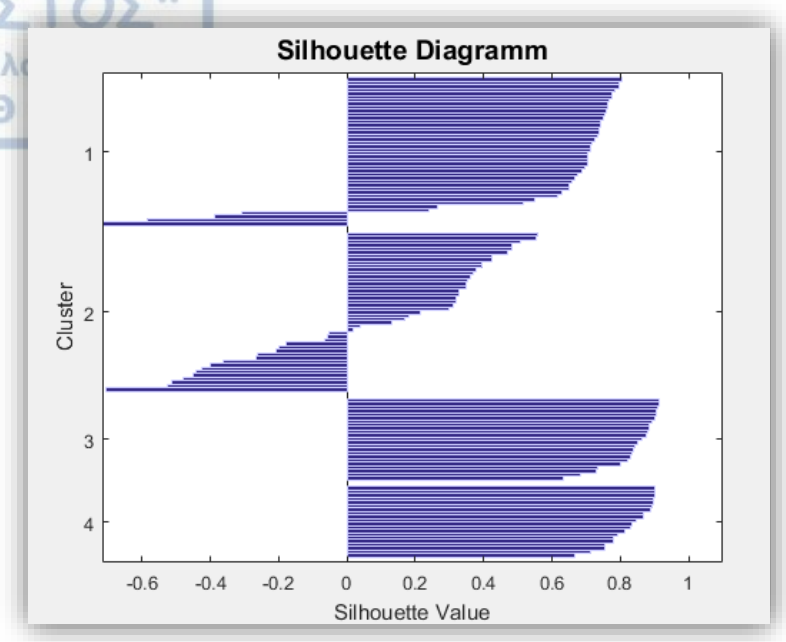
Η ποσοτική αξιολόγηση του διαχωρισμού είναι 62.14%, η συσχέτιση Pearson των παραγόμενων κοινοτήτων με τις πραγματικές είναι 89.77% και η Jaccard ομοιότητα είναι 55.41% καθώς 97 κόμβοι αντιστοιχίστηκαν σωστά.

Από την άλλη, τα αποτελέσματα του αλγορίθμου Greedy agglomerative modularity maximization method (ModulMax3) είναι στον Πίνακα (5.4).

1η Κοινότητα	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 22, 23, 24, 25, 26, 27, 28, 29, 30, 33, 34, 35, 38, 39, 40
2η Κοινότητα	21, 31, 32, 36, 37, 41, 42, 43, 44, 45, 46, 47, 48, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66, 67, 68, 69, 70, 106, 107, 109, 110, 127, 128
3η Κοινότητα	49, 50, 51, 63, 71, 74, 76, 78, 79, 80, 81, 84, 85, 87, 89, 90, 91, 92, 95, 98, 101, 104, 105, 108, 112, 113, 115, 116, 117, 118, 121, 123, 125, 130
4η Κοινότητα	72, 73, 75, 77, 82, 83, 86, 88, 93, 94, 96, 97, 99, 100, 102, 103, 106, 107, 109, 111, 112, 114, 119, 120, 122, 124, 126, 129

ΠΙΝΑΚΑΣ (5.4) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο Greedy*

Η συσχέτιση Pearson των παραγόμενων κοινοτήτων με τις πραγματικές είναι 79.77%, η Jaccard ομοιότητα 49.42% αφού 91 κόμβοι ταξινομήθηκα σωστά. Ακόμη, το διάγραμμα της ποσοτικής αξιολόγησης του modularity maximization αλγορίθμου για το δεύτερο δίκτυο με την μέθοδο Silhouette είναι στην Εικόνα (5.9) από όπου και οπτικά διαφαίνεται ότι ο αλγόριθμος δεν πραγματοποίησε καλή συσταδοποίηση.

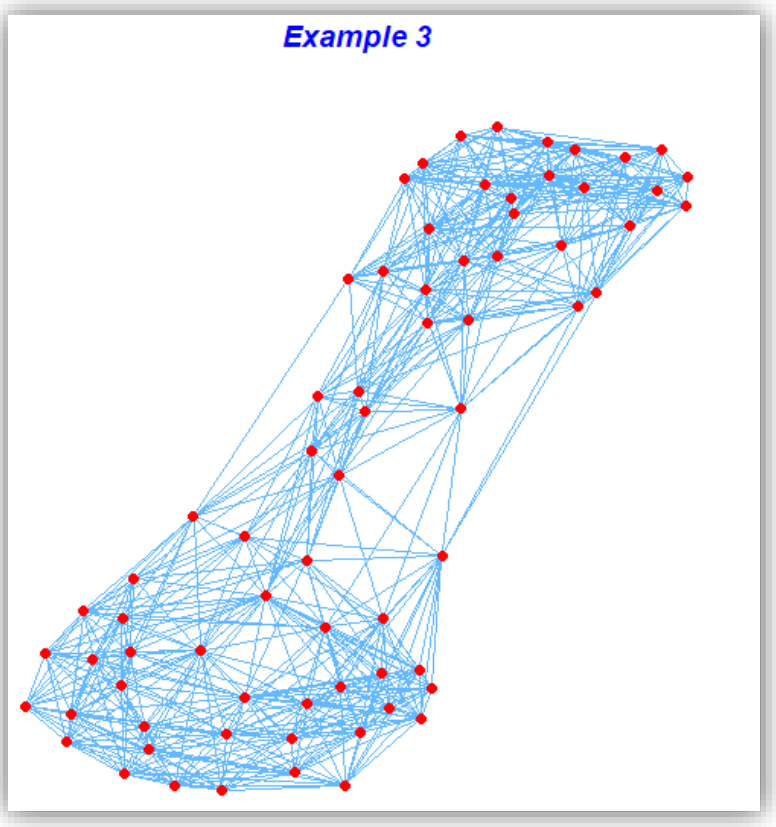


ΕΙΚΟΝΑ (5.9) *Το διάγραμμα Silhouette της Greedy διαμέρισης*

Και η ποσοτική αξιολόγηση του διαχωρισμού είναι 49.24%, δηλαδή πολύ χαμηλότερη από του K-Means που ήταν 62.14%.

5.1.3 (Adj3: 70x70)

Το τρίτο παράδειγμα αναφέρεται σε δίκτυο με 70 κόμβους και 632 ακμές. Το δίκτυο κατασκευάστηκε από γεννήτρια σημείων στο δισδιάστατο χώρο, όπου τα πρώτα 40 στοιχεία επιλέγονται στην πρώτη κλάση και τα επόμενα 30 στη δεύτερη. Τα σημεία των κλάσεων ορίζονται αρχικά στο επίπεδο με την πρώτη κλάση να έχει κέντρο (0,0) και η δεύτερη (2.5,2.5) αντίστοιχα. Τα σημεία σε κάθε κλάση δίνονται από κανονική δισδιάστατη κατανομή με πίνακα συν διασπορών $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Τέλος, για κάθε σημείο εντοπίζονται οι 15 κοντινότεροι γείτονες και με βάση τα γειτονικά σημεία ορίζεται ο πίνακας γειτνίασης που αντιπροσωπεύει το κανονικό δίκτυο που απεικονίζεται στην Εικόνα (5.10).



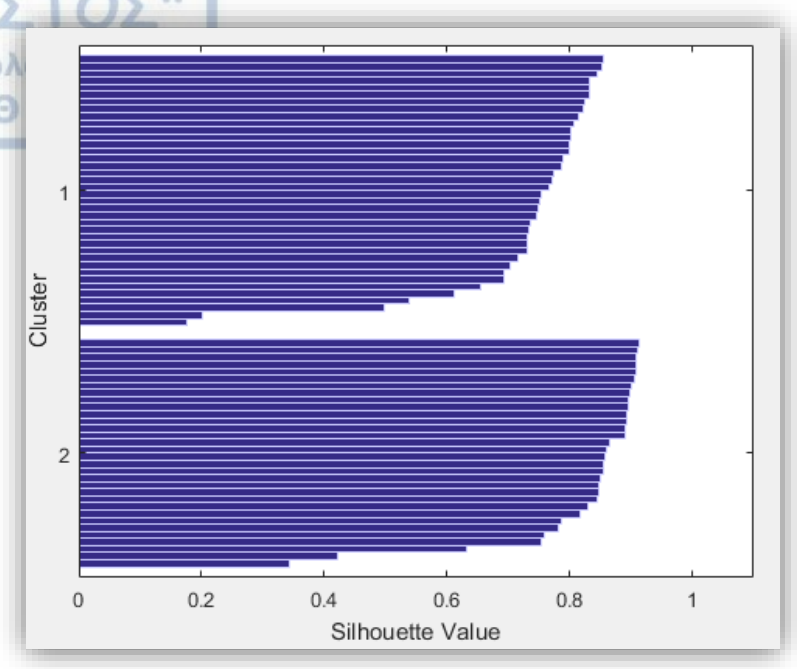
ΕΙΚΟΝΑ (5.10) *Απεικόνιση του δικτύου Adj3*

Οι εντοπισμένες κοινότητες με τον K-Means αλγόριθμο είναι στον Πίνακα (5.5).

1η Κοινότητα	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 38, 39, 40,
2η Κοινότητα	36, 37, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70,

ΠΙΝΑΚΑΣ (5.5) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο K-Means*

Η συσχέτιση Pearson με την πραγματική αλήθεια των κοινοτήτων είναι 94.37% και η Jaccard ομοιότητα 89.33% καθώς 68 κόμβοι ταξινομήθηκαν σωστά και το διάγραμμα Silhouette της K-means για το τρίτο δίκτυο βρίσκεται στην Εικόνα (5.11).



ΕΙΚΟΝΑ (5.11) *Το διάγραμμα Silhouette της K-means συσταδοποίησης*

Και η ποσοτική αξιολόγηση του διαχωρισμού που απεικονίζεται στο διάγραμμα είναι αρκετά καλή και είναι 77.13%.

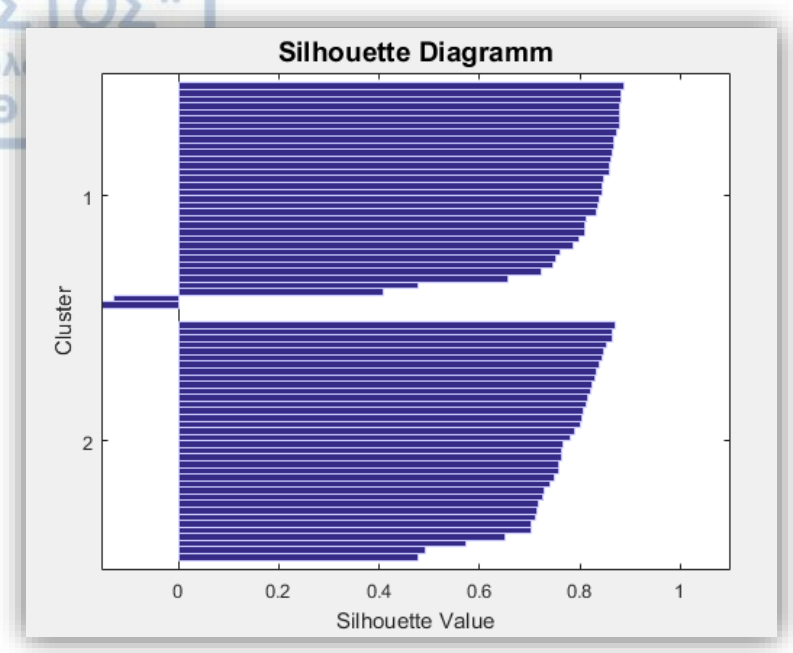
Αντίθετα, τα αποτελέσματα των κοινοτήτων του αλγορίθμου Greedy agglomerative modularity maximization method είναι στον Πίνακα (5.6).

1η Κοινότητα	1, 2, 3, 4, 5, 6, 7, 8 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 38, 39, 40,
2η Κοινότητα	21, 32, 36, 37, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70,

ΠΙΝΑΚΑΣ (5.6) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο Greedy*

Η συσχέτιση Pearson με την πραγματική αλήθεια των κοινοτήτων είναι 89.11%, η Jaccard ομοιότητα 80.22% καθώς 66 κόμβοι ταξινομήθηκαν σωστά και η συσχέτιση μεταξύ των δύο μεθόδων εντοπισμού κοινοτήτων είναι 94.43%.

Το διάγραμμα Silhouette της modularity maximization για το τρίτο δίκτυο βρίσκεται στην Εικόνα (5.12).

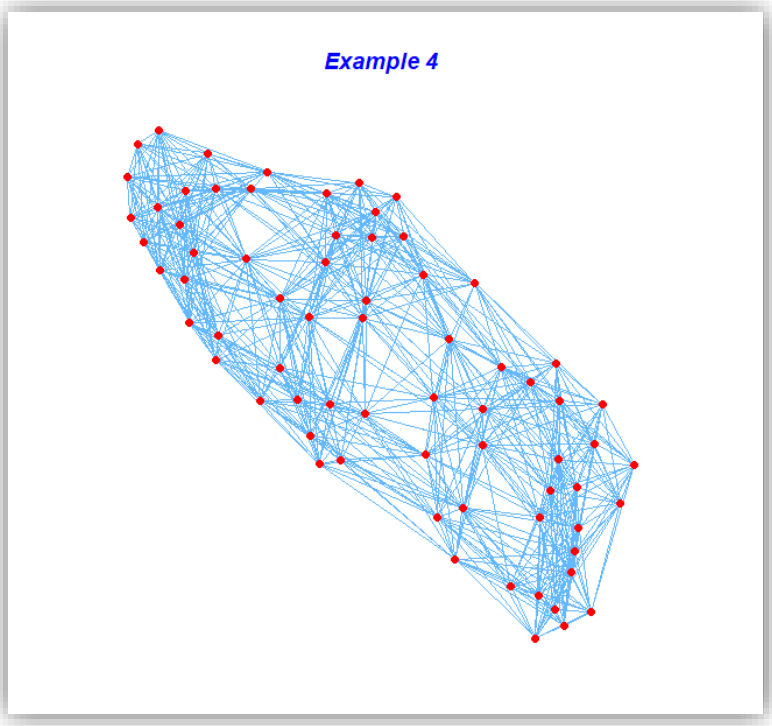


ΕΙΚΟΝΑ (5.12) *Το διάγραμμα Silhouette της Greedy διαμέρισης*

Και η ποσοτική αξιολόγηση του διαχωρισμού είναι 75.60%, δηλαδή λίγο χαμηλότερη από αυτή της K-Means που ήταν 77.13%.

5.1.4 (Adj4: 70x70)

Το τέταρτο παράδειγμα αναφέρεται σε δίκτυο με 70 κόμβους και 703 ακμές το οποίο κατασκευάστηκε κι αυτό μέσω γεννήτριας σημείων από κανονική δισδιάστατη κατανομή. Τα πρώτα 40 στοιχεία επιλέχθηκαν στην πρώτη κλάση και τα υπόλοιπα 30 στη δεύτερη. Η πρώτη κλάση έχει κέντρο (0,0) και η δεύτερη (2.5,2.5) με πίνακα συν διασπορών $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Επιπρόσθετα. οι τιμές υψώνονται στο τετράγωνο με αποτέλεσμα να αλλάζει η κατανομή από κανονική σε X^2 κατανομή με 1 βαθμό ελευθερίας. Για την κατασκευή του δικτύου εντοπίζονται οι 15 κοντινότεροι γείτονες και με βάση τα γειτονικά σημεία ορίζεται ο πίνακας γειτνίασης που αντιπροσωπεύει το δίκτυο της Εικόνα (5.13).



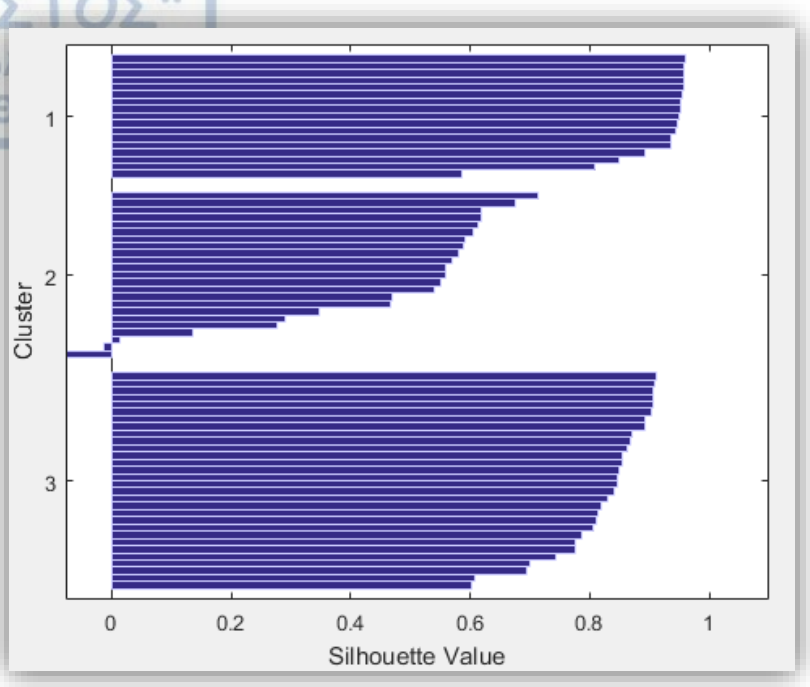
ΕΙΚΟΝΑ (5.13) *Απεικόνιση του δικτύου Adj4*

Οι εντοπισμένες κοινότητες με τον K-Means αλγόριθμο είναι στον Πίνακα (5.7).

1η Κοινότητα	1, 2, 3, 5, 7, 9, 11, 13, 14, 15, 19, 20, 21, 24, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40, 46, 69
2η Κοινότητα	4, 6, 8, 10, 16, 17, 18, 22, 25, 26, 27, 28, 29, 31
3η Κοινότητα	12, 23, 41, 42, 43, 44, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 70

ΠΙΝΑΚΑΣ (5.7) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο K-Means*

Η συσχέτιση Pearson με την πραγματική αλήθεια των κοινοτήτων είναι 80.84%, η Jaccard ομοιότητα 56.63% καθότι 52 από τους κόμβους ταξινομήθηκαν σωστά και το διάγραμμα Silhouette της K-means για το τέταρτο δίκτυο βρίσκεται στην Εικόνα (5.14).



ΕΙΚΟΝΑ (5.14) *Το διάγραμμα Silhouette της K-means συσταδοποίησης*

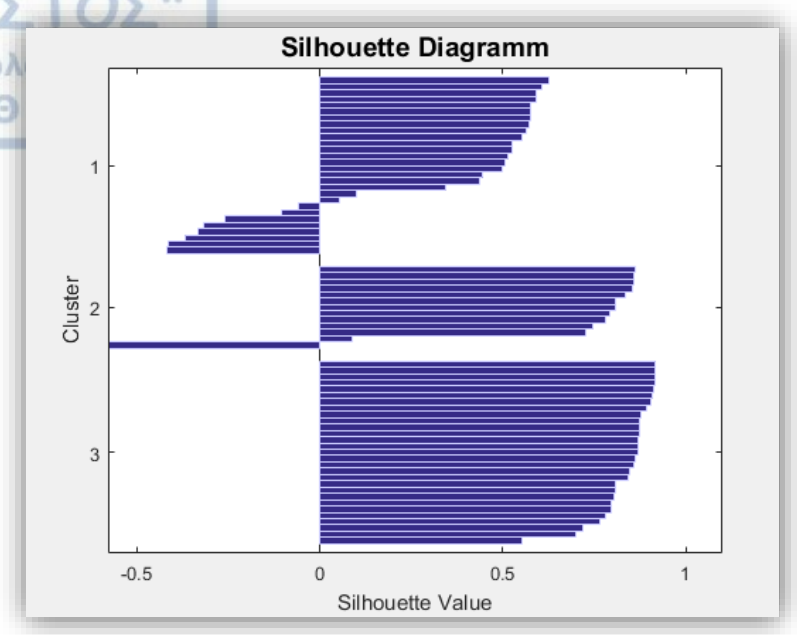
Και η ποσοτική αξιολόγηση της συσταδοποίησης είναι 72.12%.

Τα αποτελέσματα του αλγορίθμου μεγιστοποίησης modularity (ModulMax3) είναι στον Πίνακα (5.8).

1η Κοινότητα	1, 2, 4, 5, 6, 7, 8, 10, 13, 14, 15, 16, 17, 18, 25, 26, 27, 28, 29, 31, 32, 33, 36, 37, 45, 63, 69
2η Κοινότητα	3, 9, 19, 20, 21, 22, 24, 30, 34, 35, 38, 39, 40 , 41, 42, 43, 44, 46, 48, 49, 50, 61
3η Κοινότητα	11, 12, 23, 47, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 62, 63, 64, 65, 66, 67, 68, 70

ΠΙΝΑΚΑΣ (5.8) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο Greedy*

Η συσχέτιση Pearson με την πραγματική αλήθεια είναι 72.5% και η Jaccard ομοιότητα 40.35% καθότι 44 από τους κόμβους ταξινομήθηκαν σωστά και η συσχέτιση μεταξύ των δύο μεθόδων είναι 84.39%. Ακόμη, το διάγραμμα Silhouette της modularity maximization βρίσκεται στην Εικόνα (5.15).

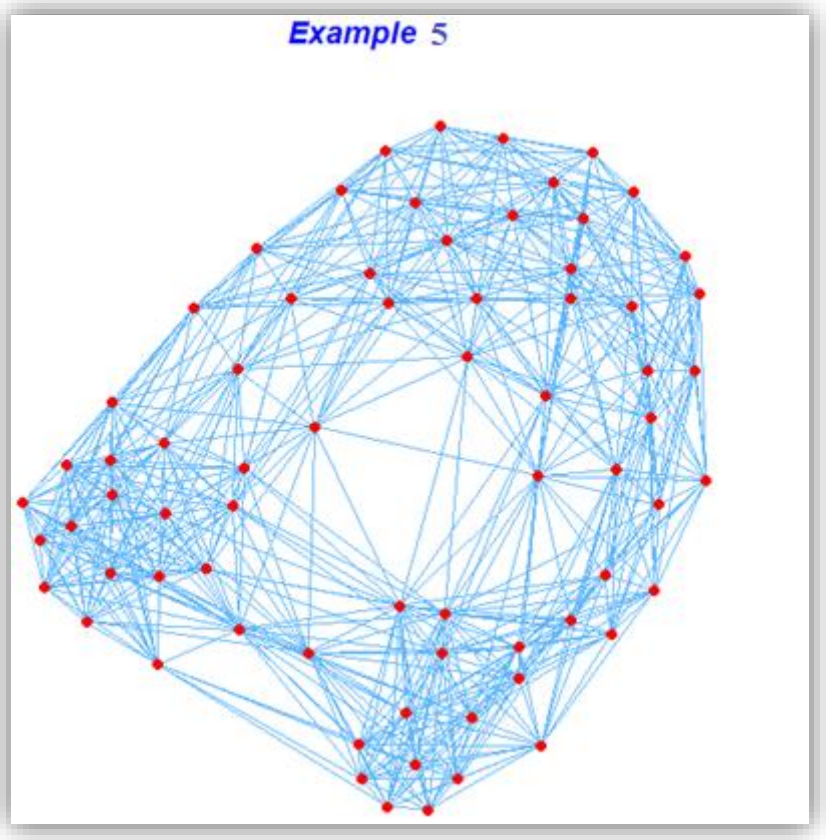


ΕΙΚΟΝΑ (5.15) *Το διάγραμμα Silhouette της Greedy διαμέρισης*

Και η ποσοτική αξιολόγηση του διαχωρισμού είναι 57.51%, όπου είναι σημαντικά χαμηλότερη αυτής της K-Means η οποία ήταν 72.82%.

5.1.5 (Adj5: 70x70)

Το πέμπτο παράδειγμα αναφέρεται σε δίκτυο με 70 κόμβους και 655 ακμές. Τα πρώτα 40 στοιχεία είναι στην πρώτη κλάση και τα επόμενα 30 στη δεύτερη όπου η πρώτη κλάση έχει κέντρο (2.5,2.5) και η δεύτερη (2.5,2.5) αντίστοιχα. Δηλαδή, οι δύο κλάσεις έχουν τα ίδιο κέντρο και παρόμοια διασπορά των σημείων, κάτι που καθιστά την δημιουργία συστάδων τυχαία. Τα σημεία σε κάθε κλάση δίνονται από κανονική δισδιάστατη κατανομή με τα δεδομένα κέντρα και πίνακα συν διασπορών $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Το κανονικό δίκτυο κατασκευάζεται όπως και τα προηγούμενα με τον αντίστοιχο πίνακα γειτνίασης και η απεικόνισή του είναι στην Εικόνα (5.16).



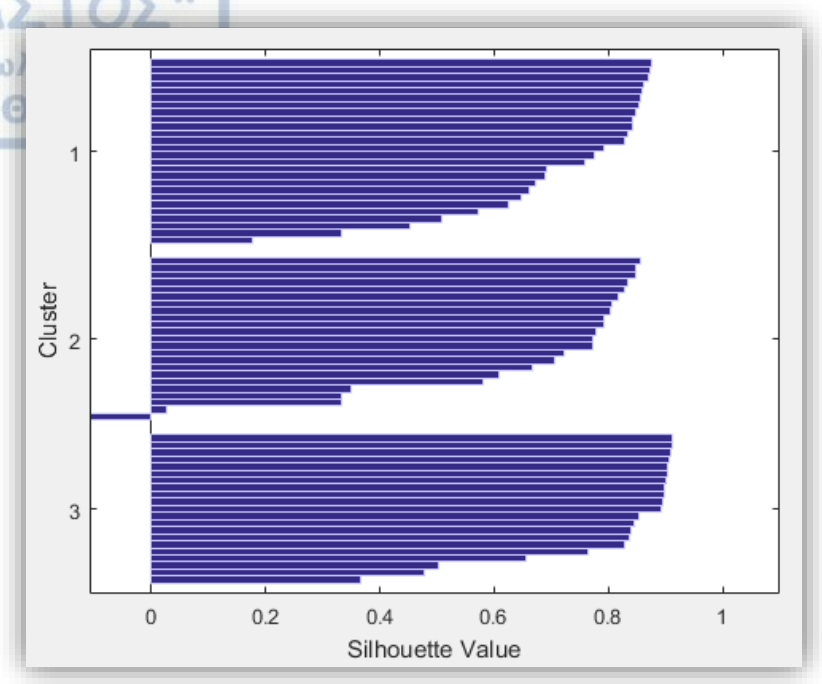
ΕΙΚΟΝΑ (5.16) *Απεικόνιση του δικτύου Adj5*

Οι εντοπισμένες κοινότητες με την K-Means είναι στον Πίνακα (5.9).

1η Κοινότητα1	1, 3, 7, 9, 10, 14, 15, 17, 21, 23, 27, 32, 33, 36, 37, 39, 40, 41, 47, 49, 56, 57, 58, 63, 68, 70
2η Κοινότητα	2, 4, 8, 25, 26, 28, 29, 31, 38, 44, 50, 51, 52, 53, 59, 61, 64, 65, 67
3η Κοινότητα2	5, 6, 11, 12, 13, 16, 18, 19, 20, 22, 24, 30, 34, 35, 42, 43, 45, 46, 48, 54, 55, 60, 62, 66, 69

ΠΙΝΑΚΑΣ (5.9) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο K-Means*

Η συσχέτιση με την πραγματική αλήθεια των κοινοτήτων είναι μόλις 10.18% και η ομοιότητα Jaccard 24.76% καθώς 28 δηλαδή λιγότεροι από τους μισούς κόμβοι ταξινομήθηκαν στην σωστή κοινότητα και το διάγραμμα Silhouette είναι στην Εικόνα (5.17).



ΕΙΚΟΝΑ (5.17) *Το διάγραμμα Silhouette της K-means συσταδοποίησης*

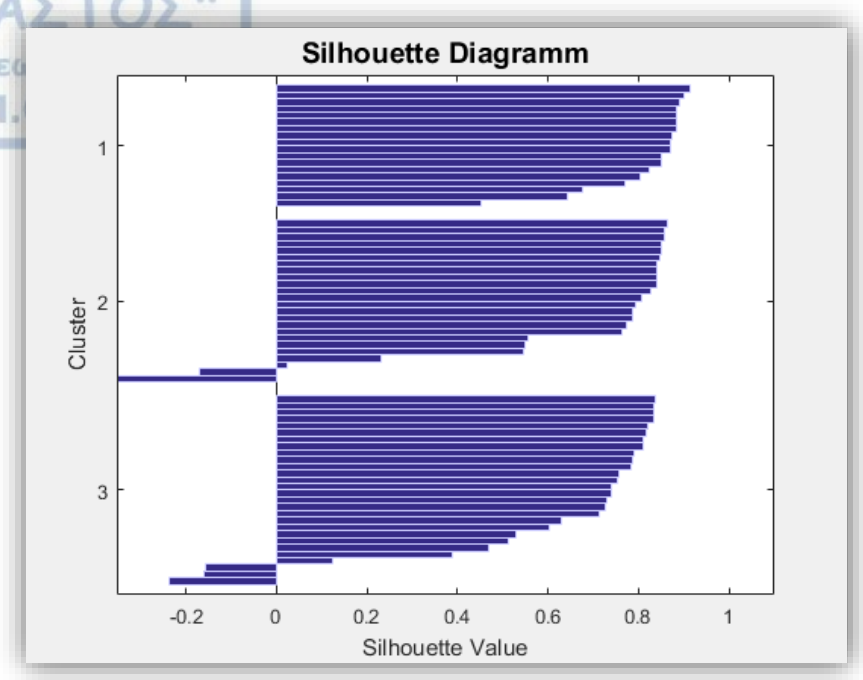
Και η ποσοτική αξιολόγηση του διαχωρισμού είναι 72.12%.

Τα αποτελέσματα του αλγορίθμου Greedy agglomerative modularity maximization method (ModulMax3) είναι στον Πίνακα (5.10).

1η Κοινότητα3	3, 9, 10, 14, 17, 21, 27, 32, 33, 36, 37, 41, 47, 49, 56, 57, 62, 64, 68, 70,
2η Κοινότητα1	1, 2, 4, 8, 15, 25, 26, 28, 29, 31, 38, 44, 50, 51, 52, 53, 59, 61, 63, 65, 67
3η Κοινότητα2	5, 6, 7, 11, 12, 13, 16, 18, 19, 20, 22, 23, 24, 30, 34, 35, 39, 40, 42, 43, 45, 46, 48, 54, 55, 58, 60, 66, 69,

ΠΙΝΑΚΑΣ (5.10) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο Greedy*

Η συσχέτιση με την πραγματική αλήθεια των κοινοτήτων είναι μόλις 9.29%, η ομοιότητα Jaccard 24.74% καθώς 24 κόμβοι ταξινομήθηκαν στην σωστή κοινότητα και συσχέτιση μεταξύ των δύο μεθόδων είναι 90.48%. Τέλος, το διάγραμμα Silhouette της modularity maximization βρίσκεται στην Εικόνα (5.18).

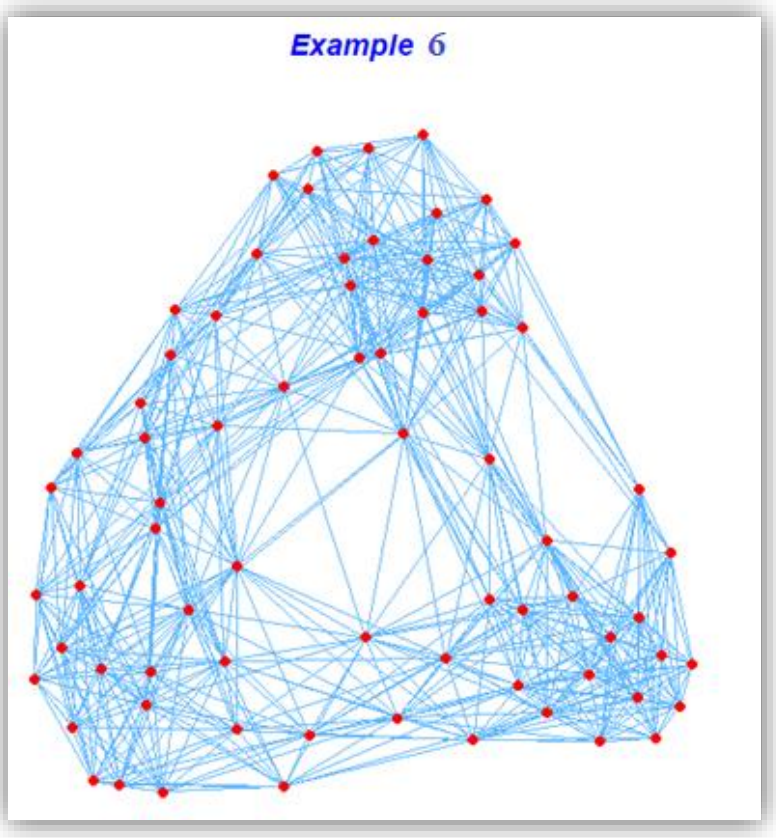


ΕΙΚΟΝΑ (5.18) *Το διάγραμμα Silhouette της Greedy διαμέρισης*

Και η ποσοτική αξιολόγηση του διαχωρισμού είναι 67.10% που είναι χαμηλότερη αυτής της K-Means η οποία ήταν 72.12%.

5.1.6 (Adj6: 70x70)

Στο έκτο παράδειγμα το δίκτυο κατασκευάζεται με 70 κόμβους και 662 ακμές. Τα στοιχεία ορίζονται στο επίπεδο από τα οποία τα 40 ανήκουν στην πρώτη κλάση και τα υπόλοιπα 30 στη δεύτερη. Η πρώτη κλάση έχει κέντρο το (2.5,2.5) και η δεύτερη το (2.5,2.5), δηλαδή οι δύο κλάσεις έχουν τα ίδιο κέντρο γεγονός που δυσκολεύει τη συσταδοποίηση. Τα σημεία σε κάθε κλάση παράγονται από κανονική δισδιάστατη κατανομή με τα δεδομένα κέντρα και πίνακα συν διασπορών $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Επιπλέον, οι τιμές υψώνονται στο τετράγωνο με αποτέλεσμα να αλλάζει η κατανομή από κανονική σε X^2 κατανομή με ένα βαθμό ελευθερίας, κάτι που κάνει τον εντοπισμό κοινοτήτων ακόμη πιο δύσκολο. Ο πίνακας γειτνίασης κατασκευάστηκε με τον ίδιο τρόπο και το κανονικό δίκτυο απεικονίζεται στην Εικόνα (5.19).



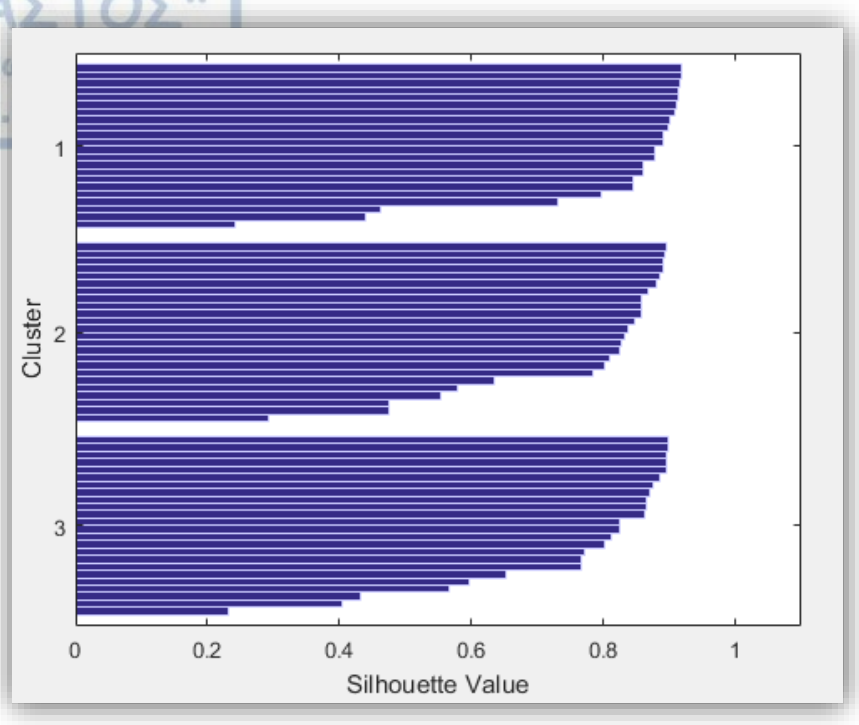
ΕΙΚΟΝΑ (5.19) *Απεικόνιση του δικτύου Adj6*

Οι εντοπισμένες κοινότητες με την K-Means είναι στον Πίνακα (5.11).

1η Κοινότητα2	1, 2, 4, 8, 15, 25, 26, 28, 29, 31, 38, 44, 50, 51, 52, 53, 59, 61, 64, 65, 66, 67, 68 69
2η Κοινότητα3	5, 6, 7, 12, 13, 16, 18, 19, 22, 24, 30, 34, 35, 42, 43, 45, 46, 54, 55, 60, 62,
3η Κοινότητα1	3, 9, 10, 11, 14, 17, 20, 21, 23, 27, 32, 33, 36, 37, 39, 40, 41, 47, 48, 49, 56, 57, 58, 63, 70,

ΠΙΝΑΚΑΣ (5.11) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο K-Means*

Η συσχέτιση με την πραγματική αλήθεια είναι 14.5% και η ομοιότητα 24.6% Jaccard καθώς 28 κόμβοι έχουν καταχωρηθεί στην σωστή κλάση και το αντίστοιχο διάγραμμα Silhouette απεικονίζεται στην Εικόνα (5.20).



ΕΙΚΟΝΑ (5.20) *Το διάγραμμα Silhouette της K-means συσταδοποίησης*

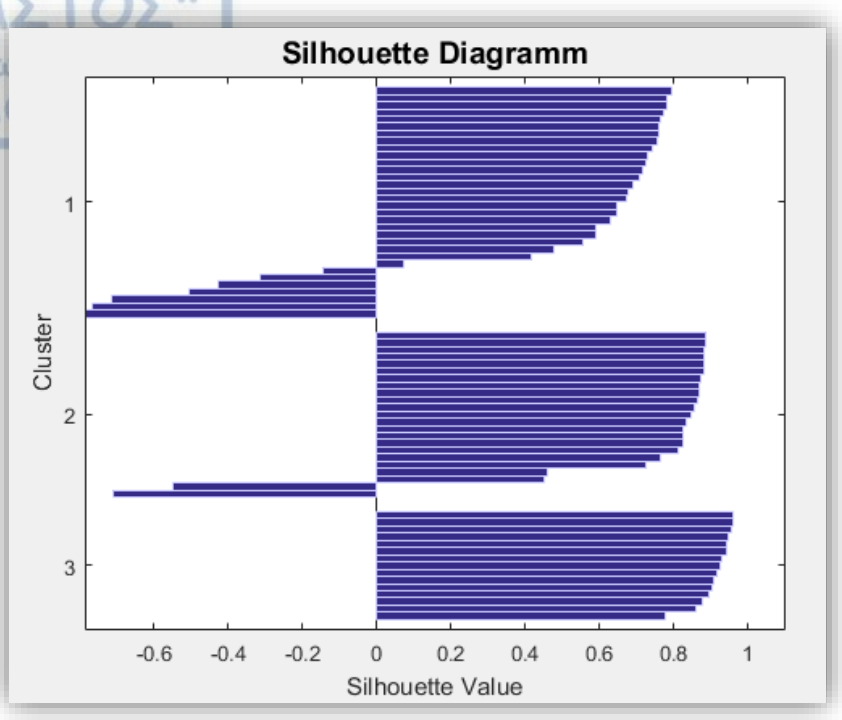
Η ποσοτική αξιολόγηση του διαχωρισμού υπολογίζεται στο 77.62%.

Τα αποτελέσματα του αλγορίθμου Greedy agglomerative modularity maximization method (ModulMax3) είναι στον Πίνακα (5.12).

1η Κοινότητα3	1, 2, 4, 8, 15, 18, 25, 26, 28, 29, 31, 38, 43, 44, 51, 52, 53, 59, 61, 64, 68
2η Κοινότητα2	5, 6, 7, 11, 12, 13, 19, 22, 24, 30, 35, 42, 45, 46, 54, 55, 60, 62, 65, 66, 67, 69
3η Κοινότητα1	3, 9, 10, 14, 16, 17, 20, 21, 23, 27, 32, 33, 34, 36, 37, 39, 40, 41, 47, 48, 49, 50, 56, 57, 58, 63, 70

ΠΙΝΑΚΑΣ (5.12) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο Greedy*

Η συσχέτιση με την πραγματική αλήθεια είναι 11.64% και η ομοιότητα Jaccard 23.8% καθώς 24 κόμβοι έχουν καταχωρηθεί στην σωστή κλάση. Επίσης, η συσχέτιση μεταξύ των δύο μεθόδων είναι 86.72% και το διάγραμμα Silhouette της modularity maximization βρίσκεται στην Εικόνα (5.21).



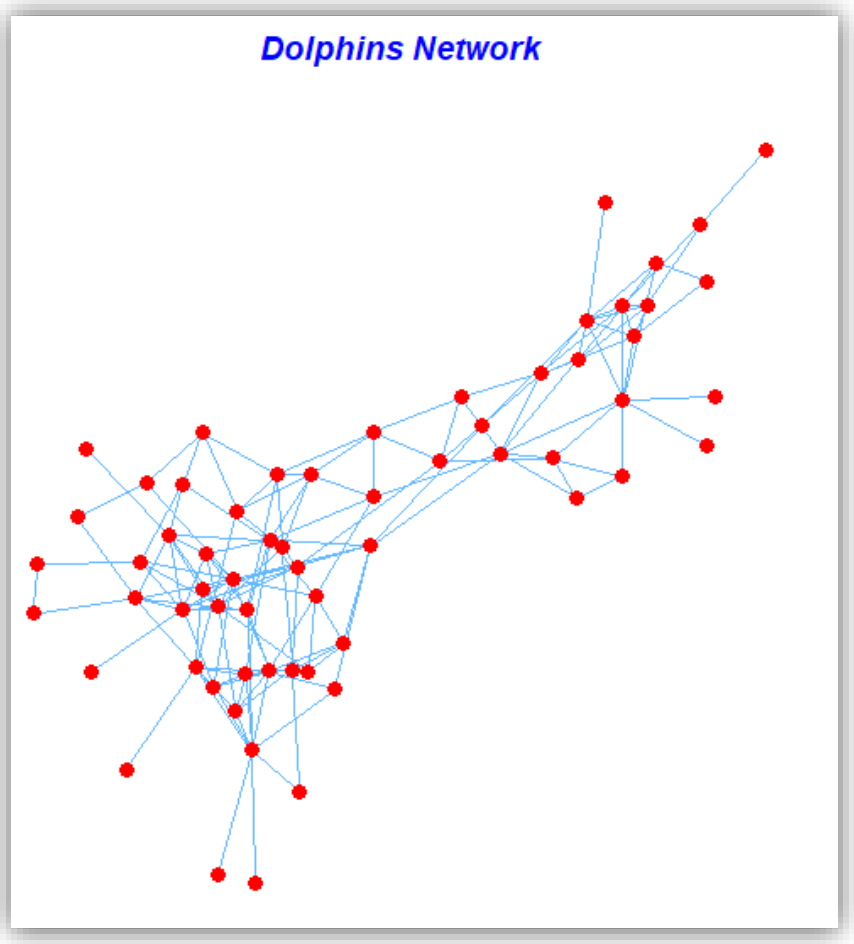
ΕΙΚΟΝΑ (5.21) *Το διάγραμμα Silhouette της Greedy διαμέρισης*

Όπου η ποσοτική αξιολόγηση του διαχωρισμού είναι 60.27%, αρκετά χαμηλότερη από την αξιολόγηση του K-Means που ήταν 77.62%.

5.1.7 Dolphin social network

Όπως πολλά ζώα έτσι και τα δελφίνια αρέσκονται στο να ταξιδεύουν κατά ομάδες οι οποίες κυμαίνονται από δύο έως τριάντα αλλά ακόμη και χίλια δελφίνια. Αυτή είναι μια στρατηγική επιβίωσης καθώς με αυτόν τον τρόπο είναι πιο δύσκολο να τους επιτεθεί κάποιο αρπακτικό. Επίσης, έχει παρατηρηθεί ότι τα δελφίνια είναι κοινωνικά και εκλεπτυσμένα ζώα αφού παρατηρούνται σε εκείνα ανθρώπινα χαρακτηριστικά όπως η συμπάθεια, η συνεργασία και ο αλτρουισμός.

Το δίκτυο «Dolphin social network» είναι ένα μη κατευθυνόμενο δίκτυο που αποτελείται από 62 κόμβους που αντιπροσωπεύουν δελφίνια μιας κοινότητας δελφινιών στην Νέα Ζηλανδία και 159 ακμές που αντιπροσωπεύουν την συσχέτιση ή αλληλεπίδραση μεταξύ αυτών. Η καταγραφή του δικτύου έγινε την περίοδο των ετών 1994-2001 στην Νέα Ζηλανδία (Lusseau, 2003). Το προκύπτουν δίκτυο απεικονίζεται στην Εικόνα (5.22).



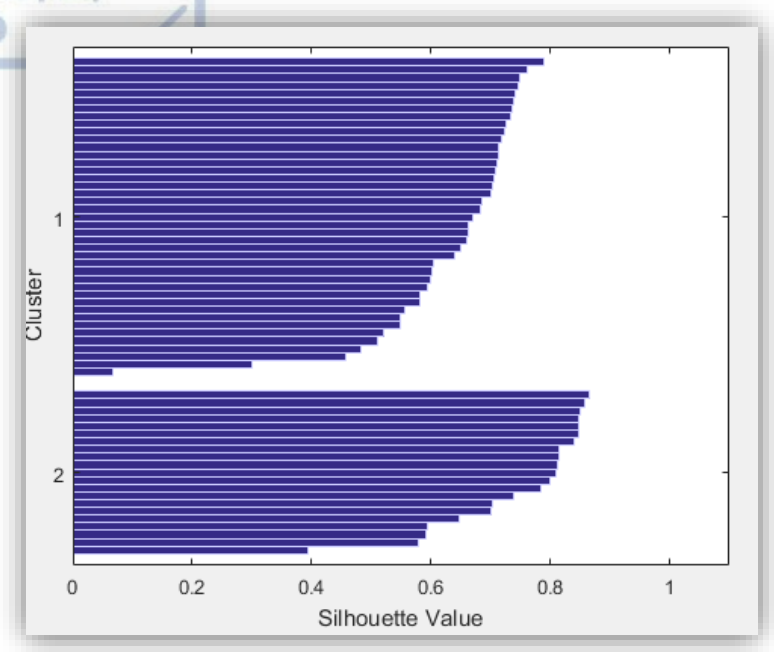
ΕΙΚΟΝΑ (5.22) *Απεικόνιση του δικτύου Dolphin social network*

Τα αρχικά επιλεγμένα κέντρα είναι οι κόμβοι 15 και 18 και οι εντοπισμένες κοινότητες με την K-Means απεικονίζονται στον Πίνακα (5.13).

1η Κοινότητα	1, 3, 4, 5, 9, 11, 12, 13, **15** 16, 17, 19, 21, 22, 24, 25, 29, 30, 31, 34, 35, 36, 37, 38, 39, 41, 43, 44, 45, 46, 47, 48, 50, 51, 52, 53, 54, 56, 59, 60, 62
2η Κοινότητα	2, 6, 7, 8, 10, 14, **18**, 20, 23, 26, 27, 28, 32, 33, 40, 42, 49, 55, 57, 58, 61

ΠΙΝΑΚΑΣ (5.13) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο K-Means*

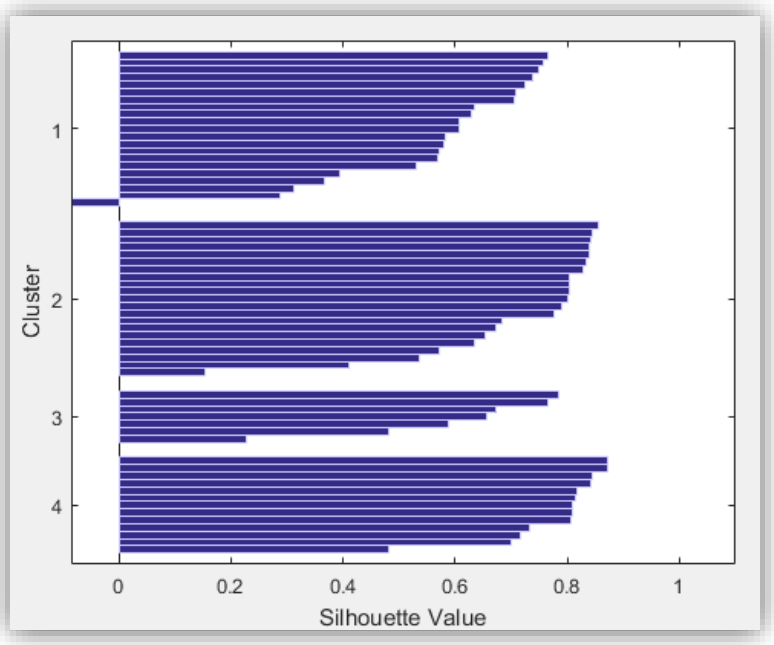
Το διάγραμμα Silhouette που δείχνει τις αποστάσεις μεταξύ των κόμβων της ίδιας κοινότητας με εκείνους των υπολοίπων κοινοτήτων βρίσκεται στην Εικόνα (5.23).



ΕΙΚΟΝΑ (5.23) *Το διάγραμμα Silhouette της K-means συσταδοποίησης*

Και η ποσοτική αξιολόγηση του διαχωρισμού με βάση τη μέθοδο Silhouette είναι 67.31%.

Επιπρόσθετα, αν ο αριθμός των κοινοτήτων επιλεγεί να είναι 4, με τον αλγόριθμο K-Means η ποσοτική αξιολόγηση του διαχωρισμού είναι 66.15% και η απεικόνιση του αντίστοιχου διαγράμματος είναι στην Εικόνα (5.24).



ΕΙΚΟΝΑ (5.24) *Το διάγραμμα Silhouette της K-means συσταδοποίησης με 4 κοινότητες*

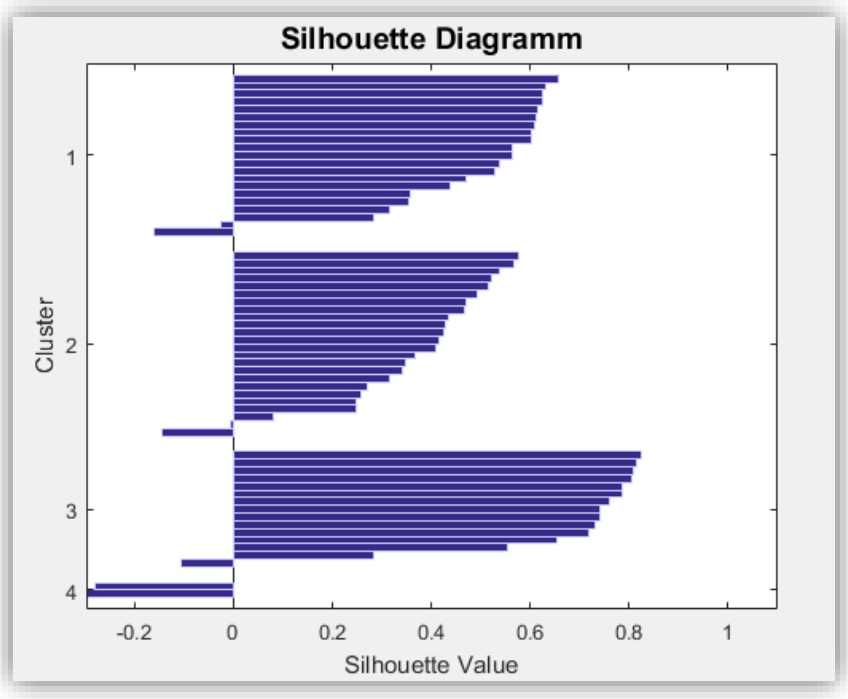
Αντίθετα, οι εντοπισμένες κοινότητες του αλγορίθμου modularity maximization (ModulMax3) απεικονίζονται στον Πίνακα (5.14).

1η Κοινότητα	1, 3, 11, 13, 15, 17, 21, 34, 35, 38, 39, 41, 43, 44, 45, 47, 48, 50, 51, 53, 54, 59, 62

2η Κοινότητα	2, 6, 7, 8, 10, 14, 18, 20, 23, 26, 27, 28, 29, 31, 32, 33, 42, 49, 55, 57, 58, 61
3η Κοινότητα	4, 5, 9, 12, 16, 19, 22, 24, 25, 30, 36, 46, 52, 56, 60,
4η Κοινότητα	37, 40

ΠΙΝΑΚΑΣ (5.14) *Οι εντοπισμένες κοινότητες με τον αλγόριθμο Greedy*

Η συσχέτιση μεταξύ των δύο μεθόδων είναι 68.05% και το διάγραμμα Silhouette της modularity maximization είναι στην Εικόνα (5.25).



ΕΙΚΟΝΑ (5.25) *Το διάγραμμα Silhouette της Greedy διαμέρισης*

Τέλος, η ποσοτική αξιολόγηση του διαχωρισμού της modularity maximization είναι 44.72%, και όπως παρατηρείται είναι πολύ μικρότερη από αυτής του K-Means αλγορίθμου που είναι 67.31%.

5.2 Τελικά αποτελέσματα

Στον Πίνακα (5.15) παρουσιάζονται συγκεντρωτικά τα αποτελέσματα της ποσοτικής αξιολόγησης με τη μέθοδο Silhouette τόσο των έξι τεχνιτών δικτύων όσο και του δικτύου των δελφινιών. Αξίζει να τονιστεί ότι σε όλες τις περιπτώσεις ο προτεινόμενος αλγόριθμος K-Means είχε καλύτερη συσταδοποίηση από τον κλασικό αλγόριθμο modularity maximization.

Example	K-Means	Modularity Maximization
Adj1	62.14%	58.82%
Adj2^2	62.14%	49.24%
Adj3	77.13%	75.60%
Adj4^2	72.12%	57.51%
Adj5	72.82%	67.10%
Adj6^2	77.62%	60.27%
Dolph	67.31%	44.72%
Average	**70.18%**	**59.04%**

ΠΙΝΑΚΑΣ (5.15) *Συγκρίσεις Αλγορίθμων με την αξιολόγηση Silhouette*

Στον Πίνακα (5.16) παρουσιάζεται η γραμμική συσχέτιση Pearson και η Jaccard ομοιότητα μεταξύ των εντοπισμένων κοινοτήτων των αντίστοιχων αλγορίθμων με αυτές της πραγματικής αλήθειας. Στα αποτελέσματα παρατηρείται υψηλότερη συσχέτιση του αλγορίθμου K-Means και στις έξι περιπτώσεις όπως και μεγαλύτερη ομοιότητα Jaccard σε πέντε από αυτές. Επίσης, από τον μέσο όρων όλων των δεικτών συμπεραίνεται ότι η απόδοση του προτεινόμενου αλγορίθμου ήταν συνολικά καλύτερη έναντι του κλασικού.

Ground truth	**K – Means Pearson**	**Modularity Pearson**	**K-Means Jaccard**	**Modularity Jaccard**
Adj1	90.09%	86.06%	55,8%	52.23%
Adj2^2	89.77%	79.77%	55,41%	49.42%
Adj3	94.37%	89.11%	89.33%	80.22%
Adj4^2	80.84%	72.50%	56.63%	40.35%
Adj5	10.18%	9.29%	24.74%	23.76%
Adj6^2	14.50%	11.64%	24.60%	23.80%
Average	**63.29%**	**58.06%**	**51.09%**	**44.96%**

ΠΙΝΑΚΑΣ (5.16) *Συσχέτιση και ομοιότητα κοινοτήτων με την πραγματική αλήθεια*

Στον Πίνακα (5.17) παρουσιάζεται ο αριθμός των κόμβων που αντιστοιχίστηκαν στην σωστή κοινότητα, σύμφωνα με την πραγματική αλήθεια που ήταν γνωστή εκ των προτέρων και τέλος παρουσιάζεται το συνολικό ποσοστό των σωστά ταξινομημένων κόμβων.

Correct Nodes	K - Means	Modularity Maximization
Adj1 (130)	99	91
Adj2^2 (130)	97	91
Adj3 (70)	68	66
Adj4^2 (70)	52	44
Adj5 (70)	28	24
Adj6^2 (70)	28	24
Average	**68.89%**	**62.97%**

ΠΙΝΑΚΑΣ (5.17) *Σωστά κατανεμημένες κοινότητες*

Συμπεραίνεται λοιπόν από τα συνολικά ποσοστά ότι ο προτεινόμενος αλγόριθμος είναι καλύτερος από τον κλασικό σε όλα τα παραδείγματα τόσο με τον δείκτη συσχέτισης Pearson όσο με τον δείκτη ομοιότητα Jaccard, αλλά και με την Silhouette αξιολόγηση.

6. Εφαρμογή του αλγορίθμου σε πραγματικό δίκτυο

6.1Δίκτυο με λέξεις κλειδιά στο Twitter

Μία φυσική καταστροφή είναι η συνέπεια ενός φυσικού κινδύνου, όπως για παράδειγμα μία ηφαιστειακή έκρηξη, ένας σεισμός, μία κατολίσθηση, ο οποίος περνάει από το στάδιο της πιθανότητας σε μία ενεργή φάση και κατά συνέπεια έχει επιπτώσεις στην καθημερινή ζωή των ανθρώπων. Ακόμη, η πηγή της φυσικής καταστροφής μπορεί να είναι η ανθρώπινη δραστηριότητα όπως για παράδειγμα στην πρόκληση μιας πυρκαγιάς ή στην πρόκληση πλημμύρας στην οποία στεγνές περιοχές καλύπτονται από ποσότητες νερού για συγκεκριμένο χρονικό διάστημα. Τέλος, η ανθρώπινη αδυναμία μπροστά στις φυσικές καταστροφές, που επιδεινώνεται από την έλλειψη προγραμματισμού ή την έλλειψη κατάλληλου συστήματος διαχείρισης έκτακτων αναγκών, οδηγεί σε οικονομικές, δομικές και ανθρώπινες απώλειες.

Η ευρεία χρησιμοποίηση των social media όπως του Twitter από φυσικά άτομα ή φορείς μπορεί επίσης να αντικατοπτρίζει περιπτώσεις φυσικών καταστροφών. Ως εκ τούτου, λόγω της άμεσης μετάδοσης πληροφοριών το Twitter μπορεί να γίνει ένα μέσο μετάδοσης απαραίτητων πληροφοριών σε μια περίπτωση φυσικής καταστροφής ώστε να γίνει έγκαιρη ενημέρωση των ανθρώπων σχετικά με αυτή.

Στο άρθρο (Andreadis et all., 2018) παρουσιάζεται μια μέθοδος εντοπισμού πλημμυρών μέσα από το Twitter στην Ιταλική επικράτεια με χρήση λέξεων-κλειδιών που εισήχθησαν στην εφαρμογή Twitter's Streaming API. Η εφαρμογή αυτή συλλέγει δεδομένα σε πραγματικό χρόνο σύμφωνα με λέξεις κλειδιά οι οποίες αφορούν σε γεγονότα πλημμυρών.

Οι λέξεις-κλειδιά που χρησιμοποιήθηκαν στην αναζήτηση βρίσκονται στον Πίνακα (6.1), όπου είναι οι Ιταλικές λέξεις που εισήχθησαν μαζί με την μετάφρασή τους στα Αγγλικά για την καλύτερή τους κατανόηση.

Keywords	English translation
alluvione	flood
alluvionevicenza	flood Vicenza
allagamento	flooding
bacchiglione	Bacchiglione
fiumepiena	full river
allertameteo	weather alert
sottopassoallagato	underpass flooded
alluvione2017	flood 2017
allertameteovicenza	weather alert Vicenza
esondazione	flooding

ΠΙΝΑΚΑΣ (6.1) *Λέξεις κλειδιά που χρησιμοποιήθηκαν στην αναζήτηση στο Twitter*

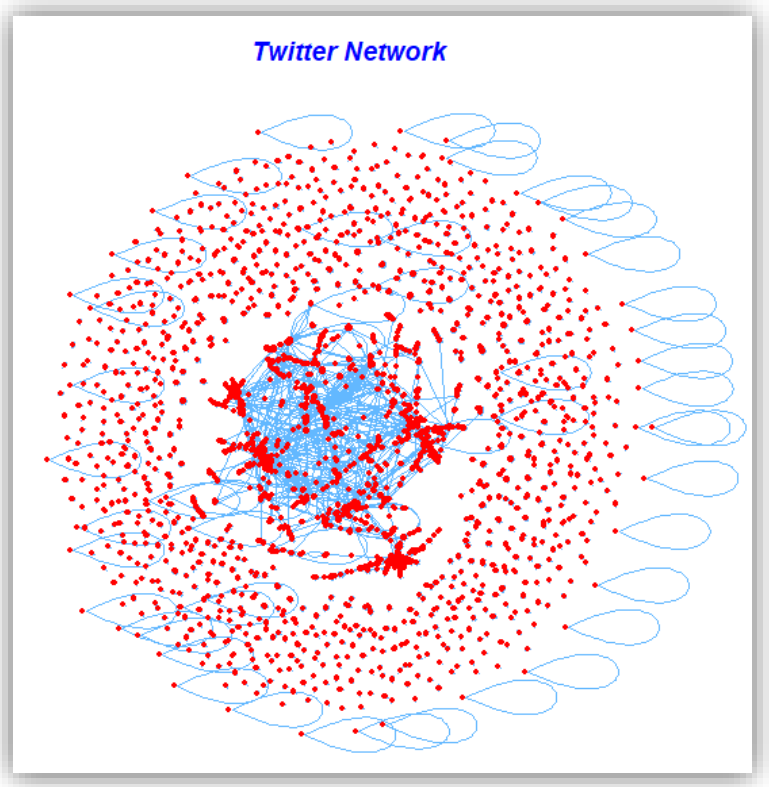
Η καταγραφή των tweets έλαβε χώρα την περίοδο από την 1η Απριλίου του 2017 μέχρι τις 31η Μαρτίου του 2018. Επίσης, τα tweets που δεν ήταν σχετικά αφαιρέθηκαν με ανθρώπινη παρέμβαση. Σημαντικό είναι να αναφερθεί ότι κατά τη διάρκεια της έρευνας παρατηρήθηκαν δύο σημαντικά γεγονότα, αλλά μόνο το ένα αφορούσε πλημμύρες καθώς το άλλο ήταν επέτειος μιας προηγούμενης πλημμύρας.

Στον Πίνακα (6.2) παρατηρούνται οι όροι των tweets που βρεθήκαν να είναι πιο χρησιμοποιημένοι στα tweets που εντοπίστηκαν.

	Non-locations			Locations	
#	Appearances	Word	English translation	Appearances	Word
1	30820	alluvione	flood	9542	Livorno
2	5418	colpire	to hit	1869	Roma
3	5388	maltempo	bad weather	1532	Firenze
4	4915	vittima	victim	999	Italia
5	4774	allagare	to flood	968	Genova
6	4100	allertameteo	weather alert	932	Valtellina
7	3751	famiglia	family	878	Toscana
8	3484	pensiero	thought	563	Milano
9	3407	tenere	to hold	414	Sardegna
10	3360	benjiefede	Benji & Fede (band)	386	Parma

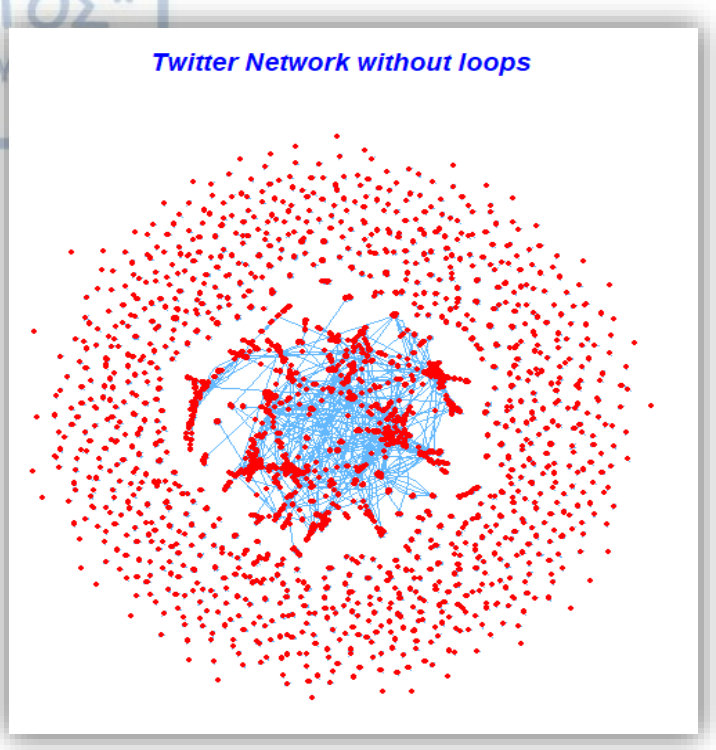
ΠΙΝΑΚΑΣ (6.2) *Οι όροι και η συχνότητα εμφάνισής τους στο Twitter*

Έτσι, χρησιμοποιώντας το top 5 των όρων, εξετάστηκε η συχνότητα τους στην περίοδο της μελέτης και κατασκευάστηκε το αντίστοιχο δίκτυο. Οι κόμβοι του δικτύου αντιπροσωπεύουν τα id των λογαριασμών που δημοσιεύσαν κάποιο tweet με μια λέξη κλειδί και οι ακμές αντιπροσωπεύουν τις αναφορές του ενός λογαριασμού σε έναν άλλον, που υπάρχουν ανάμεσα στα συγκεκριμένα tweet. Έτσι λοιπόν το δίκτυο που κατασκευάστηκε απεικονίζεται στην Εικόνα (6.1).



ΕΙΚΟΝΑ (6.1) *Απεικόνιση δικτύου Twitter*

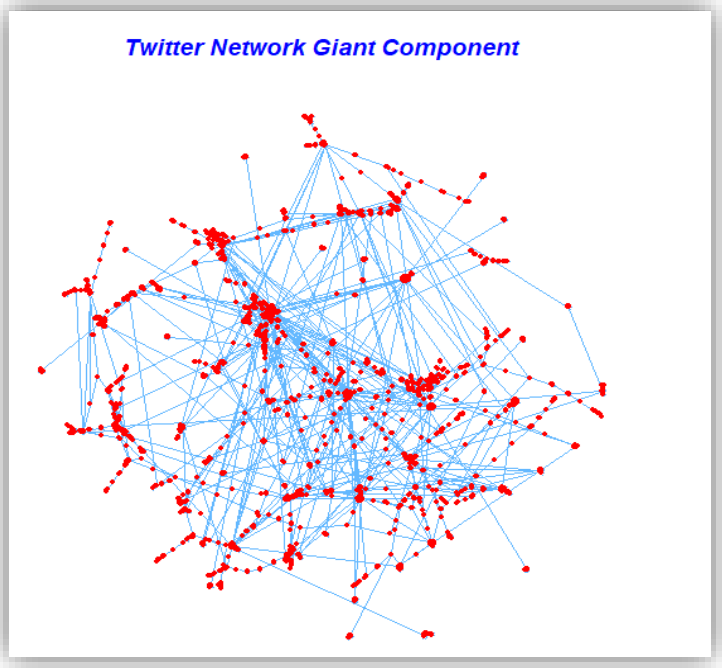
Το δίκτυο αποτελείται από 5658 μοναδικούς κόμβους-id και 6407 ακμές. Καθώς οι αναφορές ενός λογαριασμού προς τον ίδιο δεν προσθέτουν κάποια πληροφορία αφαιρέθηκαν από την ανάλυση. Έτσι δημιουργείται το δίκτυο χωρίς λούπες που απεικονίζεται στην Εικόνα (6.2).



ΕΙΚΟΝΑ (6.2) *Απεικόνιση απλού δικτύου Twitter*

Το δίκτυο χωρίς τις λούπες αποτελείται από 5658 μοναδικούς κόμβους-id και 6234 ακμές, στο οποίο εντοπίζονται 1038 συνιστώσες.

Για την εντοπισμό των κοινοτήτων έχει ουσιαστικό νόημα να αναλυθεί η μεγαλύτερη συνιστώσα η οποία αποτελείται από 2772 κόμβους και 4249 ακμές, δηλαδή αποτελεί το 48.99% του δικτύου, έχει πυκνότητα 0.0008524696, η ακτίνα της είναι 12, η διάμετρος της 23 και η οποία παριστάνεται στην Εικόνα (6.3).



ΕΙΚΟΝΑ (6.3) *Απεικόνιση κύριας συνιστώσας του δικτύου Twitter*

6.2 Βασικά μέτρα κεντρικότητας της κύριας συνιστώσας του δικτύου

Στην μελέτη των κοινωνικών δικτύων όπως το Twitter, το ενδιαφέρον επικεντρώνεται στην πληροφορία που μεταδίδεται μέσα στο δίκτυο. Ως εκ τούτου, η Betweenness κεντρικότητα η οποία μετράει το κατά πόσο ένας κόμβος βρίσκεται στο μονοπάτι μεταξύ των άλλων κόμβων, είναι αρκετά χρήσιμη για την μελέτη των σημαντικών κόμβων. Ο ορισμός της κεντρικότητας betweenness ενός κόμβου υ δίνεται από τον Τύπο (6.1).

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Τύπος (6.1)

Όπου σ_{st} είναι ο συνολικός αριθμός των συντομότερων μονοπατιών μεταξύ των κόμβων s και t και $\sigma_{st}(υ)$ είναι ο αριθμών εκείνων των μονοπατιών που διέρχονται από τον κόμβο υ.

Ακόμη, οι κεντρικότητες closeness και degree που υπολογίζουν το πόσο ένας κόμβος είναι κοντά με τους υπόλοιπους και με πόσους συνδέεται αντίστοιχα θα χρησιμοποιηθούν για τον εντοπισμό των σημαντικότερων κόμβων της κύριας συνιστώσας. Ο ορισμός της closeness κεντρικότητας ενός x κόμβου δίνεται από τον Τύπο (6.2).

$$C(x) = \frac{N}{\sum_y d(y,x)}$$

Τύπος (6.2)

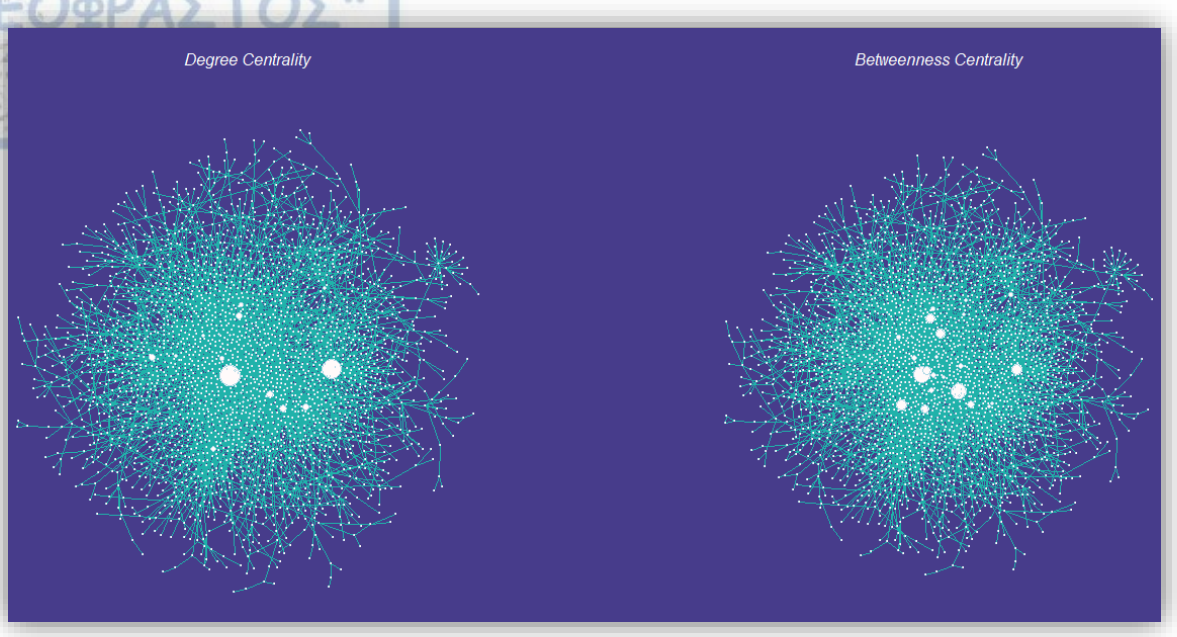
Όπου N είναι το πλήθος των κόμβων του δικτύου, και d(y,x) η απόσταση του x από οιονδήποτε άλλο κόμβο y του δικτύου, όπου απόσταση εννοούνται τα βήματα που απαιτούνται μέσα στο δίκτυο από το έναν κόμβο στον άλλον. Τέλος, ο ορισμός της degree κεντρικότητας για ένα κόμβο υ ενός απλού δικτύου δίνεται από τον Τύπο (6.3).

$$C_D(v) = \deg(v)$$

Τύπος (6.3)

Όπου deg(υ) είναι ο αριθμός των ακμών ρου κόμβου.

Στην Εικόνα (6.4) παριστάνεται η κύρια συνιστώσα του δικτύου με το μέγεθος του κάθε κόμβου να είναι ανάλογο του μέτρου της αντίστοιχης κεντρικότητας.



ΕΙΚΟΝΑ (6.4) *Η κύρια συνιστώσα του δικτύου με το μέγεθος των κόμβων ανάλογο της αντίστοιχης κεντρικότητας*

Αναλυτικότερα, θα επιγραφούν τα στατιστικά περιγραφικά μέτρα των τριών κεντρικοτήτων αλλά και οι πρώτοι 20 κόμβοι για κάθε ένα από αυτά.

6.2.1 Betweenness

Στον Πίνακα (6.3) παρατηρείται πως λίγοι κόμβοι έχουν υψηλή betweenness αλλά και παραπάνω από τους μισούς έχουν μηδενική betweenness καθώς η διάμεσος είναι 0.

Min	**1st Quart**	**Median**	**Mean**	**3rd Quart**	**Max**
0	0	0	8931	3202	813578

ΠΙΝΑΚΑΣ (6.3) *Τα περιγραφικά στατιστικά μέτρα για την Betweenness κεντρικότητα*

Στον Πίνακα (6.4) υπάρχουν οι ταξινομημένοι 20 πρώτοι κόμβοι με την betweenness κεντρικότητα, όπου ξεχωρίζουν αισθητά οι κόμβοι 3058 και 94.

3058	**94**	**3391**	**3048**	**23**	**3105**	**630**	**3137**	**3204**	**3205**
813579	749870	545310	536589	496243	475276	432954	421835	392683	298275
1028	**3140**	**441**	**4149**	**1363**	**1639**	**3176**	**3031**	**187**	**1219**
282489	272041	268913	262306	257647	230207	213998	205656	204923	187624

ΠΙΝΑΚΑΣ (6.4) *Οι 20 κόμβοι με τη μεγαλύτερη Betweeness όπου με έντονη γραφή είναι οι κόμβοι*

6.2.2 Closeness

Στον Πίνακα (6.5) τα στατιστικά μέτρα της κεντρικότητας closeness.

Min	**1st Quart**	**Median**	**Mean**	**3rd Quart**	**Max**
2.49e05	4.36e05	5.09e05	5.03e05	5.76e05	7.70e05

ΠΙΝΑΚΑΣ (6.5) *Τα περιγραφικά στατιστικά μέτρα για την Closeness κεντρικότητα*

Και στον Πίνακα (6.6) υπάρχουν οι ταξινομημένοι 20 πρώτοι κόμβοι με την closeness, όπου ξεχωρίζουν ελάχιστα οι κόμβοι 3137, 3058 και 9.

3137	**3058**	**94**	**1028**	**441**	**187**	**3048**	**3204**	**3177**	**1719**
7.70e05	7.69e05	7.57e05	7.56e05	7.28e05	7.26e05	7.24e05	7.21e05	7.19e05	7.17e05
1287	**838**	**1219**	**3140**	**1474**	**1363**	**23**	**843**	**3821**	**3040**
7.16e05	7.13e05	7.08e05	7.07e05	7.06e05	7.05e05	7.04e05	7.01e05	6.97e05	6.96e05

ΠΙΝΑΚΑΣ (6.6) *Οι 20 κόμβοι με τη μεγαλύτερη Closeness, όπου με έντονη γραφή είναι οι κόμβοι*

6.2.3 Degree

Min	**1st Quart**	**Median**	**Mean**	**3rd Quart**	**Max**
1.00	1.00	2.36	2.36	2.00	149.00

ΠΙΝΑΚΑΣ (6.7) *Τα περιγραφικά στατιστικά μέτρα για την Degree κεντρικότητα*

3058	**3048**	**94**	**3204**	**3176**	**1260**	**3105**	**3241**	**1363**	**3140**
149	147	60	57	53	50	50	42	36	34
3293	**4149**	**1639**	**1360**	**3205**	**1550**	**3391**	**3248**	**1454**	**206**
32	30	28	25	25	23	23	22	20	19

ΠΙΝΑΚΑΣ (6.8) *Οι 20 κόμβοι με τη μεγαλύτερη Degree, όπου με έντονη γραφή είναι οι κόμβοι*

Από τον Πίνακα (6.7) παρατηρείται ότι οι τα τρία τέταρτα των κόμβων έχουν βαθμό 1 ή 2, κι αυτό είναι μια πρώτη ένδειξη ότι η degree κεντρικότητα ακολουθεί προσεγγιστικά την κατανομή δύναμης (Power Law) καθώς επίσης και ότι στον Πίνακα (6.8) ξεχωρίζουν αρκετά δύο κόμβοι, οι 3158 και 3048, με τον 94 να ακολουθεί.

6.2.4 Οι σημαντικότεροι κόμβοι του δικτύου

Με τα μέτρα κεντρικότητας που περιεγράφηκαν έχουν ξεχωριστεί οι 11 πιο σημαντικοί κόμβοι της κύριας συνιστώσας και περιγράφονται στον Πίνακα (6.9).

Κόμβοι	Pagerank	Betweeness	Degree
3058	0.0208	813578	149
3048	0.0226	536589	147
94	0.0069	749870	60
3204	0.0077	392683	57
3105	0.0065	475276	50
3176	0.0072	213998	53
3140	0.0042	272041	34
1363	0.0044	257647	36
3391	0.0027	545309	23
3205	0.0033	298275	25
4149	0.0036	262306	30

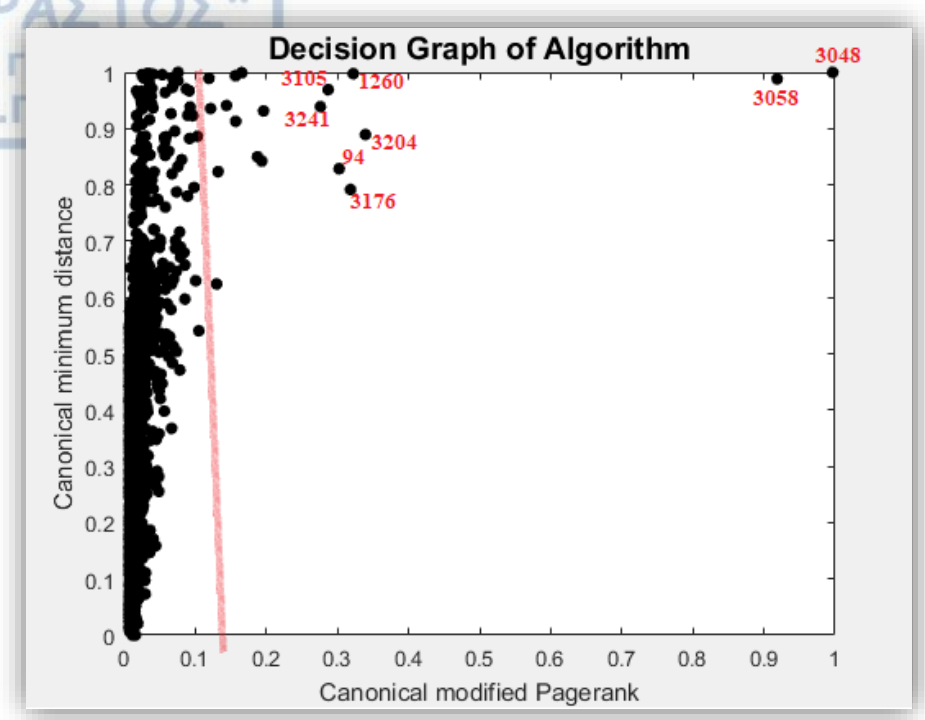
ΠΙΝΑΚΑΣ (6.9) *Οι 11 σημαντικότεροι κόμβοι του δικτύου στο Twitter*

Άρα παρατηρείται πως οι συγκεκριμένοι κόμβοι είναι ιδιαίτερα σημαντικοί για το δίκτυο και το ενδιαφέρον επικεντρώνεται σε αυτούς αλλά και στον εντοπισμό της κοινότητας στην οποία αυτοί ανήκουν, διότι κάθε σημαντικός κόμβος ενδέχεται να διαδραματίζει κύριο ρόλο στην κοινότητα στην οποία ανήκει.

6.3Εντοπισμός κοινοτήτων με τον προτεινόμενο αλγόριθμο K-Means

Στην κύρια συνιστώσα που αποτελεί το 48.99% του δικτύου ο δείκτης modularity με τη μέθοδο Fast Greedy υπολογίστηκε Q=0.87, δηλαδή αρκετά μεγαλύτερος από το 0.3, που σημαίνει ότι υπάρχει δομή κοινοτήτων στην κύρια συνιστώσα. Έτσι λοιπόν στο πραγματικό δίκτυο από το Twitter θα πραγματοποιηθεί η εφαρμογή του προτεινόμενου αλγορίθμου όπου θα εντοπιστούν οι κοινότητες. Επίσης, θα εντοπιστεί σε ποια κοινότητα ανήκουν οι πιο σημαντικοί κόμβοι και αυτό είναι αξιοσημείωτο, διότι οι χρήστες που ανήκουν σε κοινότητες και μάλιστα έχουν αρκετά σημαντική θέση σύμφωνα με τα μέτρα κεντρικότητας, γίνονται "authorities", δηλαδή αποκτούν αξιοπιστία στην πληροφορία που ανεβάζουν και επηρεάζουν σημαντικά την κοινότητα στην οποία ανήκουν καθώς επίσης και χρονικά πρώτα σε σχέση με τους υπόλοιπους κόμβους του δικτύου.

Τα δεδομένα που χρειάζεται ο αλγόριθμος είναι ο πίνακας γειτνίασης του δικτύου από τον οποίο εξάγεται ως ξεχωριστό δίκτυο η κύρια συνιστώσα. Σε αυτή, εξάγεται ο πίνακας γειτνίασης και εισάγεται στον αλγόριθμο για τον εντοπισμό των κοινοτήτων. Το πρώτο γράφημα που παράγει ο αλγόριθμος είναι το διάγραμμα απόφασης και απεικονίζεται στην Εικόνα (6.5).



ΕΙΚΟΝΑ (6.5) *Διάγραμμα απόφασης με κόκκινη γραμμή εκεί που έγινε η επιλογή των κέντρων*

Το k επιλέγεται να είναι 19, με συνέπεια να δημιουργούνται 19 κοινότητες στις οποίες τα αρχικά κέντρα είναι οι κόμβοι του Πίνακα (6.10).

94 126 1360 1363 155 1639 3048 3058 3105 3140 3176 3204 3205 3241 3248 3257 3391 4149

ΠΙΝΑΚΑΣ (6.10) *Τα αρχικά κέντρα των κοινοτήτων*

Το k επιλέχθηκε 19 σύμφωνα με το γράφημα απόφασης της Εικόνας (6.5). Από τη μέθοδο των πρώτων διαφορών εξάγεται ως k ο αριθμός 2, όμως διαισθητικά σε ένα τόσο μεγάλο δίκτυο δεν έχουν νόημα μόνο δύο κοινότητες. Έτσι, από το γράφημα, επιλέγεται ο διαχωρισμός ο οποίος θα έχει γραφικά αλλά και ουσιαστικά νόημα για την μελέτη του δικτύου. Εντούτοις από το γράφημα απόφασης παρατηρείται ότι ξεχωρίζουν σημαντικά δύο κόμβοι, οι 3048 και 3058.

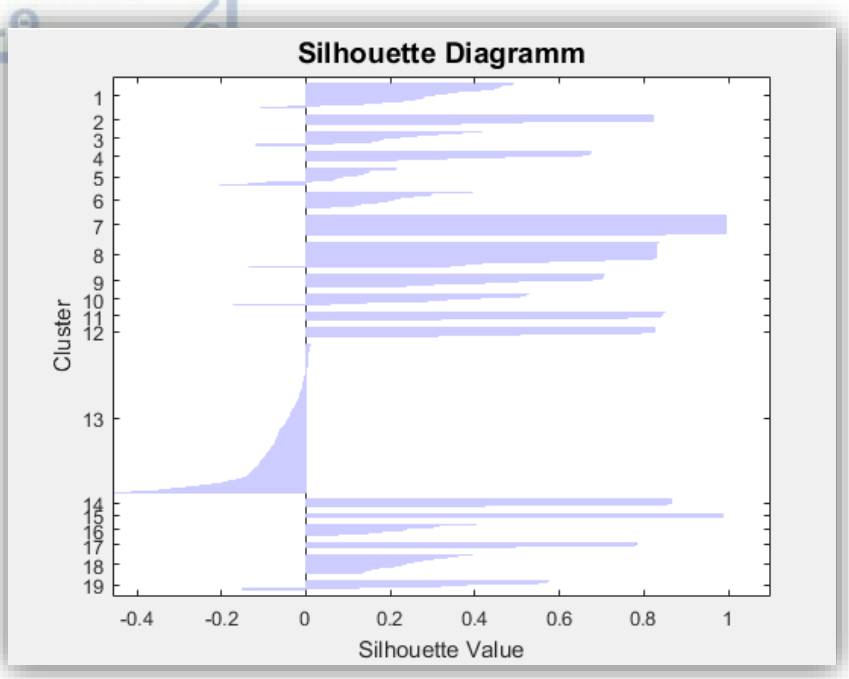
Παρατηρείται ακόμη ότι στα 19 αρχικά επιλεγμένα κέντρα ανήκουν και οι 11 σημαντικότεροι κόμβοι, άρα ο αλγόριθμος πραγματοποίησε σωστά την υπόθεση που τέθηκε, η οποία ήταν τα αρχικά κέντρα να είναι σημαντικοί και όχι απομακρυσμένοι κόμβοι.

Οι 19 εντοπισμένες κοινότητες του δικτύου με όλους τους κόμβους που περιλαμβάνουν βρίσκονται στο αρχείο (i.3) του παραρτήματος.

6.3.1 Αποτελέσματα από τις εντοπισμένες κοινότητες

Η κοινότητα με τους περισσότερους κόμβους είναι η κοινότητα 13 (1116 κόμβους δηλαδή το 40.26% των κόμβων) και αυτή με τους λιγότερους η κοινότητα 15 (23 κόμβους). Και το

διάγραμμα του δείκτη Silhouette για την ποσοτική αξιολόγηση της συσταδοποίηση είναι στην Εικόνα (6.6).



ΕΙΚΟΝΑ (6.6) *Το διάγραμμα Silhouette του δικτύου*

Η ποσοτική αξιολόγηση του διαχωρισμού είναι 25.27%, δηλαδή η συσταδοποίηση είναι θετική και για ένα τόσο μεγάλο δίκτυο, όπως φαίνεται και στην Εικόνα (6.6), είναι σχετικά καλή. Επίσης, από την Εικόνα (6.6) φαίνεται και το πόσο καλά οι κόμβοι προσαρμόζονται στην εκάστοτε κοινότητα, όπως για παράδειγμα παρατηρείται με τους κόμβους της κοινότητας 7 σε αντίθεση με αυτούς της κοινότητας 13.

6.4 Κοινότητες που ανήκουν οι σημαντικοί κόμβοι

Οι σημαντικοί κόμβοι του δικτύου λόγω τη υπόθεσης του αλγορίθμου, ανήκουν σε διαφορετικές κοινότητες και καθώς είναι και τα αρχικά κέντρα, λαμβάνουν έναν κυρίαρχο ρόλο στην κοινότητα στην οποία ανήκουν. Στην περίπτωση εξάλλου της μελέτης ενός κοινωνικού δικτύου όπως το Twitter οι λογαριασμοί με μεγάλη σημαντικότητα παίζουν ισχυρό ρόλο στην επιρροή και στην ροή πληροφορίας της κοινότητας στην οποία ανήκουν. Άλλωστε, ο κάθε κόμβος σε ένα δίκτυο ροής πληροφορίας επηρεάζει και επηρεάζεται πρώτα από τους κόμβους της κοινότητάς του.

Οι δύο πιο σημαντικοί κόμβοι του δικτυού που ξεχωρίζουν αρκετά από τους υπολοίπους όπως φαίνεται στην Εικόνα (6.5) και στον Πίνακα (6.9) είναι οι κόμβοι με αριθμό-id στον πίνακα γειτνίασης 3048 και 3058. Έπειτα ακολουθούν οι κόμβοι 94, 3204, 3105 και 3176.

Στον Πίνακα (6.11) φαίνεται σε ποια κοινότητα ανήκει κάθε ένας από τους 11 πιο σημαντικούς κόμβους του δικτύου.

Κόμβος	Κοινότητα
3058	8
3048	13
94	1
3204	12
3105	9
3176	11
3140	10
1363	4
3391	18
3205	13
4149	19

ΠΙΝΚΑΚΑΣ (6.11) *Οι κοινότητες στις οποίες ανήκουν οι σημαντικότεροι κόμβοι*

Ο κόμβος 3058 ανήκει στην κοινότητα 8 και ο κόμβος 3048 στην κοινότητα 13 όπως απεικονίζεται στον Πίνακα (6.11). Επίσης οι κόμβοι 94, 3204, 3105 και 3176 ανήκουν στις κοινότητες 1, 12, 9 και 11 αντίστοιχα.

Αυτό που παρατηρείται από τον Πίνακα (6.11) είναι ότι δύο σημαντικοί κόμβοι ανήκουν στην κοινότητα 13, η οποία είναι η κοινότητα με τους περισσότερους κόμβους που αποτελείται από 1116 κόμβους. Ο κόμβος 94 ανήκει στην κοινότητα 1 η οποία αποτελείται από 180 κόμβους, ο κόμβος 3204 στην κοινότητα 12 που αποτελείται από 68 κοινότητες. Με τον ίδιο τρόπο οι υπόλοιποι σημαντικοί κόμβοι μοιράζονται στις εκάστοτε κοινότητες όπως είναι στον Πίνακα (6.11).

Για περαιτέρω ανάλυση που είναι έξω από το σκοπό της εργασίας, μπορεί η κάθε κοινότητα να θεωρηθεί ως ξεχωριστό δίκτυο και να γίνει η ανάλυση σε κάθε ένα ξεχωριστά και να εξαχθούν επιπλέον πληροφορίες για την δομή και την λειτουργία του δικτύου.

7.Συμπεράσματα

Στην παρούσα εργασία αρχικά αντιμετωπίστηκαν τα προβλήματα του K-Means αλγορίθμου, κάτω από το πρίσμα των δικτύων, που αφορούν την επιλογή του αριθμού των συστάδων και των αρχικών κέντρων. Επιπλέον, μέσω της μεθόδου signal similarity επιτεύχθηκε η μετατροπή του δικτύου σε μετρικό n-διάστατο Ευκλείδειο χώρο.

Στα έξι τεχνητά δίκτυα που κατασκευάστηκαν για την αξιολόγηση του προτεινόμενου αλγορίθμου, ο αριθμός των κοινοτήτων που εντοπίστηκαν ήταν ο ίδιος ανάμεσα στην προτεινόμενη μέθοδο και την κλασική modularity maximization. Ωστόσο, στο πραγματικό δίκτυο dolphin social network ο K-means εντόπισε δύο κοινότητες ενώ ο κλασικός αλγόριθμος τέσσερις.

Η ποσοτική αξιολόγηση της συσταδοποίησης του προτεινόμενου αλγορίθμου K-Means μέσω της μεθόδου Silhouette, ήταν καλύτερη και στα έξι τεχνητά δίκτυα που κατασκευάστηκαν, από αυτής του κλασικού αλγορίθμου modularity maximization. Δηλαδή, κατά μέσο όρο το ποσοστό αξιολόγησης του K-Means ήταν 70%, ενώ του modularity maximization 59%. Επίσης, στο πραγματικό δίκτυο dolphin social network το ποσοστό αξιολόγησης της συσταδοποίησης του K-Means ήταν πάλι υψηλότερο (67%), έναντι του κλασικού αλγορίθμου βελτιστοποίησης του modularity (45%).

Στα έξι τεχνητά δίκτυα (Adj1, Adj2, Adj3, Adj4, Adj5 Adj6) που οι κοινότητες ήταν εκ των προτέρων γνωστές, η συσχέτιση Pearson των εντοπισμένων κοινοτήτων της K-Means με την πραγματική αλήθεια ήταν υψηλότερη της κλασικής μεθόδου modularity maximization. Συγκεκριμένα, η συσχέτιση Pearson των ανιχνεύσιμων κοινοτήτων της K-Means με τις πραγματικές ήταν κατά μέσο όρο 63%, σε αντίθεση με τις παραγόμενες κοινότητες του κλασικού αλγορίθμου που η συσχέτιση ήταν κατά μέσο όρο 58% αντίστοιχα. Επιπρόσθετα, η ομοιότητα Jaccard των εντοπισμένων κοινοτήτων με τις πραγματικές, ήταν πάλι υψηλότερη για την K-Means. Δηλαδή, η ομοιότητα Jaccard ήταν κατά μέσο όρο 51% για την K-Means σε αντίθεση με την modularity maximization που ήταν 45%. Τέλος, το ποσοστό των κόμβων που κατανεμήθηκαν σωστά ήταν 69% με την προτεινόμενη μέθοδο και 63% με τον κλασικό αλγόριθμό, κάτι που επιβεβαιώνει την επικράτηση του προτεινόμενο K-Means αλγορίθμου.

Στο πραγματικό δίκτυο του Twitter ανιχνεύτηκαν δεκαεννέα κοινότητες από τον προτεινόμενο αλγόριθμο και το ποσοστό αξιολόγησης της συσταδοποίησης υπολογίστηκε 25%. Ακόμη, εντοπίστηκαν οι έντεκα πιο σημαντικοί κόμβοι σύμφωνα με τις betweenness, closeness και degree κεντρικότητες. Επιπλέον, η πρόσθετη πληροφορία που λαμβάνεται από την ανίχνευση των κοινοτήτων, είναι ο εντοπισμός της επιρροής του σημαντικού κόμβου στους υπόλοιπους της κοινότητας που ανήκει. Ωστόσο, μπορεί να πραγματοποιηθεί περεταίρω μελέτη της εκάστοτε κοινότητας ώστε να αξιολογηθεί ο ρόλος του σημαντικού κόμβου μέσα στην κοινότητα αλλά και να αποκαλυφθούν κρυμμένα μοτίβα και σχέσεις εντός της κοινότητας.

Τα προβλήματα που εντοπίστηκαν τα οποία είναι συνυφασμένα με τη φύση του αλγορίθμου μη επιβλεπόμενης μάθησης K-Means, είναι η επιλογή του ακριβή αριθμού των κοινοτήτων, διότι παρατηρήθηκε ότι καθώς το μέγεθος του δίκτυο αυξάνει, η αυστηρή επιλογή του αριθμού γίνεται με εκτίμηση ή ανάλογα με τη φύση του δικτύου. Συνεπώς, η επιλογή για τον ακριβή αριθμό των κοινοτήτων παραμένει ένα ανοικτό πρόβλημα το οποίο λύθηκε προσεγγιστικά στην παρούσα εργασία.

Συμπερασματικά, μετά το πέρας των δοκιμών αξιολόγησης της μεθόδου K-Means προτείνεται μια καλύτερη πρόταση για εντοπισμό κοινοτήτων σε πολύπλοκα δίκτυα από αυτή της επιλογής ενός κλασικού αλγορίθμου.

Παράρτημα

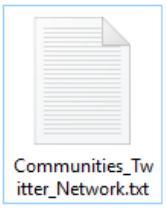
1.

ΕΙΚΟΝΑ (i.1) *Το αρχείο του κώδικα για τον εντοπισμό κοινοτήτων*

2.

ΕΙΚΟΝΑ (i.2) *Το αρχείο με την ανάλυση των κόμβων του δικτύου*

3.



ΕΙΚΟΝΑ (i.3) *Το αρχείο με τα αποτελέσματα των κοινοτήτων*

Βιβλιογραφία

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

Andreadis S., Gialampoukidis I., Fiorin R., Lombardo F., Norbiato D., Karakostas A., Ferri M., Vrochidis S., Kompatsiaris I. (2018). SOCIAL MEDIA OBSERVATIONS FOR FLOOD EVENT MONITORING. *Citizen Observatories for natural hazards and Water Management.* Venice.

Barabási, A. (2009). Scale-Free Networks: A Decade and Beyond. *Science*, 412-413.

Biswajit S., Amitabha M., Soumendu B. T., Debaprasad . (2015). Complex Networks, Communities and Clustering: A survey. *CoRR, abs/1503.06277*.

D. J. Watts , S. H. Strogatz. (1998). Collective dynamics of 'small-world' networks. *Nature*, 440-442.

David Arthur, Sergei Vassilvitskii. (2007). k-means++: The Advantages of Careful Seeding. 1027-1035.

Donetti L., Munoz M.A. (2004). Detecting Network Communities: a new systematic and efficient algorithm. *Statistical Mechanics*, P10012.

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

Franceschet, M. (2011). Pagerank: Standing on the shoulders of giants. 92-101.

Jaewon Y., Julian Mc., LeskovecJure L. (2013). Community Detection in Networks with Node Attributes. *2013 IEEE 13th International Conference on Data Mining.* Dallas, TX, USA: IEEE.

Jierui X., Stephen K., Boleslaw K. S. (2013). Overlapping Community Detection in Networks: the State of the Art and Comparative Study. *ACM Computing Surveys*, Volume 45 Issue 4.

L. Page, S. Brin, R. Motwani, T. Winograd. (1999). The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Conference*, (σσ. 161-172). Brisbane, Australia.

Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proc Biol Sc*.

Lusseau, D. (2007). Evidence for Social Role in a Dolphin Social Network. *Evolutionary Ecology*, 357-366.

M. E. J. Newman, M. Girvan. (2004). Finding and evaluating community structure in networks. *Physical Review* .

M. Girvan, M. E. J. Newman. (2002). Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 7821–7826.

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

Muhammad A. J., Muhammad S. Y., Siddique L., Junaid Q., Adeel B. (2018). Community detection in networks: A multidisciplinary review. *Network and Computer Applications*, 87-111.

N. Masuda, M.A. Porter, R. Lambiotte. (2017). Random walks and diffusion on networks. *Physics Report*, 1-58.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E 69*, 066133 .

Newman, M. E. (2006). Modularity and community structure in networks. *PNAS*, 8577-8582.

P.N. Tan, M. Steinbach, A. Karpatne, V. Kumar. (2018). *Introduction to Data Mining (Second Edition).* 2018.

Pang-Ning Tan, Michael Steinbach,Anuj Karpatne, Vipin Kumar. (2018). *Introduction to Data Mining (Second Edition).* Pearson.

Pearson, K. (1948). Early Statistical Papers. *University Press*.

Pizzuti, C. (2013). A Multiobjective Genetic Algorithm to Find. *TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, 418 - 430.

Pothen, A. (1997). Graph Partitioning Algorithms with Applications to Scientific Computing. *Parallel Numerical Algorithms*, 323-368.

Réka Albert, Albert-László Barabási. (2002). *Statistical mechanics of complex networks.* Minneapolis: American Physical Society.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 53-65.

S. Boettcher, A.G. Percus. (2001). Extremal optimization for graph partitioning. *Phys. Rev. E 64*.

S. Niwattanakul, J.Singthongchai, E. Naenudorn, S. Wanapu. (2013). Using of Jaccard Coefficient for Keywords Similarity. *International MultiConference of Engineers and Computer Scientists*, (σ. Vol I). Hong Kong.

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 467-482.

Wilson, R. (1996). *Introduction to Graph Theory.* Prentice Hall, 5 edition.

Y. Jiang, C. Jia, J. Yu. (2013). An efficient community detection method based on rank centrality. *Physica A: Statistical Mechanics and its Applications*, 2182-2194.

Y. Li, C. J. (2015). A parameter-free community detection method based on. *Physica A: Statistical Mechanics and its Applications 438*, pp. 321-334.

Yanqing Hu, M. L. (2008). Community detection by signaling on complex networks. *PHYSICAL REVIEW*, E 78, 016115.

Zachary, W. W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. *Anthropological Research*, 452-473.

Zhang L., Ye Q., Shao Y., Li C., Gao H. (2014). An Efficient Hierarchy Algorithm for Community Detection in Complex Networks. *Mathematical Problems in Engineering*.