Θεωρία Δικτύων

Εργασία 2024
Παπαδόπουλος Παναγιώτης 10697

Εισαγωγή

Η εργασία αυτή εξετάζει την υλοποίηση της συνάρτησης ποιότητας ομαδοποίησης απόστασης (*Distance Quality Function*) για την αναγνώριση και βελτίωση συμπλεγμάτων σε γραφήματα. Αρχικά, παρουσιάζεται η μαθηματική διατύπωση της συνάρτησης ποιότητας ομαδοποίησης απόστασης και υλοποιείται η διαδικασία υπολογισμού της. Η ανάλυση ξεκινά με τη μαθηματική διατύπωση της συνάρτησης και την υλοποίησή της, παρέχοντας τη θεωρητική βάση για την αξιολόγηση της ποιότητας των διαμερισμάτων σε ένα γράφημα.

Στη συνέχεια, περιγράφεται η ανάπτυξη ενός αλγορίθμου για τη μεγιστοποίηση της συνάρτησης ποιότητας απόστασης. Ο αλγόριθμος περιλαμβάνει διαδικασίες όπως η αρχικοποίηση πρωτοσυμπλεγμάτων, η επέκτασή τους με βάση συγκεκριμένα κριτήρια και η σταδιακή συγχώνευση των συμπλεγμάτων για τη βελτίωση της ποιότητας. Η μέθοδος βασίζεται σε στατιστική ανάλυση και υπολογιστικές τεχνικές για την εξασφάλιση βέλτιστων αποτελεσμάτων.

Τέλος, παρουσιάζονται συγκριτικά αποτελέσματα με τη γνωστή μέθοδο, *modularity clustering*, καταδεικνύοντας τη δυναμική του προτεινόμενου αλγορίθμου.

Υλοποίηση συνάρτησης Ποιότητας Ομαδοποίησης Απόστασης

Η συνάρτηση ποιότητας απόστασης $Q_d(G, C)$ αποσκοπεί στην αξιολόγηση της συνολικής ποιότητας ενός διαμερισμού C του γραφήματος G λαμβάνοντας υπόψη όλα τα συμπλέγματα $C_i \in C$, συγκρίνοντας τις αποστάσεις των κορυφών εντός των συμπλεγμάτων με τις αντίστοιχες αποστάσεις σε ένα τυχαίο μοντέλο γραφήματος με ορισμένα ίδια χαρακτηριστικά, τα οποία αναφέρονται παρακάτω:

- Κατανομή Βαθμών (*Degree Distribution*)
  - Το τυχαίο γράφημα έχει ίδιο αριθμό κορυφών και κάθε κορυφή έχει τον ίδιο βαθμό (αριθμό ακμών) όπως στο αρχικό γράφημα.
  - Η κατανομή βαθμών αντικατοπτρίζει τη συνολική δομή του γραφήματος. Αν δεν είναι ίδια, η σύγκριση μεταξύ του $D_V(C)$ (παρατηρούμενου γράφου) και $\overline{D_V(C)}$ (τυχαίου γραφήματος) δεν είναι δίκαιη.

- Αριθμός Κορυφών (*Number of Nodes*)
  - Ο τυχαίος γράφος έχει τον ίδιο αριθμό κορυφών $|V|$ με τον παρατηρούμενο γράφο.
  - Οι κορυφές μπορεί να μην έχουν τις ίδιες ετικέτες, αλλά ο αριθμός τους πρέπει να παραμένει σταθερός.
  - Η ύπαρξη του ίδιου αριθμού κορυφών εξασφαλίζει ότι οι αποστάσεις (παρατηρούμενες και αναμενόμενες) υπολογίζονται στο ίδιο “πλαίσιο”.

- Υποσύνολα Κορυφών στις Συστάδες ($Clusters$)

  - Οι ίδιες συμπλέγματα C_i που ορίζονται στον παρατηρούμενο γράφο G χρησιμοποιούνται και στον τυχαίο γράφο.
  - Δηλαδή, τα υποσύνολα κορυφών που ανήκουν σε κάθε σύμπλεγμα παραμένουν τα ίδια.
  - Αν τα σύνολα κορυφών στις συμπλέγματα C_i είναι διαφορετικά στο τυχαίο γράφημα, τότε το $\overline{D_V(C)}$ δεν μπορεί να συγκριθεί δίκαια με το $D_V(C)$.

- Αριθμός Ακμών ($Number\ of\ Edges$)

  - Ο τυχαίος γράφος έχει ίδιο συνολικό αριθμό ακμών με τον παρατηρούμενο γράφο.
  - Ο αριθμός ακμών επηρεάζει τις αποστάσεις στο γράφημα. Με τον ίδιο αριθμό ακμών, οι τυχαίοι γράφοι διατηρούν παρόμοια πυκνότητα με τον παρατηρούμενο.

- Συνολική Δομή ($Global\ Structure$)

  - Ο τυχαίος γράφος μπορεί να έχει παρόμοια συνολική δομή με τον παρατηρούμενο, υπό την έννοια της γενικής συνδεσιμότητας.
  - Αν και οι ακμές τοποθετούνται τυχαία, οι κορυφές με υψηλό βαθμό στο αρχικό γράφημα έχουν υψηλό βαθμό και στον τυχαίο.
  - Η συνολική συνδεσιμότητα εξασφαλίζει ότι τα τυχαία γραφήματα δεν είναι υπερβολικά διαφορετικά, επιτρέποντας χρήσιμη σύγκριση αποστάσεων.

Τα παρακάτω σημεία δεν είναι ίδια:

- Τοπολογία του Γράφου:

  - Ο τυχαίος γράφος δεν έχει την ίδια ακριβώς τοπολογία (δηλαδή το πού συνδέονται οι ακμές) με τον παρατηρούμενο.

- Διαδρομές:

  - Οι διαδρομές και οι αποστάσεις μεταξύ κορυφών στο τυχαίο γράφημα διαφέρουν από αυτές του αρχικού.

- Κοινότητα:

  - Το τυχαίο γράφημα δεν έχει “κοινότητες” ή “συμπλέγματα” που να προκύπτουν από πραγματικές σχέσεις. Είναι εντελώς τυχαίο.

Ορισμός Συνάρτησης Ποιότητας Ομαδοποίησης Απόστασης

Η συνάρτηση $Q_d(G, C)$ υπολογίζεται ως εξής:

$$Q_d(G, C) = \sum_{C_i \in C} \left(\overline{D_V(C_i)} - D_V(C_i) \right)$$

όπου:

- $D_V(C_i)$: Το άθροισμα των ζευγαριών αποστάσεων εντός της σύμπλεγμας C_i, όπως υπολογίζεται με βάση τις αποστάσεις στον όλο γράφο G.

- $\overline{D_V(C_i)}$: Το άθροισμα των ζευγαριών αποστάσεων εντός της σύμπλεγμας C_i σε ένα τυχαίο γράφο που έχει την ίδια κατανομή βαθμών με τον G, όπως υπολογίζεται με βάση τις αποστάσεις στον όλο γράφο G.

Βήματα Υλοποίησης

1. Υπολογισμός $D_V(C_i)$ (Παρατηρούμενη Απόσταση)

Για κάθε σύμπλεγμα C_i του διαμερισμού:
- Εξετάζουμε όλα τα ζεύγη κορυφών $i, j \in C_i$.
- Υπολογίζουμε την συντομότερη απόσταση $d(i, j)$ ανάμεσα στις κορυφές i και j, χρησιμοποιώντας τις διαδρομές στον πλήρη γράφο G.
- Αθροίζουμε όλες τις αποστάσεις.
- Διαιρούμε το άθροισμα με 2 για να μην μετρήσουμε κάθε απόσταση δύο φορές ($d(i, j) = d(j, i)$).

Τύπος:

$$D_V(C_i) = \frac{1}{2} \sum_{i,j \in C_i} d(i, j)$$

2. Υπολογισμός $\overline{D_V(C_i)}$ (Αναμενόμενη Απόσταση)

- Για κάθε σύμπλεγμα C_i στον γράφο G:
  - Δημιουργούμε πολλούς τυχαίους γράφους που διατηρούν την ίδια κατανομή βαθμών με τον G.
- Για κάθε τυχαίο γράφημα:
  - Υπολογίζουμε τις αποστάσεις όλων των ζευγαριών $i, j \in C_i$, χρησιμοποιώντας ολόκληρο τον τυχαίο γράφο.
  - Αν ένα ζεύγος κορυφών είναι αποσυνδεδεμένο, επιβάλλεται δυναμικό $penalty$ που υπολογίζεται ως διπλάσιο της διαμέτρου του γραφήματος.
  - Υπολογίζουμε τον μέσο όρο των αποστάσεων.

Τύπος:

$$\overline{D_V(C_i)} = \frac{1}{2} \sum_{i,j \in C_i} \mathbb{E}[d(i, j)]$$

3. Υπολογισμός του $Q_d(G, C)$

Για κάθε σύμπλεγμα C_i:
- Υπολογίζουμε τη διαφορά $\overline{D_V(C_i)} - D_V(C_i)$.
- Προσθέτουμε τη διαφορά στο συνολικό Q_d.

Το Q_d κανονικοποιείται με βάση τον αριθμό ακμών για την ικανοποίηση της κλίμακας ($scale\ invariance$).

Βασικές Αρχές Σχεδίασης:
1. Υπολογισμός αποστάσεων στο όλο γράφημα:

- Οι αποστάσεις $D_V(C_i)$ και $\overline{D_V(C_i)}$ χρησιμοποιούν τον πλήρη γράφο G για τη διατήρηση της συνολικής δομής.

2. Πολυδιάστατος υπολογισμός:
   - Υπολογίζονται πολλοί τυχαίοι γράφοι (π.χ. 10–20) για να προσεγγιστεί η $\overline{D_V(C_i)}$.
3. Δυναμικό $penalty$:
   - Αν ζεύγη κορυφών είναι αποσυνδεδεμένα, επιβάλλεται $penalty$ βασισμένο σε χαρακτηριστικά του γραφήματος (διάμετρος).
4. Συμπερίληψη όλων των κορυφών:
   - Οι αποστάσεις εντός κάθε σύμπλεγμας υπολογίζονται χρησιμοποιώντας όλες τις πιθανές διαδρομές στον γράφο G .
5. Κανονικοποίηση αποτελεσμάτων:
   - Το Q_d κανονικοποιείται με βάση τον αριθμό ακμών, για δίκαιες συγκρίσεις.

Ιδιότητες της συνάρτησης:

- $Isomorphism\ Invariance$:
  - Χρησιμοποιείται μόνο η δομή του γραφήματος, ανεξάρτητα από τις ετικέτες κορυφών.

- $Scale\ Invariance$:
  - Το Q_d κανονικοποιείται με βάση τον αριθμό ακμών.

- $Richness$:
  - Για οποιοδήποτε διαμέρισμα C ενός συνόλου κορυφών V, μπορούμε να κατασκευάσουμε ένα γράφημα G τέτοιο ώστε το διαμέρισμα C να αποτελεί το βέλτιστο διαμέρισμα που μεγιστοποιεί τη συνάρτηση Q_d. Αυτό επιτυγχάνεται με κατάλληλη διαμόρφωση του γράφου G, ώστε:
  - Οι αποστάσεις εντός των συμπλεγμάτων $D_V(C_i)$ να ελαχιστοποιούνται μέσω υψηλής συνοχής (π.χ. πλήρως συνδεδεμένες συμπλέγματα).
  - Οι αναμενόμενες αποστάσεις $\overline{D_V(C_i)}$ να μεγιστοποιούνται, καθώς τα τυχαία γραφήματα δεν διατηρούν τη δομή της συνοχής.

- $Perfectness$:
  - Σε πλήρως συνδεδεμένες συμπλέγματα, το $D_V(C_i)$ είναι το ελάχιστο δυνατό, ενώ το $\overline{D_V(C_i)}$ είναι μεγαλύτερο λόγω τυχαιότητας, μεγιστοποιώντας το $\overline{D_V(C_i)} - D_V(C_i)$.

- $Connectivity$:
  - Αποσυνδεδεμένα ζεύγη μειώνουν το Q_d μέσω δυναμικών $penalties$ και επιβραβεύονται για τη σωστή διάσπασή τους.

Υλοποίηση Αλγορίθμου Μεγιστοποίησης Συνάρτησης Ποιότητας Ομαδοποίησης Απόστασης

Η μέγιστη τιμή της συνάρτησης ποιότητας ομαδοποίησης απόστασης είναι επιθυμητή, καθώς υποδηλώνει ότι οι συμπλέγματα που προκύπτουν είναι καλύτερα ορισμένες. Συγκεκριμένα:

- Μείωση των παρατηρούμενων αποστάσεων: Οι κόμβοι εντός κάθε σύμπλεγμας είναι πιο “κοντά” ο ένας στον άλλο σε σχέση με την υπόλοιπη δομή του γραφήματος. Αυτό σημαίνει ότι οι συμπλέγματα είναι πιο συνεκτικές.

- Αύξηση των αναμενόμενων αποστάσεων: Στους τυχαίους γράφους οι αποστάσεις είναι μεγαλύτερες, γεγονός που δείχνει ότι η πραγματική δομή του γράφου δεν προκύπτει από τυχαιότητα αλλά έχει κρυφή κοινότητα ή συσχέτιση.

Ο στόχος του αλγορίθμου είναι η μέγιστη βελτιστοποίηση της συνάρτησης ποιότητας απόστασης ($Distance\ Quality\ Function$ - Q_d). Η διαδικασία ξεκινά με τη δημιουργία αρχικών συμπλεγμάτων ($proto-clusters$) και επεκτείνεται σε ένα πλήρες σύνολο συμπλεγμάτων, ακολουθώντας τη βελτιστοποίηση μέσω συγχωνεύσεων. Ακολουθούν τα βήματα που εκτελούνται:

1. Διαχωρισμός του Γράφου σε Συνδεδεμένα Συστατικά

Ο γράφος G χωρίζεται σε όλα τα συνδεδεμένα συστατικά του. Για κάθε ένα από αυτά τα συστατικά, εκτελείται ξεχωριστά ο αλγόριθμος. Αυτό διασφαλίζει ότι κόμβοι από διαφορετικά "νησιά" του γράφου δεν θα βρεθούν ποτέ στο ίδιο σύμπλεγμα, καθώς δεν υπάρχει διαδρομή που να τους συνδέει.

2. Δημιουργία Αρχικών Συστάδων ($Proto-Clusters$)

Το πρώτο βήμα του αλγορίθμου αφορά την εύρεση ενός συνόλου αρχικών συμπλεγμάτων. Αυτές οι συμπλέγματα σχηματίζονται από κόμβους υψηλής σημασίας, οι οποίοι επιλέγονται βάσει διαφόρων κριτηρίων όπως:

- Κεντρικότητα ενδιάμεσης διαδρομής ($Betweenness\ Centrality$): Κόμβοι που εμφανίζονται συχνά στις συντομότερες διαδρομές μεταξύ άλλων κόμβων.
- Κεντρικότητα εγγύτητας ($Closeness\ Centrality$): Κόμβοι που βρίσκονται πιο κοντά στο σύνολο των υπόλοιπων κόμβων.
- Πυρήνες $k-core$: Υπογράφηματα υψηλής συνοχής.
- Πυκνότητα ακμών: Κόμβοι με υψηλό αριθμό τοπικών συνδέσεων.

Εάν κανένας κόμβος δεν ικανοποιεί τα παραπάνω κριτήρια για την εκάστοτε αρχικοποίηση, εφαρμόζεται ένας μηχανισμός $failsafe$ που επιλέγει τυχαίους κόμβους. Αυτό εξασφαλίζει ότι υπάρχουν τουλάχιστον κάποιες αρχικές συμπλέγματα για περαιτέρω επεξεργασία.

Αξίζει να σημειωθεί ότι η αρχικοποίση των συμπλεγμάτων μπορεί να προκύπτει από οποιονδήποτε κανόνα, να εμπεριέχει οποιοδήποτε πλήθος κόμβων (περισσότερους του μηδενός) και οποιαδήποτε αρχική ομαδοποίησή τους και είναι ιδιαίτερα σημαντική για τον αλγόριθμο.

3. Επέκταση Συστάδων ($Cluster\ Expansion$)

Αφού εντοπιστούν τα $proto-clusters$, οι συμπλέγματα επεκτείνονται με στόχο να συμπεριλάβουν όλους τους κόμβους του γραφήματος. Εξετάζονται οι γείτονες των ήδη ανατεθειμένων κόμβων και προστίθενται στα συμπλέγματα στα οποία αυξάνεται η ποιότητα ομαδοποίησης απόστασης. Προτεραιότητα έχουν οι κυριότεροι κόμβοι, οι οποίοι θεωρούνται ως αυτοί με την μεγαλύτερη ενδιάμεση κεντρικότητα. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να μην υπάρχουν πλέον αδέσμευτοι κόμβοι. Αξίζει να σημειωθεί ότι σε αυτό το βήμα του αλγορίθμου, ο υπολογισμός της ποιότητας ομαδοποίησης απόστασης γίνεται με τη χρήση του προσεγγιστικού τύπου αποστάσεων, έτσι ώστε να αποφεύγονται οι επανειλημμένοι και δαπανηροί υπολογισμοί των αποστάσεων για όλα τα ζεύγη κόμβων, βελτιώνοντας τη συνολική απόδοση.

4. Συγχώνευση Συστάδων ($Cluster\ Merging$)

Με βάση τις τιμές του και τη συμβολή κάθε σύμπλεγμας, ο αλγόριθμος εξετάζει πιθανές συγχωνεύσεις συμπλεγμάτων. Οι συγχωνεύσεις που βελτιώνουν τη συνολική ποιότητα απόστασης πραγματοποιούνται, ενώ όσες δεν προσφέρουν βελτίωση απορρίπτονται. Για να μειωθούν οι υπολογισμοί, χρησιμοποιείται πάλι ο προσεγγιστικός τύπος για την εκτίμηση της αλλαγής στο

χωρίς να υπολογίζονται ξανά οι αποστάσεις για κάθε πιθανό ζεύγος κόμβων. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να μην είναι δυνατή η περαιτέρω βελτίωση της ποιότητας απόστασης.

5. Αξιολόγηση και Τελική Ομαδοποίηση

Μετά τη συγχώνευση, τα τελικά συμπλέγματα αξιολογούνται. Υπολογίζονται τα Q_d και $modularity$ για τη μέτρηση της ποιότητας της ταξινόμησης και την αποτελεσματικότητα του αλγορίθμου.

Ο Προσεγγιστικός Τύπος και η Χρήση του

Ο προσεγγιστικός τύπος που χρησιμοποιήθηκε στην παραπάνω υλοποίηση αφορά την εκτίμηση της αναμενόμενης απόστασης μεταξύ των κόμβων μιας κοινότητας σε έναν τυχαίο γράφο, χωρίς να χρειάζεται η πραγματική δημιουργία τυχαίων γράφων. Αυτός ο τύπος βασίζεται στην προσέγγιση $ERCQ$ ($Expected\ Random\ Clustering\ Quality$), η οποία εκτιμά την αναμενόμενη απόσταση με βάση την κατανομή βαθμών του γράφου.

Ο τύπος που χρησιμοποιήθηκε είναι:

$$\text{expected\_distance} = \frac{\text{total\_degree}}{n \cdot (n-1)}$$

όπου:

- n: Ο αριθμός των κόμβων στη συστάδα.
- total_degree: Το άθροισμα των βαθμών των κόμβων στη συστάδα.

Το αποτέλεσμα διαιρείται επιπλέον με το 2 για να διατηρηθεί η συμμετρία στις αποστάσεις, καθώς κάθε ζεύγος κόμβων μετριέται δύο φορές.

Πλεονεκτήματα & Περιορισμοί του Προσεγγιστικού Τύπου

Πλεονεκτήματα:

1. Μείωση Υπολογιστικού Κόστους:
   Σε αντίθεση με την κλασική μέθοδο, όπου θα έπρεπε να δημιουργήσουμε πολλούς τυχαίους γράφους και να υπολογίσουμε τις αποστάσεις, η $ERCQ$ προσφέρει μια γρήγορη εκτίμηση χωρίς την ανάγκη προσομοίωσης.
2. Ευκολία Εφαρμογής:
   Χρησιμοποιώντας μόνο τη διανομή βαθμών του γράφου, μπορούμε να προσεγγίσουμε την αναμενόμενη απόσταση με ελάχιστους υπολογισμούς.

Περιορισμοί:

1. Η ακρίβεια της εκτίμησης εξαρτάται από τη δομή του γράφου. Σε γράφους με ανομοιόμορφη διανομή βαθμών, η $ERCQ$ μπορεί να αποκλίνει περισσότερο από την πραγματική τιμή.

Χαρακτηριστικά του Αλγορίθμου

1. Διαχείριση μη συνδεδεμένων εξαρτημάτων
   - Ο αλγόριθμος εφαρμόζεται ξεχωριστά σε κάθε συνδεδεμένο συστατικό του γράφου, διασφαλίζοντας ότι οι κόμβοι που δεν συνδέονται άμεσα δεν θα ανήκουν στο ίδιο σύμπλεγμα.
2. Αρχικοποίηση με βάση τα πρωτο-συμπλέγματα
   - Τα πρωτο-συμπλέγματα ($proto - clusters$) ορίζονται με βάση χαρακτηριστικά όπως:
     - Κόμβοι με υψηλό βαθμό συνδεσιμότητας.
     - Κόμβοι με υψηλό συντελεστή σύμπλεξης.
     - Κόμβοι με υψηλή μετρησιμότητα μεταξύτητας.
3. Επεκτάσιμη ανάπτυξη συμπλεγμάτων

- Μετά τον καθορισμό των πρωτο-συμπλεγμάτων, οι μη εκχωρημένοι κόμβοι ενσωματώνονται στα συμπλέγματα χρησιμοποιώντας κριτήρια ελαχιστοποίησης της παρατηρούμενης απόστασης.

4. Προσεγγιστικός υπολογισμός της αναμενόμενης απόστασης
   - Για την αποδοτική εκτίμηση της αναμενόμενης απόστασης, χρησιμοποιείται ο προσεγγιστικός τύπος $ERCQ$ που βασίζεται στην κατανομή βαθμών του γράφου, μειώνοντας σημαντικά τους υπολογιστικούς χρόνους.
   - Η $ERCQ$ υπολογίζει την αναμενόμενη απόσταση μέσω ενός μαθηματικού τύπου που εξαρτάται αποκλειστικά από την κατανομή των βαθμών του γράφου και το μέγεθος του κάθε συμπλέγματος. Επειδή αυτός ο τύπος είναι καθοριστικός, τα αποτελέσματα είναι συνεπή κάθε φορά που εκτελείται ο αλγόριθμος στον ίδιο γράφο. Αυτό σημαίνει ότι οποιαδήποτε διακύμανση στα αποτελέσματα εξαλείφεται, επιτρέποντας την αξιολόγηση του Q_d με ακρίβεια και επαναληψιμότητα.
5. Συγχώνευση συμπλεγμάτων
   - Μετά τη φάση της αρχικοποίησης, ο αλγόριθμος επιχειρεί να συγχωνεύσει συμπλέγματα με στόχο τη βελτίωση της ποιότητας απόστασης. Ελέγχεται αν η συγχώνευση οδηγεί σε βελτίωση της συνάρτησης ποιότητας.
6. Αυτοπροσαρμοζόμενος χαρακτήρας
   - Ο αλγόριθμος είναι προσαρμοστικός και μπορεί να χρησιμοποιήσει διαφορετικά κριτήρια για την αρχικοποίηση και ανάπτυξη των συμπλεγμάτων, ανάλογα με τις ιδιαιτερότητες του γράφου.
7. Αποφυγή τοπικών βέλτιστων λύσεων
   - Με την ενσωμάτωση διαφορετικών φάσεων (πρωτο-συμπλέγματα, επέκταση, συγχώνευση), ο αλγόριθμος μειώνει την πιθανότητα να παγιδευτεί σε τοπικά μέγιστα, βελτιώνοντας την απόδοση σε μεγάλα ή σύνθετα γραφήματα.
8. Διαχείριση μονομελών συμπλεγμάτων
   - Ο αλγόριθμος λαμβάνει υπόψη τη διαχείριση κόμβων που αρχικά σχηματίζουν μονομελή συμπλέγματα και προσπαθεί να τα ενσωματώσει ή να τα συγχωνεύσει σε μεγαλύτερα σύνολα.

Έλεγχος του Αλγορίθμου ενάντια της Μεθόδου Μεγιστοποίησης *Modularity*

Από τα δεδομένα γραφήματα χρησιμοποιήθηκε αυτό με τίτλο *Twitter Interaction Network for the US Congress* λόγω των περιορισμών των πόρων, καθώς μεγαλύτερα γραφήματα αύξαναν κατα πολύ τους χρόνους εκτέλεσης. Παρακάτω παραθέτονται τα αποτελέσματα για διάφορες περιπτώσεις του αλγορίθμου, με διαφορετική αρχικοποίηση κάθε φορά.

	Modularity	Distance Quality	Running Time
Modularity-based Clustering	0,398	0,5337	3,66sec
Proto Clusters with Betweenness 0.01	0,166	0,067	20min
Proto Clusters with Betweenness 0.03	0,149	0,205	53min
Proto Clusters with Betweenness 0.05	0,115	0,1878	1h 44min
Proto Clusters with Betweenness 0.1	0,062	0,093	11h 47min

	Modularity	Distance Quality	Running Time
Proto Clusters with Closeness 0.01	0,178	0,141	20min
Proto Clusters with Closeness 0.03	0,124	0,098	1h 3min
Proto Clusters with Closeness 0.05	0,153	0,184	1h 48min
Proto Clusters with Closeness 0.1	0,072	0,102	11h 32min
Proto Clusters with Spectral 5 Eigenvectors	0,024	-0,195	14,16sec
Proto Clusters with Spectral 10 Eigenvectors	0,0123	-0,098	58sec
Proto Clusters with Spectral 15 Eigenvectors	0,007	-0,083	167sec
Proto Clusters with k-core Decomposition for 3 nodes	0	-0,797	167sec
Proto Clusters with k-core Decomposition for 6 nodes	0	-0,746	11sec
Proto Clusters with k-core Decomposition for 10 nodes	0	-0,709	15sec
Proto Clusters with Clustering Coefficient of 0.01	0,03	-0,42	24min
Proto Clusters with Clustering Coefficient of 0.03	0,054	-0,008	1h 13min
Proto Clusters with Clustering Coefficient of 0.05	0,169	0,215	1h 56min
Proto Clusters with Clustering Coefficient of 0.1	0,108	0,115	5h 23min

Είναι εμφανές από τα παραπάνω δεδομένα ότι η ομαδοποίηση με κριτήριο $modularity$ φάνηκε να κυριαρχεί στο συγκεκριμένο γράφημα, με πάνω από διπλάσια τιμή ποιότητας ομαδοποίησης απόστασης από τη δεύτερη θέση, η οποία ανήκει στην ομαδοποίηση με αρχικοποίηση τους κόμβους με συντελεστή ομαδοποίησης μεγαλύτερο του 0,05.

Το γεγονός ότι η ομαδοποίηση του αλγορίθμου δεν είναι τόσο αποδοτική όσο αυτή με κριτήριο $modularity$ ως προς την ποιότητα ομαδοποίησης απόστασης δεν σημαίνει ότι δεν είναι αποδοτικός. Κατάφερε να μειώσει το $observed\ distance$ σε κάθε ομάδα του γραφήματος συγκριτικά με την τυχαία ομαδοποίηση και να εντοπίσει ουσιώδεις σχέσης μεταξύ κορυφών.

Σε διαφορετικά γραφήματα, η πρώτη ομαδοποίηση φαίνεται να υστερεί όπως φαίνεται στον παρακάτω πίνακα.

	Modularity	Distance Quality	Running Time
Modularity-based Clustering	0,188	7,16	1sec
Proto Clusters with Clustering Coefficient of 0.05	0,082	9,21	10sec

Επίλογος

Στο παρόν έργο αναλύθηκε η συνάρτηση ποιότητας ομαδοποίησης απόστασης και υλοποιήθηκε ένας αλγόριθμος για τη βελτιστοποίησή της. Η μελέτη περιελάμβανε τόσο τη θεωρητική διατύπωση της συνάρτησης, όσο και την πρακτική εφαρμογή της σε διάφορες περιπτώσεις γραφημάτων, με στόχο την εξαγωγή χρήσιμων συμπερασμάτων σχετικά με τη δομή και την κατανομή των κόμβων.

Η υλοποίηση του αλγορίθμου ανέδειξε τη σημασία της κατάλληλης αρχικοποίησης των συμπλεγμάτων, καθώς διαπιστώθηκε ότι διαφορετικές μεθοδολογίες δημιουργίας πρωτο-συμπλεγμάτων οδηγούν σε σημαντικά διαφορετικά αποτελέσματα ως προς την τελική ποιότητα του διαμερισμού. Οι μέθοδοι που βασίζονται στην κεντρικότητα των κόμβων, όπως η ενδιάμεση κεντρικότητα και ο συντελεστής ομαδοποίησης, απέδωσαν υψηλότερη ποιότητα ομαδοποίησης σε σύγκριση με άλλες πιο απλοϊκές προσεγγίσεις, όπως η τυχαία ομαδοποίηση.

Επιπλέον, η χρήση προσεγγιστικών μεθόδων για την εκτίμηση των αναμενόμενων αποστάσεων, όπως η προσέγγιση $ERCQ$, συνέβαλε στη δραματική μείωση του υπολογιστικού κόστους, χωρίς να θυσιάζεται σημαντικά η ακρίβεια των αποτελεσμάτων. Η συγκριτική αξιολόγηση της προτεινόμενης μεθόδου με την κλασική μεγιστοποίηση της διαμόρφωσης ($modularity\ maximization$) κατέδειξε ότι η συνάρτηση ποιότητας απόστασης μπορεί να αποτελέσει έναν εξίσου αποτελεσματικό δείκτη αξιολόγησης της ποιότητας των συμπλεγμάτων, προσφέροντας διαφορετική οπτική στη δομή του γράφου.

Παρά τα θετικά αποτελέσματα, η εργασία ανέδειξε και ορισμένες προκλήσεις, όπως η δυσκολία στη διαχείριση μεγάλων γραφημάτων λόγω της υπολογιστικής πολυπλοκότητας του αλγορίθμου συγχώνευσης. Επιπλέον, η ευαισθησία του αλγορίθμου στην επιλογή των πρωτο-συμπλεγμάτων υποδηλώνει την ανάγκη περαιτέρω διερεύνησης βελτιστοποιημένων στρατηγικών αρχικοποίησης.

Συνολικά, η εργασία καταλήγει στο συμπέρασμα ότι η συνάρτηση ποιότητας απόστασης παρέχει έναν εναλλακτικό και ισχυρό μηχανισμό για την ανάλυση της κοινότητας των κόμβων σε ένα γράφο, με δυνατότητα εφαρμογής σε διάφορους τομείς, όπως η κοινωνική δικτύωση, η βιοπληροφορική και η ανάλυση δεδομένων. Μελλοντικές βελτιώσεις μπορούν να επικεντρωθούν στη βελτίωση της αποδοτικότητας του αλγορίθμου και στην ανάπτυξη προσαρμοστικών μεθόδων ομαδοποίησης που να ανταποκρίνονται καλύτερα σε διαφορετικά είδη δικτύων.