2014

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ
ΜΗΧΑΝΙΚΩΝ &
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΛΓΟΡΙΘΜΟΙ ΕΝΤΟΠΙΣΜΟΥ ΚΟΙΝΟΤΗΤΩΝ (COMMUNITY DETECTION ALGORITHMS)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΣΙΤΣΙΓΙΑΝΝΗΣ ΕΜΜΑΝΟΥΗΛ

Επιβλέποντες: Παναγιώτης Μποζάνης ,Αναπληρωτής Καθηγητής
Δημήτριος Κατσαρός, Λέκτορας

Φεβρουάριος 2014
Βόλος

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή και Πρόεδρο του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών και βασικό επιβλέποντα της πτυχιακής αυτής εργασίας, κ. Παναγιώτη Μποζάνη, για την υπομονή του, τις συμβουλές του και τις παρατηρήσεις του καθ' όλη την διάρκεια της εκπόνησης αλλά και για την γενικότερη συμπαράσταση του κατά την διάρκεια των σπουδών μου. Επιπρόσθετα θα ήθελα να ευχαριστήσω τον έταιρο επιβλέποντα καθηγητή κ. Δημήτριο Κατσαρό για την βοήθειά του.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την αμέριστη συμπαράστασή τους όλα αυτά τα χρόνια και χωρίς αυτούς ο δρόμος θα ήταν περισσότερο δύσκολος και πολύ λιγότερο ευχάριστος.

Περιεχόμενα

Πρόλογος

Πολλά δίκτυα του πραγματικού κόσμου οργανώνονται σύμφωνα με δομές που αποτελούν κοινότητες. Κατά την διάρκεια των χρόνων, έχουνε πραγματοποιηθεί πολλές έρευνες με σκοπό να δημιουργηθούν μέθοδοι και αλγόριθμοι που θα αποκαλύπτουν αυτή την κρυμμένη δομή ενός δικτύου, οδηγώντας σε αυτό που αποκαλούμε σήμερα **εντοπισμό κοινοτήτων.**

Ωστόσο, ένα δίκτυο μπορεί να είναι αρκετά πολύπλοκο, περιέχοντας διαφορετικές παραλλαγές σε ένα παραδοσιακό γραφικό μοντέλο. Έτσι κάθε αλγόριθμος επικεντρώνεται σε αυτές τις παραλλαγές και δημιουργεί, άμεσα ή έμμεσα, τον δικό του ορισμό της κοινότητας και σύμφωνα με αυτό τον ορισμό εξάγει τις κοινότητες.

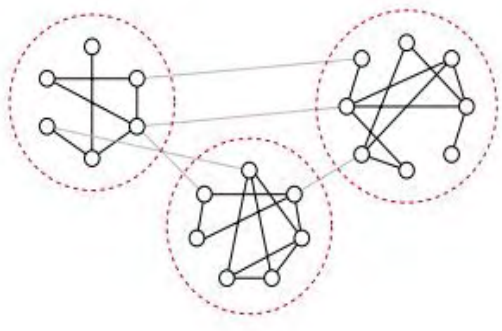
Δοσμένου ενός ορισμού της κοινότητας σε ένα κοινωνικό δίκτυο, σκοπός αυτής της εργασίας είναι να παρουσιάσει κάποιες κατηγορίες από μεθόδους εύρεσης κοινοτήτων που βασίζονται στον ορισμό τον οποίο υιοθετούν. Όσον αναφορά την διάρθρωση της ύλης σημειώνουμε ότι στο πρώτο κεφάλαιο παρουσιάζονται οι βασικές έννοιες του πολύπλοκου δικτύου και της κοινότητας και έπειτα γίνεται η κατηγοριοποίηση των αλγορίθμων και επισημαίνεται σε ποιο κεφάλαιο γίνεται η αναφορά τους.

Κεφάλαιο 1

Εισαγωγικά

1.1 Εισαγωγή

Τα πολύπλοκα δίκτυα είναι ένα πολύ δημοφιλές εργαλείο μοντελοποίησης για φαινόμενα αλληλεπίδρασης που εμφανίζονται στον πραγματικό κόσμο (για παράδειγμα στο Διαδίκτυο). Επιτρέπουν την μελέτη ενός συστήματος όπου κάθε μέλος του και κάθε σχέση του, αντιπροσωπεύεται από κόμβους και ακμές αντίστοιχα. Μια βασική λειτουργία στην ανάλυση των δικτύων είναι ο εντοπισμός κοινοτήτων. Ένας πολύ αποδεκτός ορισμός του τι είναι μια κοινότητα βασίζεται στην τοπολογία των ακμών. Σε αυτή την περίπτωση, ο ορισμός της κοινότητας διαμορφώνεται σε σχέση με τις διαφορές στις πυκνότητες των συνδέσεων σε διαφορετικά μέρη του δικτύου. Έτσι, μια κοινότητα είναι ένα υποσύνολο κόμβων με πιο πυκνές εσωτερικές συνδέσεις σε σχέση με το υπόλοιπο δίκτυο. Μια ιδιότητα, η οποία είναι κοινή σε πολλά δίκτυα, είναι η δομή των κοινοτήτων ή αλλιώς, όπως θα την αναφέρουμε και παρακάτω, η κοινοτική δομή. Η κοινοτική δομή είναι ένα σύνολο κοινοτήτων ή πιο συγκεκριμένα, ο διαχωρισμός του συνόλου των κόμβων του δικτύου(Σχήμα 1.1). Στο Διαδίκτυο, μια κοινότητα μπορεί να αντιστοιχεί σε μια ομάδα από ιστοσελίδες οι οποίες αναφέρονται στο ίδιο θέμα. Η ικανότητα εντοπισμού αυτών των ομάδων(κοινοτήτων) μπορεί να βοηθήσει στην κατανόηση της δομής των δικτύων.



Σχήμα 1.1: Ένα μικρό δίκτυο με την κοινοτική του δομή. Υπάρχουν τρεις κοινότητες (κόκκινοι κύκλοι με τις διακεκομμένες ακμές). Μέσα στις κοινότητες οι συνδέσεις είναι πιο πυκνές ενώ μεταξύ τους υπάρχουν συνδέσεις με μικρότερη πυκνότητα[1].

Ωστόσο, δύο οντότητες μπορούν να θεωρηθούν σχετικές εάν μοιράζονται μια κοινή ενέργεια παρόλο που δεν συνδέονται απευθείας. Σε αυτή την περίπτωση ο παραπάνω ορισμός είναι ανακριβής. Οι συγγραφείς του [2] αναφέρουν, ότι πολλές φορές έχει αναπτυχθεί μια προσέγγιση για τον εντοπισμό κοινοτήτων με σκοπό να αντιμετωπίσει ένα συγκεκριμένο πρόβλημα και αυτό έχει οδηγήσει στον δικό της ορισμό της κοινότητας. Σαν αποτέλεσμα υπάρχει μια πληθώρα ορισμών που έχουν οδηγήσει σε εξαιρετικές λύσεις όσο αναφορά το πρόβλημα εντοπισμού των κοινοτήτων. Οι ορισμοί που θα παρουσιαστούν σε αυτή την εργασία βασίζονται σε ένα γενικό ορισμό που προτείνεται από τους συγγραφείς του [2]:

Γενικός ορισμός (Κοινότητα): Κοινότητα σε ένα πολύπλοκο δίκτυο ονομάζεται ένα σύνολο από οντότητες που μοιράζονται κάποια στενά συσχετιζόμενα σύνολα ενεργειών/ιδιοτήτων μαζί με τις άλλες οντότητες της ίδιας κοινότητας. Εδώ, η απευθείας σύνδεση δύο οντοτήτων θεωρείται ως μία συγκεκριμένη και πολύ σημαντική ενέργεια/ιδιότητα.

Ο σκοπός ενός αλγορίθμου εντοπισμού κοινοτήτων είναι να αναγνωρίσει τέτοιες κοινότητες μέσα σε ένα δίκτυο. Το επιθυμητό αποτέλεσμα είναι μια λίστα ομαδοποιημένων οντοτήτων. Λαμβάνοντας υπόψιν τον παραπάνω ορισμό, μπορούν πλέον να μοντελοποιηθούν οι κύριες πτυχές του προβλήματος του εντοπισμού κοινοτήτων σε πολύπλοκα δίκτυα[3]:

- Ορισμοί βασισμένοι στην πυκνότητα(Density-based definitions).
- Ορισμοί βασισμένοι στην ομοιότητα των κορυφών(Vertex similarity-based definitions).
- Ορισμοί βασισμένοι στην ενέργεια(Action-based definitions).
- Ορισμοί βασισμένοι στην διάδοση επιρροής(Influence Propagation-based definitions).

Επίσης υπάρχουν πολλά χαρακτηριστικά γνωρίσματα που πρέπει να εξεταστούν στο πολύπλοκο έργο του εντοπισμού των κοινοτήτων στις δομές των γραφημάτων. Πιο κάτω παρουσιάζονται μερικά από τα χαρακτηριστικά που ένας αναλυτής θα ενδιαφερόταν να μελετήσει για τον εντοπισμό κοινοτήτων σε ένα δίκτυο. Οι ιδιότητες αυτές μπορούν να ομαδοποιηθούν σε δύο κατηγορίες. Η πρώτη κατηγορία εξετάζει τα χαρακτηριστικά της αναπαράστασης του προβλήματος, ενώ η δεύτερη τα χαρακτηριστικά της προσέγγισης του προβλήματος.

Στην πρώτη κατηγορία των χαρακτηριστικών ομαδοποιούνται όλες μαζί οι πιθανές παραλλαγές στην αναπαράσταση του αρχικού φαινομένου στον πραγματικό κόσμο. Τα πιο σημαντικά χαρακτηριστικά που εξετάζονται είναι[3]:

- Επικαλυπτόμενες Κοινότητες (Overlapping Communities).
- Κατευθυνόμενες Κοινότητες (Directed Communities).
- Σταθμισμένες Κοινότητες (Weighted Communities).
- Δυναμικές Κοινότητες (Dynamic Communities).

Η δεύτερη κατηγορία των χαρακτηριστικών συλλέγει διάφορες επιθυμητές ιδιότητες που μία προσέγγιση ενός προβλήματος θα μπορούσε να έχει. Τα χαρακτηριστικά αυτά μπορούν να καθορίσουν περιορισμούς για τα δεδομένα της εισόδου ενός προβλήματός, να βελτιώσουν την εκφραστική ισχύ των αποτελεσμάτων ή να διευκολύνουν το έργο του εντοπισμού της κοινότητας[3]:

- Απουσία Παραμέτρων (Parameter free).
- Πολυδιάστατη Είσοδος (Multidimensional input).
- Στοιχειώδης Προσέγγιση (Incremental).
- Πολυμερής Είσοδος (Multipartite input).

Στη συνέχεια οι συγγραφείς επιλέγουν να κατηγοριοποιήσουν τους αλγορίθμους εντοπισμού κοινοτήτων λαμβάνοντας υπόψιν τον ορισμό της κοινότητας από τον οποίο προήλθαν, ο οποίος εξαρτάται από τι είδους ομάδες θέλουν αυτοί οι αλγόριθμοι να εντοπίσουν.

1.2 Κατηγοριοποίηση των αλγορίθμων με βάση τον ορισμό

Στα επόμενα κεφάλαια γίνεται η παρουσίαση ορισμένων αλγορίθμων εντοπισμού κοινοτήτων. Αυτοί κατηγοριοποιούνται σε διαφορετικό κεφάλαιο ανάλογα με τον ορισμό της κοινότητας στον οποίο βασίζονται. Πιο συγκεκριμένα:

- **Ανίχνευση Γέφυρας-Bridge Detection**(Κεφάλαιο 2): Αυτό το κεφάλαιο περιλαμβάνει προσεγγίσεις εντοπισμού κοινοτήτων που βασίζονται στην ιδέα ότι οι κοινότητες είναι πυκνά τμήματα ενός γραφήματος και μεταξύ αυτών υπάρχουν λιγότερες ακμές, οι οποίες αν αφαιρεθούν μπορούν να διασπάσουν το δίκτυο. Αυτές οι ακμές είναι αποκαλούνται «γέφυρες» και τα τμήματα του δικτύου που προκύπτουν από την αφαίρεσή τους είναι οι επιθυμητές κοινότητες.

- **Διάχυση-Diffusion**(Κεφάλαιο 3): Σε αυτό το κεφάλαιο υπάρχουν οι προσεγγίσεις εντοπισμού κοινοτήτων που βασίζονται στην ιδέα ότι οι κοινότητες είναι ομάδες κόμβων οι οποίες μπορούν να επηρεαστούν από την διάχυση μιας συγκεκριμένης ιδιότητας ή πληροφορίας μέσα στο δίκτυο.

- **Εγγύτητα-Closeness**(Κεφάλαιο 4): Μια κοινότητα μπορεί επίσης να οριστεί σαν μια ομάδα από οντότητες καθεμία από την οποία μπορεί να φτάσει την άλλη μέσα από πολύ λίγες αναπηδήσεις(λίγα βήματα) πάνω από τις ακμές του γραφήματος, ενώ οι οντότητες έξω από την κοινότητα είναι σημαντικά μακριά.

- **Δομή-Structure**(Κεφάλαιο 5): Μια άλλη προσέγγιση εντοπισμού κοινοτήτων που παρουσιάζεται σε αυτό το κεφάλαιο είναι ο ορισμός της κοινότητας

ακριβώς σαν μια πολύ ακριβής και σχεδόν αμετάβλητη δομή από ακμές. Οι αλγόριθμοι που παρουσιάζονται σε αυτό το κεφάλαιο και ακολουθούν αυτή την προσέγγιση ορίζουν κάποια είδη δομών και μετά προσπαθούν να τις βρούνε αποτελεσματικά μέσα στο γράφημα.

- **Ομαδοποίηση συνδέσεων-Link Clustering**(Κεφάλαιο 6): Σε αυτό το κεφάλαιο αντί να ομαδοποιηθούν οι κόμβοι, η προσέγγιση που χρησιμοποιείται ισχυρίζεται ότι η σχέση(δηλαδή η σύνδεση), μεταξύ των κόμβων είναι αυτή που ανήκει σε μια κοινότητα και όχι οι κόμβοι. Συνεπώς, η μέθοδος που παρουσιάζεται σε αυτό το κεφάλαιο ομαδοποιεί τις ακμές του δικτύου και έτσι οι κόμβοι ανήκουν στο σύνολο των κοινοτήτων των ακμών τους.

Επιπλέον οι συγγραφείς του [2,3] προτείνουν και τρεις επιπλέον κατηγορίες. Αυτές είναι:

- **Χαρακτηριστικό απόσταση- Feature Distance**: Σε αυτή την κατηγορία συγκεντρώνονται όλες οι μέθοδοι εντοπισμού κοινοτήτων που ξεκινάνε από την υπόθεση ότι η κοινότητα αποτελείται από οντότητες που μοιράζονται ένα πολύ ακριβές σύνολο από χαρακτηριστικά, με όμοιες τιμές, δηλαδή ορίζοντας το χαρακτηριστικό απόσταση, οι οντότητες είναι κοντά η μία στην άλλη.

- **Εσωτερική πυκνότητα-Internal Density**: Σε αυτή την κατηγορία υπάρχουν εκείνες οι μέθοδοι που ορίζουν τον εντοπισμό μιας κοινότητας σαν μια διαδικασία απευθείας ανακάλυψης των πυκνών περιοχών του δικτύου.

- **Meta Clustering**: Υπάρχει ένας αριθμός μεθόδων εντοπισμού κοινοτήτων οι οποίοι δεν έχουν ένα ακριβή ορισμό για τα χαρακτηριστικά των κοινοτήτων που θέλουν να βρουν. Αντί αυτού, ορίζουν διάφορες λειτουργίες και αλγορίθμους για να συνδυάσουν τα αποτελέσματα διαφόρων προσεγγίσεων εντοπισμού κοινοτήτων και μετά χρησιμοποιούν ένα ορισμό της κοινότητας για τα αποτελέσματά τους. Εναλλακτικά, αφήνουν τον αναλυτή να ορίσει την δικιά του αντίληψη της κοινότητας και να την αναζητήσει μέσα στο γράφημα.

1.3 Βιβλιογραφία

[1] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109–1120, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, 3Department of Physics, Cornell University, Ithaca, NY 14853–2501(2003)

[2] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

[3] Ευστάθιος Ραπτοδήμος, Αλγόριθμοι Εντοπισμού Κοινοτήτων, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Φεβρουάριος 2014.

Κεφάλαιο 2
Ανίχνευση Γέφυρας (Bridge Detection)

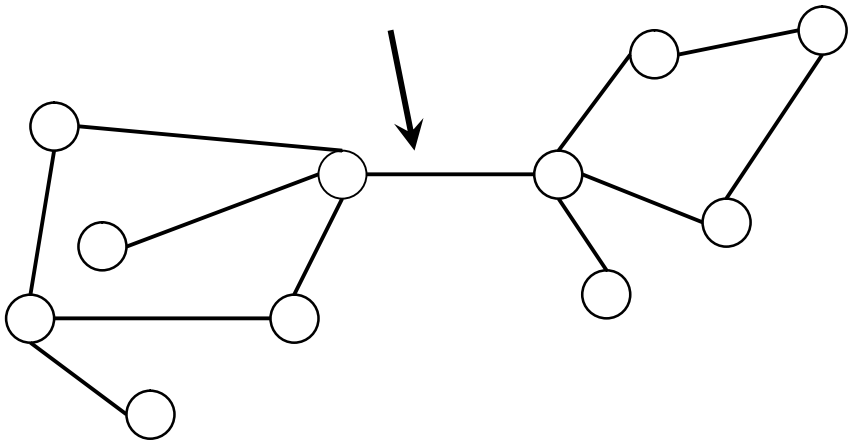
2.1 Εισαγωγή

Ο ορισμός της κοινότητας για τους αλγορίθμους σε αυτό το κεφάλαιο είναι :

Ορισμός 1 (Απομονωμένη Κοινότητα) Μια απομονωμένη κοινότητα σε ένα πολύπλοκο δίκτυο είναι ένα μέρος του δικτύου που λαμβάνεται με την αφαίρεση των αραιών γεφυρών από την δομή, που συνδέουν τα πυκνά μέρη του δικτύου.

Συνήθως οι προσεγγίσεις σε αυτή την κατηγορία εφαρμόζουν την ακόλουθη διαδικασία :

Διαδικασία 1 Κατάταξε τους κόμβους και τις ακμές στο δίκτυο σύμφωνα με ένα μέτρο της συνεισφοράς τους στη διατήρηση του δικτύου συνδεδεμένο και στη συνέχεια αφαίρεσε αυτές τις γέφυρες ή απέφυγε την επέκταση της κοινότητας με το να τις συμπεριλάβεις.



Σχήμα 2.1: Παράδειγμα ενός γραφήματος που μπορεί να διαχωριστεί προσδιορίζοντας μια «γέφυρα»[13].

Η γέφυρα που προσδιορίζεται στο Σχήμα 2.1 είναι ένα τέλειο παράδειγμα μιας ακμής για αφαίρεση, με σκοπό να αποσυνθέσουμε το δίκτυο σε μη συνδεόμενα μέρη που αντιπροσωπεύουν τις κοινότητές μας. Ο κύριος στόχος αυτών των προσεγγίσεων είναι το πώς θα βρούμε αυτές τις γέφυρες (που μπορεί να είναι είτε κόμβοι είτε ακμές) μέσα στο δίκτυο. Η πιο δημοφιλής προσέγγιση σε αυτή την κατηγορία είναι να χρησιμοποιήσουμε το μέτρο **centrality**(κεντρικότητα). Δεν γίνονται καθόλου υποθέσεις για την εσωτερική πυκνότητα των προσδιοριζόμενων ομάδων.

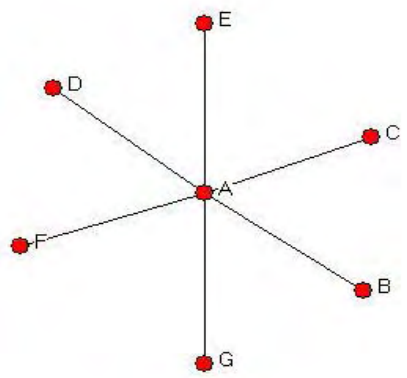
2.2 Οι διάφορες πτυχές της δύναμης και η σύνδεσή της με την έννοια centrality [1]

Στην ανάλυση κοινωνικών δικτύων, το μέτρο **centrality**(κεντρικότητα) είναι ένα μέγεθος που ορίζεται προκειμένου να επιτύχει μια ποσοτική αξιολόγηση της δομικής δύναμης μιας οντότητας σε ένα δίκτυο. Η έννοια της δύναμης είναι στενά συνδεδεμένη με την έννοια centrality. Η προσέγγιση ενός δικτύου τονίζει πως η δύναμη είναι εγγενώς σχεσιακή. Μια οντότητα δεν έχει δύναμη από μόνη της, έχει δύναμη γιατί μπορεί να κυριαρχήσει πάνω σε άλλες.

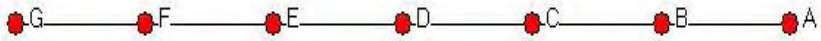
Οι αναλυτές των δικτύων συχνά περιγράφουν τον τρόπο που μια οντότητα προστίθεται σε ένα σχεσιακό δίκτυο επιβάλλοντας περιορισμούς στην οντότητα και προσφέροντας στην οντότητα ευκαιρίες. Οντότητες που αντιμετωπίζουν λιγότερους περιορισμούς και έχουν περισσότερες ευκαιρίες από άλλες, είναι σε ευνοϊκότερες δομικές θέσεις. Έχοντας ευνοϊκότερη θέση σημαίνει ότι μια οντότητα μπορεί να αποσπάσει καλύτερες τιμές σε συναλλαγές, να έχει μεγαλύτερη επιρροή και αυτή η οντότητα θα είναι εστία σεβασμού και προσοχής από ότι άλλες οντότητες σε λιγότερο ευνοϊκές θέσεις.

Τι εννοούμε με τις φράσεις «έχει ευνοϊκότερη θέση», έχει «περισσότερες ευκαιρίες» και «λιγότερους περιορισμούς»; Δεν υπάρχουν σωστές και τελικές απαντήσεις σε αυτά τα δύσκολα ερωτήματα. Ωστόσο, η ανάλυση των δικτύων έχει συμβάλει σημαντικά στη παροχή ορισμών που είναι ακριβής καθώς και σαφή μέτρα των διάφορων και διαφορετικών προσεγγίσεων για την αντίληψη της δύναμης που συνδέεται με τις θέσεις σε δομές κοινωνικών σχέσεων.

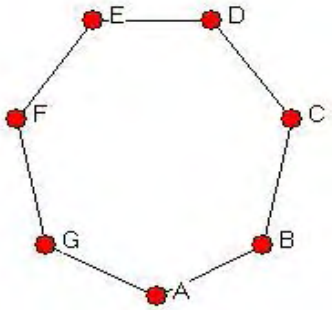
Για να καταλάβουμε τις προσεγγίσεις που η ανάλυση των δικτύων χρησιμοποιεί για να μελετήσει την δύναμη, είναι αρχικά χρήσιμο να σκεφτούμε κάποια πολύ απλά συστήματα. Οι συγγραφείς του [1] θεωρούν τρία απλά γραφήματα δικτύων που απεικονίζονται στα παρακάτω σχήματα(Σχήμα 2.2, Σχήμα 2.3, Σχήμα 2.4) και ονομάζονται «star», «line» και «circle» αντίστοιχα.



Σχήμα 2.2: «star».



Σχήμα 2.3: «line».



Σχήμα 2.4: «circle».

Μπορούμε εύκολα να παρατηρήσουμε ότι ο κόμβος Α έχει μια πιο ωφελημένη δομική θέση στο δίκτυο «star» , αν το δίκτυο περιγράφει μια σχέση όπως η ανταλλαγή πόρων ή ο διαμερισμός πόρων. Γιατί όμως ο κόμβος Α έχει «καλύτερη» θέση από τους άλλους κόμβους; Τι γίνεται με την θέση του κόμβου Α στο δίκτυο «line»; Το γεγονός ότι βρίσκεται στο τέλος της γραμμής αποτελεί μειονέκτημα ή πλεονέκτημα; Είναι πραγματικά όλοι οι κόμβοι στο δίκτυο «circle» στη ίδια δομική θέση;

2.2.1 Προσέγγιση degree (Βαθμός)

Πρέπει να σκεφτούμε γιατί η δομική θέση μπορεί να είναι μειονέκτημα ή πλεονέκτημα για ένα κόμβο. Ας επικεντρωθούμε στον κόμβο A και στο γιατί έχει πλεονέκτημα στο δίκτυο «star».

Degree: Στο δίκτυο «star», ο κόμβος A έχει περισσότερες ευκαιρίες και εναλλακτικές από τους άλλους κόμβους. Εάν ο κόμβος D επιλέξει να μην παράσχει στον A ένα πόρο, ο κόμβος A έχει ένα αριθμό από άλλα μέρη από όπου μπορεί να τον πάρει. Ωστόσο αν ο D επιλέξει να μην κάνει ανταλλαγή με τον A, τότε ο D δεν θα μπορεί να κάνει καθόλου ανταλλαγή. Όσο περισσότερα δεσίματα ένας κόμβος έχει, τόσο περισσότερο δύναμη είναι πιθανόν να έχει. Στο δίκτυο «star», ο κόμβος A έχει βαθμό (degree) έξι, ενώ όλοι οι άλλοι κόμβοι έχουνε βαθμό ένα. Αυτή η λογική αποτελεί βάση για μέτρα centrality και δύναμης που βασίζονται στον βαθμό των κόμβων. Κόμβοι που έχουν περισσότερα δεσίματα έχουνε και περισσότερες επιλογές. Η αυτονομία τους κάνει λιγότερο εξαρτημένους από άλλους κόμβους και έτσι πιο δυνατούς.

Ας θεωρήσουμε τώρα το δίκτυο «circle» σε σχέση με τον βαθμό. Κάθε κόμβος έχει τον ίδιο αριθμό εναλλακτικού συνεταίρου με τον οποίο μπορεί να πραγματοποιήσει συναλλαγή ή με άλλα λόγια έχει τον ίδιο βαθμό. Έτσι όλες οι θέσεις είναι το ίδιο πλεονεκτικές ή μειονεκτικές.

Στο δίκτυο «line» το θέμα είναι λίγο πιο περίπλοκο. Οι κόμβοι στην άκρη της γραμμής (κόμβοι A και D) έχουν δομικό μειονέκτημα, αλλά όλοι οι άλλοι είναι προφανώς ίσοι. Γενικά, κόμβοι που είναι πιο κεντρικά στη δομή, με την λογική ότι έχουν μεγαλύτερο βαθμό και περισσότερες συνδέσεις, τείνουν να έχουν πιο ωφελημένες θέσεις, άρα περισσότερη δύναμη.

2.2.2 Προσέγγιση closeness (Εγγύτητα)

Closeness: Ο δεύτερος λόγος για τον οποίο ο κόμβος A είναι πιο ισχυρός από τους άλλους κόμβους στο δίκτυο «star» είναι γιατί ο κόμβος A είναι πιο κοντά στους περισσότερους κόμβους από ότι κάθε άλλος κόμβος. Η δύναμη μπορεί να ασκηθεί από απευθείας συναλλαγή. Επιπλέον η δύναμη έρχεται από το γεγονός ότι ένας κόμβος μπορεί να είναι «σημείο αναφοράς», μέσω του οποίου οι άλλοι κόμβοι κρίνουν τον εαυτό τους καθώς επίσης και από το γεγονός ότι μπορεί να είναι το κέντρο της προσοχής όπου η προβολή του είναι ορατή από ένα μεγάλο αριθμό από κόμβους. Κόμβοι που είναι ικανοί να φτάσουν άλλους κόμβους με πιο σύντομα μονοπάτια ή που είναι πιο προσβάσιμοι από άλλους κόμβους μέσω συντομότερων μονοπατιών έχουνε πιο πλεονεκτικές θέσεις. Αυτό το δομικό πλεονέκτημα μπορεί να μεταφραστεί σε δύναμη. Στο δίκτυο «star», ο κόμβος A έχει απόσταση ένα από

όλους τους άλλους κόμβους. Οι υπόλοιποι κόμβοι έχουνε απόσταση δύο από όλους τους άλλους κόμβους (εκτός του Α). Αυτή η λογική του δομικού πλεονεκτήματος αποτελεί την βάση των προσεγγίσεων που τονίζουν την κατανομή του closeness και της απόστασης σαν πηγή ενέργειας.

Ας θεωρήσουμε το δίκτυο «circle» σύμφωνα με το closeness(εγγύτητα) του κόμβου. Κάθε κόμβος έχει διαφορετικό μήκος μονοπατιού από τους άλλους κόμβους αλλά όλοι οι κόμβοι έχουνε ίδια κατανομή του closeness και ως εκ τούτου θα είναι ίσοι σύμφωνα με την δομική τους θέση. Στο δίκτυο «line», ο μεσαίος κόμβος D είναι πιο κοντά σε όλους τους άλλους κόμβους από ότι το σύνολο C,E , το σύνολο B,F , και το σύνολο A,G. Και πάλι οι κόμβοι στα άκρα της γραμμής ή περιφερειακά είναι σε μειονεκτική θέση.

2.2.3 Προσέγγιση betweenness

Betweenness: Ο τρίτος λόγος που ο κόμβος Α είναι σε πλεονεκτική θέση στο δίκτυο «star» είναι διότι ο κόμβος Α βρίσκεται μεταξύ κάθε άλλου ζευγαριού κόμβων και κανένας κόμβος δεν βρίσκεται μεταξύ του Α και άλλων κόμβων. Εάν ο Α θέλει να επικοινωνήσει με τον F, ο Α μπορεί απλά να το κάνει. Εάν ο F θέλει να επικοινωνήσει με τον Β μπορεί να το κάνει, μέσω όμως του Α. Αυτό δίνει στον κόμβο Α την ιδιότητα του μεσάζοντα στην επικοινωνία μεταξύ των άλλων κόμβων, εξάγει «χρεώσεις» για τις «υπηρεσίες», απομονώνει κόμβους ή αποτρέπει την επικοινωνία μεταξύ άλλων κόμβων. Κατά αυτό τον τρόπο, η τρίτη πτυχή του δομικού πλεονεκτήματος μιας θέσης είναι το γεγονός ότι βρίσκεται μεταξύ άλλων κόμβων.

Στο δίκτυο «circle», κάθε κόμβος βρίσκεται μεταξύ κάθε άλλου ζευγαριού κόμβων. Βασικά υπάρχουν δύο μονοπάτια που ενώνουν κάθε ζευγάρι κόμβων και κάθε τρίτος κόμβος βρίσκεται στο ένα μονοπάτι αλλά όχι στο άλλο. Έτσι όλοι οι κόμβοι είναι το ίδιο σε πλεονεκτική ή μειονεκτική θέση. Στο δίκτυο «line», τα ακραία σημεία A,G δεν βρίσκονται μεταξύ κανενός ζευγαριού και δεν έχουνε την δύναμη του «μεσάζοντα». Οι κόμβοι πιο κοντά στο κέντρο της αλυσίδας βρίσκονται σε περισσότερα μονοπάτια μεταξύ ζευγαριών και είναι πάλι σε πλεονεκτικότερη θέση.

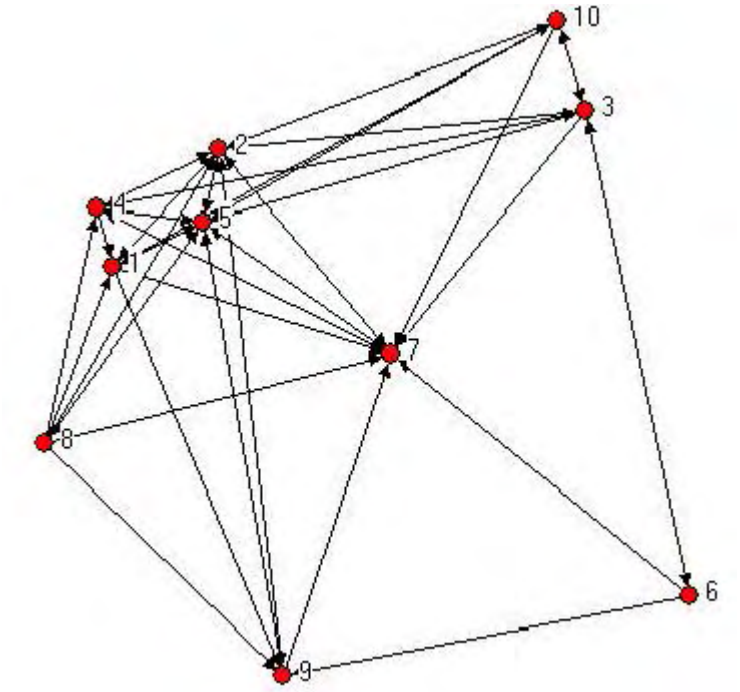
Οι αναλυτές των δικτύων είναι πιο πιθανόν να περιγράφουν τις προσεγγίσεις τους σύμφωνα με την ιδέα του centrality(κεντρικότητα) παρά με την έννοια της δύναμης. Κάθε μία από τις προσεγγίσεις που αναφέρθηκαν πιο πάνω (degree, closeness, betweenness) περιγράφουν τις θέσεις των κόμβων σύμφωνα με το πόσο κοντά είναι στο «κέντρο» της δράσης του δικτύου, ωστόσο οι ορισμοί του τι σημαίνει να βρίσκεσαι στο «κέντρο» διαφέρουν. Είναι πιο σωστό, να περιγράψουμε τις προσεγγίσεις των δικτύων με αυτό τον τρόπο δηλαδή με μέτρα centrality, παρά με μέτρα δύναμης.

Σύμφωνα με όσα αναφέρθηκαν πιο πάνω υπάρχει ένας αριθμός μέτρων που έχουνε οριστεί για να συλλάβουμε την έννοια της δύναμης μιας οντότητας σε ένα δίκτυο. Αυτός περιλαμβάνει: **Degree centrality , Closeness centrality , Betweenness centrality**.

2.2.4 Degree centrality

Degree centrality: Κόμβοι που έχουνε περισσότερα δεσίματα με άλλους κόμβους μπορεί να είναι σε πλεονεκτικότερη θέση. Επειδή έχουνε περισσότερα δεσίματα, μπορεί να έχουνε εναλλακτικούς τρόπους να ικανοποιήσουνε τις ανάγκες τους και ως εκ τούτου, είναι λιγότερο εξαρτημένοι από άλλους κόμβους. Επίσης μπορούνε να έχουνε πρόσβαση και να απευθύνονται σε περισσότερες πηγές του δικτύου συνολικά. Ακόμη, είναι συχνά μεσάζοντες σε συναλλαγές και είναι ικανοί να έχουνε οφέλη από αυτές. Έτσι ένα πολύ απλό, αλλά παράλληλα πολύ αποτελεσματικό, μέτρο για το centrality ενός κόμβου και της πιθανής του δύναμης, είναι ο βαθμός του.

Σε αυτό το σημείο αξίζει να σημειωθεί ότι σε μη κατευθυνόμενα δεδομένα, οι κόμβοι διαφέρουν ο ένας με τον άλλον από το πόσες συνδέσεις έχουν. Σε κατευθυνόμενα δεδομένα, ωστόσο, είναι σημαντικό να ξεχωρίσουμε το centrality που βασίζεται στο αριθμό των ακμών που φτάνουν σε έναν κόμβο από το centrality που βασίζεται στον αριθμό των ακμών που φεύγουν από έναν κόμβο. Εάν ένας κόμβος λαμβάνει πολλές συνδέσεις συνήθως λέγεται ότι είναι ένας «εξέχων» κόμβος ή έχει ένα υψηλό κύρος. Για αυτό τον λόγο πολλοί κόμβοι αναζητούν να κατευθύνουν συνδέσεις σε αυτούς κάτι το οποίο θα υποδείξει και την σημαντικότητά τους. Κόμβοι που έχουνε ασυνήθιστα υψηλό εξωτερικό βαθμό είναι αυτοί οι κόμβοι που είναι ικανοί να έχουνε συναλλαγές με πολλούς άλλους ή αυτοί που κρατάνε πολλούς άλλους ενήμερους μέσω της προβολής τους. Κόμβοι που παρουσιάζουν υψηλό εξωτερικό βαθμό συνήθως αποκαλούνται ισχυροί κόμβοι. Στο Σχήμα 2.5, απλά μετρώντας τις συνδέσεις που εισέρχονται σε ένα κόμβο και αυτές που εξέρχονται από έναν κόμβο μπορούμε να συμπεράνουμε ότι κάποιοι κόμβοι είναι πιο «κεντρικοί» (π.χ. 2,5,7). Ακόμα γίνεται εμφανές, από το δίκτυο που παρουσιάζεται στο Σχήμα 2.5, ότι μπορούμε να έχουμε ένα σύνολο από κεντρικούς κόμβους, σε αντίθεση με το δίκτυο «star». Μπορούμε να δούμε το centrality σαν ένα χαρακτηριστικό των κόμβων σαν συνέπεια των θέσεών τους. Μπορούμε επίσης να δούμε πόσο «centralized» είναι το γράφημα στο σύνολό του – πόσο άνιση είναι η κατανομή του centrality.



Σχήμα 2.5 : Απεικόνιση της ανταλλαγής πληροφοριών μεταξύ οργανισμών που συντελείται στο πεδίο της κοινωνικής πρόνοιας [1].

2.2.5 Closeness centrality

Closeness centrality: Μπορεί κάποιος να επικρίνει τα μέτρα του degree centrality επειδή λαμβάνουν υπόψιν μόνο τα άμεσα δεσίματα που έχει κάποιος κόμβος ή τα δεσίματα των γειτόνων του κόμβου και όχι τα έμμεσα δεσίματα με όλους τους άλλους κόμβους. Ένας κόμβος μπορεί να έχει συνδέσεις με ένα μεγάλο αριθμό από άλλους κόμβους αλλά αυτοί οι κόμβοι να είναι αποσυνδεδεμένοι από το δίκτυο στο σύνολό τους. Σε αυτή την περίπτωση ο κόμβος μπορεί να είναι σχετικά κεντρικός αλλά μόνο σε μια τοπική γειτονιά.

Η προσέγγιση του closeness centrality δίνει έμφαση στην απόσταση ενός κόμβου από όλους τους άλλους στο δίκτυο, εστιάζοντας στην απόσταση του κάθε κόμβου από όλους τους άλλους. Όσο πιο κοντά είναι μια οντότητα σε μια άλλη τόσο περισσότερη δύναμη έχει.

2.2.6 Betweenness centrality

Betweenness centrality: Ας υποθέσουμε ότι θέλουμε να επιδράσουμε σε κάποιον στέλνοντάς του πληροφορία ή να κάνουμε μια συμφωνία για την ανταλλαγή πόρων. Για να μπορέσουμε να μιλήσουμε μαζί του όμως, θα πρέπει να το κάνουμε μέσω ενός μεσάζοντα. Για παράδειγμα, έστω ότι θέλουμε να πείσουμε τον πρύτανη του πανεπιστημίου για την αγορά ενός ηλεκτρονικού υπολογιστή. Σύμφωνα με τους γραφειοκρατικούς ιεραρχικούς κανόνες θα πρέπει να προωθήσουμε την αίτησή μας στον κοσμήτορα και έπειτα σε ένα ειδικό σύμβουλο του πρύτανη. Καθένας από αυτούς τους δύο ανθρώπους μπορεί να καθυστερήσει την αίτησή μας ή να αποτρέψει την προώθησή της. Αυτό δίνει στους ανθρώπους που βρίσκονται μεταξύ αυτού που πραγματοποιεί την αίτηση και του πρύτανη, δύναμη αναφορικά με αυτόν που έκανε την αίτηση. Γενικότερα, μπορούμε να πούμε ότι η πιο σημαντική οντότητα στο δίκτυο είναι η οντότητα που είναι παρούσα στη πλειοψηφία των μονοπατιών μεταξύ όλων των άλλων οντοτήτων. Περισσότερες οντότητες εξαρτώνται από αυτή για να πραγματοποιήσει τις συνδέσεις με άλλες οντότητες και ως εκ τούτου έχει περισσότερη δύναμη.

Θα θεωρήσουμε δύο μεθόδους που βασίζονται στο παραδοσιακό ορισμό του betweenness centrality : στο πολύ αρχικό edge betweenness αλγόριθμο για την εύρεση κοινοτήτων, ο οποίος πρόσφατα αποτέλεσε το επίκεντρο περαιτέρω εξελίξεων που είχε ως αποτέλεσμα μια γενική προσέγγιση που χρησιμοποιεί split betweenness με σκοπό να επιτύχει ένα επικαλυπτόμενο πλαίσιο εύρεσης κοινοτήτων. Στη συνέχεια θα θεωρήσουμε δύο εναλλακτικές μεθόδους οι οποίες προσπαθούν να εντοπίσουν τις γέφυρες επεκτείνοντας την κοινοτική δομή και υπολογίζοντας μια συνάρτηση καταλληλότητας της κοινότητας.

2.3 Αλγόριθμος Girvan-Newman [2]

Σε αυτό το σημείο επικεντρωνόμαστε σε μια ιδιότητα που παρατηρείται σε πολλά δίκτυα, την ιδιότητα της δομής μιας κοινότητας, στην οποία οι κόμβοι ενός δικτύου είναι σφιχτά συνδεδεμένοι σε ομάδες, μεταξύ των οποίων υπάρχουν χαλαρές συνδέσεις. Ας θεωρήσουμε ένα κοινωνικό δίκτυο και πιο συγκεκριμένα ένα δίκτυο με φιλίες ή γενικότερα ένα δίκτυο γνωριμιών μεταξύ των ανθρώπων. Από την εμπειρία μας είναι εύκολο να αντιληφθούμε ότι μέσα σε αυτά τα δίκτυα υπάρχουν κοινότητες. Μπορούμε να δούμε σχηματικά τέτοιες κοινότητες στο Σχήμα 2.6 όπου υπάρχουν υποσύνολα κορυφών μέσα στα οποία οι συνδέσεις μεταξύ των κορυφών είναι πυκνές και σε άλλες περιπτώσεις είναι λιγότερο πυκνές. Η ικανότητα να εντοπίζεται η δομή μιας κοινότητας σε ένα δίκτυο έχει πρακτικές εφαρμογές. Κοινότητες σε ένα κοινωνικό δίκτυο μπορεί να αντιπροσωπεύουν πραγματικές κοινωνικές ομάδες που έχουνε τα ίδια ενδιαφέροντα ή τις ίδιες γνώσεις. Ακόμα κοινότητες στο διαδίκτυο μπορεί να αντιπροσωπεύουν ιστοσελίδες με συναφή

θέματα. Η δυνατότητα να εντοπίζονται αυτές οι κοινότητες βοηθάει στην κατανόηση και στην αξιοποίηση των δικτύων πιο αποτελεσματικά.

Παραθέτουμε σε αυτό το σημείο μια προσέγγιση για την ανίχνευση μιας κοινότητας. Αντί να προσπαθήσουμε να κατασκευάσουμε ένα μέτρο που να μας λέει ποιες ακμές είναι πιο κεντρικές στις κοινότητες, επικεντρωνόμαστε σε αυτές τις ακμές που είναι λιγότερο κεντρικές δηλαδή για ακμές που είναι πιο «ανάμεσα» στις κοινότητες.

Καταρχήν, οι συγγραφείς του [2] αναφέρονται στην έννοια vertex betweenness που είναι ένα μέτρο centrality(κεντρικότητας) και επιρροής των κόμβων στα δίκτυα. Έχοντας αρχικά προταθεί από τον Freeman[10,11], το betweenness centrality μιας κορυφής i έχει οριστεί ως ο αριθμός των συντομότερων μονοπατιών μεταξύ ζευγαριών άλλων κορυφών που διαπερνάνε το i. Είναι ένα μέτρο της επιρροής ενός κόμβου πάνω στη ροή της πληροφορίας μεταξύ άλλων κόμβων, ειδικότερα σε περιπτώσεις όπου η πληροφορία που ρέει πάνω από ένα δίκτυο ακολουθεί πρωταρχικά το συντομότερο διαθέσιμο μονοπάτι.

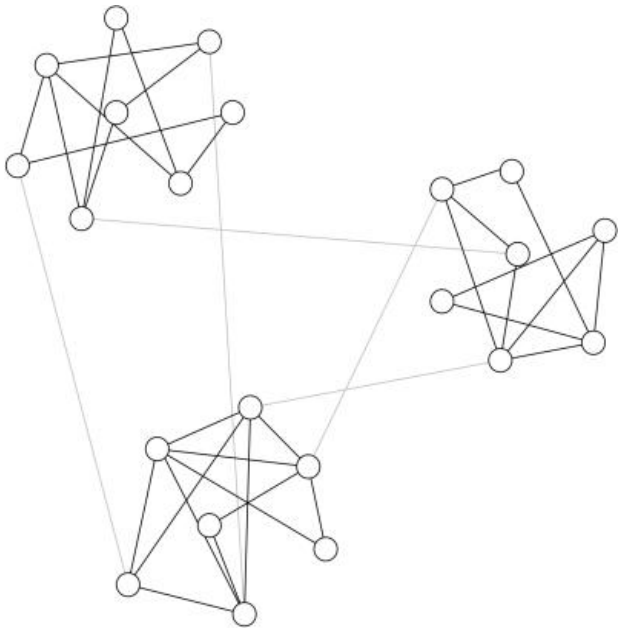
Για να βρούνε οι αναλυτές ποιες ακμές σε ένα δίκτυο πιο πολύ «ανάμεσα» άλλων ζευγαριών από κορυφές γενικοποιούνε το μέτρο του betweenness centrality του Freeman σε ακμές και ορίζουνε το edge betweenness μιας ακμής ως τον αριθμό των συντομότερων μονοπατιών μεταξύ ζευγαριών από κορυφές που περνάνε από αυτή. Εάν υπάρχουν περισσότερα από ένα συντομότερο μονοπάτι μεταξύ ενός ζευγαριού κορυφών, στο κάθε μονοπάτι δίνεται ίδιο βάρος έτσι ώστε το συνολικό βάρος όλων των μονοπατιών να είναι ίσο. Εάν ένα δίκτυο περιέχει κοινότητες ή ομάδες που είναι χαλαρά συνδεδεμένες μέσω κάποιων, ελάχιστων και μεταξύ των ομάδων, ακμών τότε όλα τα συντομότερα μονοπάτια μεταξύ διαφορετικών κοινοτήτων θα πρέπει να περνάνε από μία από αυτές τις ελάχιστες ακμές. Κατά αυτό τον τρόπο οι ακμές που ενώνουν κοινότητες θα έχουνε υψηλό edge betweenness. Όπως υποστηρίζουν οι αναλυτές, με την αφαίρεση αυτών των ακμών είναι εφικτό να διαχωρίσουμε μια ομάδα από την άλλη και έτσι να αποκαλυφθεί η υποκείμενη δομή της κοινότητας στο γράφημα. Κοιτώντας πάλι το Σχήμα 2.6 έχουμε ένα παράδειγμα όπου το χρώμα των ακμών αναλογεί στο edge betweenness. Όπως μπορούμε να παρατηρήσουμε έχουμε πάρει τις πιο υψηλές edge betweenness τιμές από τις ακμές μεταξύ των κοινοτήτων.

Ο αλγόριθμος (Girvan-Newman Algorithm) για την εύρεση κοινοτήτων έχει την ακόλουθη μορφή:

Αλγόριθμος 2.1 Girvan-Newman Algorithm

1. Υπολόγισε το betweenness για όλες τις ακμές στο δίκτυο.
2. Αφαίρεσε την ακμή με το υψηλότερο betweenness.
3. Υπολόγισε ξανά το betweenness για όλες τις ακμές που επηρεάστηκαν από την αφαίρεση.

4. Επανέλαβε από το βήμα 2 μέχρι να μην μείνει καμία ακμή.



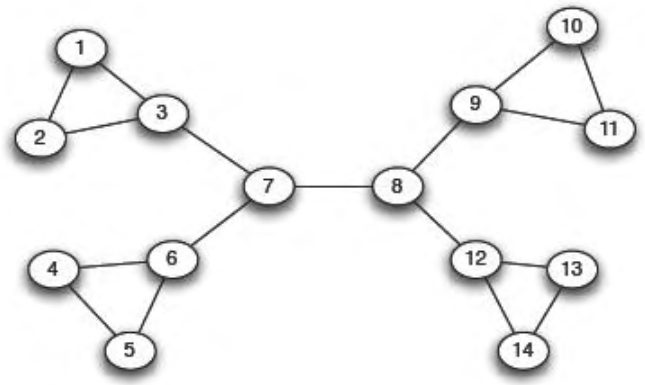
Σχήμα 2.6: Μια σχηματική αναπαράσταση ενός δικτύου με δομή κοινότητας. Σε αυτό το δίκτυο υπάρχουν τρεις κοινότητες με πυκνά συνδεδεμένες κορυφές (μαύρες γραμμές) και συνδέσεις που είναι λιγότερο πυκνές (γκρι γραμμές), μεταξύ τους [2].

Ο παραπάνω αλγόριθμος είναι από τους πρώτους που κατασκευάστηκαν. Προηγουμένως η κλασική προσέγγιση κατασκεύαζε κοινότητες προσθέτοντας τις πιο ισχυρές ακμές σε ένα αρχικά άδειο σύνολο κορυφών. Εδώ πέρα κατασκευάζουμε κοινότητες αφαιρώντας σταδιακά ακμές από το αρχικό γράφημα.

Όπως αναφέρουν οι συγγραφείς του [2] ο υπολογισμός του edge betweenness γίνεται με τον γρήγορο αλγόριθμο του Newman[12] ο οποίος υπολογίζει το betweenness σε χρόνο O(nm).

2.3.1 Παράδειγμα [14]

Ας υποθέσουμε τον παρακάτω γράφημα των δεκατεσσάρων κόμβων (Σχήμα 2.7) και ας δοκιμάσουμε να εφαρμόσουμε τον αλγόριθμο.



Σχήμα 2.7: Το αρχικό γράφημα
των δεκατεσσάρων κόμβων.

Εφαρμόζοντας το πρώτο βήμα του αλγορίθμου θα υπολογίσουμε το betweenness κάθε ακμής (edge betweenness) μετρώντας ουσιαστικά την ροή της πληροφορίας ανάμεσα σε όλα τα ζευγάρια των κόμβων που χρησιμοποιούνε αυτή την ακμή. Αρχικά για δύο κόμβους x και y που συνδέονται μέσω ενός μονοπατιού θέτουμε 1 μονάδα "ροής" . Κατά αυτό τον τρόπο στο πρώτο βήμα θα ισχύει:

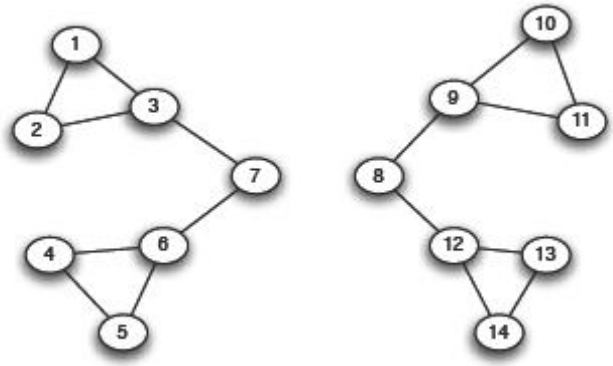
- Ας δούμε την ακμή που συνδέει τους κόμβους 7 και 8 και ας την ονομάσουμε ***7-8***. Για κάθε ζευγάρι κόμβων ανάμεσα στους [1-7] και [8-14] που χρησιμοποιεί την **7-8** υπολογίζουμε την ροή. Λίγο αναλυτικότερα, το ζευγάρι **1-8** μας δίνει ροή 1, το ίδιο και το ζευγάρι **1-9**. Για το ζευγάρι **1-10** έχουμε τρία μονοπάτια, μέσω των κόμβων 3,7,8,9, μέσω των κόμβων 2,3,7,8,9,11 και μέσω των κόμβων 3,7,8,9,11, όμως κάθε φορά διαλέγουμε το συντομότερο μονοπάτι. Κάτι αντίστοιχο συμβαίνει και με το ζευγάρι **1-11**. Συνεχίζοντας και με τους κόμβους 12,13 και 14 θα έχουμε παρόμοια αποτελέσματα. Μέχρι στιγμής, έχοντας κάνει τα παραπάνω βήματα, η συνολική ροή μας είναι 7. Ακολουθώντας την ίδια διαδικασία για κάθε ζευγάρι κόμβων ανάμεσα στους [1-7] και [8-14] που χρησιμοποιεί την **7-8** θα πάρουμε μια συνολική ροή που θα είναι ίση με 49.

- Για την συνέχεια ας δούμε την ακμή **3-7**. Για κάθε ζευγάρι κόμβων ανάμεσα στους [1-3] και [4-14] που χρησιμοποιεί την **3-7** θα υπολογίσουμε και εδώ την ροή. Ακολουθώντας το ίδιο σκεπτικό με πριν βρίσκουμε ότι η συνολική ροή είναι ίση με 33. Την ίδια συνολική ροή, ακολουθώντας την ίδια διαδικασία, θα έχουμε για τις ακμές **6-7**, **8-9,** και **8-12**.

- Κοιτώντας την ακμή **1-3,** θα υπολογίσουμε την ροή για κάθε ζευγάρι κόμβων ανάμεσα στους [1] και [3-14] (όχι τον κόμβο 2, γιατί δεν υπάρχει σύντομο μονοπάτι μεταξύ αυτού του κόμβου και κανενός άλλου που να χρησιμοποιεί την ακμή **1-3).** Η συνολική ροή θα είναι ίση με 12. Τα ίδια αποτελέσματα θα έχουμε και για τις ακμές **2-3**, **4-6**, **5-6**, **9-10**, **9-11**, **12-13** και **12-14**.

- Για την ακμή **1-2** έχουμε ότι η συνολική ροή πληροφορίας είναι ίση με 1 και αυτό γιατί η ακμή **1-2** δεν βρίσκεται σε κανένα σύντομο μονοπάτι μεταξύ ενός ζεύγους κόμβων. Το μόνο ζεύγος κόμβων που χρησιμοποιεί την ακμή **1-2** είναι οι ίδιοι οι κόμβοι 1 και 2 που συνδέονται μεταξύ τους με το συντομότερο μονοπάτι. Ακριβώς το ίδιο συμβαίνει και με τις ακμές **4-5**, **10-11** και **13-14**.

Έτσι υπολογίσαμε στο πρώτο βήμα το betweenness των ακμών του γραφήματος και στο βήμα 2 αφαιρούμε την ακμή με το μεγαλύτερο. Στο παράδειγμά μας είναι η ακμή **7-8**. Έτσι μετά το βήμα 2 το αρχικό μας γράφημα παίρνει την μορφή που φαίνεται στο Σχήμα 2.8.



Σχήμα 2.8: Βήμα 2 ($1^{η}$ επανάληψη).

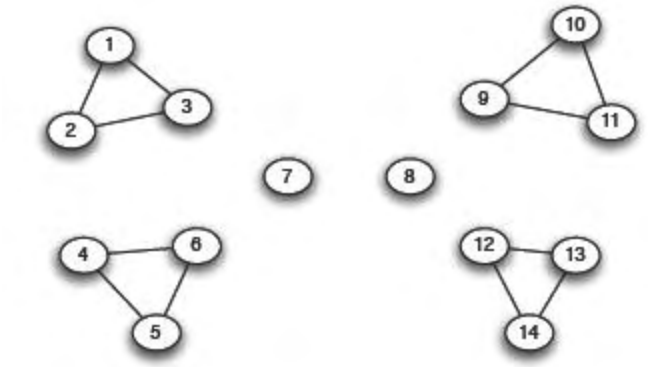
Πηγαίνοντας στο βήμα 3 υπολογίζουμε πάλι το betweenness των ακμών στον νέο μας πλέον γράφημα.

- Για κάθε ζευγάρι κόμβων ανάμεσα στους [1-3] και [4-7] που χρησιμοποιεί την ακμή **3-7** υπολογίζουμε την ροή. Όπως πριν, βρίσκουμε ότι η συνολική ροή της ακμής είναι ίση με 12. Το ίδιο συμβαίνει και με τις ακμές **6-7**, **8-9** και **8-12**.

- Για την ακμή **1-3**, υπολογίζουμε την ροή για κάθε ζευγάρι κόμβων ανάμεσα στους [1] και [3-7] (όχι τον κόμβο 2). Η συνολική ροή είναι ίση με 5. Το ίδιο ισχύει και για τις ακμές **2-3**, **4-6**, **5-6**, **9-10**, **9-11**, **12-13** και **12-14**.

- Για την ακμή **1-2** έχουμε ότι η συνολική ροή πληροφορίας είναι ίση με 1 για τους λόγους που εξηγήσαμε και προηγουμένως. Το ίδιο συμβαίνει και με τις ακμές **4-5**, **10-11** και **13-14**.

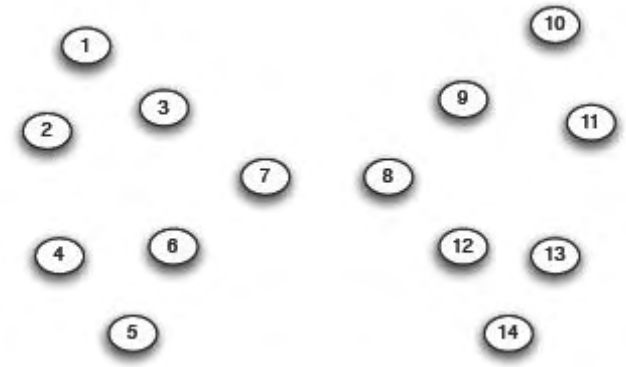
Το βήμα 4 μας οδηγεί πάλι στο βήμα 2 για μία δεύτερη επανάληψη. Έτσι αφαιρούμε τις ακμές με το μεγαλύτερο betweenness (μεγαλύτερη ροή). Σε αυτή την περίπτωση είναι οι ακμές **3-7**, **6-7**, **8-9** και **8-12**. Έτσι μετά το βήμα 2 της $2^{ης}$ επανάληψης προκύπτει το γράφημα όπως φαίνεται στο Σχήμα 2.9.



Σχήμα 2.9: Βήμα 2 ($2^{η}$ επανάληψη).

Συνεχίζοντας στο βήμα 3 της $2^{ης}$ επανάληψης θα υπολογίσουμε τις ροές των ακμών που έχουν απομείνει.

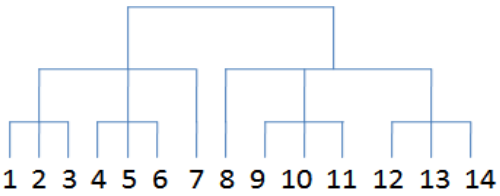
- Για την ακμή **1-3**, είναι προφανές πλέον ότι η συνολική ροή είναι 1. Το ίδιο θα ισχύει για όλες τις υπόλοιπες ακμές. Έτσι τελικά, θα καταλήξουμε σε ένα γράφημα χωρίς ακμές όπως φαίνεται και στο Σχήμα 2.10. Η εφαρμογή του αλγορίθμου για αυτό το γράφημα τελειώνει στο βήμα 4 της $2^{ης}$ επανάληψης αφού δεν έχει μείνει καμία ακμή.



Σχήμα 2.10: Βήμα 4 ($2^{ης}$ επανάληψης)
και τέλος του αλγορίθμου.

Οι συγγραφείς του [2] αναφέρουν ότι ένας παραδοσιακός τρόπος για τον εντοπισμό της δομής των κοινοτήτων σε ένα δίκτυο είναι μέσω ενός δέντρου(tree) το οποίο οι αναλυτές καλούνε δενδρόγραμμα (dendrogram). Στο κατώτερο επίπεδο αυτού του δέντρου απεικονίζονται οι κόμβοι του δικτύου και σε κάθε ανώτερο επίπεδο υπάρχουν συνεχώς αυξανόμενα τμήματα(συνδεδεμένα υποσύνολα κορυφών) που αντιπροσωπεύουν τις κοινότητες. Στο Σχήμα 2.11 απεικονίζεται το

δενδρόγραμμα του εντοπισμού της κοινοτικής δομής του δικτύου σύμφωνα με τον αλγόριθμο που παρουσιάστηκε.



Σχήμα 2.11: Δενδρόγραμμα στο οποίο απεικονίζεται η ιεραρχική ομαδοποίηση του δικτύου μας.

2.3.2 Πολυπλοκότητα

Ο αλγόριθμος Girvan-Newman υπολογίζει το betweenness για όλες τις m ακμές σε ένα γράφημα n κορυφών σε χρόνο O(nm). Πρόσφατα έχει προταθεί μια επιτάχυνση για παράλληλα συστήματα τα οποία είναι γραμμικά [3]. Εφόσον ο υπολογισμός πρέπει να επαναληφθεί κάθε φορά που αφαιρείται μια ακμή, η χειρότερη περίπτωση της χρονικής πολυπλοκότητας ολόκληρου του αλγορίθμου είναι $O(m^2n)$.

2.4 Αλγόριθμος Cluster-Overlap Newman - Girvan

Ο προηγούμενος αλγόριθμος υπέθετε πως οι κοινότητες είναι διασπασμένες, τοποθετώντας κάθε κορυφή σε μία μόνο ομάδα. Ωστόσο ,στον πραγματικό κόσμο οι κοινότητες είναι συνήθως επικαλυπτόμενες. Για αυτό τον λόγο θα παρουσιάσουμε τον Cluster–Overlap Newman Girvan (CONGA) αλγόριθμο, που βασίζεται στον Girvan- Newman αλγόριθμο που εξετάσαμε πριν αλλά έχει επεκταθεί ώστε να εντοπίζει επικαλυπτόμενες κοινότητες. Ο CONGA αλγόριθμος προσθέτει την δυνατότητα του διαχωρισμού κορυφών μεταξύ των κοινοτήτων, βασιζόμενος στη καινούργια ιδέα του split betweenness[4].

2.4.1 Splitting Vertices

Στον Girvan- Newman αλγόριθμο η βασική λειτουργία είναι η αφαίρεση μιας ακμής. Οι συγγραφείς του [4] προσθέτουνε μια καινούργια λειτουργία, την διάσπαση μιας κορυφής. Αν διασπαστεί μια κορυφή, έστω v, αυτή διαχωρίζετε πάντα σε δύο κορυφές v_1 και v_2. Οι ακμές που είχαν την κορυφή v στην άκρη τους να ανακατευθύνονται στην v_1 ή στην v_2 έτσι ώστε οι v_1 και v_2 να έχουν τουλάχιστον μια ακμή. Διασπώντας επαναλαμβανόμενα, μια κορυφή v μπορεί τελικά να διασπαστεί στο πολύ $d_{(v)}$ κορυφές, όπου $d_{(v)}$ είναι ο βαθμός της κορυφής v. Οι κορυφές διασπώνται αυξητικά κατά την διάρκεια της ομαδοποίησης. Αυτή η δυαδική διάσπαση ταιριάζει με τον Girvan- Newman αλγόριθμο γιατί, όπως αφαιρούσαμε μια ακμή, η διάσπαση μιας κορυφής μπορεί να προκαλέσει κάθε ομάδα να διασπαστεί σε δύο.

2.4.2 Split betweenness

Το σημαντικό σημείο με τον Cluster –Overlap Newman Girvan (CONGA) αλγόριθμο είναι η έννοια του split betweenness. Αυτό μας δίνει την δυνατότητα να αποφασίσουμε:

1. Πότε να διασπαστεί μια κορυφή, αντί να αφαιρεθεί μια ακμή.
2. Ποια κορυφή θα διασπαστεί.
3. Πως θα διασπαστεί η κορυφή.

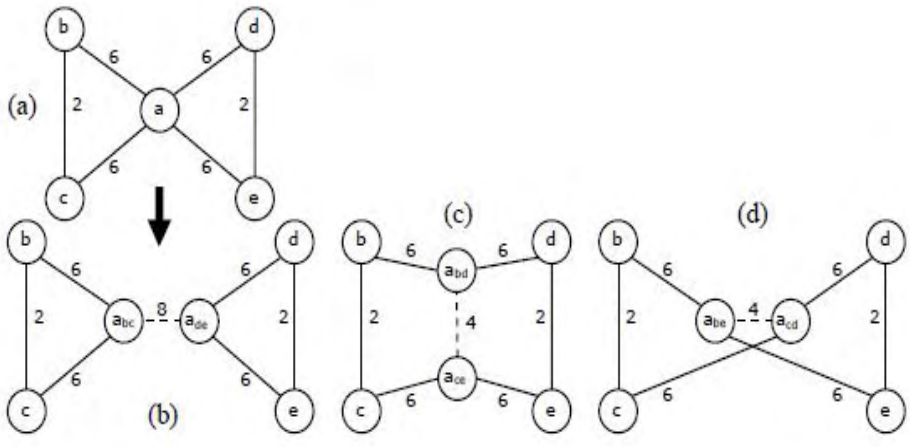
Μια κορυφή v θα πρέπει να διασπαστεί μόνο σε δύο κορυφές v_1 και v_2 ,αν αυτές οι δύο κορυφές ανήκαν σε διαφορετικές ομάδες. Μπορούμε να το εξακριβώσουμε αυτό μετρώντας τον αριθμό των συντομότερων μονοπατιών που θα περνούσαν ανάμεσα από τις v_1 και v_2 εάν ενώνονταν από μια ακμή. Έτσι, εάν υπήρχαν περισσότερα συντομότερα μονοπάτια στην ακμή $\{v_1, v_2\}$ από ότι σε κάθε άλλη πραγματική ακμή, η κορυφή θα έπρεπε να διασπαστεί. Διαφορετικά, μια ακμή θα πρέπει να αφαιρεθεί, ως συνήθως. Αυτή είναι η βάση της μεθόδου της διάσπασης μιας κορυφής που προτείνουν οι αναλυτές και έχει ως εξής:

Για κάθε διάσπαση μιας κορυφής v σε v_1 και v_2 προσθέτουνε μια φανταστική ακμή ανάμεσα στις v_1 και v_2. Αν u είναι γείτονας της v_1 και w γείτονας της v_2 , όλα τα συντομότερα μονοπάτια που περνούσαν από την v μέσω των ακμών {u,v}, {v,w} τώρα περνάνε από τις $\{u,v_1\}$, $\{v_1,v_2\}$, $\{v_2,w\}$. Η φανταστική ακμή έχει μηδενικό κόστος, γιατί τα μήκη των μονοπατιών που την διασχίζουν δεν έχουν αλλάξει και επιπλέον δεν έχουν δημιουργηθεί καινούργια συντομότερα μονοπάτια. Μονοπάτια που ξεκινούν από την κορυφή v δεν διασχίζουν αυτή την ακμή. Υπολογίζουμε το

betweenness της φανταστικής ακμής $\{v_1, v_2\}$, έστω $c_B(\{v_1, v_2\})$. Υπάρχουν $2^{d(v)-1}$ -1 τρόποι να διασπαστεί μια κορυφή v σε δύο. Καλούμε την διάσπαση που μεγιστοποιεί το $c_B(\{v_1, v_2\})$ την καλύτερη διάσπαση της v και την μέγιστη τιμή του $c_B(\{v_1, v_2\})$ την ονομάζουμε split betweenness της κορυφής v.

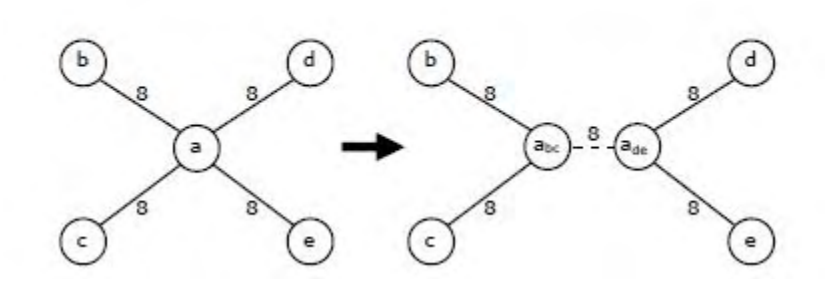
Τροποποιώντας τον Girvan- Newman αλγόριθμο, σε κάθε βήμα λαμβάνεται υπόψιν το split betweenness κάθε κορυφής καθώς και το edge betweenness κάθε ακμής. Εάν το μέγιστο split betweenness είναι μεγαλύτερο από το μέγιστο edge betweenness η αντίστοιχη κορυφή διασπάται. Οι φανταστικές ακμές δεν προστίθενται στο δίκτυο αλλά χρησιμοποιούνται για τον υπολογισμό του split betweenness. Συνοπτικά το edge betweenness μιας ακμής, όπως αναφέραμε στον προηγούμενο αλγόριθμο, ορίζεται ως ο αριθμός των συντομότερων μονοπατιών μεταξύ ζευγαριών από κορυφές που περνάνε από αυτή. Το split betweenness μιας κορυφής v ορίζεται ως ο αριθμός των συντομότερων μονοπατιών που θα περάσουν μεταξύ των δύο τμημάτων της κορυφής v αν αυτή χωριστεί.

Υπάρχουν διάφοροι τρόποι να χωρίσουμε μια κορυφή σε δύο μέρη. Ο καλύτερος διαχωρισμός είναι αυτός που μεγιστοποιεί το split betweenness. Το Σχήμα 2.12(a) δείχνει ένα δίκτυο που αποτελείται από δύο επικαλυπτόμενες ομάδες: {a, b, c} και {a, d, e}. Οι ετικέτες στις ακμές δηλώνουν το edge betweenness (τα συντομότερα μονοπάτια έχουν μετρηθεί και προς τις δύο κατευθύνσεις). Το Σχήμα 2.12(b) μας δείχνει τον καλύτερο διαχωρισμό της κορυφής a στις a_{bc} και a_{de} με την φανταστική ακμή (betweenness 8) να δηλώνεται με διακεκομμένη ακμή. Τα σχήματα 2.12(c), 6.12(d) δείχνουν δύο άλλους πιθανούς διαχωρισμούς της κορυφής a. Σε αυτά τα δύο σχήματα η φανταστική ακμή έχει χαμηλότερο betweenness, τιμή 4, αποδεικνύοντας ότι ο διαχωρισμός στο Σχήμα 2.12(b) είναι ο καλύτερος διαχωρισμός και το split betweenness της κορυφής a είναι 8. Επειδή είναι μεγαλύτερο από οποιοδήποτε edge betweenness, η κορυφή a πρέπει πράγματι να χωριστεί.



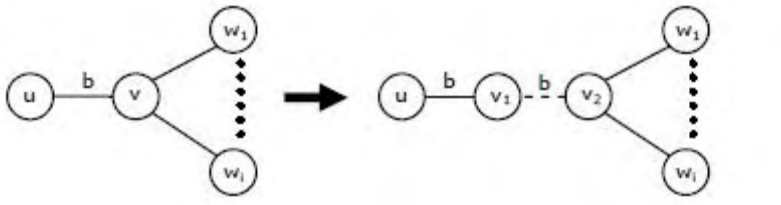
Σχήμα 2.12: (a) Δίκτυο **(b)** Καλύτερος διαχωρισμός για την κορυφή a **(c),(d)** Άλλοι διαχωρισμοί για την κορυφή a [4].

Στο Σχήμα 2.13 έχουμε ένα δίκτυο που δεν μπορεί να παρουσιάσει ομαδοποίηση. Οποιαδήποτε από τις τέσσερις διασπάσεις της a είναι η καλύτερη διάσπαση. Το split betweenness της a είναι 8, που είναι ίδιο με το betweenness κάθε ακμής. Παρόλα αυτά, εξορισμού, αφαιρείται οποιαδήποτε ακμή αντί να διασπαστεί η a.



Σχήμα 2.13: Καλύτερη διάσπαση της κορυφής a.
Το split betweenness της a είναι 8 [4].

Η μέθοδος αυτή δεν θα διασπάσει ποτέ μια κορυφή σε v_1 και v_2 έτσι ώστε η v_1 να έχει μόνο ένα γείτονα, τον u. Αυτό συμβαίνει γιατί το betweenness της ακμής $\{v_1, v_2\}$ θα μπορούσε να είναι ίδιο με αυτό της ακμής $\{u, v_1\}$, όπως φαίνεται στο Σχήμα 2.14. Έτσι, θα ήταν προτιμότερο η αφαίρεση της {u, v} από ότι η διάσπαση της v. Σαν αποτέλεσμα αυτού, οι κορυφές με βαθμό λιγότερο από 4 δεν διασπώνται ποτέ. Γενικότερα, τώρα υπάρχουν $2^{d(v)-1}$-d(v)-1 τρόποι να διασπαστεί μια κορυφή σε δύο.



Σχήμα 2.14: Η κορυφή v δεν θα
διασπαστεί σε κορυφές με βαθμό 1 [4].

2.4.3 Vertex betweenness και Split betweenness

Όταν διασπάται μια κορυφή v σε v_1 και v_2 το split betweenness είναι ο αριθμός των συντομότερων μονοπατιών μεταξύ των γειτόνων της v_1 και των γειτόνων της v_2 δια μέσω των κορυφών v_1 και v_2. Εξορισμού αυτός δεν είναι μεγαλύτερος από τον

αριθμό των συντομότερων μονοπατιών που περνάνε μέσω της κορυφής, δηλαδή από το vertex betweenness της v. Το vertex betweenness $c_B(v)$ μπορεί να υπολογιστεί από το edge betweenness από την παρακάτω εξίσωση:

$$c_B(v) = \frac{1}{2} \sum_{e \in \Gamma(v)} c_B(e) - (n-1) \qquad (1)$$

όπου Γ(v) είναι το σύνολο των ακμών προσπίπτουσες στην v και n ο αριθμός των κορυφών. Επιπλέον όπως αναφέρουν οι αναλυτές μπορούμε να χρησιμοποιήσουμε το vertex betweenness σαν ένα όριο στο split betweenness. Έτσι, αν η τιμή του vertex betweenness δεν είναι μεγαλύτερη από το μέγιστο edge betweenness δεν χρειάζεται να υπολογίσουμε το split betweenness της v.

2.4.4 Υπολογισμός του split betweenness

Για να υπολογίσουμε το split betweenness και την καλύτερη διάσπαση μιας κορυφής v, πρώτα υπολογίζουμε το pair betweenness της v. Το pair betweenness της κορυφής v για την ακμή {u,w}, όπου u και w γείτονες της v και u≠w, είναι ο αριθμός των συντομότερων μονοπατιών που διασχίζουν τις ακμές {u,v} και {v,w}. Το vertex betweenness της v είναι το άθροισμα όλων των pair betweenness της v.

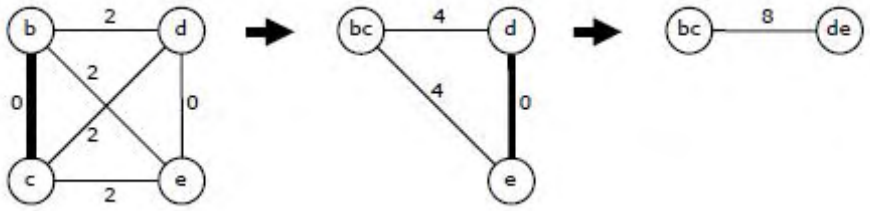
Μπορούμε να παρουσιάσουμε το pair betweenness της v, βαθμού k, με μία k-clique στην οποία κάθε κορυφή παίρνει μια ετικέτα από ένα γείτονα της v και κάθε ακμή {u,w} παίρνει ετικέτα από την τιμή του pair betweenness της v για την ακμή {u,w}. Για να βρούμε τη καλύτερη διάσπαση της v ακολουθούμε τα εξής βήματα:

1. Διάλεξε την ακμή {u,w} με την μικρότερη τιμή.
2. Ένωσε u και w σε μια κορυφή, uw.
3. Για κάθε κορυφή x μέσα στην κλίκα, αντικατέστησε τις ακμές {u,x}, με τιμή b_1 , {w,x} με τιμή b_2 με μια καινούργια ακμή {uw,x} και με τιμή $b_1 + b_2$.
4. Επανέλαβε από το βήμα 1 k-2 φορές (συνολικά).

Οι ετικέτες στις εναπομείναντες δύο κορυφές δείχνουνε την διάσπαση και η τιμή στην ακμή είναι το split betweenness.

Αυτός ο αλγόριθμος δεν μας εγγυάται ότι θα βρει την καλύτερη διάσπαση. Για να το καταφέρουμε αυτό θα έπρεπε να δοκιμάσουμε όλες τις ακμές στο βήμα 1 σε κάθε επανάληψη, κάτι που θα απαιτούσε εκθετικό χρόνο. Η «άπληστη» μέθοδος που παρουσιάστηκε είναι πιο αποτελεσματική και στην πράξη βρίσκει συνήθως την καλύτερη διάσπαση ή μια κοντινή προσέγγιση σε αυτή. Στο Σχήμα 2.15 φαίνεται η

εύρεση της καλύτερης διάσπασης της a. Υπάρχουν k-2 φάσεις. Με μαύρο χρώμα δείχνουμε τις ακμές που διαλέγουμε σε κάθε φάση.



Σχήμα 2.15: Η καλύτερη διάσπαση της κορυφής a του σχήματος 2.12 [4].

Ο CONGA [5] περιλαμβάνει μια σειρά από βήματα, καθένα από τα οποία αφαιρεί μια ακμή από το δίκτυο ή χωρίζει μια κορυφή σε δύο:

Αλγόριθμος 2.2 Cluster –Overlap Newman Girvan (CONGA) Algorithm

---

1. Υπολόγισε το edge betweenness των ακμών και το split betweenness των κορυφών.
2. Αφαίρεσε την ακμή με το μέγιστο edge betweenness ή χώρισε μια κορυφή με το μέγιστο split betweenness, αν είναι μεγαλύτερο.
3. Υπολόγισε ξανά το edge betweenness και το split betweenness.
4. Επανέλαβε από το βήμα 2 μέχρι να μην μείνει καμία ακμή.

Στον CONGA αλγόριθμο το δίκτυο συμπεριφέρεται σαν μια κοινότητα θεωρώντας πως είναι συνδεδεμένο. Μετά από μία ή περισσότερες επαναλήψεις, το βήμα 2 αναγκάζει το δίκτυο να χωριστεί σε δύο τμήματα (κοινότητες). Οι κοινότητες επαναλαμβανόμενα χωρίζονται σε δύο μέχρι να μείνουνε μεμονωμένες κοινότητες.

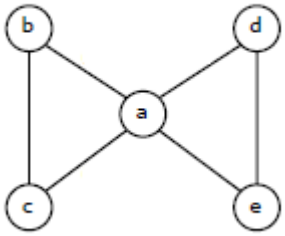
2.4.5 Υπολογισμός του pair betweenness

Υπάρχει μια επιβάρυνση, τόσο σε χρόνο όσο και σε χώρο στον υπολογισμό του pair betweenness κατά την διάρκεια του edge betweenness. Τις περισσότερες φορές αυτή η πληροφορία δεν χρειάζεται γιατί μπορούμε να διαπιστώσουμε μέσω του vertex betweenness αν μια κορυφή δεν πρέπει να διασπαστεί. Σύμφωνα με τους συγγραφείς του [4] ο παραπάνω αλγόριθμος μπορεί να γίνει ως εξής:

Αλγόριθμος 2.3 Complete Cluster –Overlap Newman Girvan (CONGA) Algorithm

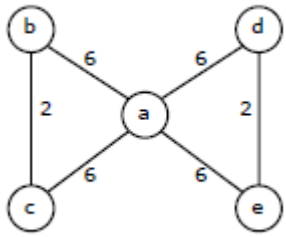
1. Υπολόγισε το edge betweenness όλων των ακμών στο δίκτυο.
2. Υπολόγισε το vertex betweenness των κορυφών, μέσω του edge betweenness, χρησιμοποιώντας την εξίσωση (1).
3. Βρες ένα υποψήφιο σύνολο κορυφών των οποίων το vertex betweenness είναι μεγαλύτερο από το μέγιστο edge betweenness.
4. Αν το σύνολο δεν είναι άδειο, υπολόγισε το pair betweenness των υποψήφιων κορυφών και στη συνέχεια υπολόγισε το split betweenness των υποψηφίων κορυφών.
5. Αφαίρεσε την ακμή με το μέγιστο edge betweenness ή διάσπασε την κορυφή με το μεγαλύτερο split betweenness (αν είναι μεγαλύτερο).
6. Υπολόγισε πάλι το edge betweenness για τις ακμές που έχουν μείνει αφότου αφαιρέσαμε την ακμή ή διασπάσαμε την κορυφή.
7. Επανέλαβε από το βήμα 2 μέχρι να μην μείνει καμία ακμή.

Μπορούμε τώρα να δούμε ολοκληρωμένα την εφαρμογή του αλγορίθμου στο δίκτυο του σχήματος 2.12 (a) και όπως αυτό παρουσιάζεται Σχήμα 2.16.



Σχήμα 2.16: Αρχικό δίκτυο [4].

Στο βήμα 1 υπολογίζεται το edge betweenness όλων των ακμών (τα συντομότερα μονοπάτια υπολογίζονται και προς τις δύο κατευθύνσεις). Για την ακμή {b, c} είναι προφανές ότι η τιμή του edge betweenness είναι 2. Το ίδιο ισχύει και για την ακμή {d, e}. Για την ακμή {b, a} θα αναζητήσουμε τα συντομότερα μονοπάτια ανάμεσα στα ζευγάρια των κόμβων [b] και [a,d,e] (όχι του c). Εύκολα καταλήγουμε στην τιμή του edge betweenness το οποίο θα είναι ίσο με 6. Με αντίστοιχο τρόπο βρίσκουμε την ίδια τιμή του edge betweenness και για τις ακμές {a, d}, {c, a} και {a, e}. Στο Σχήμα 2.17 απεικονίζεται το αρχικό δίκτυο με τις τιμές του edge betweenness να αναγράφεται σε κάθε ακμή.

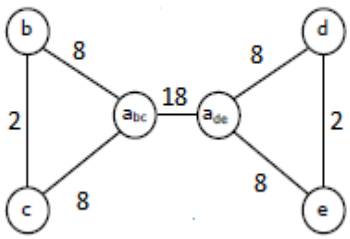


Σχήμα 2.17: Δίκτυο με τις τιμές του betweenness σε κάθε ακμή.

Στο βήμα 2 υπολογίζεται το vertex betweenness μέσω της εξίσωσης (1). Έτσι θα έχουμε:

- $c_B(b) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(6+2) - (5-1) = \frac{1}{2}8 - 4 = 4 - 4 = 0$
  όπου $\Gamma(v) = \{\{b,a\},\{b,c\}\}$.
- $c_B(c) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(6+2) - (5-1) = \frac{1}{2}8 - 4 = 4 - 4 = 0$
  όπου $\Gamma(v) = \{\{c,b\},\{c,a\}\}$.
- $c_B(d) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(6+2) - (5-1) = \frac{1}{2}8 - 4 = 4 - 4 = 0$
  όπου $\Gamma(v) = \{\{d,a\},\{d,e\}\}$.
- $c_B(e) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(6+2) - (5-1) = \frac{1}{2}8 - 4 = 4 - 4 = 0$
  όπου $\Gamma(v) = \{\{e,d\},\{e,a\}\}$.
- $c_B(a) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(6+6+6+6) - (5-1) = \frac{1}{2}24 - 4 = 12 - 4 = 8$
  όπου $\Gamma(v) = \{\{a,b\},\{a,d\},\{a,c\},\{a,e\}\}$.

Πηγαίνοντας στο βήμα 3 βρίσκουμε ότι μόνο η κορυφή a έχει vertex betweenness μεγαλύτερο από το μέγιστο edge betweenness. Στο βήμα 4 υπολογίζεται το pair betweenness και εν τέλει το split betweenness όπως δείξαμε πιο πάνω. Έτσι στο βήμα 5 αποφασίζουμε να διασπάσουμε την ακμή a και το δίκτυο μας γίνεται όπως φαίνεται στο Σχήμα 2.18.

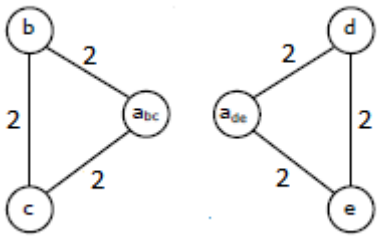


Σχήμα 2.18: Το δίκτυο μετά το βήμα 5 της 1ης επανάληψης.

Στο βήμα 5 υπολογίζεται εκ νέου το edge betweenness στο δίκτυο που προέκυψε. Για τις ακμές {b ,c} και {d ,e} η τιμή του edge betweenness θα είναι πάλι 2. Για την ακμή {b, a_{bc}} υπολογίζονται τα συντομότερα μονοπάτια ανάμεσα στα ζευγάρια των κόμβων [b] και [a_{bc}, a_{de} ,d ,e] αυτή την φορά. Η τιμή του edge betweenness θα είναι 8. Το ίδιο θα ισχύει και για τις ακμές {c, a_{bc}}, {d, a_{de}} και {e, a_{de}}. Για την ακμή {a_{bc}, a_{de}} υπολογίζεται το edge betweenness από τον αριθμό των συντομότερων μονοπατιών ανάμεσα στα ζευγάρια των κόμβων [a, c, a_{bc}] και [a_{de}, d, e] και αυτό θα είναι ίσο με 18. Συνεχίζοντας στο βήμα 2 της 2ης επανάληψης υπολογίζουμε εκ νέου το vertex betweenness των κορυφών. Θα έχουμε:

- $c_B(b) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(2+8) - (6-1) = \frac{1}{2}10 - 5 = 0$
  όπου $\Gamma(v) = \{\{b, a_{bc}\}, \{b, c\}\}$.
- $c_B(c) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(2+8) - (6-1) = \frac{1}{2}10 - 5 = 0$
  όπου $\Gamma(v) = \{\{c, b\}, \{c, a_{bc}\}\}$.
- $c_B(d) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(2+8) - (6-1) = \frac{1}{2}10 - 5 = 0$
  όπου $\Gamma(v) = \{\{d, a_{de}\}, \{d, e\}\}$.
- $c_B(e) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(2+8) - (6-1) = \frac{1}{2}10 - 5 = 0$
  όπου $\Gamma(v) = \{\{e, a_{de}\}, \{e, d\}\}$.
- $c_B(a_{bc}) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(8+8+18) - (6-1) = \frac{1}{2}30 - 5 = 10$
  όπου $\Gamma(v) = \{\{b, a_{bc}\}, \{c, a_{bc}\}, \{a_{bc}, a_{de}\}\}$.
- $c_B(a_{de}) = \frac{1}{2}\sum_{e\in\Gamma(v)} c_B(e) - (n-1) = \frac{1}{2}(8+8+18) - (6-1) = \frac{1}{2}30 - 5 = 10$
  όπου $\Gamma(v) = \{\{d, a_{de}\}, \{e, a_{de}\}, \{a_{bc}, a_{de}\}\}$.

Στο βήμα 3 ψάχνουμε για ένα σύνολο κορυφών των οποίων το vertex betweenness είναι μεγαλύτερο από το μέγιστο edge betweenness. Κάτι τέτοιο δεν ισχύει σε αυτό το σημείο και έτσι δεν χρειάζεται να μεταβούμε στο βήμα 4. Στο βήμα 5 αφαιρούμε την ακμή με το μεγαλύτερο edge betweenness που αυτή είναι η {a_{bc} ,a_{de}}. Στο Σχήμα 2.19 παρουσιάζεται η νέα μορφή που πήρε το δίκτυο.



Σχήμα 2.19: Το δίκτυο μετά το βήμα 6 της 2ης επανάληψης.

Στο βήμα 6, το edge betweenness όλων των ακμών θα είναι τώρα ίσο με 2. Στην τρίτη επανάληψη, εύκολα μπορούμε να υπολογίσουμε το vertex betweenness όλων

των κορυφών το οποίο θα είναι ίσο με μηδέν. Παρομοίως, δεν υπάρχει κορυφή με vertex betweenness μεγαλύτερο από το μέγιστο edge betweenness οπότε μπορούμε να μεταβούμε στο βήμα 5 και να αφαιρέσουμε αυτή την φορά όλες τις ακμές αφού έχουνε την ίδια τιμή για το edge betweenness. έτσι, καταλήξαμε στο βήμα 7 της τρίτης επανάληψης όπου πλέον δεν έχει μείνει καμία ακμή στο δίκτυο μας και ο αλγόριθμος σταματάει.

2.4.6 Πολυπλοκότητα

Ο Girvan- Newman αλγόριθμος είχε χρονική πολυπλοκότητα στην χειρότερη περίπτωση $O(m^2n)$ όπου m είναι ο αριθμός των ακμών και n ο αριθμός των κορυφών. Στον CONGA αλγόριθμο κάθε κορυφή διασπάται κατά μέσο όρο σε 2m/n κορυφές και έτσι ο αριθμός των κορυφών μετά τη διάσπαση είναι O(m). Ο αριθμός των επαναλήψεων συνεχίζει να είναι O(m) και ο αριθμός των ακμών δε έχει αλλάξει. Έτσι καταλήγουμε ότι ο αλγόριθμος στην χειρότερη περίπτωση έχει χρονική πολυπλοκότητα $O(m^3)$. Στο ref [6] παρουσιάζεται μια βελτιωμένη έκδοση του CONGA που ονομάζεται CONGO (CONGA Optimized). Αυτός ο αλγόριθμος μειώνει την χρονική πολυπλοκότητα σε O(nlogn).

2.5 Αλγόριθμος L-Shell

Η βασική ιδέα του L-Shell αλγορίθμου είναι η επέκταση μιας κοινότητας όσο δυνατόν το περισσότερο, σταματώντας την επέκταση όταν η δομή του δικτύου δεν επιτρέπει οποιαδήποτε περαιτέρω αύξηση, που σημαίνει ότι βρέθηκαν οι γέφυρες. Οι αναλυτές του [7] προτείνουν έναν αλγόριθμο που αποτελείται από ένα l–shell (μια ομάδα από l κορυφές που σκοπό έχουνε να μεγαλώσουν και να καταλάβουν μια κοινότητα) που εξαπλώνεται εξωτερικά από μια αρχική κορυφή. Όσο οι κοντινότεροι γείτονες της αρχικής κορυφής και οι επόμενοι κοντινότεροι γείτονες και ούτω καθεξής, επισκέπτονται, μέσα από ένα l-shell, δύο ποσότητες υπολογίζονται: ο προκύπτων βαθμός και ο συνολικός προκύπτων βαθμός. Ο προκύπτων βαθμός μιας κορυφής ορίζεται ως ο αριθμός των ακμών που συνδέουν αυτή την κορυφή με κορυφές που το l-shell δεν έχει ακόμα επισκεφθεί καθώς επεκτάθηκε από τα προηγούμενα(l-1),(l-2),…-shells. Σημειώνεται σε αυτό το σημείο, ότι ακμές μεταξύ κορυφών μέσα στο ίδιο l-shell δεν συνεισφέρουν στον προκύπτων βαθμό.

Οι συγγραφείς του [7] ορίζουν την ακόλουθη σημειογραφία για τον προκύπτων βαθμό και τον συνολικό προκύπτων βαθμό:

- $K_i^e(j)$ = προκύπτων βαθμός της κορυφής i από ένα shell που άρχισε στη κορυφή j .

- K_j^l = συνολικός προκύπτων βαθμός από ένα shell βάθους l που άρχισε από τη κορυφή j. (1)

Ο συνολικός προκύπτων βαθμός ενός l-shell είναι το άθροισμα των προκύπτων βαθμών όλων των κορυφών στην ηγετική ακμή του l-shell. Ο συνολικός προκύπτων βαθμός σε βάθος l δεν είναι απαραίτητα ο αριθμός των κορυφών σε βάθος $l+1$. Σε βάθος 0, ο συνολικός προκύπτων βαθμός είναι ακριβώς ο βαθμός της αρχικής κορυφής. Σε βάθος l, είναι ο συνολικός αριθμός των ακμών από κορυφές σε βάθος l συνδέονται με κορυφές σε βάθος >l.

Από τις ισότητες στην (1) έχουμε:

$$K_j^0 = K_j$$
$$K_j^l = \sum_{i \in S_j^l} k_i^e(j)$$

όπου S_j^l είναι η ηγετική ακμή του l-shell, που είναι το σύνολο των κορυφών l βήματα μακριά από την κορυφή j.

Επιπρόσθετα ορίζεται η αλλαγή στο συνολικό προκύπτων βαθμό:

$$\Delta K_j^l = \frac{K_j^l}{K_j^{l-1}}$$

για ένα shell σε βάθος l αρχίζοντας από την κορυφή j.

Ο αλγόριθμος λειτουργεί επεκτείνοντας ένα l-shell εξωτερικά από μια αρχική κορυφή j και συγκρίνοντας την αλλαγή στο συνολικό προκύπτων βαθμό με ένα κατώφλι a. Όταν:

$$\Delta K_j^l < \alpha$$

το l-shell σταματάει να μεγαλώνει και όλες οι κορυφές που καλύπτονται από shells βάθους $\leq l$ απαριθμούνται σαν μέλη της κοινότητας της κορυφής j.

Οι αναλυτές περιγράφουν τον αλγόριθμο με τα παρακάτω βήματα. Για μια αρχική κορυφή j:

Αλγόριθμος 2.4 L-Shell algorithm

---

1. Ξεκίνησε ένα l-shell , $l=0$, για την κορυφή j (πρόσθεσε την j στη λίστα των μελών της κοινότητας) και υπολόγισε K_j^0.
2. Εξάπλωσε το l-shell ,$l=1$, πρόσθεσε τους γείτονες της j στη λίστα και υπολόγισε K_j^1.
3. Υπολόγισε ΔK_j^1. Εάν $\Delta K_j^1 < a$,τότε μια κοινότητα έχει βρεθεί. Σταμάτα τον αλγόριθμο.

4. Διαφορετικά επανέλαβε από το βήμα 2 για το επόμενο l-shell, μέχρι το a να ξεπεραστεί ή ολόκληρο συνδεδεμένο μέρος έχει προστεθεί στη λίστα της κοινότητας.

Παρακάτω παρουσιάζεται ο αλγόριθμός σε ψευτοκώδικα [7]:

L-Shell algorithm- pseudo code

$s \in V$; // *s is the starting vertex*
$K_s^{d-1} \longleftarrow 1$;
$Q \longleftarrow$ empty queue; // *search queue*
$C \longleftarrow$ empty queue; // *community queue*
enqueue $s \longrightarrow Q$;
$K_s^d \longleftarrow emerging\,(Q, C, G(V, E))$;
$\Delta K_s^d \longleftarrow \frac{K_s^d}{K_s^{d-1}}$;

while $\Delta K_s^d > \alpha$ **do**
  $K_s^{d-1} \longleftarrow K_s^d$;
  foreach $q \in Q$ **do**
    dequeue $q \longleftarrow Q$;
    enqueue $q \longrightarrow C$;
    enqueue $neighbors(q) \longrightarrow Q$;
  end
  $K_s^d \longleftarrow emerging\,(Q, C, G(V, E))$;
  $\Delta K_s^d \longleftarrow \frac{K_s^d}{K_s^{d-1}}$;
end

Εφόσον τείνουν να υπάρχουν πολλές διασυνδέσεις μέσα σε μια κοινότητα, όταν ένα l-shell μεγαλώνει εξωτερικά από μια αρχική κορυφή μέσα στην κοινότητα, ο συνολικός προκύπτων βαθμός τείνει να αυξάνει. Όταν το l-shell φτάσει στα «σύνορα» της κοινότητας, ο αριθμός των προκύπτων ακμών θα μειωθεί απότομα. Αυτό συμβαίνει γιατί σε αυτό το σημείο, οι μόνες προκύπτουσες ακμές είναι αυτές που συνδέουν την κοινότητα με το υπόλοιπο γράφημα ,που είναι σε αριθμό λιγότερες από αυτές μέσα στην κοινότητα. Με άλλα λόγια, η υπόθεση είναι ότι μια κοινότητα είναι μια δομή στην οποία ο συνολικός προκύπτων βαθμός δεν μπορεί να αυξηθεί σημαντικά, που σημαίνει ότι οι κορυφές που βρίσκονται στα «σύνορα» της κοινότητας έχουνε λίγες ακμές έξω από αυτή και αυτές οι ακμές είναι οι γέφυρες μεταξύ διαφορετικών κοινοτήτων.

Ωστόσο όπως αναφέρουν οι αναλυτές στο [7] η μέθοδος αυτή δεν είναι τέλεια. Το l-shell είναι πιθανό να «ξεχειλίσει» στην κοινότητα που ανιχνεύει. Αυτό εξαρτάται από το πως η αρχική κορυφή είναι τοποθετημένη στο γράφημα. Άμα είναι πιο κοντά (ή εξίσου κοντά) σε κάποια μη κοινοτική κορυφή ή κορυφές, από ότι σε κοινοτικές κορυφές, το l-shell μπορεί να εξαπλωθεί ανάμεσα σε δύο ή περισσότερες κοινότητες την ίδια στιγμή (membership overlap). Για να μετριάσουνε αυτή την συνέπεια, προτείνουν το τρέξιμο του αλγορίθμου N φορές, χρησιμοποιώντας κάθε κορυφή σαν αρχική κορυφή και έπειτα κάποιος μπορεί να αποφασίσει ποιες κορυφές ανήκουν σε ποιες κοινότητες.

2.5.1 Απόκτηση γενικής πληροφορίας

Ο αλγόριθμος 2.4 είναι μια μέθοδος ώστε μια κορυφή να προσδιορίσει σε ποια κοινότητα είναι μέλος. Εξετάζοντας αυτές τις τοπικά ορισμένες λίστες μέλους, κάποιος μπορεί να αποκτήσει μια ιδέα της γενικότερης δομής του δικτύου. Οι συγγραφείς του [7] προτείνουν μια μέθοδο, χρησιμοποιώντας ένα πίνακα μέλους (membership matrix), που σκοπό έχει την απόκτηση μιας γενικότερης εικόνας του δικτύου αλλά και της αποφυγής της επικάλυψης (membership overlap) κατά την διάρκεια του διαχωρισμού.

Για μια αρχική κορυφή j ο αλγόριθμος 2.4 μπορεί να δώσει ένα διάνυσμα v_j μεγέθους N, όπου το $i^{στό}$ στοιχείο είναι 1 εάν η κορυφή i είναι μέλος της κοινότητας της αρχικής κορυφής και 0 εάν δεν είναι. Αυτά τα διανύσματα συγκεντρώνονται σε ένα N×N πίνακα μέλους (membership matrix):

$$\mathrm{M} = (v_1 \mid v_2 \mid ... \mid v_N)^{\mathrm{T}}$$

όπου η $j^{οστή}$ γραμμή περιέχει τα αποτελέσματα χρησιμοποιώντας την κορυφή j σαν την αρχική κορυφή του αλγορίθμου. Αυτό αποτελεί ένα καλό τρόπο ώστε κάποιος να μπορεί να δει τα αποτελέσματα ξεκινώντας τον αλγόριθμο από πολλαπλές κορυφές.

Επιπλέον οι συγγραφείς ορίζουν την απόσταση (Distance) μεταξύ των γραμμών i και j του πίνακα μέλους σαν τον συνολικό αριθμό των διαφορών μεταξύ των στοιχείων τους:

$$Dis\tan ce(i,j) = n - \sum_{k=1}^{n} \delta(M_{ik}, M_{jk})$$

όπου $\delta(M_{ik}, M_{jk}) = 1$ εάν $M_{ik} = M_{jk}$ και 0 διαφορετικά.

Στη συνέχεια οι αναλυτές εφαρμόζουν έναν απλό αλγόριθμο ταξινόμησης. Για την $i^{οστή}$ γραμμή:

1. Βρες την απόσταση (Distance(i,j)) για όλες τις γραμμές j>i.
2. Διάλεξε την γραμμή που έχει την μικρότερη απόσταση με την γραμμή i (ονόμασέ την γραμμή k) και ενάλλαξέ την με την γραμμή i+1. Αυτό απαιτεί την ανταλλαγή των γραμμών i+1 και k, και την ανταλλαγή των στηλών i+1 και k. Οι στήλες ανταλλάσσονται γιατί η εναλλαγή των γραμμών είναι ισοδύναμη με την επαναρίθμηση των εμπλεκόμενων κορυφών και έτσι η αρίθμηση θα πρέπει να είναι σύμφωνη σε όλο τον πίνακα M.
3. Επανέλαβε από την γραμμή i+1.

Όπως αναφέρεται στο [7] το αποτέλεσμα της ταξινόμησης είναι ένας πίνακας μέλους ο οποίος είναι πιο ενδεικτικός όσο αναφορά την δομή. Οι αναλυτές παραθέτουν ένα παράδειγμα ενός πολύ γνωστού δικτύου (Zachary Karate Club) και

παρουσιάζουν τον πίνακα μέλους M πριν και μετά την ταξινόμηση για μια δοσμένη τιμή του a.

2.5.2 Βρίσκοντας την ιεραρχία των υπο-κοινοτήτων

Η ταξινόμηση του πίνακα μέλους (membership matrix) μας βοηθάει να δούμε τα αποτελέσματα από όλα τα διαφορετικά τρεξίματα του τοπικού μας αλγορίθμου αλλά όπως αναφέρουν οι συγγραφείς δεν είναι αρκετός για να προσδιορίσουμε πως κάθε υποκοινότητα σχετίζεται με τις μεγαλύτερες κοινότητες. Για αυτό τον λόγο εισάγουν μια επιπλέον διαδικασία την οποία εφαρμόζουν στον M με σκοπό να δημιουργήσουν ένα δενδρόγραμμα που αναπαριστά την δομή των κοινοτήτων. Η διαδικασία έχει ως εξής:

Για κάθε γραμμή i, υπολογίζεται μια αθροιστική απόσταση της γραμμής, CD_i και ισχύουν:

$$CD_1 = 0 \text{ και } CD_i = Distance(i, i-1) + CD_{i-1} = \sum_{j=2}^{i} Distance(j-1, j) \ .$$

Απεικονίζοντας σε ένα γράφημα τον αριθμό της γραμμής σε σχέση με την αθροιστική απόσταση CD_i μπορούμε να έχουμε μια συλλογή από σημεία που η τιμή τους αυξάνεται και οδηγούνε σε διακριτές ομάδες που υποδηλώνουν τα μέλη κάθε κοινότητας. Οι συγγραφείς του [7] παρουσιάζουν τέτοιες γραφικές αναπαραστάσεις μελετώντας διάφορα δίκτυα. Σημειώνεται ότι ο αριθμός της γραμμής i είναι ο καινούργιος ταξινομημένος αριθμός i για αυτή την κορυφή.

Με σκοπό την παρουσίαση ενός δενδρογράμματος της κοινοτικής δομής οι αναλυτές προτείνουν την παρακάτω διαδικασία:

1. d ⟵ 1.
2. Υπολόγισε την απόσταση Distance(i-1,i) για όλα τα i=2…n.
3. Διάλεξε την μικρότερη απόσταση (Distance) και ονόμασέ την D_{min}.
4. C_d ⟵ άδεια ουρά (empty queue).
5. Βάλε στην ουρά την 1η κορυφή ⟶ C_d.
6. Για i=2…n:

    6.1 Εάν Distance(i-1,i) > D_{min}:
        6.1.1 d ⟵ d+1.
        6.1.2 C_d ⟵ άδεια ουρά (empty queue).
    6.2 Βάλε στη ουρά την iοστή κορυφή ⟶ C_d.

7. Επανέλαβε από το 3 για την επόμενη μικρότερη απόσταση (Distance) μέχρις ότου όλες οι αποστάσεις να έχουνε χρησιμοποιηθεί.

Διαισθητικά, μπορούμε να πούμε ότι κινούμαστε κατά μήκος των γραμμών του πίνακα M και ομαδοποιούμε όλες εκείνες τις κορυφές των οποίων οι γραμμές είναι πιο κοντά η μία στην άλλη σε σχέση με την D_{min} έως ότου φτάσουμε σε μια γραμμή η οποία είναι πιο μακριά από την D_{min}. Έπειτα, ξεκινάει ένα καινούργιο σύνολο, ομαδοποιώντας τις ακόλουθες κορυφές μέχρι πάλι να βρεθεί μια γραμμή πιο μακριά από την D_{min} και ούτω καθεξής. Αυτή η διαδικασία επαναλαμβάνεται χρησιμοποιώντας την επόμενη μικρότερη απόσταση (Distance) σαν D_{min}. Ομαδοποιώντας τις γραμμές του M με αυτόν τον τρόπο είναι σαν να ομαδοποιούμε τις κορυφές ενός γραφήματος σε μια ιεραρχία υπο-κοινοτήτων. Αυτές οι ομαδοποιήσεις χρησιμοποιούνται για την παραγωγή δενδρογραμμάτων που παρουσιάζουν την κοινοτική δομή.

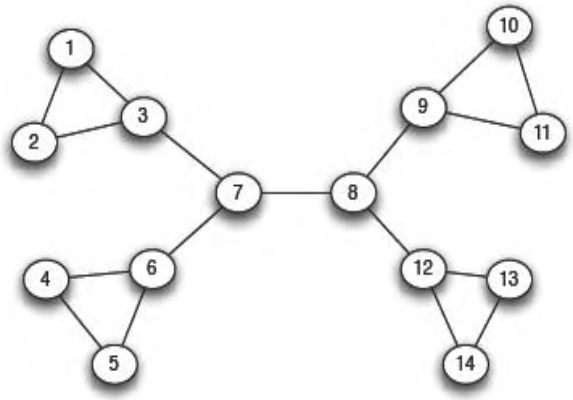
2.5.3 Η παράμετρος a

Ο αλγόριθμος βασίζεται σε μια απλή παράμετρο a που ελέγχει πότε θα σταματήσει η εξάπλωση του l-shell. Όταν το a=0, το l-shell δεν θα σταματήσει ως ότου όλο το συνδεόμενο κομμάτι έχει επισκεφθεί. Όσο το a αυξάνεται σε μέγεθος, τα l-shell θα τείνουν να σταματάνε να αυξάνονται νωρίτερα, ως ότου τελικά να μην εξαπλώνονται πέρα από την αρχική κορυφή και το τελικό αποτέλεσμα να είναι N μοναδικές κορυφές. Αυτό εγγυημένα θα συμβεί όταν $a>k_{max}$ όπου k_{max} είναι ο μεγαλύτερος βαθμός στο δίκτυο.

Κάποιοι μπορούνε να σκεφτούνε το a σαν ένα μέτρο της «φιλικής προδιάθεσης» της αρχικής κορυφής εάν θέλουμε να χρησιμοποιήσουμε μια αναλογία από τα κοινωνικά δίκτυα. Όταν το a είναι μικρό ($a<<1$), τα l-shell θα εξαπλωθούνε πολύ στο δίκτυο. Αυτό μπορεί να μας υποδείξει κορυφές που είναι πιθανό να συμπεριλάβουν άλλες κορυφές στις αντίστοιχες κοινότητες ή όπως σε ένα κοινωνικό δίκτυο, άνθρωποι που καλοδέχονται τους γείτονες τους. Παρόμοια, όταν το a είναι μεγάλο ($a>>1$), τα l-shell θα σταματήσουν να αυξάνονται αμέσως. Αυτό μπορεί να είναι ενδεικτικό κορυφών που δεν θέλουν να δεχτούν άλλες κορυφές στην κοινότητα τους ή αντίστοιχα ανθρώπων που είναι απρόθυμοι να δεχτούν άμεσους γείτονες στις κοινότητες τους και προτιμούν να είναι μόνοι. Με αυτή την λογική η παράμετρος a μπορεί να θεωρηθεί σαν ένα μέτρο «φιλικής προδιάθεσης» ή κοινωνικές αποδοχής.

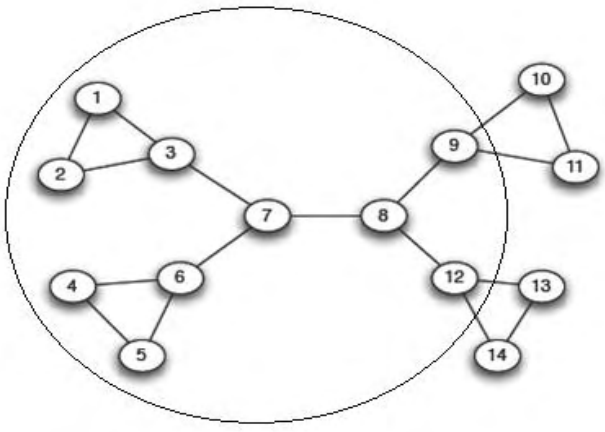
2.5.4 Παράδειγμα

Μπορούμε να δοκιμάσουμε να εφαρμόσουμε τον αλγόριθμό σε ένα ήδη γνωστό μας δίκτυο όπως φαίνεται στο Σχήμα 2.20.



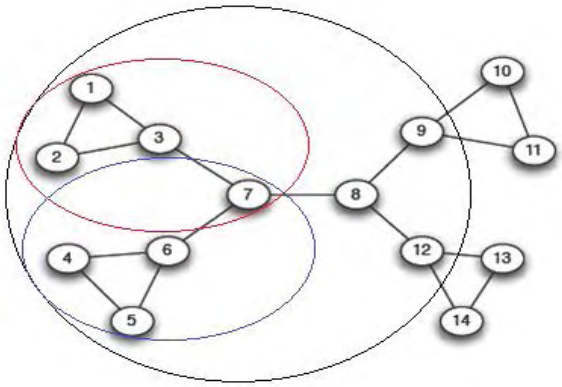
Σχήμα 2.20: Το αρχικό μας δίκτυο.

Ας υποθέσουμε a=1,2 και ότι έχουμε ως αρχική κορυφή την 7. Στο βήμα 1 του αλγορίθμου δημιουργούμε ένα l-shell, όπου l=0. Τοποθετούμε την κορυφή 7 στη λίστα με τα μέλη της κοινότητας και υπολογίζουμε το K_7^0 που θα είναι ίσο με 3. Εξαπλώνουμε το l-shell για l=1, προσθέτουμε τους γείτονες της κορυφής 7 στη λίστα της κοινότητας (τώρα η κοινότητας μας περιέχει τις κορυφές [7,3,6,8]) και υπολογίζουμε το K_7^1, το οποίο θα είναι ίσο με το άθροισμα των προκύπτων βαθμών των κορυφών l βήματα από την αρχική κορυφή 7 (στην περίπτωσή μας l=1). Οι κορυφές αυτές είναι οι 3,6,8 δηλαδή οι γείτονες της κορυφής 7. Έτσι $K_7^1 = 6$. Εν συνεχεία βρίσκουμε την αλλαγή στο προκύπτων βαθμό και θα έχουμε ότι $\Delta K_7^1 = \frac{K_7^1}{K_7^0} = \frac{6}{3} = 2$ >a . Εξαπλώνουμε και άλλο το l-shell, l=2, τοποθετούμε τους γείτονες των κορυφών στις οποίες εξαπλωθήκαμε πριν ,(3,6,8), στη λίστα της κοινότητας (η κοινότητα μας περιέχει τις κορυφές [7,3,6,8,1,2,4,5,9,12] όπως φαίνεται και στο Σχήμα 2.21) και στη συνέχεια θα έχουμε ότι $K_7^2 = 2$. Τώρα θα ισχύει $\Delta K_7^2 = \frac{K_7^2}{K_7^1} = \frac{2}{4} = 0.5$ <a και ο αλγόριθμος σταματάει.



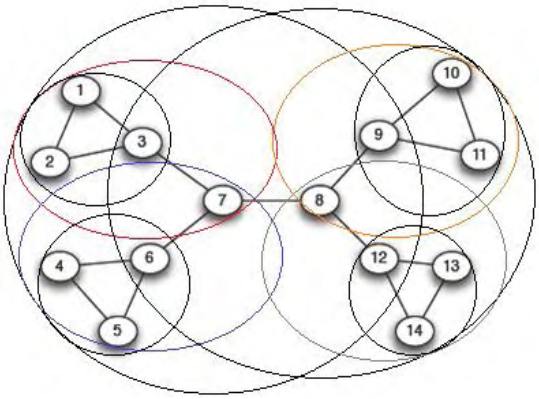
Σχήμα 2.21: Η κοινότητα (μέσα σε κύκλο) που εντοπίσαμε με αρχική κορυφή την 7.

Με αρχική κορυφή την 3 και με παρόμοιο συλλογισμό θα είχαμε ότι $K_3^0 = 3$ με την λίστα της κοινότητας μας να αποτελείται από την κορυφή 3. Για l=1, η λίστα της κοινότητας μας θα αποτελούνταν από τις κορυφές [1,2,3,7] και $K_3^1 = 2$. Έτσι θα ισχύει ότι $\Delta K_3^1 = \frac{K_3^1}{K_3^0} = \frac{2}{3} = 0.6$ <a και ο αλγόριθμος σταματάει εδώ. Κάτι αντίστοιχο θα συνέβαινε αν είχαμε ως αρχική κορυφή την 6 με την λίστα της κοινότητας μας εδώ πέρα να αποτελείται από τις κορυφές [4,5,6,7]. Στο Σχήμα 2.22 μπορούμε να το δούμε σχηματικά.



Σχήμα 2.22: Οι κοινότητες που εντοπίστηκαν με αρχικές κορυφές την 3 και 6.

Με αρχική κορυφή την 1 θα είχαμε ότι $K_1^0 = 2$ και η λίστα θα αποτελούνταν από την κορυφή 1 αρχικά. Για l=1, η λίστα θα ήταν η [1,2,3] και $K_1^1 = 1$. Έτσι θα ισχύει $\Delta K_1^1 = \frac{K_1^1}{K_1^0} = \frac{1}{2} = 0.5$ <a και ο αλγόριθμος σταματάει. Το ίδιο αποτέλεσμα θα είχαμε ένα διαλέξουμε την κορυφή 2. Για την επιλογή των κορυφών 4 και 5 θα είχαμε την λίστα [4,5,6]. Λόγω της συμμετρικότητας του δικτύου μας για τις κορυφές 8 έως 14 θα είχαμε τις λίστες [3,6,7,8,9,10,11,12,13,14], [8,9,10,11], [8,12,13,14], [9,10,11] και [12,13,14]. Στο Σχήμα 2.23 παρουσιάζονται οι κοινότητες που εντοπίστηκαν τρέχοντας τον αλγόριθμο για κάθε κορυφή ξεχωριστά.



Σχήμα 2.23: Οι κοινότητες που εντοπίστηκαν τρέχοντας τον αλγόριθμο N φορές για κάθε κορυφή ξεχωριστά.

Όπως μπορούμε εύκολα να παρατηρήσουμε τα l-shell έχουν εξαπλωθεί ανάμεσα σε δύο ή περισσότερες κοινότητες την ίδια στιγμή (membership overlap). Σύμφωνα και με όσα συζητήθηκαν πιο πάνω μπορούμε να παρουσιάσουμε ένα πίνακα μέλους (membership matrix):

$$M = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \end{array} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Τρέχοντας τον αλγόριθμο ταξινόμησης θα έχουμε:

- Για την γραμμή i=1

$$D(1,2) = 14 - 14 = 0 \quad D(1,9) = 14 - 7 = 7$$
$$D(1,3) = 14 - 13 = 1 \quad D(1,10) = 14 - 8 = 6$$
$$D(1,4) = 14 - 8 = 6 \quad D(1,11) = 14 - 8 = 6$$
$$D(1,5) = 14 - 8 = 6 \quad D(1,12) = 14 - 6 = 8$$
$$D(1,6) = 14 - 7 = 7 \quad D(1,13) = 14 - 8 = 6$$
$$D(1,7) = 14 - 7 = 7 \quad D(1,14) = 14 - 8 = 6$$
$$D(1,8) = 14 - 3 = 11$$

Η γραμμή 2 (k=2) έχει με την μικρότερη απόσταση με την γραμμή i=1, και έτσι θα την εναλλάξουμε με την γραμμή i+1=2, που είναι ουσιαστικά η εναλλαγή της 2 με τον ίδιο της τον εαυτό. Το ίδιο θα συμβεί και με την στήλη 2. Έτσι ο πίνακας θα παραμείνει ο ίδιος.

- Για την γραμμή i=2

$$D(2,3) = 14 - 13 = 1 \quad D(2,10) = 14 - 8 = 6$$
$$D(2,4) = 14 - 8 = 6 \quad D(2,11) = 14 - 8 = 6$$
$$D(2,5) = 14 - 8 = 6 \quad D(2,12) = 14 - 6 = 8$$
$$D(2,6) = 14 - 7 = 7 \quad D(2,13) = 14 - 8 = 6$$
$$D(2,7) = 14 - 4 = 10 \quad D(2,14) = 14 - 8 = 6$$
$$D(2,8) = 14 - 3 = 11$$
$$D(2,9) = 14 - 7 = 7$$

Εδώ θα έχουμε ότι η γραμμή(και στήλη) k=3 θα πρέπει να εναλλαχθεί με την γραμμή (και στήλη) i+1=3. Ο πίνακας και σε αυτή την περίπτωση θα παραμείνει ίδιος.

- Για την γραμμή i=3

$D(3,4) = 14 - 7 = 7$ $D(3,11) = 14 - 7 = 7$
$D(3,5) = 14 - 7 = 7$ $D(3,12) = 14 - 5 = 9$
$D(3,6) = 14 - 8 = 6$ $D(3,13) = 14 - 7 = 7$
$D(3,7) = 14 - 8 = 6$ $D(3,14) = 14 - 7 = 7$
$D(3,8) = 14 - 4 = 10$
$D(3,9) = 14 - 7 = 7$
$D(3,10) = 14 - 7 = 7$

Η μικρότερη τιμή είναι η 6 και διαλέγουμε την μία από τις δύο. Έστω ότι διαλέγουμε την γραμμή k=7 και την εναλλάσσουμε με την i+1=4. Το ίδιο θα συμβεί και με τις στήλες 7 και 4. Το αποτέλεσμα φαίνεται στον νέο πίνακα M:

$$M = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \end{array} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

- Για την γραμμή i=4

$D(4,5) = 14 - 7 = 7$ $D(4,12) = 14 - 3 = 11$
$D(4,6) = 14 - 8 = 6$ $D(4,13) = 14 - 1 = 13$
$D(4,7) = 14 - 7 = 7$ $D(4,14) = 14 - 1 = 13$
$D(4,8) = 14 - 4 = 10$
$D(4,9) = 14 - 6 = 8$
$D(4,10) = 14 - 5 = 9$
$D(4,11) = 14 - 5 = 9$

Η μικρότερη τιμή είναι πάλι η 6 με εναλλαγή των γραμμών(στηλών) k=6 και i+1=5 καταλήγοντας στον πίνακα M:

$$M = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \end{array} \begin{array}{c} \begin{array}{cccccccccccccc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccccccccccccc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccccccccccccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccccccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccccccccccccc} 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccccccccccccc} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccccccccccccc} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccccccccccccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \\ \begin{array}{cccccccccccccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \end{array} \\ \begin{array}{cccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array} \\ \begin{array}{cccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array} \end{array}$$

- Για την γραμμή i=5

$$D(5,6) = 14 - 13 = 1 \quad D(5,13) = 14 - 7 = 7$$
$$D(5,7) = 14 - 13 = 1 \quad D(5,14) = 14 - 7 = 7$$
$$D(5,8) = 14 - 4 = 10$$
$$D(5,9) = 14 - 8 = 6$$
$$D(5,10) = 14 - 7 = 7$$
$$D(5,11) = 14 - 7 = 7$$
$$D(5,12) = 14 - 5 = 9$$

Η μικρότερη τιμή είναι η 1 και έστω ότι γίνεται η εναλλαγή των γραμμών(στηλών) k=6 και i+1=6 όπου ο πίνακας παραμένει ο ίδιος.

- Για την γραμμή i=6

$$D(6,7) = 14 - 14 = 0 \quad D(6,14) = 14 - 8 = 6$$
$$D(6,8) = 14 - 3 = 11$$
$$D(6,9) = 14 - 7 = 7$$
$$D(6,10) = 14 - 8 = 6$$
$$D(6,11) = 14 - 8 = 6$$
$$D(6,12) = 14 - 6 = 8$$
$$D(6,13) = 14 - 8 = 6$$

Και σε αυτή την περίπτωση ο πίνακας θα παραμείνει ίδιος λόγω της εναλλαγής των γραμμών(στηλών) k=7 και i+1=7.

- Για την γραμμή i=7

$$D(7,8) = 14 - 3 = 11$$
$$D(7,9) = 14 - 7 = 7$$
$$D(7,10) = 14 - 8 = 6$$
$$D(7,11) = 14 - 8 = 6$$
$$D(7,12) = 14 - 6 = 8$$
$$D(7,13) = 14 - 8 = 6$$
$$D(7,14) = 14 - 8 = 6$$

Η μικρότερη τιμή είναι η 6 και έστω ότι πραγματοποιούμε την εναλλαγή των γραμμών(στηλών) k=11 και i+1=8 δίνοντας τον πίνακα M:

$$M = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \end{array} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

- Για την γραμμή i=8

$$D(8,9) = 14 - 11 = 3$$
$$D(8,10) = 14 - 14 = 0$$
$$D(8,11) = 14 - 8 = 6$$
$$D(8,12) = 14 - 8 = 6$$
$$D(8,13) = 14 - 8 = 6$$
$$D(8,14) = 14 - 8 = 6$$

Εδώ εναλλάσσουμε την γραμμή(στήλη) k=10 με την i+1=9 και έχουμε τον πίνακα M:

$$M = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \end{array} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Για τις υπόλοιπες τιμές του i θα έχουμε:

- Για i=9:

$D(9,10) = 14 - 11 = 3$
$D(9,11) = 14 - 8 = 6$
$D(9,12) = 14 - 8 = 6$
$D(9,13) = 14 - 8 = 6$
$D(9,14) = 14 - 8 = 6$

- Για i=10:

$D(10,11) = 14 - 9 = 5$
$D(10,12) = 14 - 7 = 7$
$D(10,13) = 14 - 7 = 7$
$D(10,14) = 14 - 7 = 7$

- Για i=11:

$D(11,12) = 14 - 10 = 4$
$D(11,13) = 14 - 8 = 6$
$D(11,14) = 14 - 8 = 6$

- Για i=12:

$D(12,13) = 14 - 12 = 2$
$D(12,14) = 14 - 12 = 2$

- Για i=13:

$D(13,14) = 14 - 14 = 0$

Για όλες τις παραπάνω περιπτώσεις ο πίνακας δεν θα αλλάξει αφού προκύπτει εναλλαγή μιας γραμμής με τον εαυτό της. Ο πίνακάς μας τελικά θα έχει την μορφή:

$$M = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \end{array} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Πρέπει να προσέξουμε τώρα ότι οι τιμές αριστερά του πίνακα χρησιμοποιούνται μόνο για την αρίθμηση. Λόγω των εναλλαγών η γραμμή 4 αναπαριστά την κορυφή 7 και η γραμμή 7 την κορυφή 4 και ούτω καθεξής. Για τις γραμμές που δεν υπήρξε εναλλαγή ,για παράδειγμα την 1, η γραμμή αναπαριστά και την κορυφή. Τώρα σύμφωνα με όσα συζητήθηκαν στο 2.5.2 μπορούμε να υπολογίσουμε την αθροιστική απόσταση CD_i.

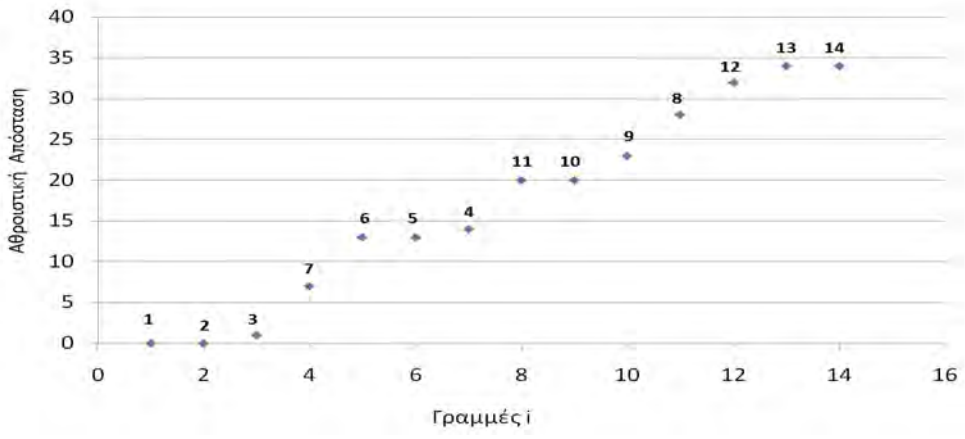
Αρχικά υπολογίζουμε τις αποστάσεις:

$$D(1,2) = 14 - 14 = 0 \qquad D(8,9) = 14 - 14 = 0$$
$$D(2,3) = 14 - 13 = 1 \qquad D(9,10) = 14 - 11 = 3$$
$$D(3,4) = 14 - 8 = 6 \qquad D(10,11) = 14 - 9 = 5$$
$$D(4,5) = 14 - 8 = 6 \qquad D(11,12) = 14 - 10 = 4$$
$$D(5,6) = 14 - 13 = 1 \qquad D(12,13) = 14 - 12 = 2$$
$$D(6,7) = 14 - 14 = 0 \qquad D(13,14) = 14 - 14 = 0$$
$$D(7,8) = 14 - 8 = 6$$

και έπειτα τις αθροιστικές αποστάσεις βάσει των παραπάνω:

$$CD_1 = 0 \qquad CD_8 = 14 + 6 = 20$$
$$CD_2 = 0 \qquad CD_9 = 20 + 0 = 20$$
$$CD_3 = 0 + 1 = 1 \qquad CD_{10} = 20 + 3 = 23$$
$$CD_4 = 1 + 6 = 7 \qquad CD_{11} = 23 + 5 = 28$$
$$CD_5 = 7 + 6 = 13 \qquad CD_{12} = 28 + 4 = 32$$
$$CD_6 = 13 + 1 = 14 \qquad CD_{13} = 32 + 2 = 34$$
$$CD_7 = 14 + 0 = 14 \qquad CD_{14} = 32 + 0 = 34$$

Μπορούμε να αναπαραστήσουμε σε ένα γράφημα τον αριθμό της γραμμής σε σχέση με την αθροιστική απόσταση CD_i. Κατά αυτό τον τρόπο έχουμε μια συλλογή από σημεία(Σχήμα 2.24) που η τιμή τους αυξάνεται και οδηγούνε σε διακριτές ομάδες που υποδηλώνουν τα μέλη κάθε κοινότητας. Πιο συγκεκριμένα για το παράδειγμά μας:



Σχήμα 2.24: Συλλογή από σημεία που η τιμή τους αυξάνεται και οδηγούνε σε διακριτές ομάδες που υποδηλώνουν τα μέλη κάθε κοινότητας.

Για την κατασκευή του δενδρογράμματος (Σχήμα 2.25) που θα μας δώσει την ιεραρχική δομή των κοινοτήτων μας ακολουθούμε τα βήματα που περιγράφτηκαν στο 2.5.2.

Αρχικά στο πρώτο βήμα θέτουμε d=1. Οι αποστάσεις του βήματος 2 έχουνε βρεθεί πιο πάνω.

- $1^{η}$ επανάληψη ,D_{min}=0 και C_1=[1]. Για το βήμα 6 θα έχουμε:

C_1= [1, 2], C_2= [3], C_3= [4], C_4= [5], C_5= [6, 7], C_6= [8, 9], C_7= [10], C_8= [11], C_9= [12], C_{10}= [13, 14].

- $2^{η}$ επανάληψη ,D_{min}=1 και C_1=[1]:

C_{11}= [1, 2, 3], C_{12}= [4], C_{13}= [5, 6, 7], C_{14}= [8, 9], C_{15}= [10], C_{16}= [11], C_{17}= [12], C_{18}= [13, 14].

- $3^{η}$ επανάληψη ,D_{min}=2 και C_1=[1]:

C_{19}= [1, 2, 3], C_{20}= [4], C_{21}= [5, 6, 7], C_{22}= [8, 9], C_{23}= [10], C_{24}= [11], C_{25}= [12, 13, 14].

- $4^{η}$ επανάληψη ,D_{min}=3 και C_1=[1]:

C_{26}= [1, 2, 3], C_{27}= [4], C_{28}= [5, 6, 7], C_{29}= [8, 9, 10], C_{30}= [11], C_{31}= [12, 13, 14].

- $5^{η}$ επανάληψη ,D_{min}=4 και C_1=[1]:

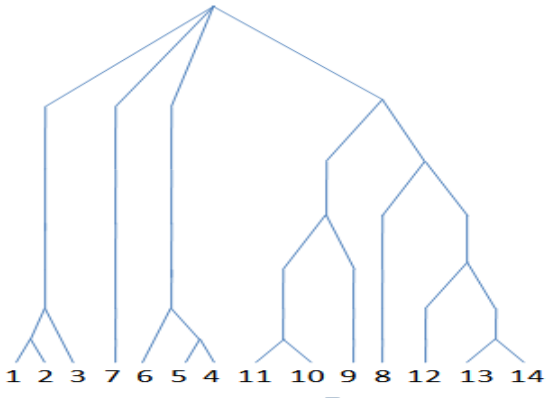
C_{32}= [1, 2, 3], C_{33}= [4], C_{34}= [5, 6, 7], C_{35}= [8, 9, 10], C_{36}= [11, 12, 13, 14].

- $6^{η}$ επανάληψη ,D_{min}=5 και C_1=[1]:

C_{37}= [1, 2, 3], C_{38}= [4], C_{39}= [5, 6, 7], C_{40}= [8, 9, 10, 11, 12, 13, 14].

- $7^{η}$ επανάληψη ,D_{min}=6 και C_1=[1]:

C_{41}= [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14].



Σχήμα 2.25: Δενδρόγραμμα στο οποίο απεικονίζεται η ιεραρχική ομαδοποίηση του δικτύου μας.

2.5.5 Πολυπλοκότητα

Όπως αναφέρουν οι αναλυτές στο [7] η διαδικασία της ταξινόμησης μπορεί να είναι υπολογιστικά ακριβή. Για την εύρεση της απόστασης Distance(i,j) απαιτείται κόστος O(N). Όταν ξεκινάει ο αλγόριθμος ταξινόμησης υπάρχουν N-1 αποστάσεις (Distances) να βρεθούνε και έτσι η πρώτη ταξινόμηση κοστίζει O(N(N-1)) ~ $O(N^2)$.

Αυτό επαναλαμβάνεται για την δεύτερη γραμμή και έχει κόστος O(N(N-2)) ~ O(N^2) και συνεχίζεται για κάθε επιπλέον γραμμή. Εφόσον υπάρχουν N γραμμές, το συνολικό κόστος θα είναι:

$$\sum_{i=1}^{N} N(N-i) = N(N^2 - \frac{1}{2}N(N+1)) = O(N^3).$$

2.6 Αλγόριθμος Internal–External Degree (Εσωτερικού-Εξωτερικού Βαθμού)

Η υπόθεση που κάνουν οι αναλυτές του [8] πίσω από αυτόν τον αλγόριθμο είναι ότι οι κοινότητες είναι ουσιαστικά τοπικές δομές και περιλαμβάνουν τους κόμβους που ανήκουν στις ομάδες συν, το πολύ, μια εκτεταμένη γειτονία από αυτούς.

Εδώ πέρα μια κοινότητα είναι ένα υπογράφημα που προσδιορίζεται από την μεγιστοποίηση της καταλληλότητας των κόμβων του. Η μορφή της καταλληλότητας δίνεται από την ακόλουθη έκφραση:

$$f_G = \frac{k_{in}^G}{\left(k_{in}^G + k_{out}^G\right)^a}, \quad (1)$$

όπου k_{in}^G και k_{out}^G είναι οι συνολικοί εσωτερικοί και εξωτερικοί βαθμοί των κόμβων ενός υπογραφήματος G και a είναι μια θετική παράμετρος με πραγματική τιμή που ελέγχει το μέγεθος των κοινοτήτων. Ο εσωτερικός βαθμός ενός υπογραφήματος είναι ίσος με το διπλάσιο αριθμό των εσωτερικών συνδέσεων του υπογραφήματος. Ο εξωτερικός βαθμός είναι ο αριθμός των συνδέσεων που ενώνουν κάθε μέλος του υπογραφήματος με το υπόλοιπο γράφημα. Ο στόχος είναι να προσδιοριστεί ένα υπογράφημα, ξεκινώντας από ένα κόμβο Α, έτσι ώστε η ένταξη ενός καινούργιου κόμβου ή η αφαίρεση ενός κόμβου από το υπογράφημα να ελαχιστοποιούσε το f_G. Αποκαλείται αυτό το υπογράφημα, **φυσική κοινότητα** του κόμβου Α.

Επίσης εισάγεται από τους συγγραφείς του [8] η ιδέα της καταλληλότητας ενός κόμβου. Έχοντας μια συνάρτηση καταλληλότητας, η καταλληλότητα ενός κόμβου Α αναφορικά με ένα υπογράφημα G, f_G^{A}, ορίζεται ως η μεταβολή της καταλληλότητας του υπογραφήματος G με και χωρίς τον κόμβο Α. Δηλαδή:

$$f_G^A = f_{G+\{A\}} - f_{G-\{A\}}. \quad (2)$$

Στην ισότητα (2), ο συμβολισμός $G+\{A\}$ ($G-\{A\}$) δηλώνει το υπογράφημα που έχει επιτευχθεί από το υπογράφημα G με τον κόμβο Α μέσα (έξω).

Η φυσική κοινότητα του κόμβου Α προσδιορίζεται από την ακόλουθη διαδικασία. Οι αναλυτές υποθέτουν ότι το υπογράφημα G περιλαμβάνει τον κόμβο Α. Αρχικά, το G προσδιορίζεται από τον κόμβο Α (k_{in}^{G}=0). Κάθε επανάληψη του αλγορίθμου αποτελείται από τα ακόλουθα βήματα:

Αλγόριθμος 2.5 Internal-External Degree Algorithm

---

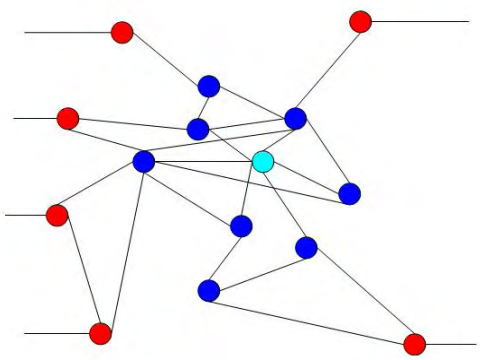
1. Εκτελείται ένας κύκλος σε όλους τους γειτονικούς κόμβους του G, που δεν συμπεριλαμβάνονται στο G.
2. Ο γείτονας με την μεγαλύτερη καταλληλότητα προστίθεται στο G, δίνοντας ένα μεγαλύτερο υπογράφημα G'.
3. Η καταλληλότητα κάθε κόμβου του G' υπολογίζετε ξανά.
4. Εάν ένας κόμβος βρεθεί με αρνητική καταλληλότητα, αφαιρείτε από το G', δίνοντας ένα νέο υπογράφημα G''.
5. Αν προκύψει το 4^{ο} βήμα, επανέλαβε από το 3^{ο}, διαφορετικά επανέλαβε από το 1^{ο} για το υπογράφημα G''.

Η διαδικασία σταματάει όταν οι κόμβοι που εξετάζονται στο βήμα 1 έχουνε όλοι αρνητική καταλληλότητα(Σχήμα 2.26), που σημαίνει ότι οι εξωτερικές τους ακμές είναι όλες γέφυρες.

Κάλυψη ενός γραφήματος ορίζεται ως ένα σύνολο από ομάδες έτσι ώστε κάθε κόμβος ανατίθεται σε τουλάχιστον μια ομάδα. Αυτό αποτελεί μια επέκταση της παραδοσιακής ιδέας του διαμερισμού γράφων (όπου κάθε κόμβος ανήκει σε μια μοναδική κοινότητα) που σκοπό έχει να εξηγήσει την πιθανή επικάλυψη κοινοτήτων. Στην περίπτωσή μας, ο εντοπισμός μιας κάλυψης ισοδυναμεί με την εύρεση της φυσικής κοινότητας κάθε κόμβου στο γράφημα. Ένας ξεκάθαρος τρόπος για να το πετύχουνε οι αναλυτές αυτό είναι να εφαρμόσουνε την παραπάνω διαδικασία για κάθε κόμβο ξεχωριστά. Αυτό ωστόσο, όπως αναφέρουν, είναι υπολογιστικά ακριβό. Οι φυσικές κοινότητες πολλών κόμβων συχνά συμπίπτουν και ως εκ τούτου ο περισσότερος υπολογιστικός χρόνος σπαταλάτε στη ανακάλυψη των ίδιων συστατικών μερών ξανά και ξανά. Ένας οικονομικός τρόπος για να την αποφυγή του παραπάνω προβλήματος έχει ως εξής:

1. Διάλεξε ένα κόμβο Α τυχαία.
2. Εντόπισε την φυσική κοινότητα του κόμβου Α.
3. Διάλεξε τυχαία έναν κόμβο Β που δεν έχει ακόμα ανατεθεί σε κάποια ομάδα.
4. Εντόπισε την φυσική κοινότητα του Β, ερευνώντας όλους τους κόμβους ανεξάρτητα από την πιθανή τους συμμετοχή σε άλλες ομάδες.
5. Επανέλαβε από το βήμα 3.

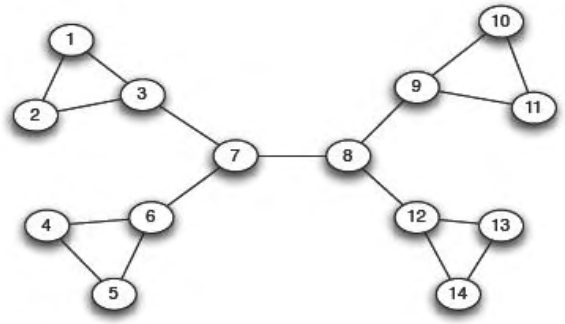
Ο αλγόριθμος σταματάει όταν όλοι οι κόμβοι έχουν ανατεθεί σε τουλάχιστον μια ομάδα. Η παραπάνω πρόταση αιτιολογείται από το εξής επιχείρημα. Οι κόμβοι κάθε κοινότητας είτε επικαλύπτονται από άλλες κοινότητες είτε όχι. Μια κοινότητα εντοπίστηκε σύμφωνα με ένα συγκεκριμένο κόμβο. Εάν κάποιος διαλέξει μεταξύ κάποιου άλλου κόμβου μπορεί να εντοπίσει την ίδια κοινότητα ή μια από τις πιθανές επικαλυπτόμενες κοινότητες. Το ίδιο θα μπορούσε να συμβεί εάν κάποιος ξεκινούσε από κόμβους που είναι έξω από την κοινότητα και κοντά της, χωρίς να επικαλύπτονται από αυτή. Με αυτό τον τρόπο κάποιος θα μπορούσε να ανακτήσει όλα τα συστατικά μέρη του γραφήματος χωρίς να χρειάζεται να ξεκινήσει από κάθε κόμβο. Την ίδια στιγμή, οι επικαλυπτόμενοι κόμβοι θα καλύπτονται κατά την διάρκεια κατασκευής της κάθε κοινότητας στην οποία ανήκουν αφού είναι πιθανό να περιλαμβάνουν κόμβους που ήδη έχουν ανατεθεί σε άλλα μέρη. Όπως αναφέρουν οι συγγραφείς του [8], εκτενή αριθμητικά τεστ έχουνε δείξει ότι η απώλεια στην ακρίβεια είναι ελάχιστη αν κάποιος αποφασίσει να προχωρήσει σύμφωνα με τα παραπάνω αντί να βρει την φυσική κοινότητα όλων των κόμβων.



Σχήμα 2.26: Σχηματικό παράδειγμα μιας κοινότητας για ένα κόμβο (γαλάζιος κόμβος στο σχήμα) σύμφωνα με τον ορισμό. Οι μπλε κόμβοι είναι τα άλλα μέλη της ομάδας και έχουν θετική καταλληλότητα στην ομάδα, ενώ οι κόκκινοι κόμβοι έχουν όλοι αρνητική καταλληλότητα σε σχέση με την ομάδα[8].

Η παράμετρος a συντονίζει την λύση της μεθόδου μας. Καθορίζοντας το a σημαίνει ότι θέτουμε την κλίμακα με την οποία κοιτάμε το δίκτυο. Μεγάλες τιμές του a παράγουν πολύ μικρές κοινότητες, αντίθετα μικρές τιμές μας δίνουν μεγάλα τμήματα. Εάν η τιμή του a είναι αρκετά μικρή, όλοι οι κόμβοι καταλήγουν στην ίδια ομάδα, η οποία είναι το ίδιο το δίκτυο. Οι αναλυτές έχουν βρει ότι στις περιπτώσεις για $a<0.5$, υπάρχει μόνο μια κοινότητα ενώ για $a>2$ κάποιος αποκαλύπτει τις μικρότερες κοινότητες. Μια φυσιολογική είναι η $a=1$, αφού είναι η αναλογία του εξωτερικού βαθμού προς το συνολικό βαθμό της κοινότητας. Αυτό αντιστοιχεί στο επονομαζόμενο αδύναμο (weak) ορισμό της κοινότητας που παρουσιάστηκε από τον Radicchi[[9]. Τις περισσότερες περιπτώσεις η κάλυψη που βρίσκεται για $a=1$ είναι σχετική και έτσι δίνει χρήσιμες πληροφορίες για την πραγματική κοινοτική δομή του επικείμενου γραφήματος.

2.6.1 Παράδειγμα



Σχήμα 2.27: το αρχικό μας δίκτυο.

Ας δούμε τώρα πως λειτουργεί ο αλγόριθμος στο γνωστό μας δίκτυο όπως αυτό φαίνεται στο Σχήμα 2.27.

- Για a=1

Διαλέγουμε τυχαία τον κόμβο 1 και αρχικά το γράφημα μας (G) προσδιορίζεται από τον κόμβο αυτόν($k_{in}=0$) και έτσι $G=\{1\}$. Στο βήμα 1 ανατρέχουμε στους γειτονικούς κόμβους που δεν συμπεριλαμβάνονται στο γράφημα, κόμβοι 2 και 3. Υπολογίζουμε την καταλληλότητα των κόμβων στο βήμα 2:

- $$f_G^2 = f_{G+\{2\}} - f_{G-\{2\}} = \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^a} - \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^a} = \frac{2}{2+2} - 0 = \frac{2}{4} = 0.5$$
- $$f_G^3 = f_{G+\{3\}} - f_{G-\{3\}} = \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^a} - \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^a} = \frac{2}{2+3} - 0 = \frac{2}{5} = 0.4$$

και αυτός με την μεγαλύτερη καταλληλότατα προστίθεται στο G δίνοντας $G' = \{1,2\}$. Στο βήμα 3 υπολογίζουμε την καταλληλότητα των κόμβων του G':

- $$f_{G'}^1 = f_{G'+\{1\}} - f_{G'-\{1\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} = \frac{2}{2+2} - 0 = \frac{2}{4} = 0.5$$
- $$f_{G'}^2 = f_{G'+\{2\}} - f_{G'-\{2\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} = \frac{2}{2+2} - 0 = \frac{2}{4} = 0.5$$

όπου οι καταλληλότητες είναι θετικές και έτσι οδηγούμαστε στο βήμα 5 που με την σειρά του μας οδηγεί στο βήμα 1 και στην δεύτερη επανάληψη του αλγορίθμου. Ο μοναδικός γείτονας του νέου μας γραφήματος είναι ο κόμβος 3:

- $$f_{G'}^{3} = f_{G'+\{3\}} - f_{G'-\{3\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^{a}} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^{a}} = \frac{6}{6+1} - \frac{2}{2+2} = \frac{6}{7} - \frac{1}{2} = \frac{5}{14}$$

και αφού είναι ο μοναδικός γείτονας προστίθεται το νέο γράφημα $G^{"} = \{1,2,3\}$ και έχουμε:

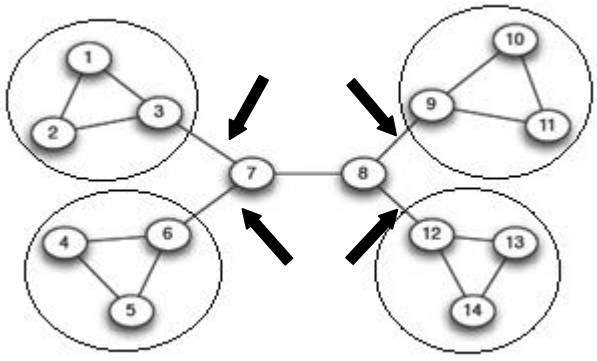
- $$f_{G'}^{1} = f_{G^{"}+\{1\}} - f_{G^{"}-\{1\}} = \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} - \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} = \frac{6}{6+1} - \frac{2}{2+3} = \frac{6}{7} - \frac{2}{5} = \frac{16}{35} = f_{G^{"}}^{2}$$

- $$f_{G'}^{3} = f_{G^{"}+\{3\}} - f_{G^{"}-\{3\}} = \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} - \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} = \frac{6}{6+1} - \frac{2}{2+2} = \frac{6}{7} - \frac{1}{2} = \frac{5}{14}$$

Οδηγούμαστε πάλι στο βήμα 1 και συναντάμε ένα γείτονα, τον κόμβο 7.

- $$f_{G'}^{7} = f_{G^{"}+\{7\}} - f_{G^{"}-\{7\}} = \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} - \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{'}})^{a}} = \frac{8}{8+2} - \frac{6}{6+1} = \frac{8}{10} - \frac{6}{7} = -\frac{4}{70}$$

Παρατηρούμε η καταλληλότητα είναι αρνητική και επειδή είναι ο μοναδικός γείτονας ο αλγόριθμος σταματάει έχοντας εντοπίσει την κοινότητα που μας υποδεικνύει το υπογράφημα $G^{"} = \{1,2,3\}$. Εν συνεχεία διαλέγοντας έναν κόμβο που δεν έχει μπει σε κάποια ομάδα συνεχίζουμε μέχρι όλοι οι κόμβοι να έχουν ανατεθεί σε τουλάχιστον μια ομάδα. Διαλέγοντας με την σειρά τους κόμβους 4,10,13 και ακολουθώντας την ίδια διαδικασία θα έχουμε τις κοινότητες που θα μας υποδεικνύουν τα γραφήματα $G^{"} = \{4,5,6\}$, $G^{"} = \{9,10,11\}$, $G^{"} = \{12,13,14\}$ αντίστοιχα, με τις γέφυρες να αποτελούνται από τις ακμές {3,7}, {6,7},{8,9},{8,12}, όπως φαίνεται και στο Σχήμα 2.28.



Σχήμα 2.28: Με κύκλο απεικονίζονται οι κοινότητες στο δίκτυο με γέφυρες τις ακμές {3,7}, {6,7},{8,9},{8,12}.

Στη συνέχεια διαλέγουμε τυχαία έναν κόμβο ο οποίος δεν έχει ανατεθεί σε κάποια κοινότητα, έστω τον 7 ερευνώντας όλους τους κόμβους ανεξάρτητα από την πιθανή συμμετοχή τους σε άλλες κοινότητες. Αρχικά το γράφημα προσδιορίζεται από τον κόμβο 7 και $G = \{7\}$. Ανατρέχοντας στους γειτονικούς κόμβους βρίσκουμε τις καταλληλότητες:

- $f_G^3 = f_{G+\{3\}} - f_{G-\{3\}} = \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^a} - \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^a} = \frac{2}{2+4} - 0 = \frac{1}{3} = f_G^6 = f_G^8$

και αφού οι καταλληλότητες είναι ίσες προστίθενται στο γράφημα $G' = \{3,6,7,8\}$. Υπολογίζουμε και πάλι τις καταλληλότητες όλων των κόμβων:

- $f_{G'}^3 = f_{G'+\{3\}} - f_{G'-\{3\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} = \frac{6}{6+6} - \frac{4}{4+5} = \frac{1}{2} - \frac{4}{9} = \frac{1}{18} = f_{G'}^6 = f_{G'}^8$

- $f_{G'}^7 = f_{G'+\{7\}} - f_{G'-\{7\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} = \frac{6}{6+6} - 0 = \frac{1}{2}$

Αφού είναι όλες θετικές πηγαίνουμε στο βήμα 1 και εξετάζουμε τις γείτονες του νέου μας γραφήματος, τους κόμβους 1,2,4,5,9 και 12, οι οποίοι θα έχουν τις εξής καταλληλότητες:

- $f_{G'}^1 = f_{G'+\{1\}} - f_{G'-\{1\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} = \frac{8}{8+6} - \frac{6}{6+6} = \frac{1}{14} = f_{G'}^2 = f_{G'}^4 = f_{G'}^5$

- $f_{G'}^9 = f_{G'+\{9\}} - f_{G'-\{9\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} = \frac{8}{8+7} - \frac{6}{6+6} = \frac{1}{30} = f_{G'}^{12}$

Προσθέτουμε στο νέο μας γράφημα G'' τους κόμβους με την μεγαλύτερη καταλληλότητα και θα έχουμε $G'' = \{1,2,3,4,5,6,7,8\}$. Στο βήμα 3 υπολογίζουμε εκ νέου τις καταλληλότητες του G'':

- $f_{G''}^1 = f_{G''+\{1\}} - f_{G''-\{1\}} = \frac{k_{in}^{G''}}{(k_{in}^{G''} + k_{out}^{G''})^a} - \frac{k_{in}^{G''}}{(k_{in}^{G''} + k_{out}^{G''})^a} = \frac{18}{18+2} - \frac{14}{14+4} = \frac{11}{90} = f_{G''}^2 = f_{G''}^4 = f_{G''}^5$
- $f_{G''}^3 = f_{G''+\{3\}} - f_{G''-\{3\}} = \frac{k_{in}^{G''}}{(k_{in}^{G''} + k_{out}^{G''})^a} - \frac{k_{in}^{G''}}{(k_{in}^{G''} + k_{out}^{G''})^a} = \frac{18}{18+2} - \frac{12}{12+5} = \frac{33}{170} = f_{G''}^6 = f_{G''}^7$
- $f_{G''}^8 = f_{G''+\{8\}} - f_{G''-\{8\}} = \frac{k_{in}^{G''}}{(k_{in}^{G''} + k_{out}^{G''})^a} - \frac{k_{in}^{G''}}{(k_{in}^{G''} + k_{out}^{G''})^a} = \frac{18}{18+2} - \frac{16}{16+1} = -\frac{7}{170}$

Η καταλληλότητα του κόμβου 8 είναι αρνητική και αφαιρείται από το G'' δίνοντας $G''' = \{1,2,3,4,5,6,7\}$ και το βήμα 5 αυτή την φορά μας οδηγεί στο βήμα 3 όπου οι καταλληλότητες υπολογίζονται ξανά:

- $f_{G'''}^1 = f_{G'''+\{1\}} - f_{G'''-\{1\}} = \frac{k_{in}^{G'''}}{(k_{in}^{G'''} + k_{out}^{G'''})^a} - \frac{k_{in}^{G'''}}{(k_{in}^{G'''} + k_{out}^{G'''})^a} = \frac{16}{16+1} - \frac{12}{12+3} = \frac{16}{17} - \frac{12}{15} = \frac{36}{255}$
  $= f_{G'''}^2 = f_{G'''}^4 = f_{G'''}^5$

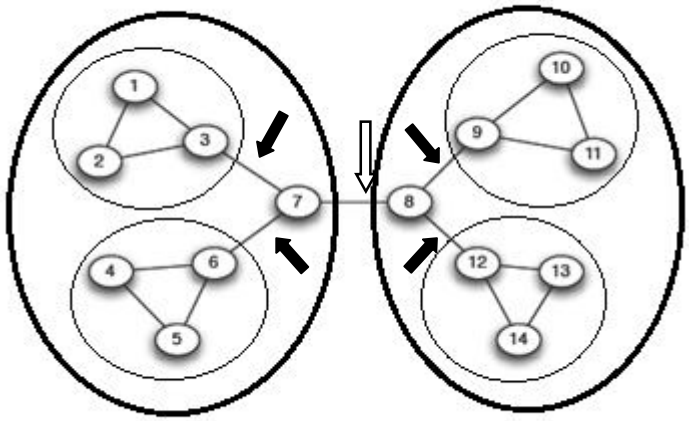
- $f_{G''}^{3} = f_{G''+\{3\}} - f_{G''-\{3\}} = \frac{k_{in}^{G'''}}{(k_{in}^{G'''} + k_{out}^{G'''})^{a}} - \frac{k_{in}^{G'''}}{(k_{in}^{G'''} + k_{out}^{G'''})^{a}} = \frac{16}{16+1} - \frac{10}{10+4} = \frac{16}{17} - \frac{10}{14} = \frac{54}{238} = f_{G''}^{6}$

- $f_{G''}^{7} = f_{G''+\{7\}} - f_{G''-\{7\}} = \frac{k_{in}^{G'''}}{(k_{in}^{G'''} + k_{out}^{G'''})^{a}} - \frac{k_{in}^{G'''}}{(k_{in}^{G'''} + k_{out}^{G'''})^{a}} = \frac{16}{16+1} - \frac{12}{12+2} = \frac{16}{17} - \frac{12}{14} = \frac{20}{238}$

Αυτή την φορά το βήμα 5 μας οδηγεί στο βήμα 1 όπου βρίσκουμε τον γείτονα του G''', τον κόμβο 8 όπου και υπολογίζουμε την καταλληλότητα του:

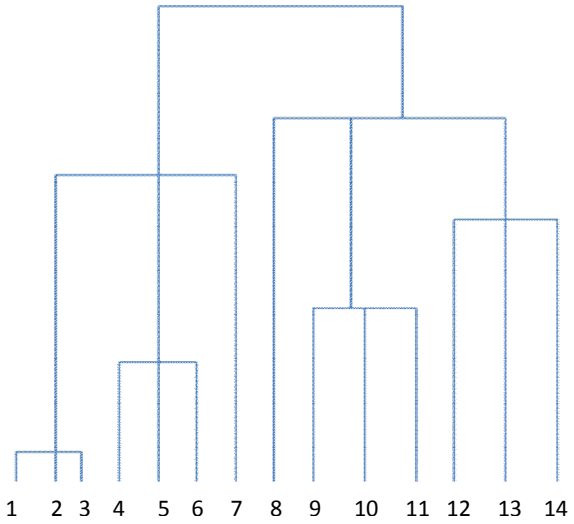
- $f_{G''}^{8} = f_{G''+\{8\}} - f_{G''-\{8\}} = \frac{k_{in}^{G'''}}{(k_{in}^{G'''} + k_{out}^{G'''})^{a}} - \frac{k_{in}^{G'''}}{(k_{in}^{G'''} + k_{out}^{G'''})^{a}} = \frac{18}{18+2} - \frac{16}{16+1} = \frac{18}{20} - \frac{16}{17} = -\frac{14}{340}$

Και επειδή είναι και ο μοναδικός κόμβος που εξετάζεται ο αλγόριθμος σταματάει, έχοντας εντοπίσει την κοινότητα που μας υποδεικνύει το $G''' = \{1,2,3,4,5,6,7\}$ και με γέφυρα την ακμή {7,8}. Τώρα μας έχει μείνει μόνο ο κόμβος 8 ο οποίος δεν έχει ανατεθεί σε καμιά ομάδα. Εκτελώντας την ίδια διαδικασία και λόγω συμμετρίας του δικτύου μας, ο κόμβος 8 θα συμπεριληφθεί στην κοινότητα που υποδεικνύει το γράφημα $G''' = \{8,9,10,11,12,13,14\}$ και με γέφυρα πάλι την ακμή {7,8}. Οι προκύπτουσες κοινότητες παρουσιάζονται στο Σχήμα 2.29.



Σχήμα 2.29: Με έντονο μαύρο κύκλο απεικονίζονται οι κοινότητες στις οποίες ανήκουν οι κόμβοι 7 και 8 και πως αυτές επικαλύπτονται με τις προηγούμενες. Με βέλη προσδιορίζονται οι γέφυρες.

Είναι εύκολο να συγκρίνει και να παρατηρήσει κανείς πως οι κοινότητες και οι γέφυρες που εντοπίστηκαν με αυτό τον αλγόριθμο συμπίπτουν με αυτές που εντοπίσαμε με τον Girvan-Newman αλγόριθμο. Στο Σχήμα 2.30 φαίνεται η ιεραρχική δομή των κοινοτήτων μας:



Σχήμα 2.30: Δενδρόγραμμα στο οποίο απεικονίζεται η ιεραρχική ομαδοποίηση του δικτύου μας για a=1.

- Για a=1,5

Αυξάνοντας την τιμή του a αναμένουμε να εντοπίσουμε μικρότερες κοινότητες. Διαλέγοντας να ξεκινήσουμε πάλι από τον κόμβο 1 ($G=\{1\}$) και ακολουθώντας τα βήματα του αλγορίθμου όπως πριν, θα έχουμε:

Οι γείτονες του κόμβου 1 είναι οι 2,3. Υπολογίζουμε τις καταλληλότητες αυτών των κόμβων:

- $f_G^2 = f_{G+\{2\}} - f_{G-\{2\}} = \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^a} - \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^a} = \frac{2}{(2+2)^{\frac{3}{2}}} - 0 = \frac{2}{4^{\frac{3}{2}}} = 0.25$
- $f_G^3 = f_{G+\{3\}} - f_{G-\{3\}} = \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^a} - \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^a} = \frac{2}{(2+3)^{\frac{3}{2}}} - 0 = \frac{2}{5^{\frac{3}{2}}} = 0.178$

και έτσι το υπογράφημά μας σε αυτό το σημείο θα είναι το $G' = \{1,2\}$. Οι καταλληλότητες αυτών των κόμβων είναι:

- $f_{G'}^1 = f_{G'+\{1\}} - f_{G'-\{1\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} = \frac{2}{(2+2)^{\frac{3}{2}}} - 0 = \frac{2}{4^{\frac{3}{2}}} = 0.25 = f_{G'}^2$

Ο μοναδικός γείτονας του G' είναι ο κόμβος 3 με καταλληλότητα:

- $f_{G'}^3 = f_{G'+\{3\}} - f_{G'-\{3\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^a} = \frac{6}{(6+1)^{\frac{3}{2}}} - \frac{2}{(2+2)^{\frac{3}{2}}} = \frac{6}{7^{\frac{3}{2}}} - \frac{2}{4^{\frac{3}{2}}}$

  $= 0.07$

και έτσι $G'' = \{1,2,3\}$ όπου θα ισχύουν:

- $f_{G^{'}}^{1} = f_{G^{"}+\{1\}} - f_{G^{"}-\{1\}} = \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} - \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} = \frac{6}{(6+1)^{\frac{3}{2}}} - \frac{2}{(2+3)^{\frac{3}{2}}} = \frac{6}{7^{\frac{3}{2}}} - \frac{2}{5^{\frac{3}{2}}} =$

  $0.15 = f_{G^{'}}^{2}$

- $f_{G^{'}}^{3} = f_{G^{"}+\{3\}} - f_{G^{"}-\{3\}} = \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} - \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} = \frac{6}{(6+1)^{\frac{3}{2}}} - \frac{2}{(2+2)^{\frac{3}{2}}} = \frac{6}{7^{\frac{3}{2}}} - \frac{2}{4^{\frac{3}{2}}}$

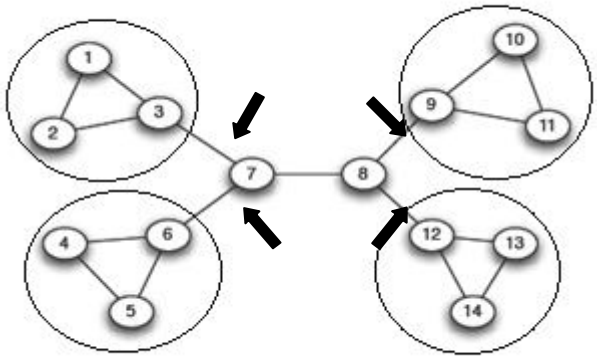
  $= 0.07$

Οδηγούμαστε στο βήμα 1 όπου γείτονας πλέον είναι ο κόμβος 7 και θα ισχύει:

- $f_{G^{'}}^{7} = f_{G^{"}+\{7\}} - f_{G^{"}-\{7\}} = \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} - \frac{k_{in}^{G^{"}}}{(k_{in}^{G^{"}} + k_{out}^{G^{"}})^{a}} = \frac{8}{(8+2)^{\frac{3}{2}}} - \frac{6}{(6+1)^{\frac{3}{2}}} = \frac{8}{10^{\frac{3}{2}}} - \frac{2}{7^{\frac{3}{2}}}$

  $= -0.07$

και επειδή είναι ο μοναδικός κόμβος με αρνητική καταλληλότητα ο αλγόριθμος σταματάει. Έτσι έχουμε εντοπίσει την κοινότητα που μας υποδεικνύει το υπογράφημα $G^{"} = \{1,2,3\}$. Διαλέγοντας με την σειρά τους κόμβους 4,10,13 (το ίδιο θα συνέβαινε αν διαλέγαμε οποιονδήποτε άλλο κόμβο εκτός από τους 7 και 8, με τέτοια σειρά πάντα ώστε να μην διαλέξουμε κόμβο που έχει μπει σε ομάδα) οι κοινότητες που θα εντοπίζαμε είναι αυτές που φαίνονται στο Σχήμα 2.31.



Σχήμα 2.31: Το δίκτυο με τις κοινότητες και τις γέφυρες για a=1.5.

Παρατηρούμε ότι οι κοινότητες και οι γέφυρες που εντοπίστηκαν είναι ίδιες με αυτές του σχήματος για a=1. Συνεχίζοντας διαλέγουμε τον κόμβο 7 και $G = \{7\}$. Με γείτονες τους κόμβους 3,6,8 θα έχουμε:

- $f_{G}^{3} = f_{G+\{3\}} - f_{G-\{3\}} = \frac{k_{in}^{G}}{(k_{in}^{G} + k_{out}^{G})^{a}} - \frac{k_{in}^{G}}{(k_{in}^{G} + k_{out}^{G})^{a}} = \frac{2}{(2+4)^{\frac{3}{2}}} - 0 = \frac{2}{6^{\frac{3}{2}}} = 0.13 = f_{G}^{6}$

  $= f_{G}^{8}$

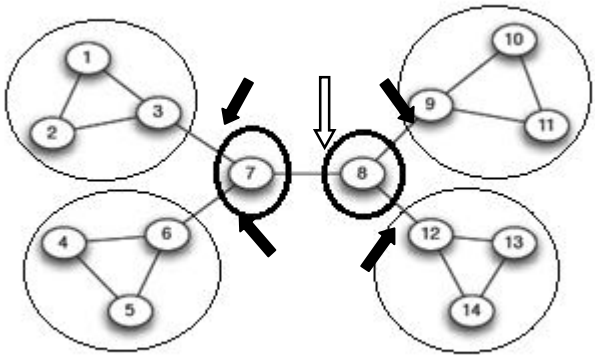
και το νέο υπογράφημα θα είναι το $G^{'} = \{3,6,7,8\}$. Οι καταλληλότητες των κόμβων θα είναι:

- $$f_{G'}^{3} = f_{G'+\{3\}} - f_{G'-\{3\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^{a}} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^{a}} = \frac{6}{(6+6)^{\frac{3}{2}}} - \frac{4}{(4+5)^{\frac{3}{2}}} = \frac{6}{12^{\frac{3}{2}}} - \frac{4}{9^{\frac{3}{2}}}$$

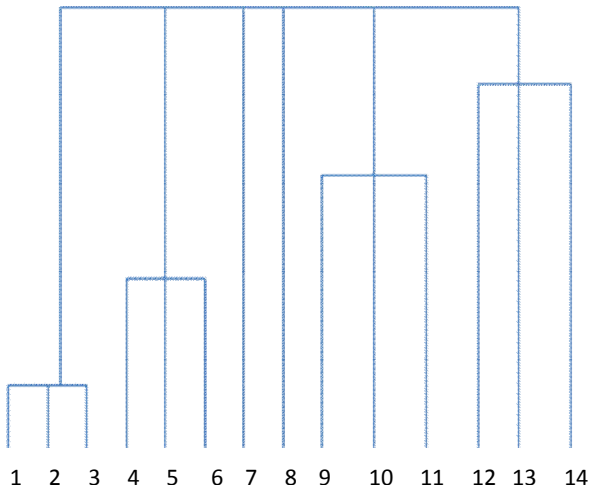
  $$= 0.038 = f_{G'}^{6} = f_{G'}^{8}$$

- $$f_{G'}^{7} = f_{G'+\{7\}} - f_{G'-\{7\}} = \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^{a}} - \frac{k_{in}^{G'}}{(k_{in}^{G'} + k_{out}^{G'})^{a}} = \frac{6}{(6+6)^{\frac{3}{2}}} - 0 = \frac{6}{12^{\frac{3}{2}}} = 0.1443$$

Οι κόμβοι με τις αρνητικές πολυπλοκότητες αφαιρούνται από το δίκτυο και επειδή αυτό συμβαίνει σε όλους τους γείτονες του κόμβου 7 ο αλγόριθμος σταματάει με $G'' = \{7\}$. Το ίδιο θα συμβεί και με τον κόμβο 8 όπου $G'' = \{8\}$. Παρατηρούμε ότι αυτοί οι κόμβοι δεν μπόρεσαν να ενταχθούν σε κάποια ομάδα αλλά μπορούμε να θεωρήσουμε ότι αποτελούν μια κοινότητα, που μόνο στοιχείο έχουν τον ίδιο τους τον εαυτό, μη μπορώντας να εντάξουν και άλλους κόμβους στην ομάδα τους. Η κοινότητα στην οποία άνηκε ο κόμβος 7 έγινε μικρότερη αποτελούμενη μόνο από ένα κόμβο, λόγω της αύξησης της τιμής του a. Οι ακμές {3,7}, {6,7},{8,9},{8,12} καθώς και η {7,8} αποτελούν τις γέφυρες. Στο Σχήμα 2.32 απεικονίζονται όσα εξετάσαμε και στο Σχήμα 2.33 φαίνεται η ιεραρχική δομή των κοινοτήτων μας.



Σχήμα 2.32: Οι κοινότητες στο δίκτυο. Με έντονο μαύρο κύκλο οι κοινότητες των κόμβων 7 και 8.



Σχήμα 2.33: Δενδρόγραμμα στο οποίο απεικονίζεται η ιεραρχική ομαδοποίηση του δικτύου μας για a=1,5.

2.6.2 Πολυπλοκότητα

Όπως αναφέρουν οι συγγραφείς του [8], είναι δύσκολο να υπολογιστεί η υπολογιστική πολυπλοκότητα του αλγορίθμου αφού εξαρτάται από το μέγεθος των κοινοτήτων και την έκταση των επικαλύψεών τους, οι οποίες με την σειρά τους εξαρτώνται από το δίκτυο που μελετάται καθώς και από την παράμετρο a. Ο χρόνος για να κατασκευαστεί μια κοινότητα με s κόμβους φτάνει περίπου να είναι $O(s^2)$ λόγω των επαναλαμβανόμενων βημάτων. Άρα μια πρόχειρη εκτίμηση της πολυπλοκότητας για μια σταθερή τιμή του a είναι $O(n_c <s^2>)$, όπου n_c είναι ο αριθμός των συστατικών μερών από την κάλυψη που επιτεύχθηκε και $<s^2>$ η δεύτερη στιγμή του μεγέθους της κοινότητας. Το τετράγωνο είναι αποτέλεσμα της επανεξέτασης όλων των κόμβων μια κοινότητας για να ελεγχθεί η καταλληλότητα μετά από κάθε κίνηση. Η πολυπλοκότητα στη χειρότερη περίπτωση είναι $O(n^2)$, όπου n είναι ο αριθμός των κόμβων του δικτύου, όταν οι κοινότητες είναι μεγέθους ισάξιου με n. Γενικότερα, δεν ισχύει αυτή η περίπτωση και έτσι οι περισσότερες εφαρμογές του αλγορίθμου τρέχουν πιο γρήγορα και σχεδόν γραμμικά όταν οι κοινότητες είναι μικρές. Η συνολική πολυπλοκότητα του αλγορίθμου, για να πραγματοποιήσουμε μια ολοκληρωμένη ανάλυση του δικτύου, εξαρτάται επίσης από τον αριθμό των τιμών του a που απαιτούνται για την επίλυση της ιεραρχικής του δομής. Η ιεραρχία από τις καλύψεις μπορεί να παρουσιαστεί με τον καλύτερο δυνατό τρόπο μέσω των μεγάλων τιμών του a που χρησιμοποιούνται για το τρέξιμο του αλγορίθμου. Εάν το δίκτυο έχει μια ιεραρχική δομή, όπως συνήθως συμβαίνει στα πραγματικά συστήματα, ο αριθμός των καλύψεων μεγαλώνει ως προς logn. Σε αυτή την περίπτωση, ο αριθμός των διαφορετικών τιμών του a που απαιτούνται για την επίλυση της ιεραρχίας είναι logn και ολόκληρη η ανάλυση μπορεί να πραγματοποιηθεί πολύ γρήγορα. Επίσης, κάθε επανάληψη του αλγορίθμου, για μια συγκεκριμένη τιμή του a, είναι ανεξάρτητη από τις άλλες. Έτσι ο υπολογισμός μπορεί εύκολα να γίνει παράλληλα τρέχοντας διαφορετικές τιμές του a σε κάθε υπολογιστή. Εάν ένα μεγάλο μέρος των πόρων του συστήματος δεν είναι διαθέσιμο, ένας οικονομικός τρόπος για να προχωρήσουμε με τον αλγόριθμό μας, είναι να ξεκινήσουμε με μια μεγάλη τιμή του a, με την οποία ο αλγόριθμος ολοκληρώνεται σε πολύ μικρό χρόνο και να χρησιμοποιήσουμε την τελική κάλυψη σαν αρχική δομή για το τρέξιμο του αλγορίθμου με μια ελαφρώς λιγότερη τιμή του a. Οι συγγραφείς του [8] καταλήγουνε πως για τα ιεραρχικά δίκτυα η διαδικασία που προτείνουν έχει χρονική πολυπλοκότητα χειρότερης περιπτώσεως ίση με n^2logn.

2.7 Βιβλιογραφία

[1] R. A. Hanneman, and M. Riddle, Introduction to social network methods, Berkeley, CA, University of California; University of California Press, 2005.

[2] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc Natl Acad Sci U S A 99 (2002), 7821.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, The pagerank citation ranking: bringing order to the web, Stanford, CA, 1998.

[4] S. Gregory, An algorithm to find overlapping community structure in networks, Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007), Warsaw, Poland, Springer-Verlag, 2007, 91–102.

[5] S. Gregory, Finding overlapping communities using disjoint community detection algorithms, Complex Networks: CompleNet 2009, Berlin, Heidelberg, Springer-Verlag, 2009, 47–61.

[6] Gregory, S.: A Fast Algorithm to Find Overlapping Communities in Networks. In: PKDD 2008. LNAI, vol. 5211, pp. 408—423. Springer, Heidelberg (2008).

[7] J. Bagrow and E. Bollt, A local method for detecting communities, Phys Rev E 72 (2005), 46–108.

[8] A. Lancichinetti, S. Fortunato, and J. Kertesz, Detecting the overlapping and hierarchical community structure of complex networks, New J Phys 11 (2009), 033015.

[9] Radicchi F, Castellano C, Cecconi F, Loreto V and Parisi D 2004 Proc. Natl. Acad. Sci. USA 101 2658.

[10] S. Wasserman and K. Faust, Social Network Analysis. Cambridge University Press, Cambridge (1994).

[11] L. Freeman, A set of measures of centrality based upon betweenness. Sociometry 40, 35{41 (1977).

[12] M. E. J. Newman, Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. Phys. Rev. E 64, 016132 (2001).

[13] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

[14] http://www.ismll.uni-hildesheim.de/lehre/cmie-11w/script/lecture5.pdf

Κεφάλαιο 3

Διάχυση (Diffusion)

3.1 Εισαγωγή

Η διάχυση είναι μια διαδικασία στην οποία οι κορυφές ή οι ακμές ενός γραφήματος χαρακτηρίζονται «κατειλημμένες» ή «μη κατειλημμένες». Μπορεί να χρησιμοποιηθεί μια γενίκευση της διαδικασίας της διάχυσης που χρησιμοποιείται για την εύρεση κοινοτήτων σε πολύπλοκα δίκτυα, σύμφωνα με τον παρακάτω ορισμό:

Ορισμός 2 (Κοινότητα Διάχυσης) Μια κοινότητα διάχυσης σε ένα πολύπλοκο δίκτυο είναι ένα σύνολο από κόμβους, οι οποίοι έχουν ομαδοποιηθεί λόγω της διάδοσης της ίδιας ενέργειας ή πληροφορίας μέσα στο δίκτυο.

Ο ορισμός της διαδικασίας που ακολουθείται από τους αλγορίθμους σε αυτή την κατηγορία είναι:

Διαδικασία 2 Εκτέλεσε μια διαδικασία διάχυσης στο δίκτυο ακολουθώντας ένα συγκεκριμένο σύνολο από κανόνες διάδοσης και μετά ομαδοποίησε οποιουσδήποτε κόμβους που καταλήγουν στην ίδια κατάσταση.

Σύμφωνα με τον παραπάνω ορισμό, μια κοινότητα μπορεί επίσης να οριστεί ως ένα σύνολο από οντότητες που επηρεάζονται από ένα σταθερό σύνολο πηγών. Αυτό είναι σημαντικό γιατί, αλγόριθμοι που δεν έχουν λεπτομερώς αναπτυχθεί ως προσεγγίσεις για διαμερίσεις γράφων μπορούν επίσης να θεωρηθούν ως μέθοδοι εύρεσης κοινοτήτων. Βασικά, ο ορισμός του προβλήματος επικαλύπτεται με ένα άλλο γνωστό πρόβλημα της εξόρυξης δεδομένων: της διάδοσης της επιρροής και της ροής της πληροφορίας [1].

Οι κλασικοί αλγόριθμοι εύρεσης κοινοτήτων, βασισμένοι στη διάχυση, που θα παρουσιάσουμε εδώ είναι: τεχνική διάδοσης ετικέτας (label propagation technique), και ένας αλγόριθμος που θεωρεί το αρχικό γράφο σαν ένα ηλεκτρικό κύκλωμα, με τις ακμές να είναι αντιστάσεις.

3.2 Αλγόριθμος Label Propagation

Η βασική ιδέα πίσω από τον label propagation αλγόριθμο είναι η εξής. Οι αναλυτές του [2] υποθέτουν ότι ένας κόμβος x έχει γείτονες $x_1, x_2, \ldots, x_k$ και καθένας από αυτούς τους γείτονες φέρει μια ετικέτα που υποδηλώνει την κοινότητα στην οποία ανήκει. Στη συνέχεια, ο x καθορίζει την δικιά του κοινότητα βασιζόμενος στις ετικέτες των γειτόνων του. Υποθέτουν ακόμα, ότι κάθε κόμβος στο δίκτυο διαλέγει να γίνει μέλος μιας κοινότητας στην οποία ανήκει ο μέγιστος αριθμός των γειτόνων του, με τα δεσίματα να σπάνε ομοιόμορφα και τυχαία. Με άλλα λόγια κάθε κόμβος ενημερώνει την ετικέτα του αντικαθιστώντας την με την ετικέτα που χρησιμοποιείται από το μεγαλύτερο αριθμό των γειτόνων. Εάν περισσότερες από μια ετικέτες χρησιμοποιούνται από τον ίδιο μέγιστο αριθμό γειτόνων, τότε μια από αυτές επιλέγεται τυχαία. Μετά από κάποιες επαναλήψεις η ίδια ετικέτα θα έχει συσχετιστεί με όλα τα μέλη της κοινότητας. Οι αναλυτές στη συνέχεια, αναφέρουν ότι αρχικοποιούνε κάθε κόμβο με μοναδικές ετικέτες και αφήνουν τις ετικέτες να διαδοθούνε στο δίκτυο. Καθώς οι ετικέτες διαδίδονται, πυκνά συνδεδεμένες ομάδες κόμβων φτάνουν στην πλειοψηφία τους σε μια μοναδική ετικέτα (Σχήμα 3.1). Όταν πολλές τέτοιες ομάδες δημιουργηθούν στο δίκτυο, συνεχίζουν να επεκτείνονται εξωτερικά έως ότου είναι εφικτό να το κάνουν. Στο τέλος της διαδικασίας της διάχυσης, κόμβοι που έχουνε τις ίδιες ετικέτες ομαδοποιούνται μαζί σαν μια κοινότητα.

Η διαδικασία εκτελείται επαναληπτικά, όπου σε κάθε βήμα, κάθε κόμβος ενημερώνει την ετικέτα του βασιζόμενος στις ετικέτες των γειτόνων του. Η διαδικασία ενημέρωσης μπορεί να είναι σύγχρονη ή ασύγχρονη. Χρησιμοποιείται ασύγχρονη ενημέρωση όπου:

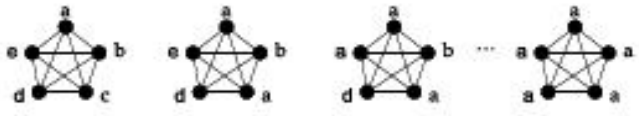
$$C_x(t) = f\left(C_{x_{i1}}(t), \ldots, C_{x_{im}}(t), C_{x_{i(m+1)}}(t-1), \ldots, C_{x_{ik}}(t-1)\right).$$

$C_x(t)$ είναι η ετικέτα του κόμβου x την χρονική στιγμή t και $x_{i1}, \ldots, x_{im}$ είναι οι γείτονες του x που έχουν ήδη ενημερωθεί στην τρέχουσα επανάληψη ενώ $x_{i(m+1)}, \ldots, x_{ik}$ είναι οι γείτονες που δεν έχουν ενημερωθεί στην τρέχουσα επανάληψη. Η σειρά με την οποία όλοι οι κόμβοι του δικτύου ενημερώνονται σε κάθε επανάληψη επιλέγεται τυχαία. Παρατηρήστε ότι, ενώ μπορεί να έχουμε n διαφορετικές ετικέτες στην αρχή του αλγορίθμου, ο αριθμός των ετικετών μειώνεται με τις επαναλήψεις, καταλήγοντας μόνο, σε τόσες μοναδικές ετικέτες όσες είναι και οι κοινότητες.

Ο label propagation αλγόριθμος περιγράφεται από τα ακόλουθα βήματα:

Αλγόριθμος 3.1 label propagation algorithm – RAK algorithm (Asynchronous version)

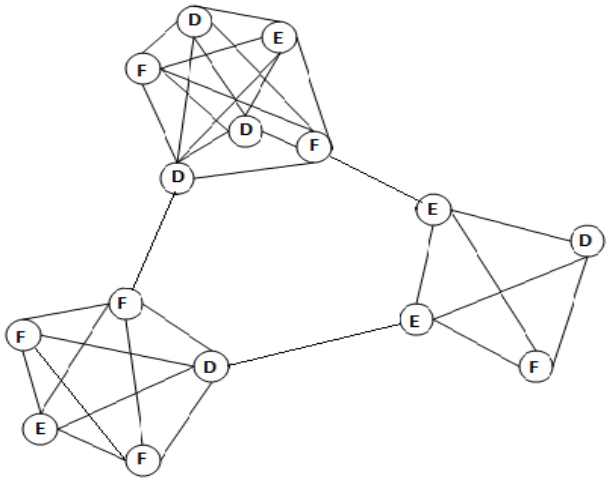
1. Αρχικοποίησε τις ετικέτες σε όλους τους κόμβους στο δίκτυο. Για ένα δοσμένο κόμβο x, $C_x(0) = x$.
2. Θέσε $t = 1$.
3. Ταξινόμησε τους κόμβους στο δίκτυο με τυχαία σειρά και τοποθέτησέ τους στο X.
4. Για κάθε $x \in X$ που επιλέγεται με αυτή την σειρά, έχουμε ότι $C_x(t) = f\left(C_{x_{i1}}(t),...,C_{x_{im}}(t),C_{x_{i(m+1)}}(t-1),...,C_{x_{ik}}(t-1)\right)$. Η f εδώ επιστρέφει την ετικέτα που προκύπτει από την μέγιστη συχνότητα μεταξύ των γειτόνων και τα δεσίματα να σπάνε ομοιόμορφα και τυχαία.
5. Εάν κάθε κόμβος έχει μια ετικέτα που ο μέγιστος αριθμός των γειτόνων του έχει, σταμάτα τον αλγόριθμο. Διαφορετικά θέσε $t = t + 1$ και πήγαινε στο βήμα 3.



Σχήμα 3.1: Οι κόμβοι ενημερώνονται ο ένας μετά τον άλλον καθώς προχωράμε από αριστερά προς τα δεξιά. Λόγω της υψηλής πυκνότητας των κόμβων (η μεγαλύτερη δυνατή σε αυτή την περίπτωση) όλοι οι κόμβοι αποκτούν την ίδια ετικέτα [2].

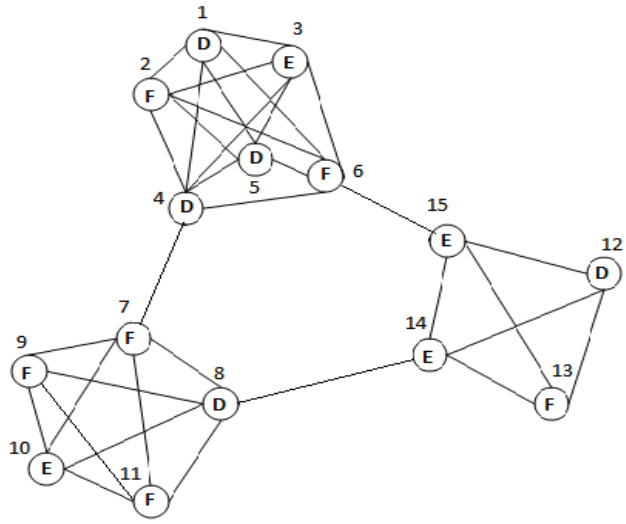
3.2.1 Παράδειγμα

Ας δοκιμάσουμε να εφαρμόσουμε τον παραπάνω αλγόριθμο στο δίκτυο του σχήματος 3.2. Εφαρμόζοντας το βήμα 1 του αλγορίθμου την χρονική στιγμή t=0 κάθε κόμβος έχει αρχικοποιηθεί με ετικέτες οι οποίες αποτελούνται από τα γράμματα D,F,E.



Σχήμα 3.2: Το αρχικό μας δίκτυο αρχικοποιημένο με τις ετικέτες [9].

Στο βήμα 2 θέτουμε t=1 και στο βήμα 3 ταξινομούμε τους κόμβους με τυχαία σειρά όπως φαίνεται στο Σχήμα 3.3. Οι αριθμοί πάνω από κάθε κόμβο μας υποδεικνύουν την σειρά που έγινε η ταξινόμηση. Οι κόμβοι τοποθετούνται σε ένα σύνολο Χ.



Σχήμα 3.3: Το δίκτυο με την ταξινόμηση.

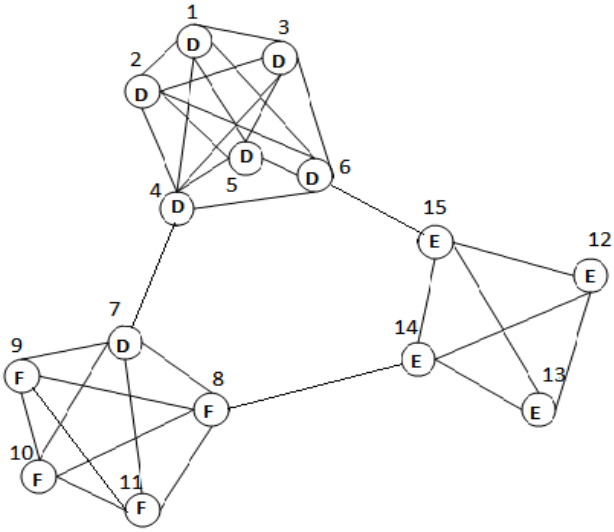
Από το σύνολο Χ διαλέγουμε κάθε κόμβο με την σειρά που τους ταξινομήσαμε και υπολογίζουμε την ετικέτα κάθε κόμβου την χρονική στιγμή t=1. Με κόκκινο χρώμα είναι οι κόμβοι οι οποίοι δεν έχουν ενημερωθεί στην τρέχουσα επανάληψη.

Για τον κόμβο που επιλέχθηκε πρώτος στην ταξινόμηση:

- $C_1(1) = f(F, E, D, D, F) = D$

Η συνάρτηση f επιστρέφει την ετικέτα με την μεγαλύτερη συχνότητα μεταξύ των γειτόνων. Ακολουθώντας με την σειρά τους κόμβους θα έχουμε:

- $C_2(1) = f(D, E, D, D, F) = D$
- $C_3(1) = f(D, D, D, D, F) = D$
- $C_4(1) = f(D, D, D, D, F) = D$
- $C_5(1) = f(D, D, D, D, F) = D$
- $C_6(1) = f(D, D, D, D, D) = D$
- $C_7(1) = f(D, D, F, E, F) = D$
- $C_8(1) = f(D, F, E, F, E) = F$
- $C_9(1) = f(D, F, E, F) = F$
- $C_{10}(1) = f(D, F, F, F) = F$
- $C_{11}(1) = f(D, F, F, F) = F$
- $C_{12}(1) = f(F, E, E) = E$
- $C_{13}(1) = f(E, E, E) = E$
- $C_{14}(1) = f(D, E, E, E) = E$
- $C_{15}(1) = f(F, E, E, E) = E$

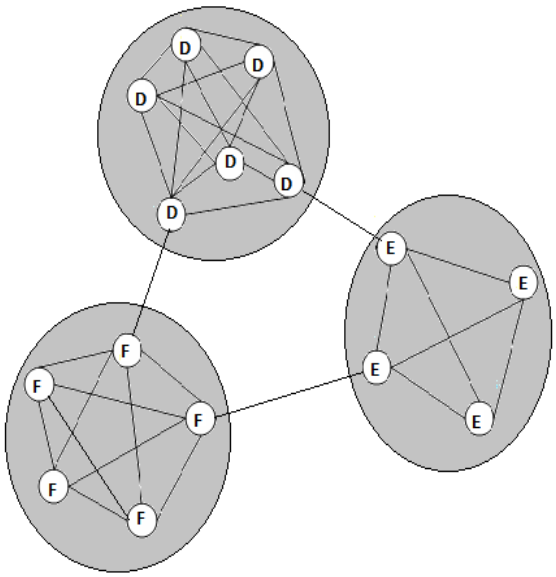


Σχήμα 3.4: Το δίκτυο στο τέλος της 1ης επανάληψης.

Στο βήμα 5 εξετάζουμε αν κάθε κόμβος φέρει μια ετικέτα που ο μέγιστος αριθμός των γειτόνων του έχει. Κάτι τέτοιο δεν ισχύει για τον κόμβο που είχε ταξινομηθεί στη θέση 7 (Σχήμα 3.4) και οδηγούμαστε ξανά στο βήμα 3, για τα t=2. Οι κόμβοι ταξινομούνται πάλι και έστω ότι διαλέγουμε την ίδια σειρά. Πραγματοποιώντας την ίδια διαδικασία όλοι οι κόμβοι θα διατηρήσουν την ετικέτα τους εκτός από τον κόμβο στη θέση 7 της ταξινόμησης όπου θα έχουμε:

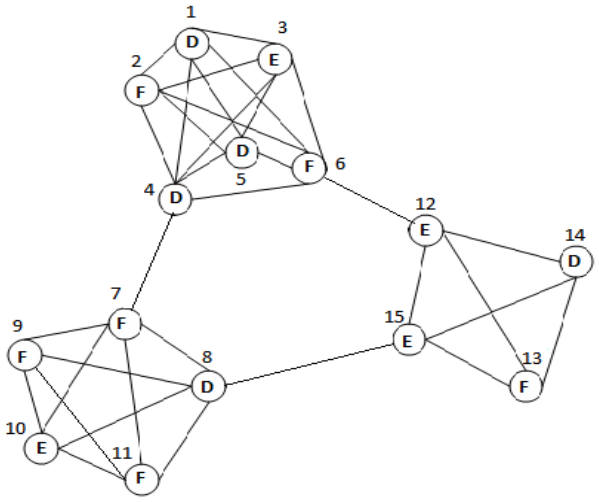
- $C_7(2) = f(D, F, F, F, F) = F$

Ο αλγόριθμος σταματάει στην δεύτερη επανάληψη αφού κάθε κόμβος φέρει την ετικέτα την οποία έχει ο μέγιστος αριθμός των γειτόνων. Στο Σχήμα 3.5 μπορούμε να δούμε τις κοινότητες που εντοπίστηκαν μετά το τέλος του αλγορίθμου.



Σχήμα 3.5: Μέσα σε γκρι κύκλο απεικονίζονται οι κοινότητες του δικτύου.

Ωστόσο, σε αυτό το σημείο οι αναλυτές δικτύων αναφέρουν ορισμένα προβλήματα που προκύπτουν με την ασύγχρονη ενημέρωση των ετικετών[3]. Πρώτον, σε κάθε επανάληψη διαλέγεται τυχαία η σειρά που γίνεται η ενημέρωση. Αυτό κάνει τον αλγόριθμο ασταθή, γιατί ένα διαφορετικό τρέξιμο μπορεί να δώσει ένα διαφορετικό αποτέλεσμα ανάθεσης ετικετών. Δεύτερον, αρκετές φορές δημιουργεί λανθασμένα μια πάρα πολύ μεγάλη κοινότητα (“monster” community) και πολλές μικρές κοινότητες [4]. Αυτό το πρόβλημα έγκειται στο γεγονός ότι, λόγω της ασύγχρονης φύσης του αλγορίθμου, κατά την διάρκεια των αρχικών βημάτων, η τυχαιότητα της ανταλλαγής των κορυφών τείνει να ευνοεί την εξάπλωση κάποιων ετικετών σε σχέση με κάποιες άλλες. Κάποιες κοινότητες δεν έχουν αρκετά ισχυρές συνδέσεις ώστε να αποτρέψουν την υπερχείλιση των ετικετών. Μπορούμε να δούμε ένα φαινόμενο υπερχείλισης στο δικό μας παράδειγμα εάν η τυχαιότητα της ταξινόμησή μας αλλάξει όπως φαίνεται και στο Σχήμα 3.6.

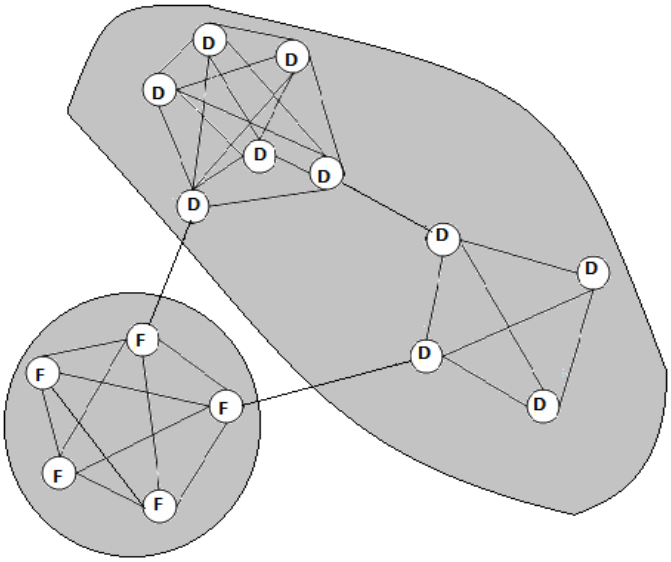


Σχήμα 3.6: Το δίκτυο με διαφορετική ταξινόμηση.

Ακολουθώντας την ίδια διαδικασία στην 1η επανάληψη θα έχουμε τις εξής αλλαγές:

- $C_{12}(1) = f(D, F, D, E) = D$
- $C_{13}(1) = f(D, D, E) = D$
- $C_{14}(1) = f(D, D, E) = D$
- $C_{15}(1) = f(D, D, D, F) = D$

και στην 2η επανάληψη θα ισχύει πάλι $C_7(2) = f(D, F, F, F, F) = F$. Έτσι θα έχουμε εντοπίσει τις κοινότητες του σχήματος 3.7.



Σχήμα 3.7: Οι κοινότητες του δικτύου λόγω υπερχείλισης (flooding).

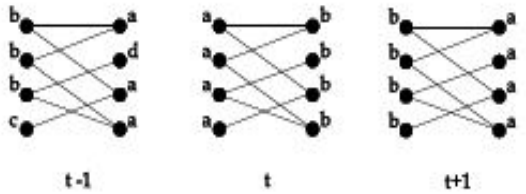
Διάφορα πειράματα αποδεικνύουν ότι η σύγχρονη πτυχή του αλγορίθμου μειώνει την πιθανότητα δημιουργίας πολύ μεγάλων κοινοτήτων χωρίς όμως να τις αποτρέπει εξολοκλήρου[4].

Αλγόριθμος 3.2 label propagation algorithm – RAK algorithm (Synchronous version)

---

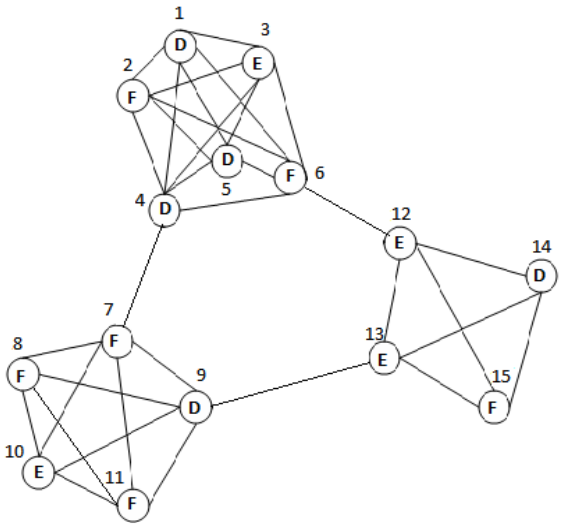
1. Αρχικοποίησε τις ετικέτες σε όλους τους κόμβους στο δίκτυο. Για ένα δοσμένο κόμβο x, $C_x(0) = x$.
2. Θέσε $t = 1$.
3. Για κάθε $x \in X$, έχουμε ότι $C_x(t) = f\left(C_{x_1}(t-1), \ldots, C_{x_k}(t-1)\right)$. Η f εδώ επιστρέφει την ετικέτα που προκύπτει από την μέγιστη συχνότητα μεταξύ των γειτόνων και τα δεσίματα να σπάνε ομοιόμορφα και τυχαία.
4. Εάν κάθε κόμβος έχει μια ετικέτα που ο μέγιστος αριθμός των γειτόνων του έχει, σταμάτα τον αλγόριθμο. Διαφορετικά θέσε $t = t + 1$ και πήγαινε στο βήμα 3.

Η διαφορά με τον προηγούμενο αλγόριθμο είναι ότι εδώ σε κάθε επανάληψη κάθε κόμβος ανανεώνει την ετικέτα του με βάση τις ετικέτες των γειτόνων του στην προηγούμενη επανάληψη, t-1. Το πλεονέκτημα αυτού του αλγορίθμου είναι ότι κάθε βήμα της διάχυσης μπορεί να γίνει παράλληλα, αφού δεν υπάρχουν εξαρτήσεις ανάμεσα στις ετικέτες που ανήκουν στο ίδιο βήμα. Αυτό είναι πολύ σημαντικό όταν τα δίκτυα είναι μεγάλα. Κάτι τέτοιο δεν ισχύει στη ασύγχρονη πτυχή του αλγορίθμου όπου κάθε κόμβος δεν μπορεί να υπολογίσει την δικιά του ετικέτα πριν οι γείτονες που προηγούνται σύμφωνα με την ταξινόμηση που επιλέξαμε δεν έχουν ολοκληρώσει τον δικό τους υπολογισμό. Από την άλλη πλευρά, ο σύγχρονος αλγόριθμος παρουσιάζει ένα πρόβλημα όταν στο δίκτυο μας έχουμε υπογραφήματα που είναι διμερή ή περίπου διμερή στην δομή τους ή κάποια άλλα είδη γραφημάτων[2][3]. Αυτό οδηγεί σε ταλάντωση των ετικετών όπως μπορούμε να δούμε στο Σχήμα 3.8. Η ασύγχρονη προσέγγιση μειώνει σημαντικά το φαινόμενο της ταλάντωσης, ωστόσο πάσχει και αυτή από τα προβλήματα που αναφέρθηκαν πιο πάνω.



Σχήμα3.8: Παράδειγμα διμερούς δικτύου στο οποίο τα σύνολα των ετικετών των δύο μερών είναι αποσυνδεδεμένα. Σε αυτή την περίπτωση λόγω των επιλογών που έγιναν από τους κόμβους στο βήμα t, οι ετικέτες των κόμβων ταλαντεύονται μεταξύ του a και του b [2].

Ας δούμε σε αυτό το σημείο πως η σύγχρονη προσέγγιση στη διάδοση των ετικετών επιδρά στο γράφημα του σχήματος 3.2. Στο Σχήμα 3.9 παρουσιάζετε ξανά το δίκτυο με τις ετικέτες αρχικοποιημένες στους κόμβους (t=0). Οι αριθμοί αυτή την φορά υποδεικνύουν την ονομασία των κόμβων και όχι την τυχαία ταξινόμησή τους όπως έγινε με την ασύγχρονη προσέγγιση.

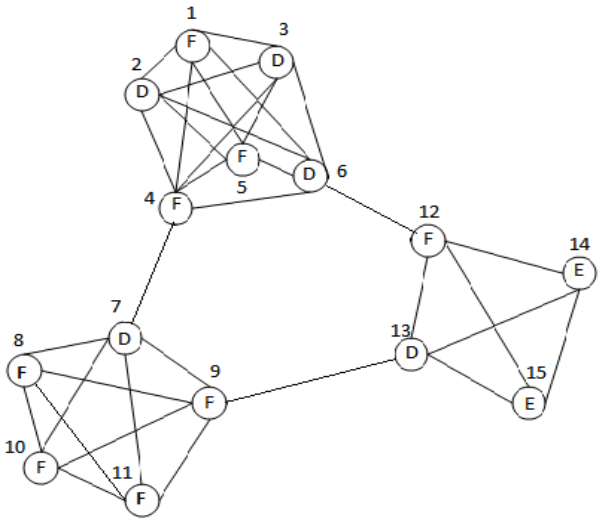


Σχήμα 3.9: Το δίκτυο με τις αρχικοποιημένες ετικέτες και αριθμούς που υποδεικνύουν την ονομασία σε κάθε κόμβο.

Για t=1 στο βήμα 3 έχουμε τα εξής αποτελέσματα:

- $C_4(1) = f(D,F,E,D,F,F) = F$
- $C_5(1) = f(D,F,E,D,F) = F$
- $C_2(1) = f(D,E,D,D,F) = D$
- $C_1(1) = f(F,E,D,D,F) = F$
- $C_3(1) = f(D,F,D,D,F) = D$
- $C_6(1) = f(D,F,E,D,D,E) = D$
- $C_{12}(1) = f(F,E,D,F) = F$
- $C_9(1) = f(F,F,E,F,E) = F$
- $C_{14}(1) = f(E,E,F) = E$
- $C_{15}(1) = f(E,E,D) = E$
- $C_{13}(1) = f(D,E,D,F) = D$
- $C_7(1) = f(D,F,D,E,F) = D$
- $C_8(1) = f(F,D,E,F) = F$
- $C_{11}(1) = f(F,F,D,E) = F$
- $C_{10}(1) = f(F,F,D,F) = F$

και καταλήγουμε στο Σχήμα 3.10.

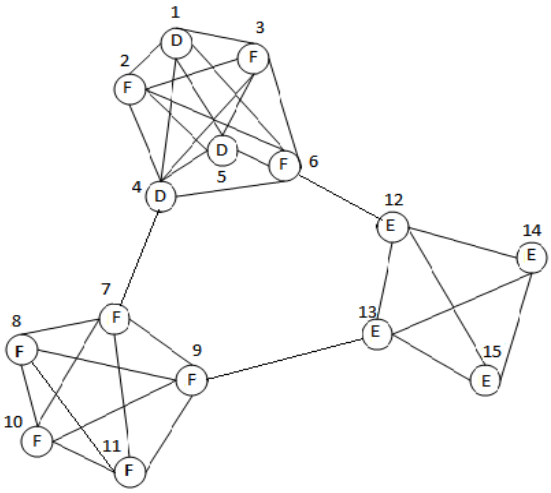


Σχήμα 3.10: Το δίκτυο στο τέλος της 1^{ης} επανάληψης.

Προφανώς σε αυτό το σημείο δεν φέρουν όλοι οι κόμβοι ετικέτα που ο μέγιστος αριθμός των γειτόνων τους έχει και έτσι προχωράμε στη $2^{η}$ επανάληψη, για t=2. Η σειρά με την οποία θα ενημερώσουμε τους κόμβους δεν έχει ξανά σημασία αφού, όπως και πριν, η ενημέρωση κάθε κόμβου είναι ανεξάρτητη. Έτσι θα έχουμε:

- $C_8(1) = f(D,F,F,F) = F$
- $C_{10}(1) = f(D,F,F,F) = F$
- $C_{11}(1) = f(D,F,F,F) = F$
- $C_9(1) = f(D,F,F,F,D) = F$
- $C_7(1) = f(F,F,F,F,F) = F$
- $C_{12}(1) = f(D,D,E,E) = E$
- $C_{13}(1) = f(F,F,E,E) = E$
- $C_{14}(1) = f(F,D,E) = E$
- $C_{15}(1) = f(F,D,E) = E$
- $C_2(1) = f(F,D,F,F,D) = F$
- $C_1(1) = f(D,D,F,F,D) = D$
- $C_3(1) = f(F,D,F,F,D) = F$
- $C_4(1) = f(F,D,D,F,D,D) = D$
- $C_5(1) = f(F,D,D,F,D) = D$
- $C_6(1) = f(F,D,D,F,F,F) = F$

Έτσι θα έχουμε ένα νέο δίκτυο ενημερωμένο με τα καινούργιες ετικέτες όπως αυτό φαίνεται στο Σχήμα 3.11.



Σχήμα 3.11: Το δίκτυο στο τέλος της $2^{ης}$ επανάληψης.

Μετά το τέλος της $2^{ης}$ επανάληψης, οι κόμβοι 1-6 δεν φέρουν ακόμα ετικέτα που ο μέγιστος αριθμός των γειτόνων του έχει. Έτσι θα οδηγηθούμε στη $3^{η}$ επανάληψη όπου μετά τους υπολογισμούς οι κόμβοι 1,4,5 θα φέρουν την ετικέτα F και οι κόμβοι 2,3,6 την ετικέτα D. Καταλήξαμε πάλι στα αποτελέσματα του σχήματος 3.10 για αυτούς τους κόμβους. Παρατηρούμε ότι αν προσπαθήσουμε να προχωρήσουμε την διαδικασία θα έχουμε μια συνεχή ταλάντωση των δύο αυτών ετικετών ανάμεσα σε αυτούς τους κόμβους, φαινόμενο που περιγράψαμε πριν. Σημαντικό ρόλο σε αυτό έπαιξε και η τυχαιότητα που διαλέξαμε ανάμεσα στις ετικέτες των κόμβων 1 και 5 στην πρώτη επανάληψη. Εάν είχαμε διαλέξει την ετικέτα D σε έναν από τους δύο ή και στους δύο θα καταλήγαμε στα ίδια αποτελέσματα με αυτά του σχήματος 3.5.

Στον RAK αλγόριθμο, η ετικέτα ενός κόμβου ταυτίζεται με μια μοναδική κοινότητα στην οποία ο κόμβος ανήκει. Εάν οι κοινότητες επικαλύπτονται, κάθε κόμβος μπορεί να ανήκει σε περισσότερες από μία κοινότητες. Οι συγγραφείς του [5] εξηγούν και σχεδιάζουν έναν αλγόριθμό για αυτή την περίπτωση που ονομάζουν COPRA Algorithm ((Community Overlap PRopagation Algorithm).

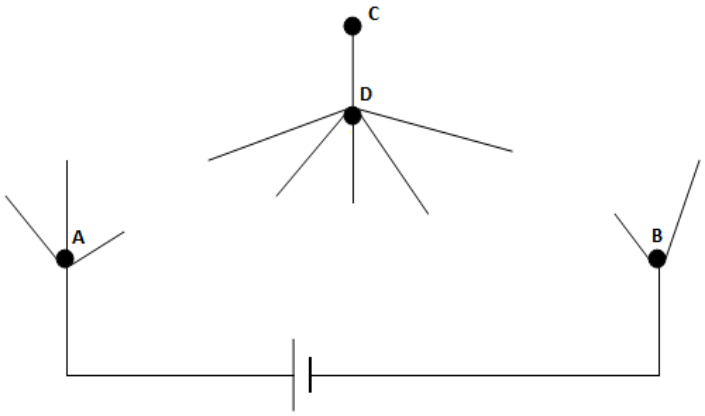
3.2.2 Πολυπλοκότητα

Η χρονική πολυπλοκότητα του αλγορίθμου είναι σχεδόν γραμμική (quasilinear). Η αρχικοποίηση κάθε κόμβου με τις μοναδικές ετικέτες παίρνει χρόνο $O(n)$ και κάθε επανάληψη του αλγορίθμου παίρνει γραμμικό χρόνο $O(m)$ και ο χρόνος για την επεξεργασία αποσυνδεδεμένων κοινοτήτων (κόμβοι με τις ίδιες ετικέτες ομαδοποιούνται μαζί) είναι $O(n+m)$ [2,5].

3.3 Kirchhoff

Οι συγγραφείς στο [6] ξεκινάνε την παρουσίαση της λειτουργίας του αλγορίθμου με το πιο απλό πρόβλημα, δηλαδή ξεκινάνε με πως θα διαχωρίσουνε ένα γράφημα σε δύο κοινότητες. Αρχικά θεωρούνε ένα γράφημα $G=(V,E)$. Στη συνέχεια υποθέτουν ότι γνωρίζουνε ήδη ότι οι κόμβοι Α και Β ανήκουν σε διαφορετικές κοινότητες, τις οποίες τις αποκαλούνε G_1 και G_2. Η ιδέα που περιγράφεται είναι ότι μπορούμε να θεωρήσουμε κάθε ακμή σαν ένα αντιστάτη με την ίδια αντίσταση και ακόμα μπορούμε να συνδέσουμε μια μπαταρία ανάμεσα στο Α και το Β έτσι ώστε να έχουμε σταθερές τάσεις, ας πούμε 1 και 0. Έχοντας κάνει αυτές τις υποθέσεις, μπορούμε να δούμε το γράφημα σαν ένα ηλεκτρικό κύκλωμα με ρεύμα να ρέει μέσα από τις ακμές (αντιστάτες). Λύνοντας τις εξισώσεις του Kirchhoff οι αναλυτές ισχυρίζονται ότι μπορούμε να λάβουμε την τιμή της τάσης κάθε κόμβου, που φυσικά θα κυμαίνεται μεταξύ 0 και 1. Ισχυρίζονται ακόμη, ότι μπορούμε να κρίνουμε από την τιμή της τάσης ενός κόμβου εάν ανήκει στο G_1 ή στο G_2. Πιο συγκεκριμένα, ένας κόμβος ανήκει στο G_1 εάν η τάση είναι μεγαλύτερη από ένα συγκεκριμένο κατώφλι, ας πούμε 0.5, και ανήκει στο G_2 εάν η τάση είναι μικρότερη από αυτό το κατώφλι.

Οι αναλυτές υποθέτουν αρχικά την απλή περίπτωση όπου ένας κόμβος C έχει μόνο ένα γείτονα D, έτσι λογικά ο C θα πρέπει να ανήκει στη ίδια κοινότητα με τον D όπως φαίνεται στο Σχήμα 3.12.



Σχήμα 3.12: Κόμβος με βαθμό 1 [6].

Επειδή δεν ρέει ρεύμα μέσα στην ακμή CD, τα δύο σημεία θα πρέπει να έχουν την ίδια τάση, που σημαίνει ότι ανήκουν στην ίδια κοινότητα.

Στη συνέχεια θεωρούνε την περίπτωση που ο κόμβος C συνδέεται με δύο γείτονες, D και E. Επειδή οι ακμές CD και CE έχουν την ίδια αντίσταση, θα πρέπει να έχουμε ότι $V_C = (V_D + V_E)/2$. Έτσι εάν ο D και ο E ανήκουν στην ίδια κοινότητα, για παράδειγμα V_D και V_E βρίσκονται πάνω ή κάτω από το κατώφλι, ο V_C που βρίσκεται ανάμεσα στο V_D και V_E θα πρέπει ομοίως να βρίσκεται πάνω ή κάτω από το κατώφλι, επομένως θα ανήκει στην ίδια κοινότητα όπως ο D και E. Από την άλλη πλευρά εάν ο D και E ανήκουν σε διαφορετικές κοινότητες , είναι συγκριτικά δύσκολο να πούμε σε ποια κοινότητα θα ανήκει ο C (V_C μπορεί να είναι κοντά στο κατώφλι). Αυτό είναι ακριβώς ένα πρόβλημα που προκύπτει όταν ένας κόμβος έχει συνδέσεις με περισσότερες από μία κοινότητες.

Στο τέλος οι αναλυτές θεωρούνε την πιο γενική περίπτωση: ο C συνδέεται με n γείτονες, $D_1,\ldots,D_n$. Οι εξισώσεις του Kirchhoff μας υποδεικνύουν ότι το συνολικό ρεύμα που ρέει στο C θα πρέπει να έχει άθροισμα ίσο με το 0, δηλαδή,

$$\sum_{i=1}^{n} I_i = \sum_{i=1}^{n} \frac{V_{D_i} - V_C}{R} = 0 \,, \qquad (1)$$

όπου I_i είναι το ρεύμα που ρέει από το D_i στο C. Έτσι

$$V_C = \frac{1}{n} \sum_{i=1}^{n} V_{D_i} \,. \qquad (2)$$

Αυτό σημαίνει ότι η τάση ενός κόμβου είναι ο μέσος όρος των τάσεων των γειτόνων του. Εάν η πλειοψηφία των γειτόνων του C ανήκει σε μια κοινότητα που έχει τάση

μεγαλύτερη από το κατώφλι, τότε ομοίως η V_C τείνει να ξεπερνάει το κατώφλι, και έτσι η μέθοδος αυτή τείνει να κατηγοριοποιεί τον C μέσα σε αυτή την κοινότητα.

Η μέθοδος μπορεί να επεκταθεί εύκολα για γραφήματα με βάρη. Το μόνο που έχουμε να κάνουμε είναι να θέσουμε την αγωγιμότητα του κάθε κόμβου αναλογικά με το βάρος του:

$$R_{ij} = w_{ij}^{-1} . \qquad (3)$$

Ο μέσος όρος που παρουσιάζεται στην εξίσωση (2) γίνεται αναλογικά ένας σταθμισμένος μέσος όρος.

3.3.1 Οι εξισώσεις του Kirchhoff σε γενική μορφή

Ακολουθώντας την εξίσωση (2) ,οι εξισώσεις του Kirchhoff για ένα κύκλωμα με n κόμβους μπορεί να γραφτούνε ως εξής:

$$V_1 = 1 , \qquad (4)$$

$$V_2 = 0 , \qquad (5)$$

$$V_i = \frac{1}{k_i} \sum_{(i,j)\in E} V_j = \frac{1}{k_i} \sum_{j\in G} V_j a_{ij} \text{ για } i = 3,...,n . \qquad (6)$$

όπου k_i είναι ο βαθμός του κόμβου i και a_{ij} είναι ο πίνακας γειτνίασης του γραφήματος. Χωρίς απώλεια τη γενικότητας, έχουμε δώσει ετικέτες στους κόμβους με τέτοιο τρόπο που η μπαταρία έχει συνδεθεί στο κόμβο 1 και 2 , τους οποίους και καλούμε πόλους ,σύμφωνα με τις εξισώσεις (4) και (5). Η εξίσωση (6) είναι ένα σύνολο από γραμμικές εξισώσεις $n-2$ τιμών $V_3,...,V_n$ που μπορούν να τοποθετηθούν σε μια πιο συμμετρική μορφή:

$$V_i = \frac{1}{k_i} \sum_{j=3}^{n} V_j a_{ij} + \frac{1}{k_i} a_{i1} \text{ για } i = 3,...,n \qquad (7)$$

Ορίζονται

$$V = \begin{pmatrix} V_3 \\ \vdots \\ V_n \end{pmatrix}, \quad B = \begin{pmatrix} \frac{a_{33}}{k_3} \cdots & \frac{a_{3n}}{k_3} \\ \vdots & \vdots \\ \frac{a_{n3}}{k_n} \cdots & \frac{a_{nn}}{k_n} \end{pmatrix}, \quad C = \begin{pmatrix} \frac{a_{31}}{k_3} \\ \vdots \\ \frac{a_{n1}}{k_n} \end{pmatrix}, \qquad (8)$$

Έτσι οι εξισώσεις του Kirchhoff μπορούνε να γραφτούν σε μορφή πίνακα

$$V = BV + C \qquad (9)$$

που έχει την μοναδική λύση

$$V = (I - B)^{-1} C \qquad (10)$$

Ακόμα ισχύει ότι:

$$L = \begin{bmatrix} k_3 & -a_{34} & \dots & -a_{3n} \\ -a_{43} & k_4 & \dots & -a_{4n} \\ \dots & & & \dots \\ -a_{n3} & -a_{n4} & \dots & k_n \end{bmatrix},\ D = \begin{bmatrix} a_{31} \\ . \\ . \\ . \\ a_{n1} \end{bmatrix}, \qquad (11)$$

και έτσι οι εξισώσεις του Kirchhoff μπορούν να γραφτούν στη μορφή

$$LV = D. \qquad (12)$$

Γενικά παίρνει χρόνο $O(n^3)$ για να λυθεί ένα σύνολο εξισώσεων όπως η εξίσωση (10). Ωστόσο όπως εξηγούνε οι αναλυτές το [6] μπορούμε να ελαττώσουμε το χρόνο σε $O(V+E)$.

3.3.2 Λύνοντας τις εξισώσεις του Kirchhoff σε γραμμικό χρόνο

Αρχικά οι αναλυτές θέτουνε $V_1 = 1, V_2 = \cdots = V_n = 0$ σε χρόνο $O(V)$. Ξεκινώντας από τον κόμβο 3, ενημερώνουνε διαδοχικά την τάση ενός κόμβου με τον μέσο όρο της τάσης των γειτόνων, σύμφωνα με την εξίσωση (2). Η διαδικασία ενημέρωσης τελειώνει όταν φτάσουν στον τελευταίο κόμβο n. Αποκαλούνε αυτή την διαδικασία ως γύρο. Επειδή κάθε κόμβος i έχει k_i γείτονες, ένας θα πρέπει να ξοδέψει μια ποσότητα χρόνου $O(k_i)$ για να υπολογίσει το μέσο όρο των γειτόνων που σημαίνει ότι ο συνολικός χρόνος που σπαταλάτε σε ένα γύρο είναι $O\left(\sum_{i=3}^{n} k_i\right) = O(E)$. Αφού επαναλάβουνε την διαδικασία της ενημέρωσης για ένα πεπερασμένο αριθμό γύρων, φτάνουν σε μια προσεγγιστική λύση με μια συγκεκριμένη ακρίβεια, η οποία δεν εξαρτάται από το μέγεθος του γραφήματος n αλλά εξαρτάται μόνο από τον αριθμό των επαναληπτικών γύρων. Με άλλα λόγια, για να αποκτήσουμε μια συγκεκριμένη ακρίβεια, έστω 0.01, χρειάζεται να επαναλάβουμε ,ας πούμε, 100 γύρους, χωρίς να λαμβάνουμε υπόψιν πόσο μεγάλο είναι το γράφημα. Έτσι ο συνολικός χρόνος τρεξίματος είναι $O(V+E)$.

Οι συγγραφείς παρουσιάζουν ένα παράδειγμα για δύο κοινότητες. Εξετάζουν τον αλγόριθμο σε ένα γνωστό δίκτυο (Zachary karate club[7]). Το γράφημα περιέχει αυστηρά δύο κοινότητες ίδιου μεγέθους. Δύο βασικά ερωτήματα πρέπει να απαντηθούν:

- Πως θα διαλέξουμε τους δύο πόλους έτσι ώστε να βρίσκονται σε διαφορετικές κοινότητες;
- Ποιο κατώφλι πρέπει να χρησιμοποιηθεί για να χωρίσουμε τις δύο κοινότητες;

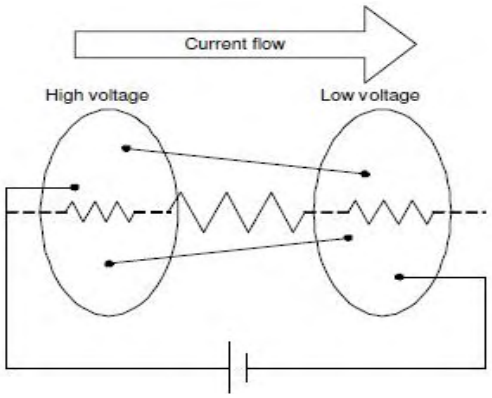
Για την πρώτη ερώτηση οι αναλυτές ισχυρίζονται πως λόγω της πυκνής σύνδεσης μέσα σε μια κοινότητα, η μέση απόσταση μεταξύ κόμβων που επιλέχτηκαν από μια κοινότητα είναι γενικά μικρότερη από την μέση απόσταση μεταξύ κόμβων διαφορετικών κοινοτήτων. Έτσι υπάρχει μεγάλη πιθανότητα δύο μακρινοί κόμβοι να βρίσκονται σε διαφορετικές κοινότητες, αποτελώντας έτσι τους πόλους. Προτείνεται στη συνέχεια μια μέθοδος για την εύρεση μακρινών ζευγαριών κόμβων η οποία επιλέγει τυχαία έναν κόμβο και ψάχνει για έναν πιο μακρινό με την διαδικασία breadth-first search. Αν προκύψουν περισσότεροι από ένας διαλέγεται τυχαία κάποιος. Στη συνέχεια εκτελείται πάλι breadth-first search για ένα μακρινό κόμβο αυτή την φορά από τον δεύτερο κόμβο που επιλέξαμε στο προηγούμενο βήμα και η διαδικασία αυτή συνεχίζεται έως ότου μετά από κάποια βήματα αναγνωριστεί ένα πολύ μακρινό ζευγάρι κόμβων.

Όσο αναφορά το δεύτερο ερώτημα, οι αναλυτές ισχυρίζονται ότι επειδή οι ακμές είναι πιο αραιές μεταξύ δύο κοινοτήτων η τοπική αντίσταση θα πρέπει να είναι μεγαλύτερη σε σύγκριση με την τοπική αντίσταση εντός των δύο κοινοτήτων. Για αυτό τον λόγο και η τάση πέφτει στην συμβολή των δύο κοινοτήτων(Σχήμα 3.13). Έτσι προτείνεται να θέσουμε το κατώφλι κοντά στο μέσο της μεγαλύτερης διαφοράς των τάσεων. Ωστόσο η μεγαλύτερη διαφορά συχνά εμφανίζεται στα δύο άκρα του φάσματος των τάσεων και αυτό το γεγονός μπορεί να οδηγήσει στο διαχωρισμό του γραφήματος σε δύο εξαιρετικά ασυμμετρικές κοινότητες, από τις οποίες η μία θα περιλαμβάνει μόνο έναν η δύο κόμβους.

Οι αναλυτές ορίζουν στη συνέχεια τον όρο «κοντά στο μέσο» (“near the middle”). Προσπαθούνε να βρούνε κοινότητες ίσου μεγέθους και ορίζουν την ανεκτικότητα (tolerance) για να περιγράψουν το εύρος των επιτρεπόμενων μεγεθών των κοινοτήτων. Για το παράδειγμα karate club που εξετάζεται στο [6] υπάρχουν 34 κόμβοι. Διαιρώντας στη μέση, βρίσκουνε δύο κοινότητες με 17 κόμβους η καθεμία. Ανεκτικότητα 0.2 σημαίνει ότι ψάχνουνε για κοινότητες με μέγεθος 17±20% που σημαίνει κοινότητες μεγέθους μεταξύ 14 και 21. Έπειτα ταξινομούν τις τιμές των τάσεων και ψάχνουν την μεγαλύτερη διαφορά κοντά στο μέσο, δηλαδή ανάμεσα στο 14 και 21. Ανάμεσα σε αυτό το διάστημα υπάρχουν 7 διαφορές τάσεων, διαλέγουν την μεγαλύτερη και βρίσκουν την τιμή στο μέσο αυτής της διαφοράς. Αυτό θα είναι και το κατώφλι (threshold). Οι συγγραφείς καταλήγουν με την επισήμανση ότι η παραπάνω μέθοδος μπορεί να μην λειτουργεί γιατί η μεγαλύτερη απόσταση μπορεί να επιτευχθεί μέσα στην ίδια κοινότητα και όχι ανάμεσα στις

κοινότητες, παρουσιάζοντας αυτή την περίπτωση στο παράδειγμα που μελετάνε (karate club).

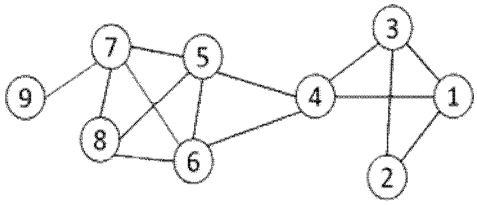
Μια άλλη μέθοδος που προτείνεται από τους αναλυτές του [6] εστιάζει στην αποφυγή του προβλήματος των δύο πόλων αντί να προσπαθήσει να το λύσει. Η ιδέα έχει να κάνει με την τυχαία επιλογή δύο κόμβων και την εφαρμογή του αλγορίθμου για το διαχωρισμό του γραφήματος σε δύο κοινότητες. Αυτή η διαδικασία επαναλαμβάνεται πολλές φορές (ο συνολικός χρόνος παραμένει $O(V+E)$). Οι αναλυτές καταλήγουν στο συμπέρασμα ότι τα μισά αποτελέσματα θα είναι σωστά ενώ τα άλλα μισά λανθασμένα. Στη συνέχεια βελτιώνουν αυτή την μέθοδο με την επιλογή δύο κόμβων οι οποίοι δεν είναι γείτονες(δεν υπάρχει ακμή μεταξύ τους). Η πιθανότητα έτσι των δύο τυχαία επιλεγμένων κόμβων να ανήκουν σε διαφορετικές κοινότητες αυξάνεται πιο πάνω από το μισό, κάτι που μας υποδεικνύει ότι η πλειοψηφία των αποτελεσμάτων είναι σωστά. Κατά αυτό τον τρόπο μπορεί κάποιος να χρησιμοποιήσει αυτή την πλειοψηφία για να εντοπίσει τις κοινότητες. Τέλος η μέθοδος εξετάζεται σε γραφήματα με n κοινότητες και συγκεκριμένα σε ένα γράφημα (US college football data) που μελετήθηκε από τους Girvan και Newman[8].



Σχήμα 3.13: Το ρεύμα ρέει από αριστερά προς στα δεξιά χτίζοντας μια διαφορά τάσης. Φυσικά σκεπτόμενοι, επειδή οι κόμβοι μέσα σε μια κοινότητα είναι πυκνά συνδεδεμένοι, οι τάσεις τους τείνουν να είναι κοντά(σε τιμή). Μεγάλη διαφορά τάσης παρατηρείται στο ενδιάμεσο μεταξύ των δύο κοινοτήτων όπου οι ακμές είναι αραιές και η τοπική αντίσταση μεγάλη [6].

3.3.3 Παράδειγμα δύο κοινοτήτων

Ας θεωρήσουμε το δίκτυο του σχήματος 3.14.



Σχήμα 3.14: Το αρχικό μας δίκτυο.

Διαλέγοντας δύο μακρινούς κόμβους για πόλους, έστω τον 1 και τον 9, θα έχουμε V_1=1 και V_9=0. Ο πίνακας γειτνίασης του παραπάνω γραφήματος είναι ο εξής:

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Οι εξισώσεις του (11) θα μας δώσουν τα εξής αποτελέσματα:

$$L = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & -1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 4 & -1 & -1 & -1 \\ 0 & 0 & -1 & -1 & 4 & -1 & -1 \\ 0 & 0 & 0 & -1 & -1 & 4 & -1 \\ 0 & 0 & 0 & -1 & -1 & -1 & 4 \end{bmatrix} \quad \text{και} \quad D = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

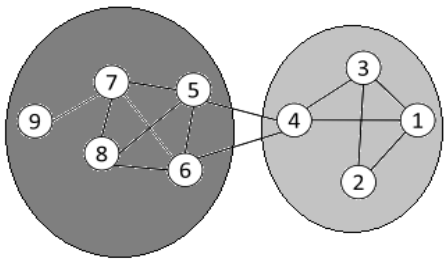
Λύνοντας την εξίσωση (12), που είναι της μορφής Ax=b, θα έχουμε:

$$V = \begin{bmatrix} 0.933 \\ 0.867 \\ 0.667 \\ 0.400 \\ 0.400 \\ 0.267 \\ 0.267 \end{bmatrix}$$

Το δίκτυο μας αποτελείται από 9 κόμβους και με ανεκτικότητα 0.2 θα βρούμε κοινότητες μεγέθους 9/2±20%, δηλαδή κοινότητες μεγέθους 4 ή 5. Ταξινομώντας τους κόμβους κατά αύξουσα σειρά έχουμε:

V_9=0, V_8=0.267, V_7=0.267, V_6=0.400, V_5=0.400, V_4=0.667, V_3=0.867, V_2=0.933, V_1=1.

Ψάχνουμε τώρα να βρούμε την μεγαλύτερη διαφορά «κοντά στο κέντρο», δηλαδή δημιουργώντας κοινότητες των τεσσάρων ή πέντε κόμβων, ψάχνουμε την διαφορά των τάσεων. Έτσι, στο παράδειγμά μας υπάρχουν δύο διαχωρισμοί κοινοτήτων που μπορούν να δημιουργηθούν (σύμφωνα πάντα με τους ταξινομημένους κόμβους). Ο ένας αποτελείται από τους κόμβους 9,8,7,6(σύνολο 4) που υποδηλώνει μια κοινότητα και τους κόμβους 5,4,3,2,1(σύνολο 5) να υποδηλώνουν την δεύτερη κοινότητα, με τις αντίστοιχες τάσεις. Ανάμεσα στους κόμβους 6 και 5 παρατηρείται μια διαφορά τάσης ίση με 0. Ο άλλος διαχωρισμός δημιουργεί μια κοινότητα από τους κόμβους 9,8,7,6,5(σύνολο 5) και μια άλλη κοινότητα από τους κόμβους 4,3,2,1(σύνολο 4). Ανάμεσα στους κόμβους 5 και 4 δημιουργείται μια διαφορά τάσης ίση με 0.267. Προφανώς η δεύτερη διαφορά είναι η μεγαλύτερη, την οποία και διαλέγουμε. Το κατώφλι θα βρίσκεται στο μέσο αυτής της διαφοράς. Προσοχή ότι δεν εννοείται το μέσο της τιμής της διαφοράς αλλά η τιμή που υπάρχει στο μέσο μεταξύ της διαφοράς τάσης που δημιουργούν οι δύο κόμβοι(4 και 5). Με λίγα λόγια αυτή θα είναι ίση με 0,400+(0,267/2)=0,400+0,133-0,533. Οι κόμβοι με τιμές τάσεις μεγαλύτερες από αυτό το κατώφλι δημιουργούν τη μια κοινότητα και οι κόμβοι με τιμές τάσεις μικρότερες από το κατώφλι δημιουργούν την δεύτερη κοινότητα. Έτσι η μια κοινότητα θα αποτελείται από τους κόμβους 1,2,3,4 και η άλλη από τους κόμβους 5,6,7,8,9. Στο Σχήμα 3.15 παρουσιάζονται σχηματικά οι κοινότητες.



Σχήμα 3.15: Οι κοινότητες που εντοπίστηκαν χρησιμοποιώντας το δίκτυο σαν ηλεκτρικό κύκλωμα.

3.4 Βιβλιογραφία

[1] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, Self-organization and identification of web communities, IEEE Comput 35 (2002), 66–71.

[2] U. N. Raghavan, R. Albert, and S. Kumara, Near linear time algorithm to detect community structures in large scale networks, Phys Rev E 76 (2007), 036106.

[3] Gennaro Cordasco and Luisa Gargano Community Detection via Semi-Synchronous Label Propagation Algorithms Dipartimento di Informatica ed Applicazioni "R.M. Capocelli"University of Salerno, Fisciano 84084, ITALY.

[4] I. X. Y. Leung, P. Hui, P. Lio, and J. Crowcroft. Towards real- time community detection in large networks. Phys. Rev. E, 79(6):1–10, Jun 2009.

[5] S. Gregory, Finding overlapping communities in networks by label propagation, New Journal of Physics 12(10) (2009), 103018.

[6] F. Wu and B. A. Huberman, Finding communities in linear time: a physics approach, Eur Phys J B 38(2) (2004), 331–338.

[7] W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research, 33, 452-473 (1977).

[8] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 8271-8276 (2002).

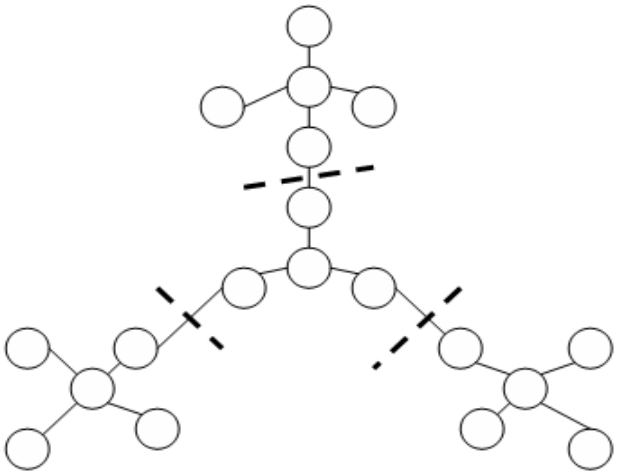
[9] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

Κεφάλαιο 4

Εγγύτητα (Closeness)

4.1 Εισαγωγή

Μια ευκολονόητη αίσθηση της κοινότητας σε ένα πολύπλοκο δίκτυο βασίζεται στην ιδέα του πόσο κοντά είναι τα μέλη της συνδεδεμένα μεταξύ τους. Μια κοινότητα είναι ένα σύνολο από άτομα που μπορούνε να επικοινωνούνε μεταξύ τους εύκολα επειδή μπορούνε να φτάσουν οποιοδήποτε άλλο μέλος, με ένα σχετικά μικρότερο αριθμό «αναπηδήσεων» από τον μέσο όρο του δικτύου. Το Σχήμα 4.1 δείχνει ένα απλό παράδειγμα αυτής της διαμόρφωσης .



Σχήμα 4.1: Παράδειγμα ενός γραφήματος που μπορεί να διαχωριστεί λαμβάνοντας υπόψιν την σχετική απόσταση , σε σχέση με τον αριθμό των ακμών, μεταξύ των κόμβων του [12].

Ο ορισμός σε αυτή την περίπτωση είναι:

Ορισμός 3 (Small World Community) Μια κοινότητα «μικρού κόσμου» (Small World Community) σε ένα πολύπλοκο δίκτυο, είναι ένα σύνολο κόμβων που μπορούν να φτάσουν σε οποιοδήποτε μέλος της ομάδας τους, διαπερνώντας συνήθως ένα πολύ

μικρό αριθμό ακμών, σημαντικά χαμηλότερο από το, κατά μέσο όρο, συντομότερο μονοπάτι στο δίκτυο.

Χρησιμοποιούμε τον όρο «μικρός κόσμος»[1] δεδομένου ότι εκφράζει την ιδέα των πολύ κοντά συνδεδεμένων κόμβων. Μια πολύ αποτελεσματική προσέγγιση που χρησιμοποιείται με αυτό τον ορισμό βασίζεται στους τυχαίους περιπάτους(**random walks**). Ένας τυχαίος περίπατος είναι μια διαδικασία στην οποία σε κάθε χρονικό βήμα ένας περιπατητής(**walker**) βρίσκεται σε μια κορυφή και κινείται προς μια κορυφή που επιλέγεται τυχαία και ομοιόμορφα από τους γείτονές του.

Διαδικασία 3 Λαμβάνοντας υπόψιν ένα δίκτυο, εκτέλεσε διάφορους τυχαίους περιπάτους και στη συνέχεια συγκέντρωσε μαζί τους κόμβους που εμφανίζονται συχνά στον ίδιο περίπατο.

4.2 Αλγόριθμος Walktrap

Η προσέγγιση Walktrap βασίζεται στη ακόλουθη διαίσθηση: οι τυχαίοι περίπατοι είναι ικανοί να αποκαλύψουν την αληθινή απόσταση μεταξύ κόμβων εξερευνώντας συχνά κόμβους στην ίδια κοινότητα. Το πρόβλημα κλειδί είναι ο ορισμός της συνάρτησης απόστασης μεταξύ δύο οποιονδήποτε κορυφών, που υπολογίζεται από την πληροφορία που δίνεται από τους τυχαίους περιπάτους στο γράφημα. Υψηλές τιμές αυτού του μέτρου σημαίνει ότι οι δύο κορυφές i και j «βλέπουν» το δίκτυο με παρόμοιο τρόπο, που σημαίνει ότι ανήκουν στην ίδια κοινότητα.

Για να ομαδοποιήσουν τις κορυφές σε κοινότητες, οι συγγραφείς του [2] παρουσιάζουν την απόσταση r μεταξύ κορυφών που συγκροτούν την κοινοτική δομή του γραφήματος. Η απόσταση θα πρέπει να είναι μεγάλη εάν οι δύο κορυφές είναι σε διαφορετικές κοινότητες και μικρή εάν βρίσκονται στην ίδια κοινότητα. Όπως προαναφέραμε αυτή θα υπολογιστεί από την πληροφορία που δίνεται από τους τυχαίους περιπάτους στο γράφημα.

Το γράφημα G συσχετίζεται με τον πίνακα γειτνίασης του, τον A έτσι ώστε $A_{ij}=1$ εάν i και j είναι συνδεδεμένοι και $A_{ij}=0$ διαφορετικά. Ο βαθμός $d(i)=\sum_j A_{ij}$ της κορυφής i είναι ο αριθμός των γειτόνων της (συμπεριλαμβανομένης και της ίδιας). Οι αναλυτές του [2] θεωρούνε μόνο γραφήματα χωρίς βάρη.

Στη συνέχεια θεωρούνε μια διακριτή διαδικασία τυχαίου περιπάτου(ή διαδικασία διάχυσης) στο γράφημα G (μια ολοκληρωμένη περιγραφή του θέματος παρουσιάζεται στο [3],[4]). Σε κάθε βήμα ένας περιπατητής είναι σε μια κορυφή και μετακινείται σε μια κορυφή που επιλέγεται τυχαία και ομοιόμορφα ανάμεσα στους γείτονες του. Η ακολουθία των επισκεπτόμενων κορυφών είναι μια αλυσίδα Markov

οι καταστάσεις της οποίας είναι οι κορυφές του γραφήματος. Σε κάθε βήμα, η πιθανότητα μετάβασης από την κορυφή i στην κορυφή j είναι $P_{ij} = \frac{A_{ij}}{d(i)}$. Αυτή ορίζει τον πίνακα μετάβασης P στη διαδικασία που εκτελείται με τυχαίους περιπάτους. Κανείς μπορεί να γράψει ότι $P = D^{-1}A$ όπου D είναι ο διαγώνιος πίνακας των βαθμών ($\forall i, D_{ii} = d(i)$ και $D_{ij} = 0$ για $i \neq j$). Η πιθανότητα για να πάμε από την κορυφή i στη j μέσα από ένα τυχαίο περίπατο μήκους t είναι P_{ij}^{t}. Ικανοποιεί δύο πολύ γνωστές ιδιότητες της διαδικασίας του τυχαίου περιπάτου που χρησιμοποιούν οι συγγραφείς:

Ιδιότητα 1 Όταν το μήκος t ενός τυχαίου περιπάτου που αρχίζει από την κορυφή i τείνει στο άπειρο, η πιθανότητα του να βρίσκεσαι στη κορυφή j εξαρτάται μόνο από τον βαθμό της κορυφής j (και όχι από την αρχική κορυφή i):

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

Ιδιότητα 2 Οι πιθανότητες για να πας από το i στο j και από το j στο i μέσα από ένα τυχαίο περίπατο σταθερού μήκους t έχουνε μια αναλογία που εξαρτάται μόνο από τους βαθμούς $d(i)$ και $d(j)$:

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

Στο [2] οι αναλυτές αποδεικνύουν τις παραπάνω ιδιότητες.

4.2.1 Μια απόσταση r για τον προσδιορισμό των ομοιοτήτων των κορυφών

Οι αναλυτές υποθέτουν τυχαίους περιπάτους σε ένα γράφημα G με ένα δοσμένο μήκος t. Χρησιμοποιούν την πληροφορία που δίνεται από όλες τις πιθανότητες P_{ij}^{t} για να πάνε από το i στο j σε t βήματα. Το μήκος t των τυχαίων περιπάτων πρέπει να είναι αρκετά μακρύ για να συλλέξουμε αρκετή πληροφορία για την τοπολογία του γραφήματος. Ωστόσο, το t δεν πρέπει να είναι πολύ μακρύ, για να αποφύγουμε το αποτέλεσμα που προβλέφθηκε στην ιδιότητα 1 όπου οι πιθανότητες θα εξαρτιόντουσαν μόνο από των βαθμών των κορυφών. Κάθε πιθανότητα P_{ij}^{t} δίνει μια πληροφορία για τις δυο κορυφές i και j, αλλά η ιδιότητα 2 μας λέει ότι η P_{ij}^{t} και η P_{ji}^{t} κωδικοποιούν ακριβώς την ίδια πληροφορία.

Τελικώς η πληροφορία για την κορυφή i που κωδικοποιείται στον P^t βρίσκεται στις n πιθανότητες $\left(P_{ik}^{\ t}\right)_{1\le k\le n}$, που δεν είναι τίποτα άλλο από την $i^{ο\sigma\tau\acute{η}}$ γραμμή του πίνακα $P^{\ t}$ που υποδηλώνεται ως $P_{i.}^{\ t}$. Για να συγκρίνουμε δύο κορυφές i και j χρησιμοποιώντας αυτά τα δεδομένα οι αναλυτές παρατηρούν ότι:

- Εάν δυο κορυφές i και j είναι στην ίδια κοινότητα, η πιθανότητα $P_{ij}^{\ t}$ θα είναι σίγουρα υψηλή. Αλλά το γεγονός ότι η $P_{ij}^{\ t}$ είναι υψηλή δεν υπονοεί απαραίτητα ότι i και j είναι στην ίδια κοινότητα.
- Η πιθανότητα $P_{ij}^{\ t}$ επηρεάζεται από τον βαθμό $d(j)$ γιατί ο περιπατητής έχει υψηλότερη πιθανότητα να πάει σε υψηλού βαθμού κορυφές.
- Δύο κορυφές της ίδιας κοινότητας τείνουν να «βλέπουν» όλες τις άλλες κορυφές με τον ίδιο τρόπο. Έτσι εάν i και j είναι στην ίδια κοινότητα, πιθανότατα θα έχουμε $\forall k, P_{ik}^{t} \simeq P_{jk}^{t}$.

Μετά από όσα αναφέρθηκαν δίνεται ο ορισμός για την απόσταση δύο κορυφών, λαμβάνοντας υπόψιν τα προηγούμενα:

Ορισμός 1 Έστω i και j δύο κορυφές στο γράφημα και

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

όπου $\|.\|$ είναι η ευκλείδεια νόρμα του $\mathbb{R}^n$.

Στη συνέχεια γενικοποιείται η απόσταση δύο κορυφών σε απόσταση μεταξύ κοινοτήτων. Οι συγγραφείς θεωρούν τυχαίους περιπάτους που αρχίζουν από μια κοινότητα. Η αρχική κορυφή επιλέγεται τυχαία και ομοιόμορφα ανάμεσα στις κορυφές της κοινότητας. Ορίζεται η πιθανότητα P_{Cj}^{t} για να πάμε από την κοινότητα C στην κορυφή j σε t βήματα:

$$P_{Cj}^{t} = \frac{1}{|C|} \sum_{i \in C} P_{ij}^{t}$$

Αυτό ορίζει ένα διάνυσμα πιθανότητας $P_{C.}^{t}$ που επιτρέπει στους αναλυτές να γενικοποιήσουν την απόσταση:

Ορισμός 2 Έστω $C_1, C_2 \subset V$ δύο κοινότητες. Ορίζουμε την απόσταση $r_{C_1C_2}$ μεταξύ αυτών των δύο κοινοτήτων :

$$r_{C_1C_2} = \left\| D^{-\frac{1}{2}} P^t_{C_1\cdot} - D^{-\frac{1}{2}} P^t_{C_2\cdot} \right\| = \sqrt{\sum_{k=1}^{n} \frac{\left(P^t_{C_1k} - P^t_{C_2k}\right)^2}{d(k)}}$$

Αυτός ο ορισμός είναι σύμφωνος με τον προηγούμενο ορισμό : $r_{ij} = r_{\{i\}\{j\}}$ και ορίζεται η απόσταση μεταξύ μιας κορυφής i και μιας κοινότητας C : $r_{iC} = r_{\{i\}C}$.

Έχοντας προτείνει την απόσταση μεταξύ κορυφών για να εντοπίσουν δομικές ομοιότητες μεταξύ τους οι συγγραφείς ισχυρίζονται τώρα ότι το πρόβλημα του εντοπισμού κοινοτήτων είναι ένα πρόβλημα συσταδοποίησης. Χρησιμοποιούνε ένα αποτελεσματικό ιεραρχικό αλγόριθμο ομαδοποίησης που μας επιτρέπει να βρούμε κοινότητες σε διαφορετικές κλίμακες. Χρησιμοποιούνε μια μέθοδο που βασίζεται στην μέθοδο του Ward (Ward's method [5]) που ταιριάζει με την απόσταση που μελετήσανε και δίνει πολύ καλά αποτελέσματα μειώνοντας τον αριθμό των υπολογισμών για την απόσταση.

Ξεκινάνε με ένα διαχωρισμό $P_1 = \{\{u\}, u \in V\}$ του γραφήματος σε n κοινότητες που είναι μειωμένο σε μια μοναδική κορυφή. Αρχικά υπολογίζονται οι αποστάσεις ανάμεσα στις παρακείμενες κορυφές. Έπειτα αυτό το τμήμα εξελίσσεται επαναλαμβάνοντας τις ακόλουθες λειτουργίες.

Αλγόριθμος 4.1 Walktrap Algorithm

Σε κάθε βήμα k :

- διάλεξε δύο κοινότητες C_1 και C_2 στο P_k σύμφωνα με ένα κριτήριο που βασίζεται στην απόσταση μεταξύ των κοινοτήτων,
- συγχώνευσε αυτές τις δύο κοινότητες σε μία καινούργια κοινότητα $C_3 = C_1 \cup C_2$ και δημιούργησε το νέο τμήμα $P_{k+1} = \left(P_k \setminus \{C_1, C_2\}\right) \cup \{C_3\}$ και
- ενημέρωσε τις αποστάσεις μεταξύ των κοινοτήτων.

Μετά από $n-1$ βήματα, ο αλγόριθμος τελειώνει και έχουμε $P_n = \{V\}$.

Κάθε βήμα ορίζει ένα διαχωρισμό P_k, το οποίο δίνει μια ιεραρχική δομή των κοινοτήτων που καλείται δενδρόγραμμα. Αυτή η δομή είναι ένα δέντρο του οποίου τα φύλλα αντιστοιχούν στις κορυφές και κάθε εσωτερικός κόμβος συσχετίζεται με μια συγχώνευση των κοινοτήτων στον αλγόριθμο.

Τα σημαντικά σημεία σε αυτό τον αλγόριθμο είναι ο τρόπος που διαλέγουμε τις κοινότητες για συγχώνευση και το γεγονός ότι οι αποστάσεις θα πρέπει να ενημερώνονται αποτελεσματικά. Επιπλέον θα πρέπει να αξιολογηθεί η ποιότητα του διαχωρισμού έτσι ώστε να γίνει η επιλογή ενός από τα P_k σαν αποτέλεσμα του αλγορίθμου.

4.2.2 Επιλογή των κοινοτήτων για συγχώνευση

Αυτή η επιλογή παίζει σημαντικό ρόλο για την ποιότητα της δομής της κοινότητας που επιτυγχάνεται. Οι συγγραφείς, για ελαττώσουν την πολυπλοκότητα, συγχωνεύουν μόνο προσκείμενες κοινότητες (κοινότητες με τουλάχιστον μια ακμή ανάμεσά τους). Αυτό η λογική ευρηματική μέθοδος (που ήδη έχει χρησιμοποιηθεί στο [6] και [7]) μειώνει σε m τον αριθμό των πιθανών συγχωνεύσεων σε κάθε στάδιο. Επιπλέον εξασφαλίζει ότι κάθε κοινότητα είναι συνδεδεμένη.

Στην αρχή οι αναλυτές διαλέγουν δύο κοινότητες για συγχώνευση σύμφωνα με την μέθοδο του Ward. Σε κάθε βήμα k συγχωνεύουν δύο κοινότητες που ελαχιστοποιούν το μέσο όρο ,σ_k, των τετραγώνων των αποστάσεων μεταξύ κάθε κορυφής και της κοινότητάς της.

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

Όμως, όπως αναφέρουν αυτή η προσέγγιση αποτελεί έναν άπληστο αλγόριθμο (greedy algorithm) και οι προσεγγιστικοί αλγόριθμοι [8, 9] είναι εκθετικοί με τον αριθμό των ομάδων που βρίσκουν και έτσι δεν ταιριάζουν με την δικιά τους προσέγγιση. Έτσι για κάθε ζευγάρι προσκείμενων κοινοτήτων $\{C_1, C_2\}$, υπολογίζουν την διακύμανση $\Delta_\sigma(C_1, C_2)$ του σ που θα είχε προκληθεί αν συγχωνευόταν οι C_1 και C_2 σε μια καινούργια κοινότητα $C_3 = C_1 \cup C_2$. Αυτή η ποσότητα εξαρτάται μόνο από τις κορυφές των C_1 και C_2 και όχι από άλλες κοινότητες ή το βήμα k του αλγορίθμου.

$$\Delta_\sigma(C_1, C_2) = \frac{1}{n}\left(\sum_{i\in C_3} r_{iC_3}^2 - \sum_{i\in C_1} r_{iC_1}^2 - \sum_{i\in C_2} r_{iC_2}^2\right) \qquad (2)$$

Τέλος συγχωνεύουν τις δύο κοινότητες που δίνουν την χαμηλότερη τιμή του Δ_σ.

4.2.3 Υπολογισμός του Δ_σ και η ενημέρωση των αποστάσεων

Το σημαντικό σημείο με αυτές τις ποσότητες είναι ότι μπορούν να υπολογιστούν αποτελεσματικά χάρη στο γεγονός ότι η απόσταση που μελετάνε οι συγγραφείς του [2] είναι μια ευκλείδεια απόσταση και έτσι είναι δυνατό να έχουμε τα εξής αποτελέσματα[10]:

Θεώρημα 1 Η αύξηση του σ μετά την συγχώνευση δύο κοινοτήτων C_1 και C_2 έχει άμεση σχέση με την απόσταση $r_{C_1C_2}$ μέσω της εξίσωσης:

$$\Delta_\sigma(C_1,C_2)=\frac{1}{n}\frac{|C_1||C_2|}{|C_1|+|C_2|}r_{C_1C_2}^2$$

Αυτό το θεώρημα δείχνει ότι χρειάζεται μόνο να ενημερωθούν οι αποστάσεις μεταξύ των κοινοτήτων για να παρθούν οι τιμές του Δ_σ. Εάν είναι γνωστά τα δύο διανύσματα $P_{C_1\bullet}$ και $P_{C_2\bullet}$ ο χρονικός υπολογισμός του $\Delta_\sigma(C_1,C_2)$ είναι $O(n)$. Το επόμενο θεώρημα που παρουσιάζεται δείχνει ότι εάν ήδη γνωρίζουμε τις τρεις τιμές $\Delta_\sigma(C_1,C_2)$, $\Delta_\sigma(C_1,C)$ και $\Delta_\sigma(C_2,C)$, μπορούμε να υπολογίσουμε το $\Delta_\sigma(C_1\cup C_2,C)$ σε ένα σταθερό χρόνο.

Θεώρημα 2 Εάν C_1 και C_2 συγχωνευτούν στο $C_3=C_1\cup C_2$ τότε για κάθε άλλη κοινότητα C:

$$\Delta_\sigma(C_3,C)=\frac{(|C_1|+|C|)\Delta_\sigma(C_1,C)+(|C_2|+|C|)\Delta_\sigma(C_2,C)-|C|\Delta_\sigma(C_1,C_2)}{|C_1|+|C_2|+|C|}.$$

Στο [2] οι συγγραφείς αποδεικνύουν τα παραπάνω θεωρήματα. Εφόσον συγχωνεύονται γειτονικές κοινότητες, χρειάζεται να ενημερωθούν οι τιμές του Δ_σ μεταξύ των γειτονικών κοινοτήτων.

4.2.4 Αξιολογώντας την ποιότητα των διαχωρισμών

Ο αλγόριθμος επιφέρει την δημιουργία μιας ακολουθίας $(P_k)_{1\le k\le n}$ διαχωρισμών σε κοινότητες. Οι αναλυτές αναφέρουν πως θέλουν να γνωρίζουν ποιοι διαχωρισμοί σε αυτή την ακολουθία λαμβάνουν καλά την κοινοτική δομή. Ένα πολύ διαδεδομένο κριτήριο είναι το modularity [6, 11].

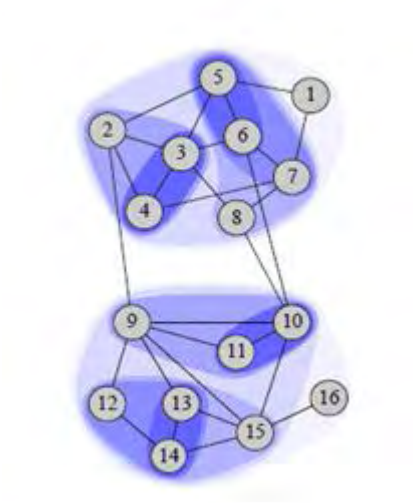
Ωστόσο, κάποιος μπορεί να σκεφτεί άλλο κριτήριο της ποιότητας ενός διαχωρισμού σε κοινότητες. Για παράδειγμα, όπως αναφέρεται από τους αναλυτές, το modularity, δεν ταιριάζει κατάλληλα για την εύρεση κοινοτήτων σε διαφορετικές κλίμακες. Το κριτήριο που παρέχεται από τους συγγραφείς για την εύρεση τέτοιων δομών έχει ως εξής. Όταν συγχωνεύονται δύο διαφορετικές κοινότητες (αναφορικά με την απόσταση r), η τιμή $\Delta_{\sigma_k}=\sigma_{k+1}-\sigma_k$ σε αυτό το βήμα είναι μεγάλη. Αντίστροφα, εάν η Δ_{σ_k} είναι μεγάλη τότε οι κοινότητες στο βήμα $k-1$ είναι σίγουρα σχετικές. Για να εντοπιστεί αυτό εισάγεται η αναλογία n_k:

$$n_k = \frac{\Delta\sigma_k}{\Delta\sigma_{k-1}} = \frac{\sigma_{k+1} - \sigma_k}{\sigma_k - \sigma_{k-1}}$$

Κάποιος μπορεί να σκεφτεί ότι οι σχετικοί διαχωρισμοί P_k συσχετίζονται με τις μεγαλύτερες τιμές του n_k. Ανάλογα με τον πλαίσιο με τον οποίο ο αλγόριθμος χρησιμοποιείται, κάποιος μπορεί να πάρει τον καλύτερο διαχωρισμό (αυτόν που το n_k είναι μέγιστο) ή να διαλέξει ανάμεσα στους καλύτερους χρησιμοποιώντας άλλο κριτήριο (το μέγεθος των κοινοτήτων για παράδειγμα).

4.2.5 Παράδειγμα

Στην συνέχεια οι αναλυτές του [2] παρουσιάζουν ένα παράδειγμα ενός γραφήματος στο οποίο έχει εφαρμοστεί ο αλγόριθμος και έχει αποκαλυφθεί η κοινοτική δομή (Σχήμα 4.2).



Σχήμα 4.2: Παράδειγμα κοινοτικής δομής μετά την εφαρμογή του αλγορίθμου Walktrap [2].

Ξεκινάμε από ένα διαχωρισμό, P_1={{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}, {11}, {12}, {13}, {14}, {15}, {16}}, του γραφήματος σε n κοινότητες όπου κάθε κοινότητα αποτελείται από μια μοναδική κορυφή. Αρχικά υπολογίζεται η απόσταση όλων των γειτονικών κορυφών. Εν συνεχεία:

- k=1

Θα πρέπει να διαλέξουμε δύο κοινότητες για συγχώνευση. Για κάθε ζευγάρι γειτονικών κοινοτήτων (κορυφών στο βήμα που βρισκόμαστε) υπολογίζεται η μεταβολή Δ_σ μέσω της εξίσωσης του θεωρήματος 1. Διαλέγουμε να συγχωνεύσουμε τις κοινότητες που μας δίνουν την χαμηλότερη τιμή του Δ_σ. Για το παράδειγμα, αυτές θα είναι το ζευγάρι των κορυφών {3,4}, δημιουργώντας έτσι ένα καινούργιο διαχωρισμό P_2={{1}, {2}, {5}, {6}, {7}, {8}, {9}, {10}, {11}, {12}, {13}, {14}, {15}, {16}, **{3,4}**}. Στο επόμενο βήμα θα πρέπει να ενημερώσουμε τις αποστάσεις μεταξύ των κοινοτήτων. Εφόσον συγχωνεύουμε γειτονικές κοινότητες, χρειάζεται να ενημερώσουμε τις τιμές του Δ_σ μεταξύ των γειτονικών κορυφών της συγχώνευσης που επιτεύχθηκε πριν μέσω της εξίσωσης του θεωρήματος 2. Τα υπόλοιπα θα παραμείνουν ως έχουν.

- k=2

Θα διαλέξουμε εκ νέου το ζευγάρι των κοινοτήτων με το χαμηλότερο Δ_σ και θα το συγχωνεύσουμε δημιουργώντας ένα νέο διαχωρισμό. Λαμβάνοντας υπόψιν και τα αποτελέσματα του προηγούμενου βήματος, οδηγούμαστε στον διαχωρισμό P_3={{1}, {2}, {5}, {6}, {7}, {8}, {9}, {10}, {11}, {12}, {15}, {16}, **{3,4}**, **{13,14}**}. Οι αποστάσεις καθώς και οι τιμές Δ_σ των γειτονικών κοινοτήτων ενημερώνονται και πάλι.

- k=3

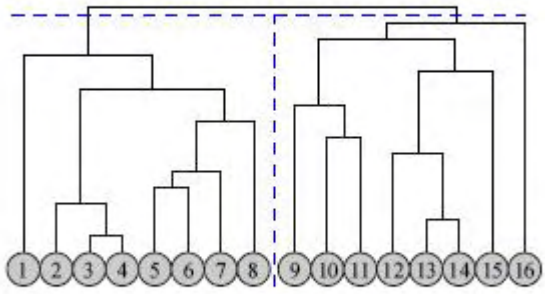
Οι κοινότητες με το χαμηλότερο Δ_σ μας δίνουν τον διαχωρισμό P_4={{1}, {5}, {6}, {7}, {8}, {9}, {10}, {11}, {12}, {15}, {16}, **{2,3,4}**, **{13,14}**}. Οι αποστάσεις καθώς και οι τιμές Δ_σ των γειτονικών κοινοτήτων ενημερώνονται εκ νέου.

Συνεχίζοντας με την ίδια συλλογιστική την διαδικασία για τα υπόλοιπα βήματα, θα έχουμε τα εξής αποτελέσματα:

- k=4: P_5={{1}, {7}, {8}, {9}, {10}, {11}, {12}, {15}, {16}, **{5,6}**, **{2,3,4}**, **{13,14}**}
- k=5: P_6={{1}, {8}, {9}, {10}, {11}, {12}, {15}, {16}, **{5,6,7}**, **{2,3,4}**, **{13,14}**}
- k=6: P_7={{1}, {8}, {9}, {10}, {11}, {15}, {16}, **{5,6,7}**, **{2,3,4}**, **{12,13,14}**}
- k=7: P_8={{1}, {8}, {9}, {15}, {16}, **{10,11}**, **{5,6,7}**, **{2,3,4}**, **{12,13,14}**}
- k=8: P_9={{1}, {9}, {15}, {16}, **{10,11}**, **{5,6,7,8}**, **{2,3,4}**, **{12,13,14}**}
- k=9: P_{10}={{1}, {15}, {16}, **{9,10,11}**, **{5,6,7,8}**, **{2,3,4}**, **{12,13,14}**}
- k=10: P_{11}={{1}, {15}, {16}, **{9,10,11}**, **{2,3,4,5,6,7,8}**, **{12,13,14}**}
- k=11: P_{12}={{1}, {16}, **{9,10,11}**, **{2,3,4,5,6,7,8}**, **{12,13,14,15}**}
- k=12: P_{13}={{16}, **{9,10,11}**, **{1,2,3,4,5,6,7,8}**, **{12,13,14,15}**}
- k=13: P_{14}={{16}, **{1,2,3,4,5,6,7,8}**, **{9,10,11,12,13,14,15}**}
- k=14: P_{15}={**{1,2,3,4,5,6,7,8}**, **{9,10,11,12,13,14,15,16}**}

- k=15: P_{16} ={**{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16}**}

Μετά από n-1=15 βήματα οδηγούμαστε σε ένα διαχωρισμό που ισοδυναμεί με την μεγαλύτερη δυνατή κοινότητα, όλο το γράφημα. Ο αλγόριθμος προξενεί μια ακολουθία διαχωρισμών $(P_k)_{1\le k\le 16}$ σε κοινότητες. Όπως είδαμε και στη παράγραφο 4.2.4, θέλουμε να γνωρίζουμε ποιοι διαχωρισμοί σε αυτή την ακολουθία λαμβάνουν καλύτερα την κοινοτική δομή. Αναφέρθηκαν δύο τρόποι: λαμβάνοντας υπόψιν το κριτήριο modularity ή το κριτήριο με την αναλογία n_k. Χρησιμοποιώντας το δεύτερο κριτήριο ο καλύτερος διαχωρισμός αποτελείται από δύο κοινότητες. Τα βήματα του αλγορίθμου και ο καλύτερος διαχωρισμός παρουσιάζονται στο δενδρόγραμμα του σχήματος 4.3.



Σχήμα 4.3: το δενδρόγραμμα που απεικονίζει την ιεραρχική δομή των κοινοτήτων και με μπλε διακεκομμένες γραμμές εμφανίζεται ο καλύτερος διαχωρισμός [2].

4.2.6 Πολυπλοκότητα

Οι συγγραφείς εξετάζουν την πολυπλοκότητα του αλγορίθμου και αποδεικνύουν ([2]) ότι η χρονική πολυπλοκότητα είναι $O(mn(H+t))$, όπου $h(C)$ ορίζεται το ύψος μιας κοινότητας και $H=h(V)$ το ύψος ολόκληρου του δέντρου. Πρακτικά, επειδή διαλέγεται μικρό t ($t=O(\log n)$) η πολυπλοκότητα καταλήγει να είναι ίση με $O(mnH)$. Τέλος αναφέρουν ότι σχεδίασαν με τέτοιο τρόπο τον αλγόριθμο ώστε να έχει πολυπλοκότητα χειρότερης περιπτώσεως ίση με $O(mn^2)$ και στην περίπτωση των σύνθετων δικτύων του πραγματικού κόσμου όπου τα δίκτυα είναι αραιά ($m=O(n)$) και τα ύψη των δενδρογραμμάτων είναι μικρά ($H=O(\log n)$) ο αλγόριθμος έχει χρονική πολυπλοκότητα ίση με $O(n^2\log n)$.

4.3 Βιβλιογραφία

[1] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393(6684) (1998), 440–442.

[2] P. Pons, and M. Latapy, Computing communities in large networks using random walks, J Graph Algor Appl 3733 (2006), 284–293.

[3] L. Lovasz. Random walks on graphs: a survey. In Combinatorics, Paul Erdos is eighty, Vol. 2 (Keszthely, 1993), volume 2 of Bolyai Soc. Math. Stud., pages 353-397. Janos Bolyai Math. Soc., Budapest, 1996.

[4] D. Aldous and J. A. Fill. Reversible Markov Chains and Random Walks on Graphs, chapter 2. Forthcoming book, http://www.stat.berkeley.edu/users/aldous/RWG/book.html.

[5] J. H. Ward. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301):236{244, 1963.

[6] M. E. J. Newman. Fast algorithm for detecting community structure in networks. Physical Review E, 69(6):066133, 2004.

[7] L. Donetti and M. A. Munoz. Detecting network communities: a new systematic and efficient algorithm. Journal of Statistical Mechanics, 2004(10):10012, 2004.

[8] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. Machine Learning, 56(1-3):9{33, 2004.

[9] W. Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In Proceedings of the thirty-fifth annual ACM Symposium on Theory of computing, STOC, pages 50{58. ACM Press, 2003.

[10] M. Jambu and Lebeaux M.-O. Cluster analysis and data analysis. North Holland Publishing, 1983.

[11] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69(2):026113, 2004.

[12] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

Κεφάλαιο 5

Δομή (Structure Definition)

5.1 Εισαγωγή

Ένας αριθμός εργασιών αντιμετωπίζουν την εύρεση κοινοτήτων με μια πολύ ισχυρή υπόθεση: για να ονομάζεται κοινότητα, μια ομάδα κορυφών πρέπει να ακολουθεί μια πολύ αυστηρή δομική ιδιότητα. Με άλλα λόγια ακολουθούν τον ακόλουθο ορισμό της κοινότητας:

Ορισμός 4 (Structure Community) Μια κοινοτική δομή σε ένα πολύπλοκο δίκτυο είναι ένα σύνολο κόμβων με έναν ακριβή αριθμό ακμών ανάμεσά τους, τα οποία διαμοιράζονται σε μια πολύ ακριβή τοπολογία που καθορίζεται από μια σειρά κανόνων. Σύνολα από κόμβους που δεν ικανοποιούν αυτούς τους δομικούς κανόνες δεν είναι κοινότητες.

Ο στόχος του αλγορίθμου εύρεσης κοινοτήτων είναι να βρει όλες τις μέγιστες δομές στο δίκτυο που ικανοποιούν τους επιθυμητούς περιορισμούς. Η αντίστοιχη διαδικασία που εφαρμόζεται σε αυτή την κατηγορία είναι απλή: βρες με ένα αποτελεσματικό τρόπο όλες τις μέγιστες δομές που ορίζονται.

Αυτή η εργασία είναι παρόμοια με ένα πολύ γνωστό πρόβλημα εξόρυξης δεδομένων στην ανάλυση δικτύων που είναι η εξόρυξη γράφου. Κάποια παραδείγματα εξόρυξης γράφων είναι [1,2,3,4]. Ωστόσο οι κλασικοί αλγόριθμοι εξόρυξης γράφων επιστρέφουν μόνο όλα τα διαφορετικά δομικά σχέδια. Στον εντοπισμό κοινοτήτων υπάρχει μόνο μια σημαντική δομή και το επιθυμητό αποτέλεσμα είναι η λίστα όλων των συνόλων των κορυφών που απαρτίζουν την συγκεκριμένη δομή στο δίκτυο.

Θα αγνοήσουμε τους αλγορίθμους που ασχολούνται καθαρά με την εξόρυξη δεδομένων και θα επικεντρωθούμε μόνο στις προσεγγίσεις που αφορούν την εύρεση κοινοτικών δομών. Οι μέθοδοι που εξετάζεται εδώ είναι η clique percolation και η biclique.

5.2 Μέθοδος Clique percolation

Όπως αναφέρουν οι συγγραφείς του [5], τα περισσότερα πραγματικά δίκτυα περιέχουν μέρη στα οποία οι κόμβοι (μονάδες) είναι περισσότερο συνδεδεμένοι μεταξύ τους σε σχέση με το υπόλοιπο δίκτυο. Τα σύνολα αυτών των κόμβων ονομάζονται συχνά συστάδες, κοινότητες, συνεκτικές ομάδες, ή μονάδες και δεν έχουν έναν ευρέως αποδεκτό και μοναδικό ορισμό. Σε αντίθεση με αυτή την ασάφεια, η παρουσία κοινοτήτων στα δίκτυα είναι ένα χαρακτηριστικό της ιεραρχικής φύσης των πολύπλοκων συστημάτων. Οι υπάρχοντες μέθοδοι για την εύρεση κοινοτήτων σε μεγάλα δίκτυα είναι χρήσιμοι εάν η κοινοτική δομή είναι τέτοια, έτσι ώστε να μπορεί να ερμηνευτεί ως προς σύνολα κοινοτήτων τα οποία είναι διαχωρισμένα. Ωστόσο, τα περισσότερα πραγματικά δίκτυα χαρακτηρίζονται από καλά ορισμένες επικαλυπτόμενες και φωλιασμένες κοινότητες. Μπορούμε να δείξουμε αυτή την κατάσταση μέσα από τις πολυάριθμες κοινότητες όπου ο καθένας από εμάς ανήκει, συμπεριλαμβάνοντας αυτές που έχουν να κάνουν με τις επιστημονικές μας δραστηριότητες ή την προσωπική ζωή (σχολείο, οικογένεια , χόμπι). Επιπλέον, μέλη των κοινοτήτων μας έχουνε τις δικιές τους κοινότητες, δίνοντας ως αποτέλεσμα ένα εξαιρετικά περίπλοκο δίκτυο από κοινότητες.

Γενικά, κάθε κόμβος i ενός δικτύου μπορεί να χαρακτηριστεί από ένα αριθμό μέλους m_i, που είναι ο αριθμός των κοινοτήτων στις οποίες ανήκει ο κόμβος. Εκ περιτροπής, δύο οποιεσδήποτε κοινότητες α και β μπορούνε να μοιράζονται $s_{\alpha,\beta}^{ov}$ κόμβους, το οποίο ορίζεται ως το επικαλυπτόμενο μέγεθος μεταξύ αυτών των κοινοτήτων. Επιπλέον, οι κοινότητες απαρτίζουν ένα δίκτυο με τις επικαλύψεις να αποτελούν τις συνδέσεις τους. Ο αριθμός αυτών των συνδέσεων της κοινότητας α μπορεί να καλείται ως ο βαθμός αυτής της κοινότητας, d_{α}^{com}. Τέλος το μέγεθος s_{α}^{com} μιας οποιασδήποτε κοινότητας α μπορεί να οριστεί ως ο αριθμός των κόμβων της.

Η βασική παρατήρηση στην οποία βασίζεται ο ορισμός της κοινότητας είναι ότι μια τυπική κοινότητα αποτελείται από αρκετά πλήρη (πλήρως συνδεδεμένα) υπογραφήματα που τείνουν να μοιράζονται πολλούς από τους κόμβους τους. Έτσι, ορίζεται μια κοινότητα, ή με πιο ακρίβεια, μια k-clique κοινότητα ως η ένωση όλων των k-clique (πλήρη υπογραφήματα μεγέθους k) που μπορούνε να φτάσουν η μία την άλλη μέσω μιας σειράς από παρακείμενες k-clique. Δύο k-clique είναι παρακείμενες όταν μοιράζονται $k-1$ κόμβους. Ο ορισμός στοχεύει να παρουσιάσει το γεγονός ότι είναι ουσιώδες το χαρακτηριστικό μιας κοινότητας να μπορούνε τα μέλη της να είναι προσβάσιμα μέσα από ένα καλά συνδεδεμένο υποσύνολο κόμβων. Υπάρχουν άλλα μέρη από ολόκληρο το δίκτυο που δεν είναι προσβάσιμα μέσω ενός συγκεκριμένου k-clique, αλλά ενδεχομένως να περιλαμβάνουν επιπλέον k-clique κοινότητες. Κατά την διάρκεια της κατασκευής, οι k-clique κοινότητες μπορεί να μοιράζονται κόμβους, που σημαίνει ότι είναι επικαλυπτόμενες. Δηλαδή μπορεί να υπάρχουν κόμβοι που ανήκουν σε μη παρακείμενες k-cliques και μπορούν να είναι προσβάσιμοι από διαφορετικά μονοπάτια καταλήγοντας σε διαφορετικές ομάδες.

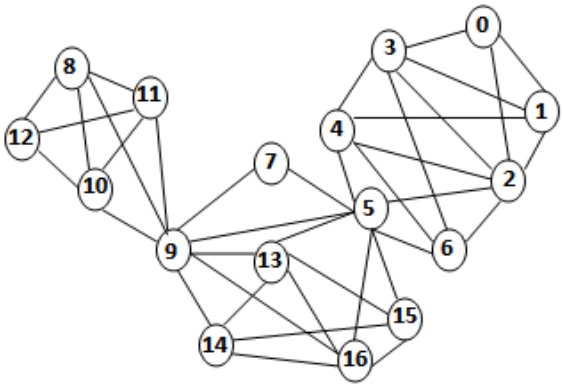
5.2.1 Μεθοδολογία

Όπως αναφέρουν οι αναλυτές στο [5], η εύρεση k-clique κοινοτήτων βασίζεται αρχικά στον εντοπισμό όλων των cliques (μέγιστο πλήρες υπογράφημα ή ισοδύναμα ένα πλήρες υπογράφημα που δεν είναι μέρος ενός μεγαλύτερου πλήρες υπογραφήματος) του δικτύου. Σε δεύτερη φάση, γίνεται η αναγνώριση των κοινοτήτων μέσω ενός πίνακα[6] που συμπληρώνεται κατά τον εντοπισμό των cliques στην πρώτη φάση (clique-clique overlap matrix).

Πιο συγκεκριμένα, αφού εντοπίσουμε όλα τα μέγιστα πλήρη υπογραφήματα συμπληρώνουμε τον clique-clique overlap πίνακα A. Σε αυτή την δομή πληροφορίας κάθε γραμμή και στήλη αντιπροσωπεύει μια clique και τα στοιχεία του πίνακα είναι ίσα με τον αριθμό των κοινών κόμβων ανάμεσα σε δύο κοινότητες (το (i,j)^{οστό} στοιχείο του A είναι ο αριθμός των κοινών κόμβων που έχουν η clique i και clique j). Τα διαγώνια στοιχεία είναι ίσα με το μέγεθος της clique. Η k-clique κοινότητες μπορούνε τώρα να βρεθούνε διαγράφοντας κάθε στοιχείο, εκτός της διαγωνίου, που είναι μικρότερο από k-1.

5.2.2 Παράδειγμα

Η παραπάνω μέθοδος εφαρμόζεται στο γράφημα που απεικονίζεται στο Σχήμα 5.1.



Σχήμα 5.1: Το αρχικό μας δίκτυο που αποτελείται από 17 κόμβους [7].

Αρχικά ψάχνουμε να βρούμε όλα τα μέγιστα πλήρη υπογραφήματα (cliques) στο δίκτυό μας. Αυτά θα είναι:

- Clique 1={0,1,2,3}
- Clique 2={1,2,3,4}
- Clique 3={2,3,4,6}
- Clique 4={2,4,5,6}
- Clique 5={9,14,13,16}
- Clique 6={13,14,15,16}
- Clique 7={13,16,5,15}
- Clique 8={12,8,10,11}
- Clique 9={10,8,9,11}

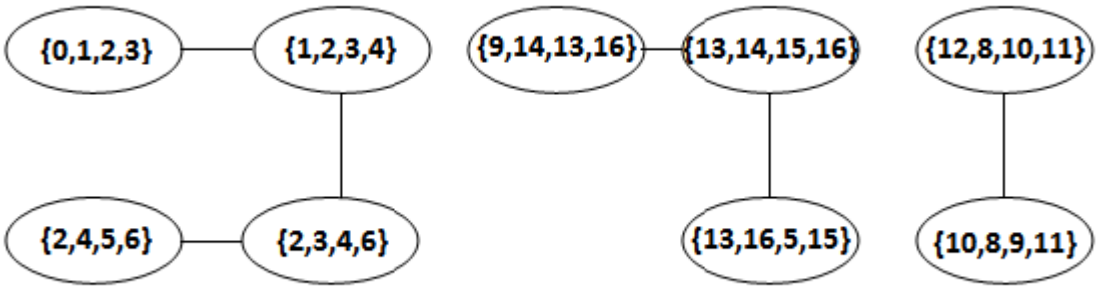
Παρατηρούμε ότι έχουμε μέγιστα πλήρη υπογραφήματα μεγέθους k=4 οπότε αναφερόμαστε σε 4-cliques. Ο clique-clique overlap πίνακας A (9×9) θα έχει την εξής μορφή:

$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{array} \begin{array}{c} \begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{array} \\ \begin{bmatrix} 4 & 3 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 3 & 4 & 3 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 3 & 4 & 3 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 3 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 3 & 4 & 3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 3 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 3 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 3 & 4 \end{bmatrix} \end{array}$$

Σε δεύτερη φάση, για την αναγνώριση των 4-clique κοινοτήτων διαγράφουμε τα στοιχεία, εκτός της διαγωνίου που είναι μικρότερα από k-1=3.

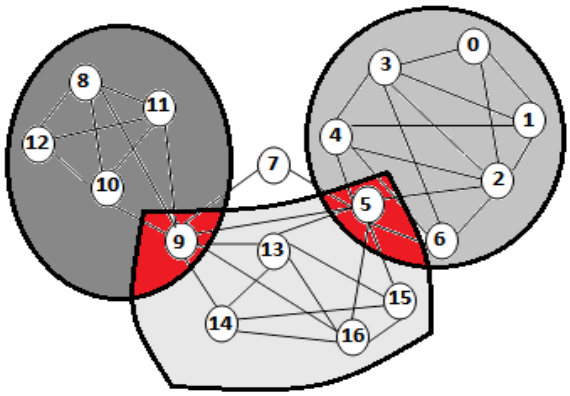
$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{array} \begin{array}{c} \begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{array} \\ \begin{bmatrix} 4 & 3 & \cancel{2} & \cancel{1} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} \\ 3 & 4 & 3 & \cancel{2} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} \\ \cancel{2} & 3 & 4 & 3 & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} \\ \cancel{1} & \cancel{2} & 3 & 4 & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} \\ \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & 4 & 3 & \cancel{2} & \cancel{0} & \cancel{1} \\ \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & 3 & 4 & 3 & \cancel{0} & \cancel{0} \\ \cancel{0} & \cancel{0} & \cancel{0} & \cancel{1} & \cancel{2} & 3 & 4 & \cancel{0} & \cancel{0} \\ \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & 4 & 3 \\ \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{1} & \cancel{0} & \cancel{0} & 3 & 4 \end{bmatrix} \end{array}$$

Αναλύοντας τον πίνακα Α, μπορούμε να δούμε ότι το στοιχείο A_{12}, όπως και το στοιχείο A_{21} (Α συμμετρικός ως προς την διαγώνιο), είναι ίσο με 3. Αυτό μας υποδεικνύει ότι οι κλίκες (cliques) 1 και 2 είναι γειτονικές αφού μοιράζονται 3 κόμβους. Με την σειρά της η 2 είναι γειτονική με την 3 (στοιχεία A_{23}=A_{32}=3) και η 3 γειτονική με την 4 (A_{34}=A_{43}=3). Όμως η 4 δεν μπορεί να φτάσει την 5. Όπως αναφέρθηκε μια k-clique κοινότητα είναι η ένωση όλων των k-clique που μπορούν να φτάσουν η μία την άλλη μέσω μιας σειράς από παρακείμενες k-cliques. Στο παράδειγμά μας εντοπίσαμε μια 4-clique κοινότητα με την ένωση των τεσσάρων γειτονικών 4- cliques, την 1,2,3 και 4. Έτσι αυτή οι κοινότητα θα αποτελείται από τους κόμβους {0,1,2,3,4,5,6}. Συνεχίζοντας την ανάλυση του πίνακα, η clique 5 είναι γειτονική με την 6 (το στοιχείο A_{56}=A_{65}=3), η 6 είναι γειτονική με την 7 (A_{67}=A_{76}=3) όχι όμως η 7 με την 8. Έτσι δημιουργείται μια άλλη 4-clique κοινότητα που αποτελείται από τους κόμβους {5,9,13,14,15,16}. Τέλος, η clique 8 είναι γειτονική με την 9 (A_{89}=A_{98}=3) και εντοπίζεται η 4-clique κοινότητα αποτελούμενη από τους κόμβους {8,9,10,11,12}. Μέσα από αυτή την διαδικασία δύο κοινότητες μπορεί να παρουσιάσουν επικάλυψη κάποιων κόμβων (στο παράδειγμά μας είναι οι κορυφές 5 και 9). Όλη την προηγούμενη πληροφορία μπορούμε να την παρουσιάσουμε και σε ένα γράφημα από κλίκες (Σχήμα 5.2) όπου κάθε συνδεόμενο κομμάτι του γραφήματος σχηματίζει μια κοινότητα.



Σχήμα 5.2: Γράφημα από κλίκες (cliques) όπου κάθε συνδεόμενο κομμάτι σχηματίζει μια κοινότητα.

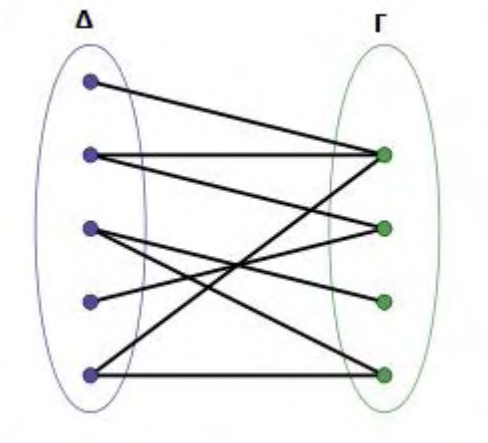
Μπορούμε να δούμε σχηματικά την δομή των κοινοτήτων που εντοπίστηκαν με την παραπάνω μέθοδο στο Σχήμα 5.3.



Σχήμα 5.3: Οι κοινότητες που ανιχνεύθηκαν με την υλοποίηση της μεθόδου clique percolation. Με κόκκινο χρώμα παρουσιάζεται η επικάλυψη (κόμβοι 5 και 9).

5.3 Μέθοδος Biclique

Ένα διμερές δίκτυο είναι ένα δίκτυο με δύο μη επικαλυπτόμενα σύνολα κόμβων Δ και Γ, όπου όλες οι συνδέσεις θα πρέπει να έχουν ένα τελικό κόμβο που θα ανήκει στο κάθε σύνολο. Ένα παράδειγμα διμερούς δικτύου φαίνεται στο Σχήμα 5.4.



Σχήμα 5.4: Παράδειγμα διμερούς γραφήματος[8].

Όπως αναφέρεται στο [9] υπάρχουν πολλά δίκτυα στον πραγματικό κόσμο που είναι διμερή. Κάποια παραδείγματα είναι:

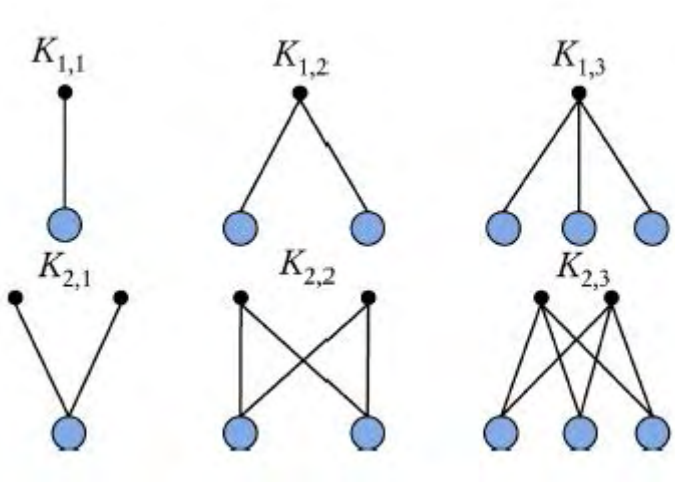
Κοινωνικά δίκτυα: ένα διμερές δίκτυο που αναπαριστά ηθοποιούς και ταινίες όπου κάθε ακμή συνδέει έναν ηθοποιό με τις ταινίες που έχει εργαστεί ή ένα διμερές δίκτυο που αναπαριστά χρήστες με ταινίες όπου κάθε ακμή συνδέει τους χρήστες με τις ταινίες που έχει παρακολουθήσει.

Δίκτυα πληροφορίας: ένα διμερές δίκτυο όπου ένας τύπος κόμβων αντιπροσωπεύει ένα έγγραφο(ιστοσελίδα, ηλεκτρονικό μήνυμα) και ο δεύτερος τύπος αντιπροσωπεύει τις λέξεις τις οποίες περιέχει.

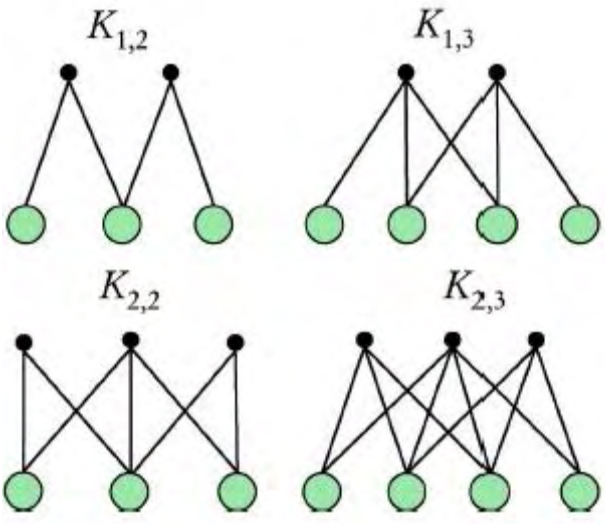
5.3.1 Διμερής κοινότητες

Σε αναλογία με τα μονομερή δίκτυα, στα διμερή δίκτυα μια κοινότητα ορίζεται ως μια κοινότητα η οποία αποτελείται από αρκετά πλήρη διμερή υπο-γραφήματα που τείνουν να μοιράζονται πολλούς από τους κόμβους τους. Μια σειρά από πλήρη διμερή γραφήματα παρουσιάζονται στο Σχήμα 5.5. Οι συγγραφείς του [9] ορίζουν μια $K_{\alpha,b}$ κλίκα ($K_{\alpha,b}$ clique) σαν ένα πλήρες υπογράφημα με a κόμβους στο Δ σύνολο

κόμβων και b κόμβους στο Γ σύνολο κόμβων. Μια $K_{\alpha,b}$ κλίκα μπορεί να είναι ίδια με ένα μέγιστο πλήρες υπογράφημα ή μπορεί να υπάρχει σε ένα υποσύνολο κόμβων ενός μέγιστου πλήρες υπογραφήματος. Γενικοποιώντας από το [5] τώρα ορίζεται μια $K_{\alpha,b}$ κοινότητα, ως η ένωση όλων των $K_{\alpha,b}$ κλικών οι οποίες μπορούν να φτάσουν η μία την άλλη μέσα από μια σειρά από γειτονικές $K_{\alpha,b}$ κλίκες. Δύο $K_{\alpha,b}$ κλίκες είναι γειτονικές εάν η επικάλυψή τους είναι τουλάχιστον μια $K_{\alpha-1,b-1}$ διμερής κλίκα. Ένας άλλος τρόπος να ειπωθεί το παραπάνω είναι ότι δύο κλίκες πρέπει να μοιράζονται τουλάχιστον a-1 κόμβους από το ένα σύνολο κόμβων και b-1 κόμβους από το άλλο σύνολο κόμβων (Σχήμα 5.6).



Σχήμα 5.5: Μέγιστα συνδεόμενα διμερή γραφήματα. Ο όρος $K_{\alpha,b}$ σημαίνει ότι το πλήρες διμερές γράφημα αποτελείται από a μαύρους κόμβους στο Δ σύνολο και b από τους μεγαλύτερους κόμβους στο Γ σύνολο[9].

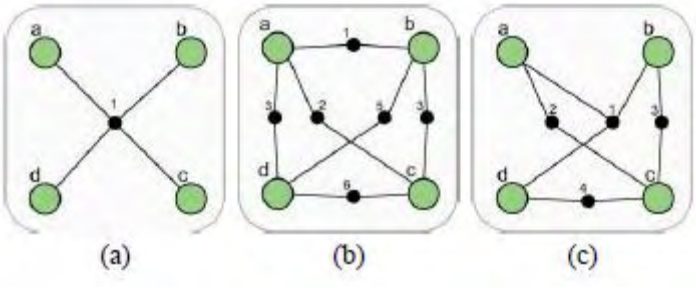


Σχήμα 5.6: Γειτονία διμερών κλικών. Δύο $K_{\alpha,b}$ κλίκες είναι γειτονικές εάν η επικάλυψή τους είναι τουλάχιστον μια $K_{\alpha-1,b-1}$ διμερής κλίκα. Οι δύο γειτονικές $K_{1,2}$ κλίκες μοιράζονται μια $K_{0,1}$ διμερή κλίκα, οι δύο γειτονικές $K_{1,2}$ κλίκες μοιράζονται μια $K_{0,2}$ κλίκα, οι δύο γειτονικές $K_{2,2}$ κλίκες επικαλύπτονται από μια $K_{1,1}$ κλίκα και οι δύο γειτονικές $K_{2,3}$ κλίκες μοιράζονται μια $K_{1,2}$ κλίκα[9].

5.3.2 Η σχέση με τις k-clique κοινότητες

Όταν ένα διμερές δίκτυο είναι διαθέσιμο, η μέθοδος εντοπισμού κοινοτήτων διμερών κλικών είναι εναλλακτική με τον k-clique αλγόριθμο που εξετάσαμε στην προηγούμενη ενότητα. Ο k-clique αλγόριθμος δεν μπορεί να αναλύσει αραιές περιοχές του δικτύου. Αυτό έχει να κάνει με το γεγονός ότι οι 2-clique κοινότητες είναι απλά τα συνδεόμενα μέρη του δικτύου και περιέχουν λίγη πληροφορία για την δομή του δικτύου. Η πρώτη σημαντική k-κλίκα έχει μέγεθος k=3. Συνδυάζοντας αυτά τα δύο γεγονότα, έχει ως αποτέλεσμα να μην μπορεί κάποιος να αναλύσει αραιές περιοχές ενός γραφήματος γιατί απλά οι κόμβοι πρέπει να έχουν τουλάχιστον δύο συνδέσεις ώστε να προχωρήσει ο διαχωρισμός σε μια 3-κλίκα.

Υποθέτοντας τα απλά διμερή δίκτυα του σχήματος 5.7, το δίκτυο του σχήματος 5.7(a) δείχνει μια περίπτωση όπου όλοι οι Δ κόμβοι συνδέονται με ένα μόνο κόμβο στο Γ σύνολο. Ένα πρακτικό παράδειγμα αυτού του μοτίβου μπορεί να βρεθεί σε ένα δίκτυο ταινίας-ηθοποιού, όπου αυτή θα ήταν η περίπτωση όταν τέσσερα άτομα παίζουν μαζί σε μια μόνο ταινία. Στο Σχήμα 5.7(b) παρουσιάζεται ένα διαφορετικό δίκτυο. Σε αυτό, όλοι οι τέσσερις κόμβοι στο Δ σύνολο ενώνονται με συνδέσεις που μεταξύ τους υπάρχουν έξι διακριτοί κόμβοι από το Γ σύνολο. Σε ένα δίκτυο ταινίας-ηθοποιού αυτό αντιστοιχεί με τέσσερις ηθοποιούς που όλοι τους έχουνε βρεθεί σε ταινία μαζί, αλλά με ακριβώς δύο κοινούς ηθοποιούς ανά ταινία. Έτσι, η έννοια αυτού του μοτίβου είναι πολύ διαφορετική από το μοτίβο του σχήματος 5.7(a). Το δίκτυο στο Σχήμα 5.7(c) βρίσκεται κάπου ανάμεσα σε αυτές τις δύο περιπτώσεις.



Σχήμα 5.7: Τρία απλά διμερή δίκτυα.

Όταν είναι διαθέσιμα διμερή δεδομένα η διμερής μέθοδος(Biclique method) είναι ικανή να εντοπίσει διακριτές δομές. Όσο αναφορά τις κλίκες, το Σχήμα 5.7(a) αντιστοιχεί σε μια $K_{4,1}$ κλίκα αποτελώντας παράδειγμα τεσσάρων ηθοποιών που συμμετέχουν στην ίδια ταινία. Το Σχήμα 5.7(b) αντιστοιχεί σε έξι γειτονικές $K_{2,1}$ κλίκες που ενώνονται σε μια $K_{2,1}$ κοινότητα. Τέλος, το Σχήμα 5.7(c) μπορεί να αναγνωριστεί σαν μια $K_{3,1}$ κλίκα και τρεις $K_{2,1}$ κλίκες. Όταν κάποιος σκεφτεί την κοινοτική δομή του σχήματος 5.7(c), όλοι οι κόμβοι περιλαμβάνονται εάν κάποιος

θέσει ένα κατώφλι στο $K_{2,1}$, αλλά εάν ανεβάσει το κατώφλι και ψάξει για $K_{3,1}$ κοινότητες, θα συμπεριλάβει μόνο τους κόμβους Δ={a,b,d} και Γ={1}.

Όπως αναφέρουν οι συγγραφείς του [9] ένα πολύ γνωστό πλεονέκτημα της μεθόδου k-clique είναι ότι επιτρέπει τον χρήστη να αλλάξει τη απόφαση στην οποία βασίζεται η παρατήρηση του δικτύου, προσαρμόζοντας το μέγεθος k της κλίκας. Μεγάλη τιμή του k, επιτρέπει τον χρήστη να παρατηρήσει δομές στις πιο πυκνές περιοχές του γραφήματος, ενώ μικρές τιμές του k επιτρέπει στον χρήστη να μελετήσει τις δομές στις πιο αραιές περιοχές του δικτύου. Στην περίπτωση των $K_{\alpha,b}$ κλικών, αυτή η ικανότητα ενισχύεται επειδή ο χρήστης έχει την ικανότητα να διαφοροποιήσει τα μεγέθη του a και b ανεξάρτητα το ένα από το άλλο. Διαφοροποιώντας τα a και b κάποιος μπορεί συστηματικά να διερευνήσει τις διαφορετικές πτυχές των σχηματισμών των κοινοτήτων μελετώντας πως κατανέμονται οι κοινότητες λόγω του μεγέθους αλλά και μέσω του οπτικού ελέγχου.

5.3.3 Εντοπισμός διμερών κοινοτήτων

Ο εντοπισμός διμερών κοινοτήτων είναι μια διαδικασία ανάλογη με αυτή που παρουσιάστηκε στην προηγούμενη ενότητα[5]. Ωστόσο κάποια βήματα είναι διαφορετικά. Σύμφωνα με τους συγγραφείς του [9] η μέθοδος εντοπισμού κοινοτήτων μεγέθους $K_{\alpha,b}$ είναι η ακόλουθη:

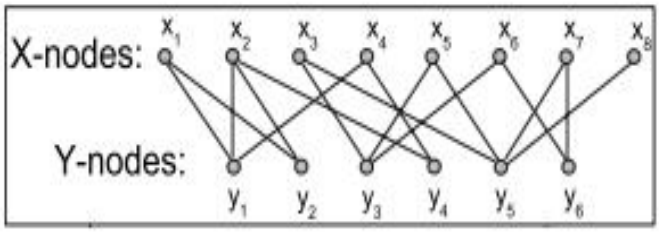
1. **Απαρίθμηση των μέγιστων διμερών κλικών**: Για τον εντοπισμό των διμερών κοινοτήτων, αρχικά θα πρέπει να βρεθούνε οι N μέγιστες διμερής κλίκες στο διμερές δίκτυο το οποίο εξετάζεται. Για αυτό τον σκοπό χρησιμοποιείται ο αλγόριθμος FIND-ALL-MAXIMAL του [10]. Χρησιμοποιώντας την λίστα από τις μέγιστες διμερής κλίκες, κατασκευάζονται δύο (N×N) συμμετρικοί πίνακες(clique-overlap matrices) L_{Δ} και L_{Γ}. Τα στοιχεία του πίνακα L_{Δ} περιέχουν πληροφορία για την επικάλυψη κλικών ανάμεσα στους κόμβους του συνόλου Δ. Κατά μήκος της διαγωνίου, αυτός ο πίνακας περιέχει τον αριθμό των κόμβων του συνόλου Δ στη μέγιστη διμερή κλίκα i. Τα στοιχεία εκτός της διαγωνίου περιέχουν τον αριθμό των κόμβων του συνόλου Δ που είναι κοινοί στη μέγιστη κλίκα i και στη μέγιστη κλίκα j. Ο πίνακας L_{Γ} είναι όμοιος, αλλά περιγράφει την επικάλυψη ανάμεσα στο σύνολο κόμβων Γ.
2. **Οριοθετώντας τους πίνακες επικάλυψης**: Η διαδικασία της οριοθέτησης περνάει από διάφορα στάδια. Το πρώτο στάδιο αξιολογεί τα διαγώνια στοιχεία. Τα διαγώνια στοιχεία μεγαλύτερα ή ίσα με το a θέτονται ίσα με ένα και όλα τα άλλα διαγώνια στοιχεία ίσα με μηδέν. Έπειτα οριοθετούνται τα εκτός διαγωνίου στοιχεία. Πρώτα, θέτονται σε μηδέν όλα τα στοιχεία των στηλών και των γραμμών που αντιστοιχούν σε ένα διαγώνιο στοιχείο με τιμή μηδέν. Έπειτα, οριοθετούνται τα εναπομείναντα στοιχεία, κρατώντας μόνο τα στοιχεία που είναι μεγαλύτερα ή ίσα με το a-1. Η ίδια διαδικασία ακολουθείται και για τον πίνακα L_{Γ}, χρησιμοποιώντας το b στην θέση του a.

Καθένας από τους οριοθετημένους πίνακες επικάλυψης (αποκαλούμενοι τώρα L_{Δ}' και L_{Γ}') περιέχει τώρα πληροφορία για την επικάλυψη για το καθένα από τα δύο σύνολα κόμβων. Για να βρεθεί πληροφορία σχετικά με την $K_{\alpha,b}$ κοινότητα δημιουργείται τώρα ο τελικός και συνολικός πίνακας επικάλυψης L αποδεχόμενοι μόνο την επικάλυψη κλικών όταν αυτή είναι παρούσα και στους δύο πίνακες. Έτσι οι αναλυτές θέτουν $L=L_{\Delta}' \wedge L_{\Gamma}'$, όπου ^ είναι η λογική λειτουργία AND. Ο συνολικός πίνακας επικάλυψης κλικών δίνει πληροφορίες για το ποιες μέγιστες κλίκες είναι γειτονικές με την έννοια $K_{\alpha-1,b-1}$.

3. **Εύρεση συνδεόμενων τμημάτων**: Το τελικό στάδιο είναι ο προσδιορισμός των συνδεόμενων τμημάτων του L. Κάθε τμήμα αντιστοιχεί σε μια κοινότητα διμερών κλικών(Biclique community). Από τις μέγιστες διμερής κλίκες που είναι μέλη της κάθε κοινότητας, εξάγονται οι κόμβοι που συμμετέχουν σε κάθε κοινότητα διμερών κλικών.

5.3.4 Παράδειγμα

Έστω ότι έχουμε διαθέσιμο το διμερές δίκτυο του σχήματος 5.8. Σε αυτό το δίκτυο υπάρχουν δύο σύνολα κόμβων, X και Y.



Σχήμα 5.8: Διμερές δίκτυο με δύο σύνολα κόμβων, X και Y[11].

Αρχικά πραγματοποιείται η εύρεση όλων των μέγιστων διμερών κλικών. Αυτό γίνεται με την εφαρμογή του αλγορίθμου που παρουσιάζεται στο [10]. Για το δίκτυο του σχήματος 5.8 η εφαρμογή του αλγορίθμου θα μας δώσει τις εξής μέγιστες διμερής κλίκες (για τις κλίκες της μορφής $K_{\alpha,b}$ θεωρούμε a για τους κόμβους που βρίσκονται στο X σύνολο και b για τους κόμβους που βρίσκονται στο Y σύνολο):

1. C_{max1}= { (**{x_1, x_2, x_4}**, **{y_1}**) }= $K_{3,1}$
2. C_{max2}= { (**{x_1, x_2}**, **{y_1, y_2}**) }= $K_{2,2}$
3. C_{max3}= { (**{x_3, x_5, x_6}**, **{y_3}**) }= $K_{3,1}$

4. C_{max4}= { (**{x_2, x_4}, {y_1, y_4}**) }= $K_{2,2}$
5. C_{max5}= { (**{x_3, x_5, x_7, x_8}, {y_5}**) }= $K_{4,1}$
6. C_{max6}= { (**{x_6, x_7}, {y_6}**) }= $K_{2,1}$
7. C_{max7}= { (**{x_2}, {y_1, y_2, y_4}**) }= $K_{1,3}$
8. C_{max8}= { (**{x_3, x_5}, {y_3, y_5}**) }= $K_{2,2}$
9. C_{max9}= { (**{x_6}, {y_3, y_6}**) }= $K_{1,2}$
10. C_{max10}= { (**{x_7}, {y_5, y_6}**) }= $K_{1,2}$

Σύμφωνα με τις παραπάνω μέγιστες διμερής κλίκες οι πίνακες επικάλυψης κλικών, L_x και L_Y θα έχουν την εξής μορφή:

$$L_X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array} \\ \begin{bmatrix} 3 & 2 & 0 & 2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 2 & 2 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 2 & 1 & 0 & 2 & 1 & 0 \\ 2 & 1 & 0 & 2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 4 & 1 & 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 2 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$L_Y = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 2 & 0 & 2 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 2 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 2 \end{bmatrix} \end{array}$$

Ας υποθέσουμε ότι ψάχνουμε για $K_{2,2}$ κοινότητα. Η διαδικασία της οριοθέτησης για τον πίνακα L_x θα θέσει τα διαγώνια στοιχεία, μεγαλύτερα η ίσα του a=2 σε ένα,

και τα υπόλοιπα σε μηδέν. Για τα εκτός διαγωνίου στοιχεία, πρώτα θέτουμε σε μηδέν όλα τα στοιχεία των στηλών και των γραμμών που αντιστοιχούν σε ένα διαγώνιο στοιχείο με τιμή μηδέν και μετά οριοθετούμε τα υπόλοιπα στοιχεία κρατώντας μόνο τα στοιχεία μεγαλύτερα ή ίσα του a-1. Η διαδικασία αυτή πραγματοποιείται και για τον πίνακα L_Y χρησιμοποιώντας το b στην θέση του a. Οι οριοθετημένοι πίνακες επικάλυψης θα έχουν τώρα την εξής μορφή:

$$L_X' = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array} \\ \begin{bmatrix} 1 & 2 & 0 & 2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 2 & 1 & 0 & 2 & 1 & 0 \\ 2 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 1 & 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$L_Y' = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Τέλος κατασκευάζετε ο συνολικός τελικός πίνακας επικάλυψης L. Αποδεχόμαστε μόνο την επικάλυψη κλικών όταν είναι παρούσα και στους δύο προηγούμενους πίνακες που αυτό σημαίνει ότι $L= L_Δ' \wedge L_Γ'$, όπου ^ είναι η λογική λειτουργία AND. Έτσι ο L θα έχει την παρακάτω μορφή:

$$L = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array} \\ \begin{bmatrix} 0 & 0 & & 0 & & & 0 & & & \\ 2\wedge 1 & 1\wedge 1 & & 1\wedge 1 & & & 1\wedge 2 & & & \\ & & 0 & & 0 & 0 & & 0 & 0 & \\ 2\wedge 1 & 1\wedge 1 & & 1\wedge 1 & & & 1\wedge 2 & & & \\ & & 0 & & 0 & 0 & & 0 & & 0 \\ & & 0 & & 0 & 0 & & & 0 & 0 \\ 0 & 0 & & 0 & & & 0 & & & \\ & & 2\wedge 1 & & 2\wedge 1 & & & 1\wedge 1 & & \\ & & 0 & & & 0 & & 0 & 0 & 0 \\ & & & & 0 & 0 & & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Αφού ψάχνουμε για μια $K_{2,2}$ κοινότητα, αυτή θα προέλθει από την ένωση των $K_{2,2}$ διμερών κλικών οι οποίες μπορούν να φτάσουν η μία την άλλη μέσα από μια σειρά από γειτονικές $K_{2,2}$ κλίκες. Δύο $K_{2,2}$ κλίκες είναι γειτονικές εάν η επικάλυψή τους είναι τουλάχιστον μια $K_{\alpha-1,b-1}=K_{1,1}$ διμερής κλίκα. Στο παράδειγμά μας οι $K_{2,2}$ διμερής κλίκες αντιστοιχούνε (και με αυτό τον τρόπο τους κατατάξαμε στον πίνακα) στους αριθμούς 2,4 και 8. Κοιτώντας τον συνολικό πίνακα επικάλυψης πίνακα η κλίκα 2 είναι γειτονική με την 4 αφού η επικάλυψή τους είναι μια $K_{1,1}$ διμερής κλίκα(το στοιχείο $L_{2,4}= L_{4,2}=1\wedge 1$). Από την άλλη πλευρά, η κλίκα 8 δεν είναι γειτονική με κάποια άλλη $K_{2,2}$ διμερής κλίκα. Έτσι στο δίκτυό μας βρίσκουμε δύο $K_{2,2}$ κοινότητες. Η πρώτη αποτελείται από τους κόμβους: (**{x_1, x_2, x_4}**, **{y_1, y_2, y_4}**) ως την ένωση των κλικών 2 και 4, και η δεύτερη αποτελείται μόνο από την κλίκα 8: (**{x_3, x_5}**, **{y_3, y_5}**) αφού δεν γειτνιάζει με κάποια άλλη. Οι συγγραφείς του [9] εξηγούνε την πολυπλοκότητα του παραπάνω αλγορίθμου.

5.4 Βιβλιογραφία

[1] X. Yan and J. Han, gspan: Graph-based substructure pattern mining, In IEEE International Conference on Data Mining,2002.

[2] M. Kuramochi and G. Karypis, Finding frequent patterns in a large sparse graph, Data Min Knowl Discov 11(3) (2005), 243–271.

[3] M. Berlingerio, F. Bonchi, B. Bringmann, and A. Gionis, Mining graph evolution rules, In ECML/PKDD (1), 2009, 115–130.

[4] S. Nijssen and J. N. Kok, A quickstart in frequent structure mining can make a difference, KDD ’04: Proceedings of the tenth ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining, New York, NY,ACM, 2004, 647–652.

[5] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005), 814–818.

[6] Everett, M. G. & Borgatti, S. P. Analyzing clique overlap. Connections 21, 49–61 (1998).

[7] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

[8] http://en.wikipedia.org/wiki/Bipartite_graph.

[9] S. Lehmann, M. Schwartz, and L. K. Hansen, Bi-clique communities, Phys Rev 78 (2008), 016108.

[10] On Enumerating All Maximal Bicliques of Bipartite Graphs, Enver Kayaaslan.

[11] http://en.wikipedia.org/wiki/Bipartite_network_projection.

Κεφάλαιο 6

Ομαδοποίηση Συνδέσεων (Link Clustering)

6.1 Εισαγωγή

Κάποιες πρόσφατες προσεγγίσεις βασίζονται στην ιδέα ότι η κοινότητα δεν είναι ένα τμήμα από κόμβους ενός δικτύου αλλά ένα τμήμα από τις συνδέσεις. Με άλλα λόγια, είναι η σχέση μεταξύ δυο οντοτήτων που ανήκει σε ένα συγκεκριμένο περιβάλλον και οι οντότητες ανήκουν σε όλες τις κοινότητες των ακμών τους (ή ένα υποσύνολο από αυτές).

Ο ορισμός της κοινότητας είναι:

Ορισμός 5 (Link Community) Μια κοινότητα συνδέσεων (link community) σε ένα πολύπλοκο δίκτυο είναι ένα σύνολο από κόμβους που μοιράζονται ένα αριθμό από σχέσεις που ομαδοποιούνται μαζί καθώς ανήκουν σε ένα συγκεκριμένο σχεσιακό περιβάλλον.

Η διαδικασία σε αυτή την κατηγορία είναι:

Διαδικασία 5 Μας δίνεται ένα σύνολο σχέσεων Μ μεταξύ ενός συνόλου οντοτήτων Ν. Εμείς ομαδοποιούμε μαζί σχέσεις που είναι παρόμοιες, που σημαίνει ότι καθιερώθηκαν μεταξύ του ίδιου συνόλου οντοτήτων, και στη συνέχεια συνδέουμε κάθε οντότητα n στις κοινότητες στις οποίες ανήκουν οι σχέσεις τους.

Η προσέγγιση αυτή υποδηλώνει επικαλυπτόμενη κατάτμηση, δεδομένου ότι ένας κόμβος ανήκει σε όλες τις κοινότητες των συνδέσεών του και μόνο σε σπάνιες περιπτώσεις όλες οι συνδέσεις ανήκουν σε μια μοναδική κοινότητα. Ένα χαρακτηριστικό που αγνοείται από αυτόν τον ορισμό της κοινότητας είναι η κατεύθυνση της σχέσης, δεδομένου ότι μια μη κατευθυνόμενη σύνδεση ανήκει σε μια μοναδική κοινότητα. Δεν υπάρχει κανένας τρόπος για να αποδώσουμε μια σχέση από το u στο v σε μια κοινότητα και μια σχέση από το v στο u σε μια άλλη κοινότητα, δεδομένου ότι και οι δυο ανήκουν στο ίδιο σχεσιακό περιβάλλον.

Η βασική προσέγγιση στο πρόβλημα του link clustering (ομαδοποίηση των συνδέσεων) είναι να ορίσουμε ένα γράφημα προβολής στο οποίο οι κόμβοι

αναπαριστούν τις συνδέσεις του αρχικού γραφήματος και να ορίσουμε την τιμή της εγγύτητας ώστε να κατανοήσουμε το πόσο κοντά είναι δυο ακμές του δικτύου. Και στις δυο περιπτώσεις, το κρίσιμο σημείο είναι η μέτρηση των σχέσεων μεταξύ των ακμών.

6.2 Μέθοδος Link modularity

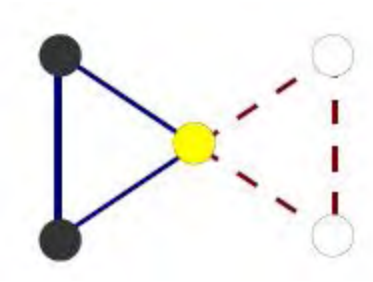
Οι συγγραφείς του [1] αναφέρουν ότι οι περισσότεροι μέθοδοι εύρεσης κοινοτήτων βρίσκουν έναν διαχωρισμό από το σύνολο των κορυφών όπου η πλειοψηφία των συνδέσεων είναι συγκεντρωμένες μέσα στις κοινότητες[2,3]. Οι κοινότητες είναι τα στοιχεία του διαχωρισμού και έτσι κάθε κόμβος σε μία και μόνο μία κοινότητα.

Το modularity είναι μια συνάρτηση ποιότητας, δηλαδή ένα ποιοτικό κριτήριο για την αξιολόγηση του πόσο καλός είναι ένας διαχωρισμός. Έχει χρησιμοποιηθεί σαν συνάρτηση ποιότητας σε πολλούς αλγορίθμους. Σε αντίθεση, η βελτιστοποίηση του modularity αποτελεί από μόνη της μια δημοφιλή μέθοδο εντοπισμού κοινοτήτων[15]. Έτσι, γνωστοί αλγόριθμοι αναζητούν την βελτιστοποίηση του modularity Q του διαχωρισμού των κόμβων του γραφήματος G[4,5,6,7,8]. Ο πιο απλός ορισμός του modularity για ένα μη κατευθυνόμενο γράφημα, που σημαίνει ότι ο πίνακας γειτνίασης A είναι συμμετρικός, είναι[9]:

$$Q(A)=\frac{1}{W}\sum_{C\in P}\sum_{i,j\in C}[A_{i,j}-\frac{k_i k_j}{W}] \qquad (1)$$

όπου $W=\sum_{i,j}A_{ij}$ και $k_i=\sum_j A_{ij}$ είναι ο βαθμός του κόμβου i. Οι δείκτες i και j τρέχουν πάνω από τους N κόμβους του γραφήματος G. Ο δείκτης C τρέχει πάνω από τις κοινότητες του διαχωρισμού P. Το modularity μετράει τον αριθμό των συνδέσεων μεταξύ όλων των ζευγαριών των κόμβων που ανήκουν στην ίδια κοινότητα και το συγκρίνει με τον αναμενόμενο αριθμό αυτών των συνδέσεων για ένα ισοδύναμο τυχαίο γράφημα στο οποίο οι βαθμοί όλων των κόμβων έχουνε μείνει ίδιοι.

Αυτή η προσέγγιση διαχωρισμού κόμβων έχει ένα μειονέκτημα όπως αναφέρουν οι συγγραφείς. Οι κόμβοι αποδίδονται σε μόνο μία κοινότητα, κάτι το οποίο μπορεί να είναι ένας μη επιθυμητός περιορισμός για δίκτυα που έχουν φτιαχτεί από επικαλυπτόμενες κοινότητες. Μόνο λίγες εναλλακτικές προσεγγίσεις έχουν προταθεί για να αποκαλυφθούνε επικαλυπτόμενες κοινότητες από κόμβους. Οι συγγραφείς προτείνουν να οριστούν οι κοινότητες σαν ένας διαχωρισμός των συνδέσεων παρά ένας διαχωρισμός του συνόλου των κορυφών. Ο κεντρικός κόμβος στο Σχήμα 6.1 είναι ένα απλό παράδειγμα.



Σχήμα 6.1: Διαχωρίζοντας τις συνδέσεις ενός δικτύου σε κοινότητες κάποιος μπορεί να αποκαλύψει επικαλυπτόμενες κοινότητες για τους κόμβους παρατηρώντας ότι ένας κόμβος ανήκει στις κοινότητες των συνδέσεών του. Σε αυτό το παράδειγμα ένας διαχωρισμός αποτελείται από τον χωρισμό των συνδέσεων σε δύο ομάδες(συμπαγής μπλε γραμμές και οι διακεκομμένες κόκκινες γραμμές). Ο κεντρικός κόμβος ανήκει στις δύο κοινότητες γιατί βρίσκεται στη διασύνδεση μεταξύ αυτών [1].

6.2.1 Δυναμική διαμόρφωση του modularity

Οι συγγραφείς αρχικά ερμηνεύουν το modularity Q (1) ως προς ένα τυχαίο περιπατητή που κινείται στους κόμβους[10,11]. Η πυκνότητα των τυχαίων περιπατητών στον κόμβο i στο βήμα n είναι $p_{i;n}$ και η δυναμική δίνεται από τον τύπο:

$$p_{i;n+1} = \sum_j \frac{A_{ij}}{k_j} p_{j;n} \qquad (2)$$

Από εδώ και στο εξής, οι συγγραφείς θεωρούνε δίκτυα που είναι μη κατευθυνόμενα (ο πίνακας γειτνίασης είναι συμμετρικός), συνδεδεμένα (υπάρχει μονοπάτι μεταξύ όλων των ζευγαριών των κόμβων), μη διμερή (δεν είναι εφικτό να χωρίσεις το δίκτυο σε δύο σύνολα κόμβων έτσι ώστε να μην υπάρχει σύνδεση μεταξύ κόμβων του ίδιου συνόλου) και απλά (χωρίς πολλαπλές συνδέσεις και βρόγχους).

Θεωρώντας έναν διαχωρισμό κόμβων P του δικτύου, επικεντρώνονται σε μια κοινότητα $C \in P$. Εάν το σύστημα είναι ισορροπημένο, δείχνουν ότι η πιθανότητα ενός τυχαίου περιπατητή να είναι στην C σε δύο επιτυχημένα χρονικά βήματα είναι:

$$\sum_{i,j \in C} \frac{A_{ij}}{k_j} \frac{k_j}{W}, \qquad (3)$$

Ενώ η πιθανότητα να βρεθούνε δύο ανεξάρτητοι περιπατητές στους κόμβους στην C είναι:

$$\sum_{i,j \in C} \frac{k_i k_j}{(W)^2}. \qquad (4)$$

Αυτή η παρατήρηση επιτρέπει στους συγγραφείς να επανερμηνεύσουν το Q ως ένα άθροισμα που αφορά τις κοινότητες, της διαφοράς των δύο αυτών πιθανοτήτων. Η ερμηνεία προτείνει μια γενίκευση του modularity που επιτρέπει τον συντονισμό της λύσης του. Πράγματι, το Q βασίζεται σε μονοπάτια μήκους ένα, αλλά μπορεί να γενικευτεί σε μονοπάτια αυθαίρετου μήκους:

$$R(A,n)=\frac{1}{W}\sum_{C\in P}\sum_{i,j\in C}[(T^n)_{ij}k_j-\frac{k_ik_j}{W}] \qquad (5)$$

όπου $T_{i,j}=A_{i,j}/k_j$. Αυτή η ποσότητα ονομάζεται η σταθερότητα του διαχωρισμού[10]. Επειδή k_j είναι το ιδιοδιάνυσμα της ιδιοτιμής ένα του T, ο συμμετρικός πίνακας $X(n)_{ij}=(T^n)_{ij}k_j$ αντιστοιχεί σε ένα χρονικά ανεξάρτητο γράφημα όπου ο βαθμός του κόμβου i είναι πάντα ίσο με k_i. Επομένως, το $R(A,n)$ μπορεί να ερμηνευτεί ως το modularity του $X(n)_{ij}$, ενός πίνακα που συνδέει όλο και περισσότερους μακρινούς κόμβους του αρχικού πίνακα γειτνιάσεως A όσο ο χρόνος n μεγαλώνει[11]. Όπως αναφέρουν οι συγγραφείς μπορεί να αποδειχθεί ότι βελτιστοποιώντας την (5) οδηγούνται σε διαχωρισμούς που φτιάχνονται από όλο και μεγαλύτερες κοινότητες όσο αυξάνεται ο χρόνος και ο βέλτιστος διαχωρισμός όταν $n\to\infty$ έχει δημιουργηθεί από δύο κοινότητες[10,11].

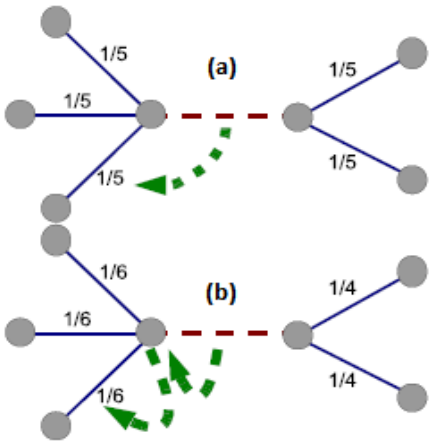
6.2.2 Διαχωρισμός συνδέσεων

6.2.2.1 Τυχαίοι περίπατοι στις συνδέσεις

Σύμφωνα με όσα προαναφέρθηκαν, οι συγγραφείς θα επικεντρωθούν σε ένα τυχαίο περιπατητή που κινείται στις ακμές του δικτύου με σκοπό να ορίσουν την ποιότητα του διαχωρισμού των συνδέσεων. Συνεπώς, αυτός ο περιπατητής θα βρίσκεται στις συνδέσεις αντί στους κόμβους σε κάθε χρονική στιγμή n και θα κινείται ανάμεσα σε γειτονικές συνδέσεις, που σημαίνει σε συνδέσεις που έχουνε ένα κόμβο κοινό. Στην περίπτωση του τυχαίου περιπάτους σε κόμβους (2), ένας περιπατητής σε έναν κόμβο i ακολουθεί μία από τις συνδέσεις του κόμβου με πιθανότητα $1/k_i$ (όλες οι συνδέσεις συμπεριφέρονται ομοιόμορφα). Ωστόσο, μια σύνδεση μεταξύ των κόμβων i και j χαρακτηρίζεται από δύο ποσότητες k_i και k_j, έτσι ένας τυχαίος περίπατος στις συνδέσεις είναι πιο διακριτικός. Στη συνέχεια οι αναλυτές επικεντρώνονται σε δύο τύπους δυναμικής προσέγγισης που εξηγούν διαφορετικά τους βαθμούς k_i και k_j.

Στην πρώτη προσέγγιση, ένας περιπατητής πραγματοποιεί άλμα με την ίδια πιθανότητα 1/($k_i + k_j$ -2) σε μία από τις συνδέσεις αφήνοντας αυτή του i και j. Όταν $k_i \neq k_j$, ο περιπατητής πηγαίνει με διαφορετική πιθανότητα ανάμεσα από τον i και j και συνεπώς αυτή η διαδικασία καλείται «link-link random walk» (Σχήμα 6.2 (a)).

Στη δεύτερη προσέγγιση, ο περιπατητής πραγματοποιεί άλμα σε έναν από τους δύο κόμβους με τους οποίους είναι συνδεδεμένος, ας πούμε τον i ,έπειτα σε μια σύνδεση η οποία συνδέεται με αυτόν τον κόμβο (εξαιρώντας την σύνδεση από την οποία προήλθε). Έτσι θα φτάσει σε μια σύνδεση, αφήνοντας τον κόμβο i , με μια πιθανότητα 1/(2(k_i -1)), και παρόμοια θα φτάσει σε μια σύνδεση που συνδέεται με τον κόμβο j με πιθανότητα 1/(2(k_j -1)). Η διαδικασία αυτή θα καλείται «link-node-link random walk». Μπορεί να δει κανείς αυτή την προσέγγιση στο Σχήμα 6.2(b). Αυτή η προσέγγιση είναι καλά ορισμένη εκτός αν η σύνδεση είναι ένα φύλο και πιο συγκεκριμένα ένα από τα άκρα της έχει βαθμό ένα, ας πούμε ο κόμβος i. σε αυτή την περίπτωση ο περιπατητής θα κάνει άλμα με πιθανότητα 1/(k_j -1) σε μια από τις συνδέσεις αφήνοντας τον j.



Σχήμα 6.2: Παράδειγμα των δύο προσεγγίσεων των τυχαίων περιπάτων. Και στις δύο περιπτώσεις οι περιπατητές βρίσκονται στις συνδέσεις του γραφήματος, εδώ αρχίζοντας από την κεντρική σύνδεση με τις κόκκινες διακεκομμένες γραμμές. Στο (a) παρουσιάζεται η «link-link random walk» προσέγγιση με τον περιπατητή να κάνει άλμα (οι πράσινες διακεκομμένες γραμμές) σε οποιαδήποτε από τις γειτονικές συνδέσεις με την ίδια πιθανότητα. Στο (b) φαίνεται η προσέγγιση «link-node-link random walk». Σε αυτή την περίπτωση ο περιπατητής μετακινείται πρώτα σε ένα γειτονικό κόμβο με την ίδια πιθανότητα και μετά κινείται σε μια καινούργια σύνδεση που διαλέγεται με την ίδια πιθανότητα ανάμεσα στις καινούργιες συνδέσεις που προσπίπτουν με αυτό τον κόμβο [1].

6.2.2.2 Προβάλλοντας τον πίνακα πρόσπτωσης (incidence matrix)

- **Bipartite structure**

Με σκοπό να μελετήσουν τους δύο τύπους του τυχαίου περιπάτου, οι αναλυτές παρουσιάζουν το δίκτυο G μέσω του πίνακα πρόσπτωσης B. Τα στοιχεία B_{ia} από τον $N \times L$ πίνακα (L είναι ο αριθμός των συνδέσεων) είναι ίσα με 1 εάν η σύνδεση a σχετίζεται με τον κόμβο i και 0 διαφορετικά. Κάποιος μπορεί να δει αυτόν τον πίνακα πρόσπτωσης ως τον πίνακα γειτνιάσεως ενός διμερούς δικτύου, $I(G)$ (Σχήμα 6.3(b)). Αυτό είναι γράφημα πρόσπτωσης του G, όπου οι δύο διαφορετικοί τύποι κόμβων αντιστοιχούν στους κόμβους και στις συνδέσεις του αρχικού γραφήματος G. Κατά την διαδικασία της κατασκευής όλη η πληροφορία του γραφήματος ενσωματώνεται στον B. Ο βαθμός k_i ενός κόμβου i και ο αριθμός των κόμβων k_a που συνδέονται με την σύνδεση a (πάντα ίσο με δύο) δίνονται από τις σχέσεις:

$$k_i = \sum_a B_{ia}\,, \qquad k_a = \sum_i B_{ia} \tag{6}$$

Ο $N \times N$ πίνακας γειτνίασης A του γραφήματος G μπορεί να δοθεί από την εξίσωση:

$$A_{ij} = \sum_a B_{ia} B_{ja} - k_i \delta_{ij}\,. \tag{7}$$

Η εξίσωση (7) μπορεί να ερμηνευτεί ως η προβολή του διμερούς γραφήματος πρόσπτωσης $I(G)$ στο μονομερές δίκτυο G[12,13]. Κατά παρόμοιο τρόπο, μπορεί να επιτευχθεί ένας πίνακας γειτνίασης για τις συνδέσεις προβάλλοντας το διμερές γράφημα πάνω στις συνδέσεις του. Οι συγγραφείς επικεντρώνονται σε δύο τύπους προβολών οι οποίοι είναι άμεσα σχετικοί με τις δύο προσεγγίσεις που παρουσιάστηκαν πιο πάνω.

- **Line graph**

Ο πιο απλός να προβληθεί ένα διμερές γράφημα είναι πάρει κάποιος όλους τους κόμβους του ενός τύπου για κόμβους του προβαλλόμενου γραφήματος. Προστίθεται μια σύνδεση ανάμεσα σε δύο κόμβους αυτού του προβαλλόμενου γραφήματος εάν αυτοί οι δύο κόμβοι είχανε τουλάχιστον ένα κόμβο, του άλλου τύπου, κοινό στο αρχικό διμερές γράφημα. Η εξίσωση (7) είναι αυτού του τύπου. Όταν εφαρμόζεται στις συνδέσεις a του γραφήματος G, ο δεύτερος τύπος των κορυφών στο διμερές γράφημα πρόσπτωσης I(G), οδηγεί στον $L \times L$ πίνακα γειτνίασης C του οποίου τα στοιχεία είναι:

$$C_{\alpha,\beta} = \sum_i B_{i\alpha} B_{i\beta} (1 - \delta_{\alpha,\beta}) . \qquad (8)$$

Είναι εύκολο να επαληθευτεί ότι αυτός ο πίνακας γειτνίασης είναι συμμετρικός και τα στοιχεία του είναι ίσα με 1 εάν δύο συνδέσεις έχουν έναν κόμβο κοινό και μηδέν διαφορετικά. Αυτός ο πίνακας γειτνίασης αντιστοιχεί σε ένα άλλο γνωστό γράφημα που συνήθως καλείται line graph του G[14] και υποδηλώνεται ως L(G) (Σχήμα 6.3(c)). Είναι ένα απλό γράφημα με L κόμβους. Κατά την κατασκευή, κάθε κόμβος i, βαθμού k_i, του αρχικού γραφήματος G αντιστοιχίζεται σε μια k_i πλήρως συνδεδεμένη κλίκα (**clique**) στο L(G). Συνεπώς έχει $\sum_i k_i (k_i - 1)/2 = O(\langle k^2 \rangle N)$ συνδέσεις.

Στη συνέχεια οι συγγραφείς εκφράζουν την δυναμική του «link-link random walk» σε σχέση με τον προβαλλόμενο πίνακα γειτνίασης C:

$$p_{a;n+1} = \sum_\beta \frac{C_{\alpha\beta}}{k_\beta} p_{\beta;n} \qquad (9)$$

όπου τώρα $p_{a;n}$ είναι η πυκνότητα των τυχαίων περιπατητών στην σύνδεση a και στο βήμα n. Επιπλέον, $k_a = \sum_\beta C_{\alpha\beta} = (k_i + k_j - 2)$ όπου i και j είναι τα άκρα της σύνδεσης a. Συνεπώς, η δυναμική προσέγγιση εξαρτάται από τους βαθμούς του i και j. Η σταθερή λύση έχει βρεθεί να είναι η $p_a^* = k_a / W$ όπου $W = \sum_{\alpha\beta} C_{\alpha\beta}$. Όταν το G είναι απλό γράφημα, τότε $W = \sum_i (k_i - 1) k_i$. Εφαρμόζοντας τα βήματα που περιγράφονται στο [11], οι αναλυτές εξάγουν μια συνάρτηση ποιότητας για τον διαχωρισμό P των συνδέσεων του γραφήματος G:

$$Q(C) = \frac{1}{W} \sum_{C \in P} \sum_{\alpha,\beta \in C} [C_{\alpha\beta} - \frac{k_\alpha k_\beta}{W}]. \qquad (10)$$

Αυτό είναι το συνηθισμένο modularity (1) για ένα γράφημα με πίνακα γειτνίασης τον C.

Όπως επισημάνθηκε, ένας απλός κόμβος i στο G οδηγεί σε μια συνδεδεμένη κλίκα $k_i(k_i - 1)/2$ συνδέσεων στο $L(G)$. Αυτό δηλώνει ότι το $L(G)$ δίνει μεγάλη σπουδαιότητα σε κόμβους μεγάλου βαθμού από το αρχικό γράφημα G. Σκοπός τώρα των αναλυτών είναι να ορίσουν ένα σταθμισμένο line graph του οποίου οι συνδέσεις κλιμακώνονται από ένα συντελεστή της μορφής $O(1/k_i)$.

- **Weighted line graph**

Οι συγγραφείς σε αυτό το σημείο έχουν ως σκοπό να εξάγουν την ποιότητα ενός διαχωρισμού συνδέσεων που σχετίζεται τον επονομαζόμενο «link-node-link» τυχαίο περίπατο. Προβάλλουν έτσι τον πίνακα πρόσπτωσης με ένα διαφορετικό τρόπο και

ορίζουν ένα διαφορετικό γράφημα $D(G)$ με ένα συμμετρικό πινάκα που δίνεται από τον τύπο:

$$D_{\alpha\beta} = \sum_{i,k_i>1} \frac{B_{ia}B_{i\beta}}{k_i - 1}(1-\delta_{\alpha\beta}). \qquad (11)$$

Το σταθμισμένο line graph έχει την ιδιότητα ότι ο βαθμός $k_a = \sum_\beta D_{\alpha\beta}$ μιας σύνδεσης a είναι ίσος με δύο (μια σύνδεση έχει πάντα δύο άκρα) εκτός εάν η a είναι φύλλο (τότε $k_a = 1$ εκτός από μια ασήμαντη περίπτωση). Το σταθμισμένο line graph του αρχικού μας δικτύου παρουσιάζεται στο σχήμα 10.3(d). Μόνο όταν το G είναι κανονικό, αυτό το σταθμισμένο line graph θα είναι ισοδύναμο (μέχρι μια γενική κλίμακα) με το αρχικό μη σταθμισμένο line graph $L(G)$.

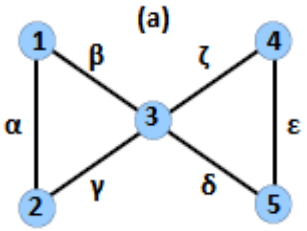
Αυτό το σταθμισμένο line graph επιτρέπει στους συγγραφείς να γράψουν την δυναμική του «link-node-link» τυχαίου περίπατου με ένα φυσικό τρόπο:

$$p_{a;n+1} = \sum_\beta \frac{D_{\alpha\beta}}{k_\beta} p_{\beta;n} \qquad (12)$$

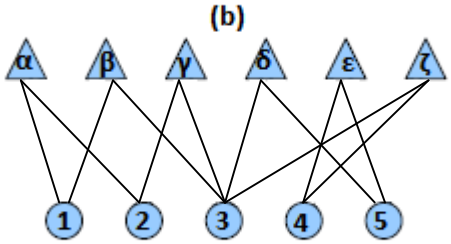
και χρησιμοποιώντας πάλι τα παραπάνω επιχειρήματα ορίζουν μια ακόμη συνάρτηση ποιότητας για τον διαχωρισμό P των συνδέσεων ενός γραφήματος:

$$Q(D) = \frac{1}{W} \sum_{C\in P} \sum_{\alpha,\beta\in C} [D_{\alpha\beta} - \frac{k_\alpha k_\beta}{W}] \qquad (13)$$

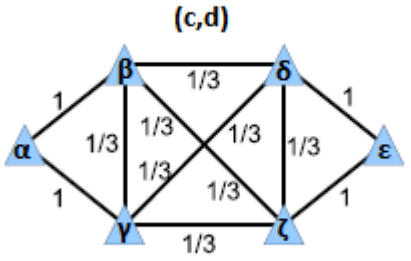
όπου $W = \sum_{\alpha\beta} C_{\alpha\beta} = 2L - L_{leaf}$ είναι ο διπλάσιος αριθμός των συνδέσεων L μείον τον αριθμό των φύλλων του αρχικού γραφήματος G, L_{leaf}. Και πάλι, αυτή είναι η ίδια συναρτησιακή μορφή με αυτή του modularity, $Q(A)$ της (1). Έχει αλλάξει μόνο ο πίνακας γειτνίασης.



$$\mathrm{A}_{ij} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array}$$



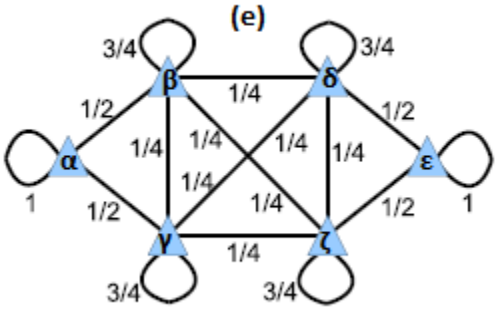
$$B_{ia} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array}$$



$$C = \begin{array}{c} \\ \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \\ \zeta \end{array} \begin{array}{c} \begin{array}{cccccc} \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

και

$$D = \begin{array}{c} \\ \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \\ \zeta \end{array} \begin{array}{c} \begin{array}{cccccc} \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1/3 & 1/3 & 0 & 1/3 \\ 1 & 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 1/3 & 0 & 1 & 1/3 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1/3 & 1/3 & 1/3 & 1 & 0 \end{bmatrix} \end{array}$$



$$E = \begin{array}{c} \\ \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \\ \zeta \end{array} \begin{array}{c} \begin{array}{cccccc} \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \end{array} \\ \begin{bmatrix} 1 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 3/4 & 1/4 & 1/4 & 0 & 1/4 \\ 1/2 & 1/4 & 3/4 & 1/4 & 0 & 1/4 \\ 0 & 1/4 & 1/4 & 3/4 & 1/2 & 1/4 \\ 0 & 0 & 0 & 1/2 & 1 & 1/2 \\ 0 & 1/4 & 1/4 & 1/4 & 1/2 & 3/4 \end{bmatrix} \end{array}$$

Σχήμα 6.3: Η πληροφορία του αρχικού γραφήματος στο (a) , όπως αυτή κωδικοποιείται με τον πίνακα γειτνίασης A της (7), έχει άλλες ισοδύναμες γραφικές αναπαραστάσεις. Στο (b) ο πίνακας πρόσπτωσης του αρχικού γραφήματος (ο B στην εξίσωση (7)) παρουσιάζεται σαν ένα διμερές δίκτυο, το γράφημα πρόσπτωσης I(G). Το line graph του αρχικού γραφήματος, L(G), είναι μη σταθμισμένη εκδοχή του γραφήματος που φαίνεται στο (c,d), με τον πίνακα γειτνίασης C της εξίσωσης (8). Η σταθμισμένη εκδοχή στο διάγραμμα (c,d) έχει έναν πίνακα γειτνίασης D της εξίσωσης (11). Το σταθμισμένο line graph με βρόγχους που φαίνεται στο (e) έχει έναν πίνακα γειτνίασης E της εξίσωσης (14). Οι κύκλοι αντιπροσωπεύουν οντότητες που σχετίζονται με κόμβους του αρχικού γραφήματος, ενώ τα τρίγωνα αντιπροσωπεύουν τις συνδέσεις του αρχικού γραφήματος [1].

6.2.2.3 Προβολή τυχαίου περιπάτου σε κόμβο

Οι τυχαίοι περίπατοι που προτάθηκαν στην προηγούμενη ενότητα ορίστηκαν σε ένα line graph και συνεπώς αποτελούνται από περιπατητές που κινούνται μεταξύ γειτονικών συνδέσεων στο αρχικό γράφημα G. Ωστόσο αυτή η διαδικασία δεν μπορεί να συσχετιστεί με τον αρχικό τυχαίο περίπατο (3) στους κόμβους του G, διότι ένας περιπατητής που κινείται στις συνδέσεις μπορεί να περάσει σε δύο ακολουθιακά βήματα μέσω του ίδιου κόμβου ενώ τέτοιοι βρόγχοι απαγορεύονται στην (3). Παρατηρώντας αυτό οι συγγραφείς προτείνουν μια εναλλακτική προσέγγιση όπου η δυναμική θα οδηγείται από τον αρχικό τυχαίο περίπατο (3) αλλά θα προβάλλεται στις συνδέσεις του δικτύου. Θεωρούν έτσι έναν περιπατητή ο οποίος δεν έχει κινηθεί ακόμα και βρίσκεται στον κόμβο i. Επιπλέον λόγω αυτής της περίπτωσης θεωρούν ακόμα ότι οι γειτονικές συνδέσεις του i συνδέονται με ένα βάρος ίσο με $1/k_i$. Έτσι προκύπτει ένας πίνακας γειτνίασης για τις συνδέσεις που δίνεται από την εξίσωση:

$$E_{\alpha\beta} = \sum_{i,k_i>0} \frac{B_{i\alpha} B_{i\beta}}{k_i}, \qquad (14)$$

που βασίζεται σε έναν, χωρίς περιορισμούς και αμερόληπτο, τυχαίο περίπατο δύο βημάτων στο διμερές γράφημα πρόσπτωσης I(G). Σε αντίθεση με τις προηγούμενες κατασκευές των line graph, C της εξίσωσης (8) και D της εξίσωσης (11), αυτό το σταθμισμένο line graph E(G) έχει βρόγχους. Αυτό απεικονίζεται στο Σχήμα 6.3(e). Όλοι οι κόμβοι στο E(G) έχουν δύναμη δύο, $\sum_{\beta} E_{\alpha\beta} = 2$, εκφράζοντας το γεγονός ότι όλες οι συνδέσεις στο αρχικό γράφημα G έχουνε δύο άκρα.

Ο πίνακας E κατασκευάζεται όταν ένας περιπατητής βρίσκεται σε ένα κόμβο και δεν έχει κινηθεί ακόμα. Η κίνηση του περιπατητή σύμφωνα με την (3) δημιουργεί έναν νέο πίνακα γειτνίασης ,E_1, που ορίζεται ως:

$$\mathrm{E}_{1;\alpha\beta} = \sum_{i,k_i>0} \frac{B_{ia}\mathrm{A}_{ij}B_{i\beta}}{k_i k_j} \qquad (15)$$

όπου E_1=ΕΕ-Ε. Το προκύπτων γράφημα παραμένει κανονικό με $k_a = \sum_{\beta} E_{1;\alpha\beta} = 2$ και έχοντας πάλι βρόγχους. Η συνάρτηση ποιότητας είναι:

$$Q(E_1) = \frac{1}{W}\sum_{C\in P}\sum_{\alpha,\beta\in C}[E_{1;\alpha\beta} - \frac{4}{W}], \qquad (16)$$

όπου $W = 2L$.

Αυτή η συνάρτηση ποιότητας είναι ιδιαίτερα ενδιαφέρουσα γιατί όπως αναφέρουν και αποδεικνύουν οι συγγραφείς έχει μια απλή σχέση με το modularity του αρχικού γραφήματος, Q(A) της εξ.(1). Πιο συγκεκριμένα καταλήγουν ότι :

$$Q(E_1;\{V_{ac}\}) = Q(A;\{V_{ic}\}) \qquad (17)$$

Έτσι βρίσκοντας τους βέλτιστους διαχωρισμούς των συνδέσεων μέσω του modularity του line graph μέσω του πίνακα γειτνίασης E_1 της εξ.(15) είναι ισοδύναμο με την βελτιστοποίηση του modularity του αρχικού γραφήματος.

Συνοψίζοντας τα παραπάνω, οι συγγραφείς προτείνουν τρεις συναρτήσεις ποιότητας, Q(C), Q(D) και $Q(E_1)$ για τον διαχωρισμό των συνδέσεων ενός γραφήματος G. Καθένας από αυτούς αντιπροσωπεύει μια διαφορετική δυναμική διαδικασία και συνεπώς εξερευνά την δομή του αρχικού γραφήματος G με διαφορετικό τρόπο. Οι βέλτιστοι διαχωρισμοί για αυτές τις συναρτήσεις ποιότητας μπορούν να βρεθούν εφαρμόζοντας απλούς αλγορίθμους βελτιστοποίησης του modularity στα line graphs που προέκυψαν προηγουμένως. Οι συγγραφείς του [1] χρησιμοποιούν δύο διαφορετικούς αλγορίθμους [6,7] και έχουν εξακριβώσει ότι και οι δύο αλγόριθμοι δίνουν συνεπή αποτελέσματα.

Κοιτώντας το γράφημα του σχήματος 6.1, η βελτιστοποίηση των τριών συναρτήσεων ποιότητας οδηγούν στο αναμενόμενο διαχωρισμό δύο τριγώνων με τιμές Q(C)=0.1, Q(D)=0.278 και $Q(E_1)$=0.167. Σε αυτή την περίπτωση ο κεντρικός κόμβος ανήκει εξίσου στις δύο κοινότητες των συνδέσεων, μια κατάσταση που μας οδηγεί στο σπάσιμο(split) του δικτύου παρά στο διαχωρισμό των κόμβων. Ο καλύτερος διαχωρισμός κόμβων μας δίνει Q(A)=0.111 όταν τρεις κόμβοι σε ένα τρίγωνο σχηματίζουν μια κοινότητα και οι υπόλοιποι δύο που απομένουν σχηματίζουν την δεύτερη κοινότητα.

Για την σύγκριση του διαχωρισμού των κόμβων και του διαχωρισμού των συνδέσεων χρησιμοποιούνται οι έννοες «boundary link» και «boundary node». Το «boundary link» ενός διαχωρισμού κόμβων είναι μια συνοριακή ακμή που συνδέει δύο κόμβους διαφορετικών κοινοτήτων. Το «boundary node» σε έναν διαχωρισμό

συνδέσεων είναι ένας συνοριακός κόμβος που συνδέεται στις ακμές δύο ή περισσότερων κοινοτήτων που αποτελούνται από συνδέσεις. Συνεπώς ο κεντρικός κόμβος του σχήματος 6.1 αποτελεί ένα συνοριακό κόμβο(«boundary node»).

6.3 Βιβλιογραφία

[1] T. S. Evans and R. Lambiotte, Line graphs, link partitions and overlapping communities, Phys Rev E 80 (2009), 016105.

[2] S. Fortunato and C. Castellano, in Encyclopedia of Complexity and System Science edited by R.A. Meyers, Springer-Verlag 2009.

[3] M.A. Porter, J.-P, Onnela, P.J. Mucha, Communities in Networks, arXiv:0902.3788.

[4] M.E.J. Newman, Phys. Rev. E 69, 066133 (2004).

[5] R. Guimera, M. Sales-Pardo and L.A.N. Amaral, Phys. Rev. E 70, 025101(R) (2004).

[6] V.D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, J. Stat. Mech., P10008 (2008).

[7] A. Noack and R. Rotta, Lecture Notes in Computer Science 5526, 257-268 (2009).

[8] J. Reichardt and S. Bornholdt, Phys. Rev. E 74, 016110 (2006).

[9] M. Girvan and M.E.J. Newman, Proc. Natl. Acad. Sci.USA 99, 7821 (2002).

[10] J.-C. Delvenne, S. Yaliraki and M. Barahona, arXiv:0812.1811.

[11] R. Lambiotte, J.-C. Delvenne and M.Barahona, arXiv:0812.1770.

[12] T. Zhou, J. Ren, M. Medo and Y.-C. Zhang, Phys. Rev. E 76, 046115 (2007).

[13] R. Lambiotte, and M. Ausloos, Phys. Rev. E 72, 066107 (2005).

[14] V.K. Balakrishnan, Schaum's Outline of Graph Theory (Mcgraw- Hill Publ. Comp., New York, 1997).

[15] Santo Fortunato, Claudio Castellano, Community Structure in Graphs, Complex Networks Lagrange Laboratory (CNLL), ISI Foundation,Torino, Italy, SMC, INFM-CNR and Dipartimento di Fisica, "Sapienza" Uni versit`a di Roma, P. le A. Moro 2, 00185 Roma, Italy(2007).

Κεφάλαιο 7
Συμπεράσματα

7.1 Επίλογος

Ο σκοπός αυτής της εργασίας ήταν να παρουσιάσει κάποιους αλγορίθμους εντοπισμού κοινοτήτων. Ο εντοπισμός κοινοτήτων είναι σημαντικός για πολλούς λόγους και συμπεριλαμβάνει την κατηγοριοποίηση των κόμβων που οδηγεί σε ομοιογενής ομάδες, ηγέτες ομάδων και κρίσιμους συνδέσμους ομάδων. Οι κοινότητες μπορεί να είναι για παράδειγμα σελίδες στο Διαδίκτυο με σχετικό θέμα[1] ή ομάδες από άτομα που σχετίζονται μεταξύ τους στα κοινωνικά δίκτυα [2] και ούτω καθεξής.

Σύμφωνα με τους συγγραφείς του [3] ορίσαμε ένα γενικό ορισμό της κοινότητας και κατηγοριοποιήσαμε τους αλγορίθμους σύμφωνα με τους πιο ειδικούς ορισμούς της κοινότητας που προήλθαν από τον γενικότερο. Για κάθε αλγόριθμο και μέθοδο που παρουσιάστηκε υπήρξαν εικόνες, παραδείγματα, η πολυπλοκότητα αλλά και πιθανές αδυναμίες του εκάστοτε αλγορίθμου, για την ευκολότερη και, όσον τον δυνατόν, εις βάθος κατανόηση του από τον αναγνώστη.

Υπάρχουν και άλλοι αλγόριθμοι σύμφωνα με το [3] που υπόκεινται στην παραπάνω κατηγοριοποίηση. Ενδεικτικά όσο αναφορά το κεφάλαιο 3 και την διάχυση (diffusion) κάποιος μπορεί να βρει και άλλες μεθόδους όπως τη μέθοδο Node Coloring[4], GuruMine[5], DegreeDiscountIC[6], στο κεφάλαιο 4(Εγγύτητα) την μέθοδο DOCS[7] και Infomap[8], στο κεφάλαιο 5(Δομή) την μέθοδο s-Plexes Enumeration[9] και EAGLE[10] και στο κεφάλαιο 6(ομαδοποίηση συνδέσεων) την μέθοδο Link Maximum Likelihood[11].

7.2 Βιβλιογραφία

[1] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, Self-organization and identification of web communities, IEEE Comput 35 (2002), 66–71.

[2] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc Natl Acad Sci U S A 99 (2002), 7821.

[3] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

[4] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, A framework for community identification in dynamic social networks, KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, ACM, 2007, 717–726.

[5] A. Goyal, F. Bonchi, and L. V. Lakshmanan, Discovering leaders from community actions, CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, New York, NY, ACM, 2008, 499–508.

[6] W. Chen, Y. Wang, and S. Yang, Efficient influence maximization in social networks, KDD '09: Proceedings of the 15th ACM SIGKDD International Conference onKnowledge Discovery and Data Mining, New York, NY, ACM, 2009, 199–208.

[7] F. Wei, W. Qian, C. Wang, and A. Zhou, Detecting overlapping community structures in networks, World Wide Web 12(2) (2009), 235–261.

[8] M. Rosvall, and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc Natl Acad Sci 105 (2008), 1118–1123.

[9] C. Komusiewicz, F. Huffner, H. Moser, and R. Niedermeier, Isolation concepts for efficiently enumerating dense subgraphs, Theor Comput Sci 410(38–40) (2009), 3640–3654.

[10] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, Detect overlapping and hierarchical community structure in networks, Physica A 388 (2009), 1706.

[11] B. Ball, B. Karrer, and M. E. J. Newman, An efficient and principled method for detecting communities in networks, ArXiv e-prints, 2011.

Συνολική Βιβλιογραφία

Κεφάλαιο 1

[1] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109–1120, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, 3Department of Physics, Cornell University, Ithaca, NY 14853–2501(2003)

[2] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

[3] Ευστάθιος Ραπτοδήμος, Αλγόριθμοι Εντοπισμού Κοινοτήτων, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Φεβρουάριος 2014.

Κεφάλαιο 2

[1] R. A. Hanneman, and M. Riddle, Introduction to social network methods, Berkeley, CA, University of California; University of California Press, 2005.

[2] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc Natl Acad Sci U S A 99 (2002), 7821.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, The pagerank citation ranking: bringing order to the web, Stanford, CA, 1998.

[4] S. Gregory, An algorithm to find overlapping community structure in networks, Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007), Warsaw, Poland, Springer-Verlag, 2007, 91–102.

[5] S. Gregory, Finding overlapping communities using disjoint community detection algorithms, Complex Networks: CompleNet 2009, Berlin, Heidelberg, Springer-Verlag, 2009, 47–61.

[6] Gregory, S.: A Fast Algorithm to Find Overlapping Communities in Networks. In: PKDD 2008. LNAI, vol. 5211, pp. 408—423. Springer, Heidelberg (2008).

[7] J. Bagrow and E. Bollt, A local method for detecting communities, Phys Rev E 72 (2005), 46–108.

[8] A. Lancichinetti, S. Fortunato, and J. Kertesz, Detecting the overlapping and hierarchical community structure of complex networks, New J Phys 11 (2009), 033015.

[9] Radicchi F, Castellano C, Cecconi F, Loreto V and Parisi D 2004 Proc. Natl. Acad. Sci. USA 101 2658.

[10] S. Wasserman and K. Faust, Social Network Analysis. Cambridge University Press, Cambridge (1994).

[11] L. Freeman, A set of measures of centrality based upon betweenness. Sociometry 40, 35{41 (1977).

[12] M. E. J. Newman, Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. Phys. Rev. E 64, 016132 (2001).

[13] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

[14] http://www.ismll.uni-hildesheim.de/lehre/cmie-11w/script/lecture5.pdf.

Κεφάλαιο 3

[1] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, Self-organization and identification of web communities, IEEE Comput 35 (2002), 66–71.

[2] U. N. Raghavan, R. Albert, and S. Kumara, Near linear time algorithm to detect community structures in large scale networks, Phys Rev E 76 (2007), 036106.

[3] Gennaro Cordasco and Luisa Gargano Community Detection via Semi-Synchronous Label Propagation Algorithms Dipartimento di Informatica ed Applicazioni “R.M. Capocelli”University of Salerno, Fisciano 84084, ITALY.

[4] I. X. Y. Leung, P. Hui, P. Lio, and J. Crowcroft. Towards real- time community detection in large networks. Phys. Rev. E, 79(6):1–10, Jun 2009.

[5] S. Gregory, Finding overlapping communities in networks by label propagation, New Journal of Physics 12(10) (2009), 103018.

[6] F. Wu and B. A. Huberman, Finding communities in linear time: a physics approach, Eur Phys J B 38(2) (2004), 331–338.

[7] W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research, 33, 452-473 (1977).

[8] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 8271-8276 (2002).

[9] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

Κεφάλαιο 4

[1] D. J. Watts and S. H. Strogatz, Collective dynamics of ‘small-world’ networks, Nature 393(6684) (1998), 440–442.

[2] P. Pons, and M. Latapy, Computing communities in large networks using random walks, J Graph Algor Appl 3733 (2006), 284–293.

[3] L. Lovasz. Random walks on graphs: a survey. In Combinatorics, Paul Erdos is eighty, Vol. 2 (Keszthely, 1993), volume 2 of Bolyai Soc. Math. Stud., pages 353-397. Janos Bolyai Math. Soc., Budapest, 1996.

[4] D. Aldous and J. A. Fill. Reversible Markov Chains and Random Walks on Graphs, chapter 2. Forthcoming book, http://www.stat.berkeley.edu/users/aldous/RWG/book.html.

[5] J. H. Ward. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301):236{244, 1963.

[6] M. E. J. Newman. Fast algorithm for detecting community structure in networks. Physical Review E, 69(6):066133, 2004.

[7] L. Donetti and M. A. Munoz. Detecting network communities: a new systematic and efficient algorithm. Journal of Statistical Mechanics, 2004(10):10012, 2004.

[8] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. Machine Learning, 56(1-3):9{33, 2004.

[9] W. Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In Proceedings of the thirty-fifth annual ACM Symposium on Theory of computing, STOC, pages 50{58. ACM Press, 2003.

[10] M. Jambu and Lebeaux M.-O. Cluster analysis and data analysis. North Holland Publishing, 1983.

[11] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69(2):026113, 2004.

[12] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

Κεφάλαιο 5

[1] X. Yan and J. Han, gspan: Graph-based substructure pattern mining, In IEEE International Conference on Data Mining,2002.

[2] M. Kuramochi and G. Karypis, Finding frequent patterns in a large sparse graph, Data Min Knowl Discov 11(3) (2005), 243–271.

[3] M. Berlingerio, F. Bonchi, B. Bringmann, and A. Gionis, Mining graph evolution rules, In ECML/PKDD (1), 2009, 115–130.

[4] S. Nijssen and J. N. Kok, A quickstart in frequent structure mining can make a difference, KDD ’04: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY,ACM, 2004, 647–652.

[5] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005), 814–818.

[6] Everett, M. G. & Borgatti, S. P. Analyzing clique overlap. Connections 21, 49–61 (1998).

[7] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

[8] http://en.wikipedia.org/wiki/Bipartite_graph.

[9] S. Lehmann, M. Schwartz, and L. K. Hansen, Bi-clique communities, Phys Rev 78 (2008), 016108.

[10] On Enumerating All Maximal Bicliques of Bipartite Graphs, Enver Kayaaslan.

[11] http://en.wikipedia.org/wiki/Bipartite_network_projection.

Κεφάλαιο 6

[1] T. S. Evans and R. Lambiotte, Line graphs, link partitions and overlapping communities, Phys Rev E 80 (2009), 016105.

[2] S. Fortunato and C. Castellano, in Encyclopedia of Complexity and System Science edited by R.A. Meyers, Springer-Verlag 2009.

[3] M.A. Porter, J.-P, Onnela, P.J. Mucha, Communities in Networks, arXiv:0902.3788.

[4] M.E.J. Newman, Phys. Rev. E 69, 066133 (2004).

[5] R. Guimera, M. Sales-Pardo and L.A.N. Amaral, Phys. Rev. E 70, 025101(R) (2004).

[6] V.D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, J. Stat. Mech., P10008 (2008).

[7] A. Noack and R. Rotta, Lecture Notes in Computer Science 5526, 257-268 (2009).

[8] J. Reichardt and S. Bornholdt, Phys. Rev. E 74, 016110 (2006).

[9] M. Girvan and M.E.J. Newman, Proc. Natl. Acad. Sci.USA 99, 7821 (2002).

[10] J.-C. Delvenne, S. Yaliraki and M. Barahona, arXiv:0812.1811.

[11] R. Lambiotte, J.-C. Delvenne and M.Barahona, arXiv:0812.1770.

[12] T. Zhou, J. Ren, M. Medo and Y.-C. Zhang, Phys. Rev. E 76, 046115 (2007).

[13] R. Lambiotte, and M. Ausloos, Phys. Rev. E 72, 066107 (2005).

[14] V.K. Balakrishnan, Schaum's Outline of Graph Theory (Mcgraw- Hill Publ. Comp., New York, 1997).

[15] Santo Fortunato, Claudio Castellano, Community Structure in Graphs, Complex Networks Lagrange Laboratory (CNLL), ISI Foundation,Torino, Italy, SMC, INFM-CNR and Dipartimento di Fisica, "Sapienza" Uni versit`a di Roma, P. le A. Moro 2, 00185 Roma, Italy(2007).

Κεφάλαιο 7

[1] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, Self-organization and identification of web communities, IEEE Comput 35 (2002), 66–71.

[2] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc Natl Acad Sci U S A 99 (2002), 7821.

[3] Michele Coscia, Fosca Giannotti, Dino Pedreschi, A classification for community discovery methods in complex networks, (Computer Science Department, University of Pisa, Pisa, Italy), (KDDLab, ISTI-CNR, Pisa, Italy), (Center for Complex Network Research, Northeastern University, Boston, USA) (2011).

[4] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, A framework for community identification in dynamic social networks, KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, ACM, 2007, 717–726.

[5] A. Goyal, F. Bonchi, and L. V. Lakshmanan, Discovering leaders from community actions, CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, New York, NY, ACM, 2008, 499–508.

[6] W. Chen, Y. Wang, and S. Yang, Efficient influence maximization in social networks, KDD '09: Proceedings of the 15th ACM SIGKDD International Conference onKnowledge Discovery and Data Mining, New York, NY, ACM, 2009, 199–208.

[7] F. Wei, W. Qian, C. Wang, and A. Zhou, Detecting overlapping community structures in networks, World Wide Web 12(2) (2009), 235–261.

[8] M. Rosvall, and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc Natl Acad Sci 105 (2008), 1118–1123.

[9] C. Komusiewicz, F. Huffner, H. Moser, and R. Niedermeier, Isolation concepts for efficiently enumerating dense subgraphs, Theor Comput Sci 410(38–40) (2009), 3640–3654.

[10] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, Detect overlapping and hierarchical community structure in networks, Physica A 388 (2009), 1706.

[11] B. Ball, B. Karrer, and M. E. J. Newman, An efficient and principled method for detecting communities in networks, ArXiv e-prints, 2011.