ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΛΓΟΡΙΘΜΟΙ ΕΝΤΟΠΙΣΜΟΥ ΚΟΙΝΟΤΗΤΩΝ

COMMUNITY DETECTION ALGORITHMS

Επιμέλεια:

Ραπτοδήμος Ευστάθιος

Επιβλέποντες καθηγητές:

Μποζάνης Παναγιώτης, Αναπληρωτής Καθηγητής Π.Θ

Κατσαρός Δημήτριος, Λέκτορας Π.Θ

Βόλος, Φεβρουάριος 2014

Ευχαριστίες

Ολοκληρώνοντας τις προπτυχιακές μου σπουδές στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, θα ήθελα να ευχαριστήσω όλους εκείνους που συνέβαλαν στην ολοκλήρωση των σπουδών.

Θα ήθελα αρχικά να ευχαριστήσω θερμά τον κ. Μποζάνη Παναγιώτη, Αναπληρωτή Καθηγητή και Πρόεδρο του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, για τις χρήσιμες συμβουλές και υποδείξεις του, καθώς και για την υποστήριξη που μου παρείχε κατά τη διάρκεια της φοίτησής μου αλλά και κατά την εκπόνηση της διπλωματικής μου εργασίας.

Στη συνέχεια, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της εργασίας κ. Κατσαρό Δημήτριο, Λέκτορα του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, για την καθοδήγησή του.

Έπειτα, θα ήθελα να ευχαριστήσω την οικογένειά μου για την αμέριστη συμπαράσταση που μου παρείχε όλα αυτά τα χρόνια των σπουδών μου.

Τέλος, θα ήθελα να ευχαριστήσω όλους τους φίλους μου οι οποίοι με υποστήριξαν όλα αυτά τα χρόνια και χωρίς αυτούς η πορεία μου στη σχολή θα ήταν πιο δύσκολη και λιγότερο ευχάριστη.

Στην οικογένειά μου

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1 ΕΙΣΑΓΩΓΗ

Η ανάγκη για συλλογή και επεξεργασία δεδομένων από τα γνωστά σε όλους μας δίκτυα, μας οδηγούν στην ενδελεχή μελέτη αυτών, και η εξαγωγή χρήσιμων πληροφοριών έχει αποδειχθεί μια τεράστια πρόκληση. Η τοπολογία αρκετών εξ' αυτών των δικτύων θυμίζει τη δομή μιας κοινότητας. Εξάλλου, πολλές ώρες δουλειάς και έρευνας έχουν δαπανηθεί ώστε να βρεθούν διάφοροι αποδοτικοί μέθοδοι και αλγόριθμοι ανάλυσης της δομής αυτών των δικτύων. Στην παγκόσμια βιβλιογραφία το παραπάνω αναφέρεται ως **εντοπισμός κοινοτήτων (community detection)**. Πολλές φορές η αναπαράσταση ενός δικτύου μπορεί να αποδειχθεί αρκετά πολύπλοκη. Έτσι, έπειτα από έρευνες χρόνων έχουν αναπτυχθεί διάφοροι αλγόριθμοι οι οποίοι εστιάζονται στις ιδιότητες του κάθε δικτύου, καθορίζοντας το δικό τους ορισμό της κοινότητας. Δεδομένου ενός επιθυμητού ορισμού κοινότητας και διαφόρων ιδιοτήτων ενός προβλήματος (μέγεθος δικτύου, κατεύθυνση ακμών κτλ), είναι σημαντικό οι προαναφερθήσαντες αλγόριθμοι να παρέχουν μια σωστή προσέγγιση του προβλήματος που μελετάται. Γι' αυτό το λόγο, η κατηγοριοποίηση των αλγορίθμων εντοπισμού κοινοτήτων κυρίως σε πολύπλοκα δίκτυα που θα παρουσιαστεί στην παρούσα διπλωματική είναι χρήσιμη και για περαιτέρω έρευνα.

Ένα πολύπλοκο δίκτυο είναι ένα γράφημα το οποίο δεν έχει κάποια τετριμμένα τοπολογικά χαρακτηριστικά που έχει ένα απλό δίκτυο, παρά μόνο κάποια χαρακτηριστικά που συναντάμε συνήθως σε πραγματικά γραφήματα. Για τον λόγο αυτό αποτελούν εφαλτήριο για την περαιτέρω εξέταση των ιδιοτήτων τέτοιων δικτύων. Ένα σημαντικό θέμα που προκύπτει από την ανάλυση πολύπλοκων δικτύων είναι ο εντοπισμός των κοινοτήτων που βρίσκονται κρυμμένες μέσα στη δομή αυτών των δικτύων.

Ο όρος *κοινότητα* εύκολα περιγράφεται ως ένα σύνολο από οντότητες στο οποίο κάθε οντότητα είναι πιο «κοντά», με την έννοια του δικτύου, με τις άλλες οντότητες εντός της κοινότητας παρά με αυτές έξω από αυτή. Σε μία κοινότητα, οι οντότητες που την αποτελούν μοιράζονται τις ίδιες ιδιότητες, καθώς αλληλεπιδρούν μεταξύ τους. Ο εντοπισμός κοινοτήτων είναι σημαντικός για πολλούς λόγους, συμπεριλαμβανομένης της κατηγοριοποίησης των κόμβων που συνεπάγεται ομογενή σύνολα. Οι κοινότητες μπορεί να αντιστοιχούν σε ομάδες σελίδων του Διαδικτύου που ασχολούνται με συναφή θέματα [1], σε ομάδες με άτομα που αλληλεπιδρούν μεταξύ τους μέσω των κοινωνικών δικτύων (Facebook, Twitter, Google+ κτλ) [2], κ.ο.κ.

Το πρόβλημα του εντοπισμού κοινοτήτων συγκαταλέγεται στην ίδια κατηγορία με το πρόβλημα της *συσταδοποίησης (clustering)*· ένα κλασικό πρόβλημα του τομέα της Εξόρυξης

Δεδομένων (Data Mining). Στην Εξόρυξη Δεδομένων η ανάλυση συστάδων χωρίζει τα δεδομένα σε κατηγορίες (συστάδες) με βάση τις πληροφορίες που βρίσκονται στα δεδομένα και που περιγράφουν τα αντικείμενα και τις σχέσεις τους. Ο στόχος είναι τα αντικείμενα μιας ομάδας να είναι όμοια (ή συσχετιζόμενα) μεταξύ τους και διαφορετικά (ή μη συσχετιζόμενα) με αντικείμενα άλλων ομάδων. Στην πραγματικότητα, ο εντοπισμός κοινοτήτων, και συνάμα η συσταδοποίηση στο τομέα της Εξόρυξης Δεδομένων, μερικές φορές αναφέρεται και ως μη εποπτευόμενη κατηγοριοποίηση (unsupervised classification). Επίσης, ο εντοπισμός κοινοτήτων αποτελεί μια εφαρμογή της Εξόρυξης Δεδομένων που έχει μελετηθεί περισσότερο πάνω στα κοινωνικά δίκτυα. Άλλες εφαρμογές, όπως η εξόρυξη γραφημάτων (graph mining) [3], βρίσκονται ακόμη σε πρώιμο στάδιο της ανάπτυξής τους. Αντίθετα ο εντοπισμός κοινοτήτων βρίσκεται σε πιο προχωρημένο στάδιο με συνεισφορές από διαφορετικά πεδία, όπως είναι η στατιστική φυσική που αποτελεί κλάδο της φυσικής που χρησιμοποιεί τις μεθόδους της θεωρίας των πιθανοτήτων και των στατιστικών στοιχείων.

Παρ' όλα αυτά, αυτό είναι μόνο ένα μέρος του προβλήματος του εντοπισμού κοινοτήτων. Σε ένα κλασικό πρόβλημα συσταδοποίησης στην Εξόρυξη Δεδομένων, υπάρχουν δεδομένα τα οποία δεν συσχετίζονται μεταξύ τους. Έτσι, σε αυτή τη γενική μορφή, το γεγονός ότι οι οντότητες είναι κόμβοι που συνδέονται μεταξύ τους μέσω ακμών, δεν αποτέλεσε αντικείμενο λεπτομερούς μελέτης. Η γειτνίαση μεταξύ των οντοτήτων σε ένα δίκτυο μπορεί να αναπαρασταθεί σε ένα γράφημα ως κόμβοι-κορυφές που συνδέονται μεταξύ τους με ακμές.

Ο περισσότερο αποδεκτός ορισμός της γειτνίασης σε ένα δίκτυο είναι βασισμένος στην τοπολογία των ακμών του. Στην περίπτωση αυτή, ο ορισμός της κοινότητας διατυπώνεται σύμφωνα με τις διαφορές στις πυκνότητες των ακμών σε διάφορα μέρη του δικτύου. Πολλά δίκτυα έχουν βρεθεί να είναι μη ομογενή, να μην αποτελούνται από μία μη διαφοροποιημένη μάζα των κορυφών, αλλά από διακριτές ομάδες. Μέσα σε αυτές τις ομάδες υπάρχουν πολλές ακμές μεταξύ των κορυφών, αλλά μεταξύ των ομάδων υπάρχουν λιγότερες ακμές. Ο σκοπός ύπαρξης αλγόριθμου εντοπισμού κοινοτήτων, σε αυτή την περίπτωση, είναι να χωρίσει τις κορυφές ενός δικτύου σε κάποιο συγκεκριμένο αριθμό k των ομάδων, μεγιστοποιώντας τον αριθμό των ακμών μέσα σε αυτές τις ομάδες και ελαχιστοποιώντας τον αριθμό των ακμών μεταξύ των κορυφών σε διαφορετικές ομάδες. Αυτές οι ομάδες αποτελούν τις επιθυμητές κοινότητες του δικτύου.

Ο ορισμός αυτός παρουσιάζει μια μικρή ασάφεια όσο η πολυπλοκότητα του δικτύου αυξάνεται, και οι ιδιότητες και η πληροφορία του κάθε κόμβου δεν είναι εμφανώς διαχωρίσιμα μεγέθη. Για παράδειγμα, δύο οντότητες μπορεί να θεωρηθούν ότι είναι κοντά μεταξύ τους άμα μοιράζονται κοινή πληροφορία ακόμη και αν δεν είναι άμεσα συνδεδεμένες. Αρκετά συχνά, μία νέα προσέγγιση πάνω στον εντοπισμό κοινοτήτων αποσκοπεί στο να αντιμετωπιστεί ένα συγκεκριμένο πρόβλημα που δεν έχει λυθεί από προηγούμενες προσεγγίσεις του θέματος, δημιουργώντας παράλληλα το δικό της ορισμό για την έννοια κοινότητα.

Εκτός του γεγονότος των πολλών ορισμών που συναντάμε για τις κοινότητες, αυτές έχουν μια σειρά από ενδιαφέρουσα χαρακτηριστικά και ιδιότητες. Μπορεί να παρουσιάζουν μια ιεραρχική διαμόρφωση των ομάδων εντός του δικτύου. Ή αλλιώς, το γράφημα μπορεί να περιλαμβάνει κατευθυνόμενες ακμές, δίνοντας έτσι σημασία προς αυτή την κατεύθυνση κατά την εξέταση των σχέσεων μεταξύ των οντοτήτων. Οι κοινότητες μπορεί να είναι δυναμικές, δηλαδή να εξελίσσονται με την πάροδο του χρόνου, ή να υπάρχουν πολλά σύνολα ατόμων που να συμπεριφέρονται ως μεμονωμένες οντότητες σε κάθε σχέση του δικτύου, σχηματίζοντας έτσι μία πυκνή κοινότητα κατά την εξέταση όλων των πιθανών σχέσεων την ίδια στιγμή.

Ως εκ τούτου, αυτή η υπερβολική αφθονία που υπάρχει στους ορισμούς έχει οδηγήσει στη δημοσίευση ενός εντυπωσιακού αριθμού λύσεων σε προβλήματα εντοπισμού κοινοτήτων. Συνεπώς είναι φυσιολογικό να υπάρχει αυτός ο μεγάλος αριθμός δημοσιεύσεων που να περιγράφουν όλες αυτές τις μεθόδους εντοπισμού κοινοτήτων.

Στην πραγματικότητα αυτό που είναι σημαντικό και πάνω στο οποίο πρέπει να βασίζεται η μελέτη μας δεν είναι τόσο το πώς μπορούν να εντοπιστούν οι κοινότητες, όσο το τι είδους κοινότητες ενδιαφερόμαστε να εντοπιστούν. Οι δημοσιεύσεις που ήδη υπάρχουν δεν ομαδοποιούν τους διαφορετικούς αλγορίθμους σύμφωνα με τον ορισμό της κάθε κοινότητας. Ωστόσο, υπάρχουν πολλοί διαφορετικοί τρόποι για να συλλάβουμε μια κοινότητα μέσα σε ένα δίκτυο, όπως ειπώθηκε και από τους Newman και Leicht σε ένα βιβλίο τους [5], όπου υποστηρίζουν ότι «όλοι οι μέθοδοι απαιτούν από εμάς να γνωρίζουμε εκ των προτέρων το τι ψάχνουμε, προτού αποφασίσουμε να το κάνουμε»· εδώ το «να γνωρίζουμε τι ψάχνουμε» προφανώς σημαίνει να καθορίσουμε πραγματικά τι είναι μία κοινότητα. Για να χρησιμοποιήσουμε μια μεταφορά, οι υπάρχουσες δημοσιεύσεις αναφέρονται σε τούβλα και υλικά τα οποία συνθέτουν ένα κτίσμα χωρίς να γίνεται καμία αναφορά για το αρχιτεκτονικό στυλ του κτίσματος. Με άλλα λόγια, ο σκοπός προγενέστερων δημοσιεύσεων είναι να απευθύνονται σε άτομα που ενδιαφέρονται να «χτίσουν» έναν νέο αλγόριθμο εντοπισμού κοινοτήτων, παρά σε άτομα που θέλουν να χρησιμοποιήσουν τις μεθόδους που παρουσιάζονται στη βιβλιογραφία. Ο σκοπός της παρούσης διπλωματικής είναι ακριβώς το δεύτερο.

Έτσι, έχει επιλεγεί να ομαδοποιηθούν οι αλγόριθμοι εντοπισμού κοινοτήτων συνυπολογίζοντας τους ορισμούς του τι είναι κοινότητα, το οποίο εξαρτάται από τα είδη των ομάδων που σκοπεύουν να δημιουργήσουν, τροποποιώντας το αρχικό δίκτυο. Για κάθε αλγόριθμο, καταγράφονται τα χαρακτηριστικά της εξόδου του αλγορίθμου, υπογραμμίζοντας έτσι για ποια χαρακτηριστικά είναι κατάλληλος ο αλγόριθμος. Θεωρούνται, επίσης, κάποια γενικά πλαίσια που παρέχουν τόσο μια προσέγγιση στον εντοπισμό κοινοτήτων όσο και μια γενική τεχνική. Αυτά μπορούν να εφαρμοστούν σε άλλους αλγορίθμους διαμέρισης γράφων προσθέτοντας νέα χαρακτηριστικά σε αυτές τις μεθόδους.

Η διπλωματική είναι οργανωμένη ως εξής: Στο Κεφάλαιο 2 παρέχεται ένας γενικός ορισμός του προβλήματος του εντοπισμού κοινοτήτων, καθώς επίσης και ο ορισμός τι είναι κοινότητα. Στο Κεφάλαιο 3 εξηγείται η κατηγοριοποίηση των αλγορίθμων με βάση τους ορισμούς της

κοινότητας. Έπειτα, στα Κεφάλαια 4, 5 και 6 παρουσιάζονται οι βασικές κατηγορίες των προσεγγίσεων δεδομένου του ορισμού ενός προβλήματος, μαζί με οτιδήποτε θεωρείται ότι είναι το πιο σημαντικό σε κάθε συγκεκριμένη κατηγορία. Τέλος, στο Κεφάλαιο 7 παρέχονται εν συντομία τα συμπεράσματα της συγκεκριμένης εργασίας και δίνεται μια πιθανή προσέγγιση σε μελλοντικές εργασίες.

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

[4] S. Fortunato, Community detection in graphs, Phys Rep 486(3–5) (2010), 175–174.

[5] M. E. J. Newman and E. A. Leicht, Mixture models and exploratory analysis in networks, Proc Natl Acad Sci 104 (2007), 9564–9569.

[6] http://en.wikipedia.org/wiki/Complex_network.

[7] Steven H. Strogatz, Exploring complex networks.

[8] Piet Van Mieghem, Delft University of Technology, The Physics of Complex Networks, July, 2011.

[9] http://en.wikipedia.org/wiki/Statistical_physics.

ΚΕΦΑΛΑΙΟ 2 ΟΡΙΣΜΟΣ ΠΡΟΒΛΗΜΑΤΟΣ

Στο Κεφάλαιο αυτό παρέχεται ένας γενικός ορισμός του προβλήματος του εντοπισμού κοινοτήτων, ο ορισμός τι είναι κοινότητα, και μερικές ιδιότητες ενός αλγόριθμου εντοπισμού κοινοτήτων.

2.1 Παρουσίαση του προβλήματος

Ας υποτεθεί ότι έχουμε ένα γράφημα G που συμβολίζεται ως εξής:

G = (V, E, L, C), όπου V είναι ένα σύνολο από επισημασμένους κόμβους, E είναι ένα σύνολο από επισημασμένες ακμές, το L είναι ένα σύνολο των ετικετών των ακμών, και C είναι ένα σύνολο των ετικετών των κόμβων. Το σύνολο E, με τη σειρά του, συμβολίζεται ως εξής: E = (u, v, l, w), όπου τα u και v είναι μία από τις κορυφές του συνόλου V, το l είναι μία από τις ετικέτες που ανήκουν στο σύνολο L, και το w είναι ένας ακέραιος αριθμός που αντιπροσωπεύει το βάρος της ακμής. Θεωρείται ότι δεδομένου ενός ζεύγους κόμβων u, v ∈ V και μιας ετικέτας l ∈ L, μόνο μία ακμή (u, v, l, w) μπορεί να υπάρχει˙ ωστόσο, η κατεύθυνση της ακμής κρίνεται από το μοντέλο, επομένως κατευθυνόμενες ακμές της μορφής (u, v, l, w) και της μορφής (v, u, l, w) θεωρούνται και είναι ξεχωριστές. Μπορεί επίσης να υποτεθεί ότι κάθε κόμβος μπορεί να επισημανθεί με μία ή περισσότερες κατηγορίες c ∈ C. Επιπλέον, κάθε ακμή, και κόμβος, μπορεί να επισημανθεί με έναν αυθαίρετο αριθμό χρονοσφραγίδων που αντιπροσωπεύουν τον χρόνο στον οποίο η ακμή εμφανίζεται και εξαφανίζεται στο δίκτυο. Οι ετικέτες ενός δεδομένου κόμβου μπορούν επίσης να αλλάζουν με την πάροδο του χρόνου. Σημειώστε ότι κόμβοι μπορούν να δημιουργήσουν ή να διαγράψουν ακμές στο δίκτυο, και/ή να αλλάξουν, εισάγουν, ή να διαγράψουν μία ή περισσότερες ετικέτες στο σύνολο της κατηγορίας τους. Τέτοια γεγονότα ονομάζονται «δράσεις» που εκτελούνται από τους κόμβους.

Με αυτό το σύνθετο μοντέλο μπορούν να αναπαρασταθούν όλες οι πιθανές παραλλαγές σε ένα γράφημα ενός σύνθετου φαινομένου πραγματικού κόσμου. Για παράδειγμα, μπορούν να μοντελοποιηθούν πολύπλοκα δίκτυα με πολλές εξαρτήσεις, λαμβάνοντας υπ’ όψιν τις ετικέτες L των ακμών ως διαφορετικές σχέσεις/διαστάσεις του δικτύου. Μπορούν επίσης να

αναπαρασταθούν απλούστερα μοντέλα, όπως για παράδειγμα μη σταθμισμένα δίκτυα, θεωρώντας ότι κάθε ακμή του δικτύου έχει εξ' ορισμού βάρος ίσο με 1 (w = 1).

Από εδώ και στο εξής χρησιμοποιείται ο συμβολισμός που παρουσιάζεται στον Πίνακα 1. Κατά τη διάρκεια της συγγραφής της διπλωματικής έχουν εισαχθεί νέα σύμβολα και συμβολισμοί στην παρουσίαση μιας συγκεκριμένης μεθόδου, όπου αυτό κρίθηκε απαραίτητο.

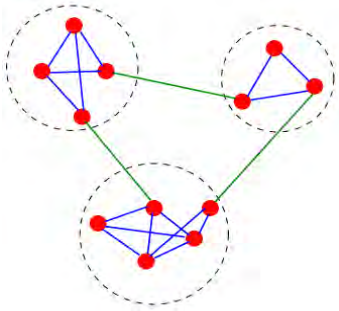
Symbol	Description
n	Number of vertices of the network
m	Number of edges of the network
k	Number of communities of the network
$\overline{K}$	Avg degree of the network
K	Max degree in the network
T	Number of action in the network
A	Max number of actions for a node
D	Number of dimensions (if any)
c	Number of vertex types (if any)
t	Number of time step (if any)

Πίνακας 1 : Πίνακας συμβόλων.

2.2 Ορισμός της κοινότητας

Σε αυτήν την ενότητα θα παρουσιαστεί ο όρος κοινότητα σε ένα πολύπλοκο δίκτυο. Με τον ορισμό αυτό δημιουργείται μία βασική ιδέα η οποία αποτελεί τη βάση πίσω από την έρευνα που πραγματοποιείται, και περιλαμβάνει όλες τις πιθανές παραλλαγές στον ορισμό που υπάρχουν στη βιβλιογραφία. Σε ένα πολύπλοκο δίκτυο, οι ακμές μεταξύ των κόμβων συχνά κατανέμονται ανομοιόμορφα, οδηγώντας στην δημιουργία ιεραρχικών δομών της κοινότητας. Ωστόσο, η εύρεση της δομής της κοινότητας σε πολύπλοκα δίκτυα έχει αποδειχθεί ότι είναι ένα δύσκολο έργο. Τί σημαίνει, όμως, ο όρος κοινότητα;

Ορισμός 2.1 (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.



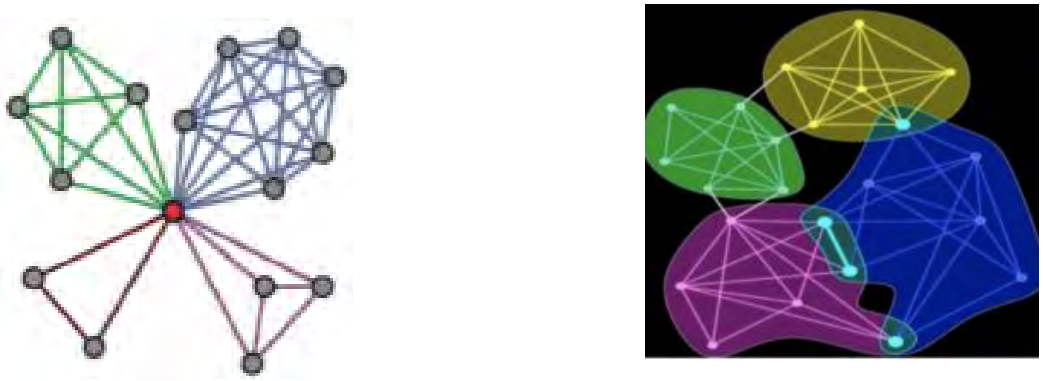
Σχήμα 2.1: Ένα απλό γράφημα με τρεις κοινότητες που βρίσκονται μέσα στις διακεκομμένες γραμμές [19].

Ο σκοπός ενός αλγορίθμου εντοπισμού κοινοτήτων είναι να εντοπίσει αυτές τις κοινότητες μέσα στο δίκτυο. Το αποτέλεσμα που προκύπτει είναι μια λίστα από σύνολα ομαδοποιημένων οντοτήτων. Έχοντας ως βάση τον παραπάνω ορισμό, μπορούν, πλέον, να μοντελοποιηθούν οι κύριες πτυχές του προβλήματος του εντοπισμού κοινοτήτων σε πολύπλοκα δίκτυα.

Σχήμα 2.2: Συστάδες/ομάδες βάσει πυκνότητας. Οι ομάδες είναι περιοχές υψηλής πυκνότητας διαχωρισμένες από περιοχές χαμηλής πυκνότητας.

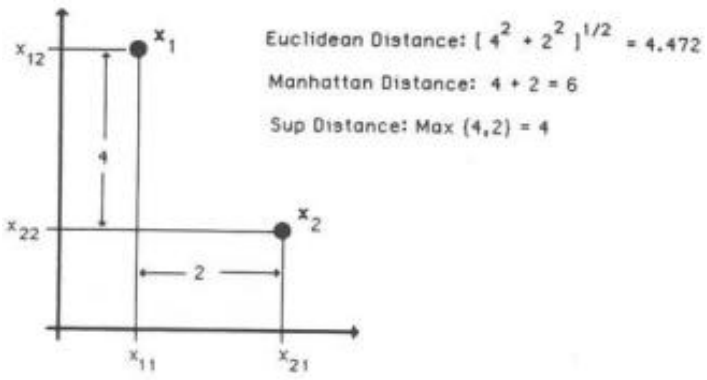
Ορισμοί βασισμένοι στην πυκνότητα (Density-based definitions). Όπως αναφέρθηκε στο Κεφάλαιο 1, ο ορισμός βασίζεται εξ' ολοκλήρου στην τοπολογία των ακμών του δικτύου. Η κοινότητα ορίζεται ως μία ομάδα (ή αλλιώς συστάδα), στην οποία υπάρχουν πολλές ακμές μεταξύ των κορυφών, αλλά μεταξύ των ομάδων υπάρχουν λιγότερες ακμές. Ο σκοπός ενός αλγορίθμου εντοπισμού κοινοτήτων είναι να χωρίσει τις κορυφές ενός δικτύου σε κάποιο k αριθμό ομάδων, μεγιστοποιώντας τον αριθμό των ακμών μέσα σε αυτές τις ομάδες και ελαχιστοποιώντας τον αριθμό των ακμών που συνδέουν κορυφές σε διαφορετικές ομάδες. Σε αυτόν τον ορισμό η σύνδεση μεταξύ δύο κορυφών θεωρείται μία συγκεκριμένη ενέργεια που μοιράζονται αυτές οι κορυφές. Ως εκ τούτου, εάν ομαδοποιηθούν οντότητες μεγιστοποιώντας τις κοινές τους ενέργειες/ιδιότητες, παράλληλα ομαδοποιούνται μεγιστοποιώντας τις ακμές μέσα στην κοινότητα. Ο εντοπισμός μιας κοινότητας είναι ακριβώς ο ίδιος, αν η δημιουργία μιας ακμής είναι η μόνη ενέργεια που καταγράφεται στην αναπαράσταση του δικτύου. Γραφικά, μία ομάδα κόμβων (ή αλλιώς συστάδα) είναι μια πυκνή περιοχή αντικειμένων, η οποία περιβάλλεται από μια περιοχή χαμηλής πυκνότητας. Στο Σχήμα 2.2 μπορείτε να δείτε ομάδες

κόμβων βάσει πυκνότητας. Επιπλέον, λαμβάνοντας υπ' όψιν διαφορετικά είδη συνόλων ενεργειών στον ορισμό, μπορεί επίσης να μοντελοποιηθεί η περίπτωση της *επικάλυψης (overlapping)*: για συγκεκριμένα σύνολα ενεργειών (π.χ συνδέσεις) ένας κόμβος ανήκει σε μία κοινότητα, ενώ για άλλα σύνολα ενεργειών ο ίδιος κόμβος ανήκει σε άλλη κοινότητα. Δηλαδή, ένα αντικείμενο μπορεί ταυτόχρονα να ανήκει σε περισσότερες από μία ομάδες. Για παράδειγμα, ένα άτομο σε ένα πανεπιστήμιο, μπορεί να είναι εγγεγραμμένος φοιτητής αλλά και εργαζόμενος, ταυτόχρονα.



Σχήμα 2.3: Overlapping κοινότητες [21].

Ορισμοί βασισμένοι στην ομοιότητα των κορυφών (Vertex similarity-based definitions). Μία ομάδα είναι ένα σύνολο από αντικείμενα, όπου κάθε αντικείμενο είναι πιο κοντά (πιο όμοιο) με το πρότυπο που ορίζει την ομάδα, από ότι με το πρότυπο οποιασδήποτε άλλης ομάδας. Θεωρείται ότι οι κοινότητες είναι ομάδες κορυφών που είναι όμοιες μεταξύ τους. Η ομοιότητα μεταξύ κάθε ζεύγους κορυφών μπορεί να υπολογιστεί σύμφωνα με ένα σημείο αναφοράς, τοπικό ή γενικό, ανεξάρτητα από το αν αυτοί συνδέονται μέσω μιας ακμής ή όχι. Κάθε κορυφή καταλήγει στη συστάδα/ομάδα, της οποίας οι κορυφές είναι πιο όμοιες σε αυτήν. Λαμβάνοντας υπ' όψιν την παρουσία ή την απουσία μια συγκεκριμένης ιδιότητας (π.χ μια ετικέτα της κορυφής) στην παρουσίαση ενός προβλήματος, μπορούν να μοντελοποιηθούν τα κριτήρια ομοιότητας, με την ομοιότητα του συνόλου των ενεργειών/ιδιοτήτων. Τα κριτήρια ομοιότητας αποτελούν τη βάση παραδοσιακών μεθόδων συσταδοποίησης, όπως είναι η ιεραρχική (hierarchical), η διαμεριστική (partitional) και η φασματική (spectral). Σαν κριτήριο ομοιότητας θα μπορούσε να χρησιμοποιηθεί η απόσταση μεταξύ ενός ζεύγους κορυφών (στην πραγματικότητα είναι ένα μέτρο ανομοιότητας, επειδή όμοιες κορυφές αναμένονται να είναι κοντά η μία στην άλλη). Δεδομένων δύο κορυφών, θα μπορούσε να χρησιμοποιηθεί ο τύπος της Ευκλείδιας απόστασης (L_2-norm), ο τύπος της απόστασης Manhattan (L_1-norm) ή η L_∞ νόρμα.



Σχήμα 2.4: Η απόσταση σαν κριτήριο ομοιότητας [22].

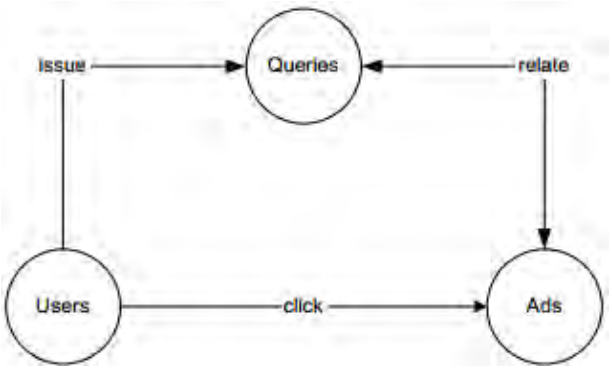
Ένα κριτήριο ομοιότητας αποτελεί και η ομοιότητα συνημιτόνου. Δεδομένων δύο διανυσμάτων A και B, η ομοιότητα συνημίτονου, θ, παριστάνεται χρησιμοποιώντας το εσωτερικό γινόμενο και το μέτρο ως εξής:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Επιπλέον, εάν το γράφημα δεν μπορεί να ενσωματωθεί στο χώρο, η ομοιότητα πρέπει αναγκαστικά να προκύπτει από τις σχέσεις γειτνίασης μεταξύ των κορυφών, ορίζοντας την απόσταση μεταξύ των κορυφών μέσω διάφορων γνωστών τύπων. Αυτό είναι ένα μέτρο ανομοιότητας που βασίζεται στην έννοια της δομικής ισοδυναμίας. Δύο κορυφές είναι δομικά ισοδύναμες αν έχουν τους ίδιους γείτονες, ακόμη και αν δεν είναι γείτονες μεταξύ τους. Ένα άλλο κριτήριο είναι ο αριθμός των ανεξάρτητων μονοπατιών ακμών (ή κορυφών) μεταξύ δύο οποιονδήποτε κορυφών. Τα ανεξάρτητα μονοπάτια δεν μοιράζονται καμία ακμή (κορυφή), και ο αριθμός τους έχει σχέση με τη μέγιστη ροή ανάμεσα σε δύο κορυφές υπό τον περιορισμό ότι κάθε ακμή μπορεί να μεταφέρει μόνο μία μονάδα ροής (max-flow/min-cut theorem). Ένα, ακόμη, κριτήριο της ομοιότητας κόμβων βασίζεται στις ιδιότητες τυχαίων διαπεράσεων των γράφων. Μία από τις ιδιότητες αυτές είναι ο χρόνος διαπέρασης μεταξύ ενός οποιουδήποτε ζεύγους κορυφών. Αυτός είναι ο μέσος αριθμός των βημάτων που απαιτούνται για μία τυχαία διαπέραση, ξεκινώντας από μία οποιαδήποτε κορυφή, να φτάσει κάποιος για πρώτη φορά σε μία άλλη κορυφή και να ξαναγυρίσει πίσω στην αρχική. Τέλος, κριτήριο αποτελεί και η πιθανότητα διαφυγής (escape probability), η οποία ορίζεται ως η πιθανότητα που έχει η διαπέραση να φτάσει στην κορυφή-στόχο προτού γυρίσει πίσω στην αρχική κορυφή.

Ορισμοί βασισμένοι στην ενέργεια (Action-based definitions). Σε αυτή την ενότητα, που κερδίζει όλο και περισσότερη προσοχή στη βιβλιογραφία, οι οντότητες μπορούν να ομαδοποιηθούν από το σύνολο των ενεργειών που εκτελούν μέσα στο δίκτυο. Για παράδειγμα θεωρήστε ένα multi-mode δίκτυο, το οποίο αποτελείται συνήθως από πολλαπλούς ετερογενείς

κοινωνικούς παράγοντες μεταξύ των οποίων διάφοροι τύποι αλληλεπιδράσεων μπορούν να συμβούν. Ένα παράδειγμα τέτοιου δικτύου είναι το online marketing [6] που φαίνεται στο Σχήμα 2.5.



Σχήμα 2.5: Online Marketing [6].

Το Σχήμα 2.5 αποτελεί ένα three-mode δίκτυο με τους εξής τρεις κοινωνικούς παράγοντες: users, queries, and online ads. Σε αυτό το δίκτυο οι χρήστες (users), που αποτελούν τον πιο σημαντικό υπό εξέταση παράγοντα, είναι συνδεδεμένοι με τα ερωτήματα (queries) και τις διαφημίσεις (ads), που και αυτοί με τη σειρά τους θεωρούνται «κοινωνικοί παράγοντες». Δύο χρήστες θεωρούνται ως τμήμα της ίδιας κοινότητας εφόσον είναι συνδεδεμένοι με τα ίδια ερωτήματα (δηλαδή εκτελούν τις ίδιες ενέργειες), ακόμη και αν δεν είναι άμεσα συνδεδεμένοι μεταξύ τους. Ο εντοπισμός κοινοτήτων που είναι βασισμένος σε αυτό τον ορισμό μπορεί να πραγματοποιηθεί είτε θεωρώντας είτε όχι την ύπαρξη μιας άμεσης σύνδεσης μεταξύ των οντοτήτων. Και οι δύο περιπτώσεις συμπεριλαμβάνονται στον ορισμό που παρουσιάζεται.

Ορισμοί βασισμένοι στην διάδοση επιρροής (Influence Propagation-based definitions). Σε ορισμένες εργασίες, έχει εισαχθεί η έννοια της «φυλής» (tribe). Η φυλή [16] ορίζεται ως ένα σύνολο οντοτήτων που επηρεάζονται από τους ίδιους ηγέτες (leaders). Ποια οντότητα, όμως, σε ένα δίκτυο πρέπει να αποτελεί έναν ηγέτη; Ο ηγέτης θα πρέπει α) για μία ενέργεια, να επηρεάζει αρκετά μεγάλο αριθμό χρηστών, β) για μία ενέργεια, να επηρεάζει αυτούς τους χρήστες σε ένα εύλογο χρονικό διάστημα, και γ) να ενεργεί ως ηγέτης σε αρκετά μεγάλο αριθμό ενεργειών. Με άλλα λόγια, ένας κόμβος είναι ηγέτης αν έχει πραγματοποιηθεί μία ενέργεια και, σε ένα επιλεγμένο χρονικό διάστημα μετά από αυτή την ενέργεια, ένας επαρκής αριθμός άλλων χρηστών πραγματοποιεί την ίδια ενέργεια. Σύμφωνα με τον ορισμό, το σύνολο των χρηστών που πραγματοποιούν συχνά τις ίδιες ενέργειες θεωρούνται ως μία κοινότητα, κάτι που οφείλεται στη επιρροή των ηγετών τους. Αμέσως παρακάτω παρουσιάζεται μια σύντομη προσέγγιση του θέματος μέσω ενός αλγορίθμου˙ του *Top Leaders Algorithm*. Η βασική ιδέα της προσέγγισής είναι εμπνευσμένη από τον ευρέως γνωστό αλγόριθμο συσταδοποίησης των K – μέσων, στον τομέα της Εξόρυξης Δεδομένων. Ωστόσο, υπάρχουν πολλές διαφορές, ιδιαίτερα

στο γεγονός ότι ενδεχομένως να εντοπιστούν ακραίες τιμές (δηλαδή, όρια). Παρόμοια με τον αλγόριθο των K - μέσων, ο συγκεκριμένος αλγόριθμος είναι ευαίσθητος στην αρχικοποίησή του, δηλαδή στην επιλογή των αρχικών k ηγετών. Η βασική ιδέα του Top Leaders Algorithm είναι αρχικά να βρει τους k ηγέτες της κοινότητας και έπειτα να καθορίσει την συμμετοχή των άλλων κόμβων της κοινότητας του δικτύου που βασίζεται στις σχέσεις τους με τους ηγέτες που έχουν εντοπιστεί. Τα βήματα του αλγόριθμου φαίνονται συνοπτικά στον Αλγόριθμο 2.1.

Αλγόριθμος 2.1 Top Leaders Algorithm

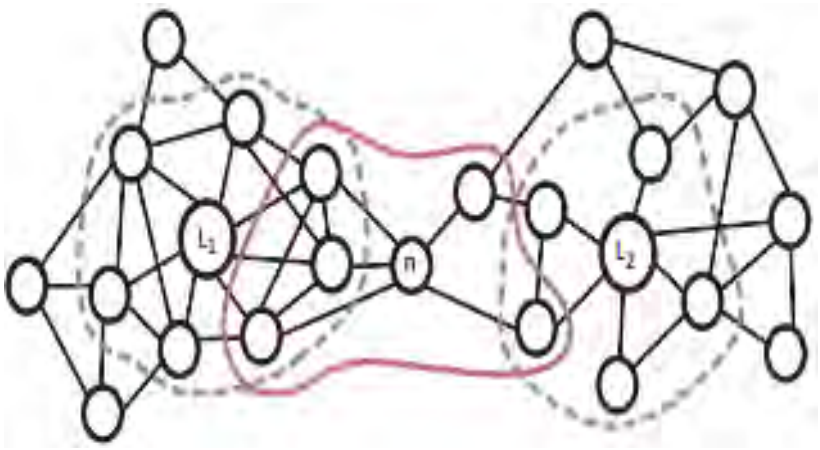
Είσοδος: Ένα δίκτυο N, και k ο αριθμός των επιθυμητών κοινοτήτων
αρχικοποίησε k ηγέτες
επανέλαβε
[εύρεση κοινοτήτων]
για όλους τους κόμβους n ∈ N **κάνε**
εάν n δεν ανήκει στους ηγέτες **τότε**
σύνδεσε τον n στην κοινότητα του ηγέτη // βλέπε Αλγόριθμο 2.2
τέλος εάν
τέλος για
[ανανέωσε την λίστα σου με τους ηγέτες]
για όλους l ∈ ηγέτες **κάνε**
l ← $\arg\max_{n \in \text{Κοινότητα(l)}}$ Κεντρικότητα (n) // Η αλλαγή των ηγετών είναι απλά η εκλογή του κόμβου με την υψηλότερη κεντρικότητα σε μια κοινότητα.
τέλος για
μέχρι να μην υπάρχει αλλαγή στους ηγέτες

Στον παρακάτω αλγόριθμο παρουσιάζεται εν συντομία η σύνδεση ενός κόμβου στην κοινότητα ενός ηγέτη όπως περιγράφεται στον Top Leaders αλγόριθμο.

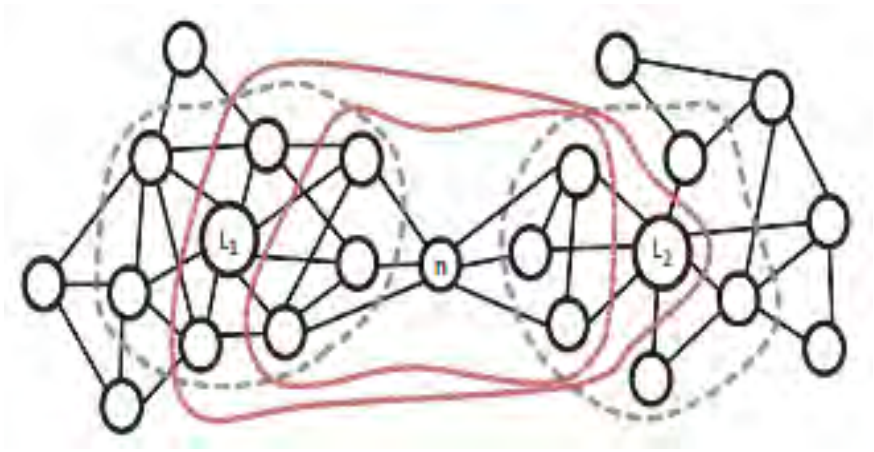
Αλγόριθμος 2.2 Σύνδεση κόμβου n στον ηγέτη του

Είσοδος: Ένα δίκτυο N, ένα κόμβος n, και ένα σύνολο από k ηγέτες
βάθος ← 1
λίστα ← ηγέτες // λίστα: λίστα με τους υποψήφιους ηγέτες
επανέλαβε
λίστα ← $\arg\max_{c \in \text{λίστα} \wedge |א(n1,d) \cap א(n2,d)| > \gamma}$ |א(n1, d) ∩ א(n2, d)| // א: γειτονιά
βάθος ← βάθος + 1
μέχρι |λίστα|≤ 1 ∨ βάθος > δ
εάν |λίστα| = 0 **τότε** [όχι υποψήφιος ηγέτης]

σύνδεσε τον n ως ακραίο σημείο
αλλιώς εάν |λίστα| > 1 **τότε** [πολλοί υποψήφιοι]
σύνδεσε τον n ως κεντρικό σημείο
αλλιώς [μόνο ένας υποψήφιος ηγέτης στη λίστα]
σύνδεσε τον n στη λίστα
τέλος εάν



α) Τομή των γειτονιών



β) Επέκταση των γειτονιών

Σχήμα 2.6: Καθορισμός κοινότητας του κόμβου n: ο θα πρέπει να συνδεθεί στην κοινότητα του ηγέτη L_1 επειδή α) ο n έχει περισσότερους κοινούς γειτόνους με τον L_1 παρά με τον L_2, β) μολονότι ο n έχει τον ίδιο αριθμό κοινών γειτόνων με τους L_1 και L_2 , έχει περισσότερους κοινούς γειτόνους με τον L_1 εάν επεκτείνουμε τα όρια της γειτονιάς του κατά ένα [7].

2.3 Χαρακτηριστικά του προβλήματος

Υπάρχουν πολλά χαρακτηριστικά γνωρίσματα που πρέπει να εξεταστούν στο πολύπλοκο έργο του εντοπισμού των κοινοτήτων στις δομές των γραφημάτων. Σε αυτή την ενότητα, παρουσιάζονται μερικά από τα χαρακτηριστικά που ένας αναλυτής θα ενδιαφερόταν να μελετήσει για τον εντοπισμό κοινοτήτων σε ένα δίκτυο.

Καταγράφονται οι βασικές ιδιότητες ενός αλγορίθμου εντοπισμού κοινοτήτων. 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.

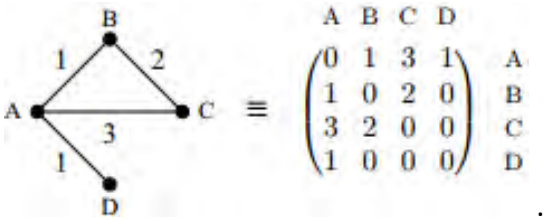
Μέσα στην πρώτη κατηγορία των χαρακτηριστικών ομαδοποιούνται όλες μαζί οι πιθανές παραλλαγές στην αναπαράσταση του αρχικού φαινομένου στον πραγματικό κόσμο. Τα πιο σημαντικά χαρακτηριστικά που εξετάζονται είναι:

- *Επικαλυπτόμενες Κοινότητες (Overlapping Communities).* Σε ορισμένα δίκτυα πραγματικού κόσμου οι κοινότητες μπορούν να μοιράζονται έναν ή περισσότερους κοινούς κόμβους. Για παράδειγμα, σε κοινωνικά δίκτυα παράγοντες μπορεί να αποτελούν μέρη διαφορετικών κοινοτήτων: εργασία, οικογένεια, φίλοι, και ούτω καθ'εξής. Όλες αυτές οι κοινότητες θα μοιράζονται ένα κοινό μέλος, και συνήθως περισσότερα από ένα από τη στιγμή που ένας συνάδελφος (εργασία) μπορεί επίσης να είναι και φίλος (φίλοι) έξω από το εργασιακό περιβάλλον. Το Σχήμα 2.6 (α) δείχνει ένα απλό παράδειγμα πιθανής επικάλυψης: ο κεντρικός κόμβος μοιράζεται από κοινού από τις δύο κοινότητες.

- *Κατευθυνόμενες Κοινότητες (Directed Communities).* Ορισμένα φαινόμενα του πραγματικού κόσμου πρέπει να αναπαρίστανται με ακμές και συνδέσεις που δεν είναι διπλής κατεύθυνσης. Αυτό συμβαίνει, για παράδειγμα, στην περίπτωση ενός web γραφήματος: ένας υπερσύνδεσμος από μία ιστοσελίδα σε άλλη είναι μιας κατεύθυνσης και η άλλη ιστοσελίδα δεν μπορεί να έχει έναν άλλον υπερσύνδεσμο που να δείχνει προς την άλλη κατεύθυνση. Το Σχήμα 2.6 (β) δείχνει ένα απλό παράδειγμα στο οποίο φαίνεται η κατεύθυνση των ακμών. Ο αριστερός κόμβος είναι συνδεδεμένος με την κοινότητα, αλλά μόνο προς μία κατεύθυνση. Αν μας ενδιέφεραν και οι δύο κατευθύνσεις, τότε ο αριστερός κόμβος θα θεωρούνταν εκτός της κοινότητας που απεικονίζεται.

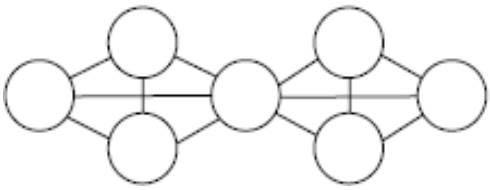
- *Σταθμισμένες Κοινότητες (Weighted Communities).* Πολλά πολύπλοκα δίκτυα είναι σταθμισμένα, δηλαδή η αλληλεπίδραση μεταξύ δύο κόμβων δεν χαρακτηρίζεται μόνο από την ύπαρξη ενός συνδέσμου αλλά και από έναν σύνδεσμο (ακμή) ο οποίος χαρακτηρίζεται από μία ακέραια συνήθως τιμή που λέγεται βάρος. Υπάρχει ένας

μεγάλος αριθμός παραδειγμάτων τα οποία παρέχουν πληθώρα στοιχείων ότι τα βάρη θα πρέπει να συμπεριλαμβάνονται σε τέτοιες αναλύσεις. Παραδείγματα είναι η ύπαρξη ισχυρών και ασθενών δεσμών μεταξύ των ατόμων στα κοινωνικά δίκτυα, ανομοιόμορφες ροές σε μεταβολικά μονοπάτια αντίδρασης, η πολυμορφία στις αλληλεπιδράσεις των θηρευτών-θηραμάτων στα τροφικά πλέγματα, διαφορετικές δυνατότητες στη μετάδοση ηλεκτρικών σημάτων σε νευρωνικά δίκτυα, η άνιση κυκλοφορία στο Διαδίκτυο ή των επιβατών σε αεροπορικά δίκτυα. Αυτά τα συστήματα μπορούν να περιγραφούν καλύτερα μέσω σταθμισμένων δικτύων. Σε πολλές περιπτώσεις τα βάρη επηρεάζουν σημαντικά τις ιδιότητες ή την λειτουργία αυτών των δικτύων. Ένα σταθμισμένο δίκτυο μπορεί να αναπαρασταθεί μαθηματικά μέσω ενός πίνακα γειτνίασης με τιμές που είναι ίσες με τα βάρη των ακμών. Για παράδειγμα, το παρακάτω σταθμισμένο δίκτυο μπορεί να αναπαρασταθεί με τον εξής πίνακα γειτνίασης:

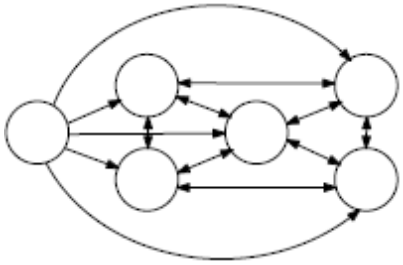


.

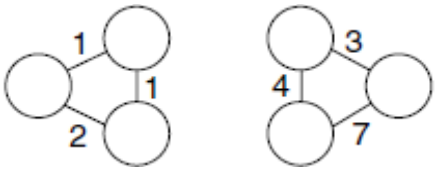
Είναι φυσικό να αναμένει κανείς ότι τα βάρη έχουν επίσης επιρροή στον σχηματισμό των κοινοτήτων, ένα θέμα που απασχολεί τη μελέτη που γίνεται. Μία ομάδα συνδεδεμένων κόμβων μπορεί να θεωρείται ως μια κοινότητα μόνο αν τα βάρη των ακμών είναι αρκετά ισχυρά, δηλαδή είναι πάνω από ένα δεδομένο όριο. Στην περίπτωση του Σχήματος 2.6 (γ), η αριστερή ομάδα μπορεί να μην είναι αρκετά ισχυρή ώστε να σχηματίσει μια κοινότητα.



(α) Overlapping Communities



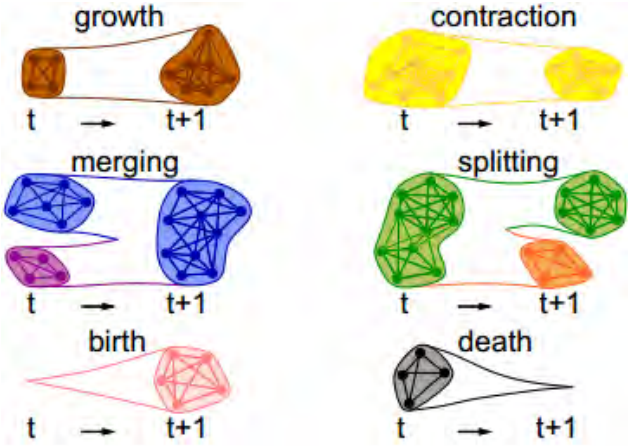
(β) Directed Community



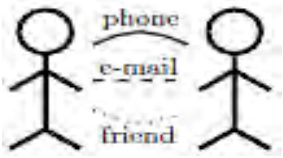
(γ) Weighted Communities

Σχήμα 2.6: Διάφορα χαρακτηριστικά κοινοτήτων.

- *Δυναμικές Κοινότητες (Dynamic Communities).* Η ανάλυση των δυναμικών κοινοτήτων βρίσκεται ακόμη σε πρώιμο στάδιο. Οι μελέτες σε αυτή την κατεύθυνση έχουν παρεμποδιστεί ως επί το πλείστον από το γεγονός ότι το πρόβλημα της συσταδοποίησης ενός γραφήματος είναι αμφιλεγόμενο για υλοποιήσεις σε ένα γράφημα. Έτσι γίνεται κατανοητό ότι οι περισσότερες προσπάθειες εξακολουθούν να επικεντρώνονται στην «στατική» εκδοχή ενός προβλήματος˙ όπου το γράφημα είναι σταθερό και δεν χρειάζεται να αλλάζει κατά την εξέταση ενός προβλήματος. Ακολουθώντας την αναπαράσταση του προβλήματος που παρουσιάστηκε στην Ενότητα 2.1, σε προβλήματα με δυναμικές κοινότητες συμβαίνει το εξής παράδοξο (άμα το ακούσει κάποιος για πρώτη φορά): ένα σύνολο ακμών έχει τη δυνατότητα να εμφανίζεται και να εξαφανίζεται κατά το δοκούν. Έτσι, οι κοινότητες θα μπορούσαν να εξελίσσονται με την πάροδο του χρόνου. Είναι χρήσιμο να ερευνηθεί πως οι κοινότητες δημιουργούνται, εξελίσσονται και πεθαίνουν. Τα κύρια φαινόμενα που εμφανίζονται κατά τη διάρκεια της ζωής μιας κοινότητας είναι (Σχήμα 2.7): γέννηση (birth), ανάπτυξη (growth), συρρίκνωση (contraction), συγχώνευση με άλλες κοινότητες (merging), διάσπαση (splitting), θάνατος (death).



Σχήμα 2.7: Πιθανά σενάρια στην εξέλιξη των κοινοτήτων [14].



Σχήμα 2.8: Παράδειγμα πολυδιάστατου δικτύου [20].

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.

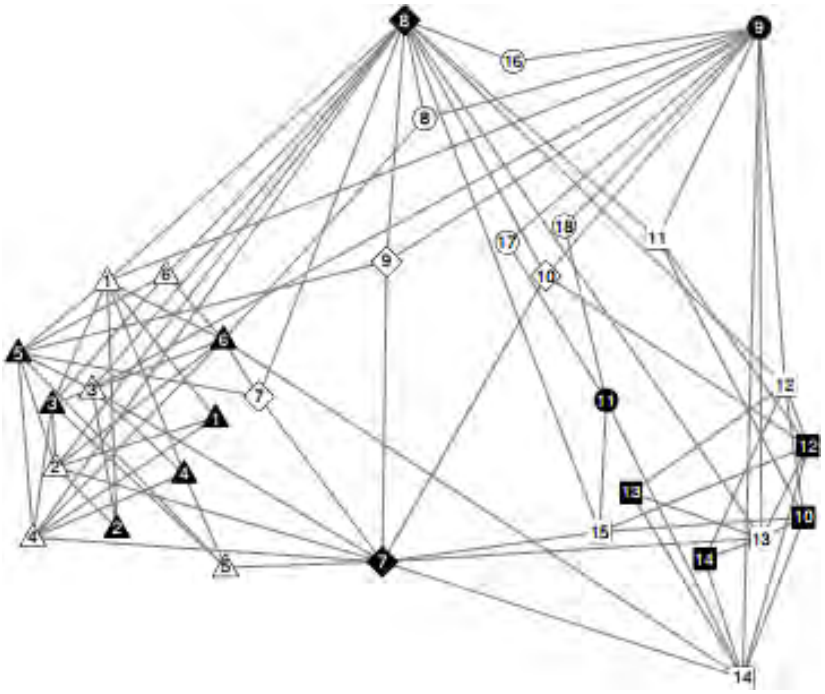
- *Απουσία Παραμέτρων (Parameter free).* Ένα επιθυμητό χαρακτηριστικό ενός αλγορίθμου, ιδίως στο τομέα της Εξόρυξης Δεδομένων, είναι η απουσία των παραμέτρων. Με άλλα λόγια, ένας αλγόριθμος πρέπει να είναι σε θέση να καθιστά σαφή τη γνώση που είναι κρυμμένη μέσα στα δεδομένα χωρίς να χρειάζεται περαιτέρω πληροφορία από τον αναλυτή σχετικά με τα δεδομένα ή το πρόβλημα (για παράδειγμα, ο αριθμός των κοινοτήτων).

- *Πολυδιάστατη Είσοδος (Multidimensional input).* Η πολυδιαστατικότητα στα δίκτυα είναι ένα ανερχόμενο θέμα στις μέρες μας. Οι κόμβοι σε ένα δίκτυο μπορεί να συνδέονται με διαφορετικής φύσεως ακμές: για παράδειγμα, ένα οποιοδήποτε ζευγάρι ατόμων μπορεί να επικοινωνήσει με διάφορα εργαλεία (τηλέφωνο, e-mail, μηνύματα, κλπ), ή σε ένα κοινωνικό δίκτυο μπορεί να συνδεθεί από μία διαφορετική σχέση (να είναι φίλοι, συνεργάτες, συγγενείς, κλπ). Ένα δίκτυο ονομάζεται *πολυδιάστατο* εάν περιέχει έναν αριθμό από διαφορετικά είδη σχέσεων που αναπτύσσονται μεταξύ των κόμβων του δικτύου. Με άλλα λόγια, ένα δίκτυο λέγεται πολυδιάστατο όταν υπάρχουν πολλές πιθανές συνδέσεις (ακμές) μεταξύ του ίδιου ζεύγους οντοτήτων (κόμβων). Το Σχήμα 2.8 απεικονίζει ένα πιθανό πολυδιάστατο δίκτυο, που αποτελείται από μόνο δύο χρήστες, όπου συνδέονται με τρεις διαστάσεις («phone», «e-mail», «friend»). Αξίζει να σημειωθεί ότι ένα πολυδιάστατο δίκτυο μπορεί να περιέχει ετερογενείς διαστάσεις όπου τα εργαλεία επικοινωνίας και οι κοινωνικές σχέσεις συνυπάρχουν μαζί. Έτσι, όταν κάποιοι ασχολούνται με πολλαπλές διαστάσεις, η έννοια της κοινότητας αλλάζει. Ο Ορισμός 2.1 που δόθηκε για την κοινότητα, αποτυπώνει αυτό το πολύπλοκο περιβάλλον παριστάνοντας την δημιουργία ή την απουσία μιας συγκεκριμένης ακμής σε μια συγκεκριμένη διάσταση με μία ενέργεια. Αυτή η έννοια της πολυδιαστατικότητας χρησιμοποιείται (με διάφορα ονόματα: πολλαπλών σχέσεων, πολυπλεξίας, και ούτω καθ' εξής) από ορισμένες προσεγγίσεις ως ένα χαρακτηριστικό της εισόδου.

- *Στοιχειώδης Προσέγγιση (Incremental).* Ένα άλλο επιθυμητό χαρακτηριστικό ενός αλγορίθμου είναι η ικανότητά του να παρέχει μία έξοδο χωρίς να πραγματοποιεί διεξοδική αναζήτηση ολόκληρης της εισόδου. Μια στοιχειώδης προσέγγιση για τον εντοπισμό κοινότητας είναι να καταταγεί ένας κόμβος σε μια κοινότητα κοιτώντας τη γειτονιά του κόμβου. Εναλλακτικά, νεοεισερχόμενοι κόμβοι τοποθετούνται σε μία από τις προγούμενες ορισθείσες κοινότητες χωρίς να ξεκινήσει η διαδικασία εντοπισμού της κοινότητας από την αρχή.

- *Πολυμερής Είσοδος (Multipartite input).* Πολλές προσεγγίσεις στον εντοπισμό κοινοτήτων λειτουργούν ακόμη και αν το δίκτυο έχει τη μορφή ενός πολυμερούς γραφήματος. Δεν είναι ασυνήθιστο να υπάρχουν δίκτυα με διαφορετικές κατηγορίες κορυφών, και ακμές που συνδέουν κορυφές μόνο τέτοιων διαφορετικών κατηγοριών. Για ένα πολυμερές δίκτυο η έννοια της κοινότητας δεν αλλάζει πολύ σε σχέση με την περίπτωση των κλασικών μονομερών γραφημάτων, καθώς παραμένει συσχετισμένη με μια μεγάλη πυκνότητα των ακμών μεταξύ των μελών της ίδιας ομάδας, με τη μόνη διαφορά ότι τα στοιχεία της κάθε ομάδας ανήκουν σε διαφορετικές κατηγορίες κορυφών. Το πολυμερές γράφημα, ωστόσο, δεν είναι εξ' ολοκλήρου ένα χαρακτηριστικό της εισόδου που ίσως να ήθελαν κάποιοι να εξετάσουν για την έξοδο. Πολλοί αλγόριθμοι χρησιμοποιούν συχνά μία (συνήθως) διμερής προβολή (ακόμη και μονομερής) ενός πολυμερούς γραφήματος προκειμένου να γίνουν αποδοτικότεροι υπολογισμοί. Βλέπε Σχήμα 2.9. Όπως και στην περίπτωση της πολυδιαστατικότητας, αυτός είναι ο λόγος για την ενσωμάτωση της πολυμερής εισόδου ως χαρακτηριστικό της προσέγγισης και όχι της εξόδου. Ο εντοπισμός κοινοτήτων σε πολυμερή δίκτυα μπορεί να έχει ενδιαφέρουσες εφαρμογές, π.χ. στο μάρκετινγκ. Τα μεγάλα δίκτυα αγορών, στα οποία οι πελάτες συνδέονται με τα προϊόντα που έχουν αγοράσει, επιτρέπουν να κατηγοριοποιηθούν οι πελάτες με βάση τους τύπους του προϊόντος που αγοράζουν πιο συχνά: αυτό θα μπορούσε να χρησιμοποιηθεί τόσο για την οργάνωση στοχευμένης διαφήμισης, καθώς και να διατυπώσει συστάσεις σχετικά με μελλοντικές αγορές. Το Σχήμα 2.10 απεικονίζει το διάσημο διμερές δίκτυο των Νότιων Γυναικών (Southern Women). Υπάρχουν 32 κορυφές, που αντιπροσωπεύουν 18 γυναίκες από την περιοχή Natchez του Μισισιπή, και 14 κοινωνικές εκδηλώσεις. Οι ακμές αντιπροσωπεύουν τη συμμετοχή των γυναικών στα γεγονότα. Από το σχήμα μπορεί κανείς να δει ότι το δίκτυο έχει μια σαφή δομή κοινότητας.

Σχήμα 2.9: Αριστερά: διμερές γράφημα. Δεξιά: η προβολή του σε μονομερές γράφημα ,όπου δύο κορυφές είναι συνδεδεμένες εάν είχαν τουλάχιστον έναν κοινό γείτονα στο διμερές γράφημα.



Σχήμα 2.10: Δομή κοινότητας σε πολυμερή δίκτυα. Αυτό το διμερές γράφημα αναφέρεται στο παράδειγμα των Νότιων Γυναικών (Southern Women). Οι γυναίκες παρουσιάζονται με τις λευκές κορυφές, ενώ οι εκδηλώσεις με τις μαύρες [19].

Βιβλιογραφία

[1] http://en.wikipedia.org/wiki/Complex_network.

[2] Steven H. Strogatz, Exploring complex networks.

[3] Eric D. Kelsic, Understanding complex networks with community finding algorithms, SURF 2005 Final Report California Institute of Technology, Pasadena, CA 91126, USA.

[4] Martin Rosvall and Carl T. Bergstrom, Maps of random walks on complex networks reveal community structure.

[5] http://en.wikipedia.org/wiki/Cosine_similarity.

[6] Lei Tang, Huan Liu, School of Computing Informatics, Arizona State University, Jianping Zhang, Zohreh Nazeri, The MITRE Corporation, Community Evolution in Dynamic Multi-Mode Networks.

[7] Reihaneh Rabbany Khorasgani, Jiyang Chen, Osmar R. Zaïane, Department of Computing Science, University of Alberta, Top Leaders Community Detection Approach in Information Networks.

[8] Sanjeev Arora, Rong Ge, Sushant Sachdeva, Grant Schoenebeck, Finding Overlapping Communities in Social Networks: Toward a Rigorous Approach.

[9] JIERUI XIE, STEPHEN KELLEY, BOLESLAW K. SZYMANSKI, Overlapping Community Detection in Networks: the State of the Art and Comparative Study.

[10] Gergely Palla, Imre Derenyi, Illes Farkas , and Tamas Vicsek, Uncovering the overlapping community structure of complex networks in nature and society.

[11] Xiang-Sun Zhang, Community Identification of Complex Network, Chinese Academy of Sciences.

[12] Riitta Toivonen, Jussi M. Kumpula, Jari Saramaki, Jukka-Pekka Onnela, Janos Kertesz, and Kimmo Kaski, The role of edge weights in social networks: modelling structure and dynamics.

[13] M. E. J. Newman, Analysis of weighted networks, Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109 and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501.

[14] R. Lambiotte, Trade, Conflict and Sentiments: Multi-relational Organisation of Large-scale Social Networks, Institute for Mathematical Sciences Imperial College London.

[15] Lior Rokach, Department of Industrial Engineering Tel-Aviv University, Oded Maimon Department of Industrial Engineering Tel-Aviv University, CLUSTERING METHODS.

[16] Amit Goyal, University of British Columbia, Vancouver, BC, Canada, Francesco Bonchi, Barcelona, Spain, Laks V. S. Lakshmanan, University of British Columbia, Vancouver, BC, Canada, Discovering Leaders from Community Actions.

[17] Liaoruo Wang and John Hopcroft, Community Structure in Large Complex Networks.

[18] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics.

[19] S. Fortunato, Community detection in graphs, Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133, Torino, I-ITALY.

[20] Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, Dino Pedreschi, Foundations of Multidimensional Network Analysis.

[21] R. Lambiotte, Community detection in complex networks, Department of Mathematics, University of Namur.

[22] Anil K. Jain, Richard C. Dubes, Michigan State University, Algorithms for Clustering Data.

ΚΕΦΑΛΑΙΟ 3 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΜΕ ΒΑΣΗ ΤΟΝ ΟΡΙΣΜΟ

3.1 Εισαγωγή

Σε αυτό το Κεφάλαιο επανεξετάζονται οι προσεγγίσεις όσον αφορά τον εντοπισμό κοινοτήτων σε πολύπλοκα δίκτυα. Σε κάθε ενότητα ομαδοποιούνται μαζί όλοι οι αλγόριθμοι που μοιράζονται τον ίδιο ορισμό για το τι είναι μία κοινότητα. Δηλαδή, πληρούνται οι ίδιες προϋποθέσεις από μία ομάδα οντοτήτων που τις επιτρέπουν να είναι συγκεντρωμένες μαζί σε μία κοινότητα.

Η κατηγοριοποίηση αυτή είναι η κύρια συνεισφορά της συγκεκριμένης διπλωματικής, και που θα πρέπει από δω και στο εξής να βοηθά οποιονδήποτε θελήσει να εξοικειωθεί περισσότερο με τους αλγορίθμους συσταδοποίησης ενός γραφήματος. Αυτό θα γίνει αποκαλύπτοντας μία πρακτική πλευρά για αυτούς τους αναλυτές που επιδιώκουν ακριβή αποτελέσματα σε αναλυτικά προβλήματά τους.

Οι προτεινόμενες κατηγορίες είναι οι ακόλουθες:

- *Χαρακτηριστικό Απόσταση (Feature Distance – Κεφάλαιο 4).* Εδώ έχουν συλλεχθεί όλες οι προσεγγίσεις που αφορούν τον εντοπισμό κοινότητας που ξεκινούν από την παραδοχή ότι η κοινότητα αποτελείται από οντότητες οι οποίες μοιράζονται ένα σύνολο επακριβών χαρακτηριστικών, με παρόμοιες τιμές (δηλαδή, ορίζοντας ως μέτρο στα χαρακτηριστικά τους την απόσταση, οι οντότητες είναι όλες σε κοντινή απόσταση μεταξύ τους). Ένα κοινό χαρακτηριστικό μπορεί να αποτελέσει μία ακμή ή ένα οποιοδήποτε γνώρισμα που συνδέεται με την οντότητα (στον ορισμό του προβλήματός: η ενέργεια). Συνήθως, αυτές οι προσεγγίσεις προτείνουν τον ορισμό της κοινότητας προκειμένου να εφαρμοστούν κλασικές τεχνικές συσταδοποίησης της Εξόρυξης Δεδομένων, όπως είναι ο αλγόριθμος του Μήκους Ελάχιστης Περιγραφής (Minimum Description Length - MDL).

- *Εσωτερική Πυκνότητα (Internal Density – Κεφάλαιο 5).* Στην ενότητα αυτή εξετάζονται οι πιο σημαντικές προσεγγίσεις που ορίζουν τον εντοπισμό κοινότητας ως μία διαδικασία που καθοδηγείται από τον άμεσο εντοπισμό των πυκνότερων περιοχών του δικτύου.

- *Συνδυαστική Συσταδοποίηση (Combined Clustering – Κεφάλαιο 6).* Σε αρκετές περιπτώσεις όπου μελετούνται αλγόριθμοι εντοπισμού κοινοτήτων, δεν υπάρχουν ξεκάθαροι ορισμοί χαρακτηριστικών της κοινότητας που πρέπει να διερευνηθούν. Αντ' αυτού, καθορίζονται διάφορες λειτουργίες και αλγόριθμοι προκειμένου να συνδυάσουν τα αποτελέσματα των διαφόρων προσεγγίσεων στον εντοπισμό κοινότητας και στη συνέχεια να χρησιμοποιηθούν οι κατάλληλοι ορισμοί ώστε να παραχθεί το επιθυμητό αποτέλεσμα. Εναλλακτικά, δίνεται η δυνατότητα στον αναλυτή να ορίσει αυτός το πως αντιλαμβάνεται την κοινότητα και να την αναζητήσει μέσα στο γράφημα.

Ενδιαφέρουσες κατηγορίες και χρήσιμες προς μελέτη αποτελούν και οι εξής:

- *Ανίχνευση Γεφυρών (Bridge Detection).* Εδώ περιλαμβάνονται οι προσεγγίσεις του εντοπισμού κοινότητας που βασίζονται στην ιδέα ότι οι κοινότητες είναι τα πυκνά μέρη του γραφήματος μεταξύ των οποίων υπάρχουν πολύ λίγες ακμές που μπορούν να «σπάσουν» το δίκτυο σε κομμάτια εάν αφαιρεθούν. Αυτές οι ακμές αποκαλούνται «γέφυρες» και οι συνιστώσες του δικτύου που προκύπτουν από την αφαίρεση των ακμών είναι οι επιθυμητές κοινότητες.

- *Διάχυση (Diffusion).* Εδώ συμπεριλαμβάνονται προσεγγίσεις στο θέμα του εντοπισμού κοινότητας που βασίζονται στην ιδέα ότι οι κοινότητες είναι ομάδες κόμβων που επηρεάζονται από τη διάχυση-διάδοση μιας συγκεκριμένης πληροφορίας στο εσωτερικό του δικτύου. Επίσης, ο ορισμός της κοινότητας μπορεί να περιοριστεί στις ομάδες που επηρεάζονται μόνο από το ίδιο ακριβώς σύνολο πηγών διάχυσης.

- *Εγγύτητα (Closeness).* Μία κοινότητα μπορεί επίσης να είναι ορισμένη ως μία ομάδα οντοτήτων όπου οποιαδήποτε από αυτές μπορεί να φτάσει τον καθένα από τους «συντρόφους» της μέσα στην κοινότητα με πολύ λίγα βήματα κατά μήκος των ακμών του γραφήματος, ενώ οι οντότητες που βρίσκονται έξω από τη κοινότητα είναι αρκετά μακριά από αυτές.

- *Δομή (Structure).* Μία άλλη προσέγγιση για τον εντοπισμό κοινότητας είναι να οριστεί η κοινότητα επακριβώς ως μία πολύ ακριβής και σχεδόν αμετάβλητη δομή ακμών. Συχνά αυτές οι δομές ορίζονται ως ένας συνδυασμός μικρότερων μοτίβων δικτύου. Οι αλγόριθμοι που βασίζονται σε αυτή τη προσέγγιση αρχικά προσδιορίζουν ορισμένα είδη δομών και έπειτα προσπαθούν αποτελεσματικά να τα βρουν μέσα στο γράφημα.

- *Συσταδοποίηση Συνδέσεων (Link Clustering).* Η κατηγορία αυτή μπορεί να θεωρηθεί ως μία προβολή του προβλήματος εντοπισμού κοινοτήτων. Αντί της συσταδοποίησης των κόμβων ενός δικτύου, οι προσεγγίσεις τέτοιου τύπου αναφέρουν ότι το σημαντικό είναι

οι σχέσεις και οι συνδέσεις που ανήκουν σε μία κοινότητα, και όχι οι κόμβοι. Ως εκ τούτου, ομαδοποιούνται οι ακμές του δικτύου, συνεπώς οι κόμβοι φέρονται να ανήκουν στο σύνολο των κοινοτήτων των ακμών τους.

Σε κάθε Κεφάλαιο αποσαφηνίζονται ποια χαρακτηριστικά, σε μία συγκεκριμένη κατηγορία εντοπισμού κοινότητας από αυτές που παρουσιάστηκαν σε προηγούμενη ενότητα, προκύπτουν φυσιολογικά, και ποια είναι εκ των πραγμάτων δύσκολο να επιτευχθούν. Δεν δημιουργείται επίσημα μία αξιωματική προσέγγιση, όπως αυτή που περιγράφεται στο άρθρο [1] για τη χωρική συσταδοποίηση. Αντ' αυτού, χρησιμοποιούνται τα χαρακτηριστικά και είναι χρήσιμο να κοιτάξει κανείς ορισμένα πειράματα, για να καταστούν πιο σαφή το σκεπτικό και οι ιδιότητες της κάθε κατηγορίας σε αυτή την κατηγοριοποίηση.

Όπου είναι δυνατόν, παρουσιάζεται ένα απλό γραφικό παράδειγμα του ορισμού που εξετάζεται εκείνη τη χρονική στιγμή. Αυτό το γραφικό παράδειγμα παρέχει τις κύριες ιδιότητες της κατηγοριοποίησης που περιγράφεται, όσον αφορά τα ισχυρά και αδύνατα σημεία, σε συγκεκριμένα χαρακτηριστικά της κοινότητας.

Ο σκοπός αυτής της διπλωματικής εργασίας είναι να επικεντρωθεί αφενός στις πιο πρόσφατες και σημαντικές προσεγγίσεις και αφετέρου στους πιο γενικούς ορισμούς της κοινότητας. Δεν θα επικεντρωθεί σε ιστορικές προσεγγίσεις. Μερικά παραδείγματα κλασικών αλγορίθμων συσταδοποίησης που δεν έχουν εκτενώς αξιολογηθεί είναι ο Kernighan–Lin αλγόριθμος [2] (του οποίου ψευδοκώδικα και ένα τρέξιμο μπορείτε να βρείτε στο άρθρο [3]) ή η κλασική προσέγγιση της φασματικής διχοτόμησης. Μία ιδιαίτερα δημοφιλής ιστορική προσέγγιση στον εντοπισμό κοινότητας είναι το blockmodeling. Σκοπός του blockmodeling είναι να δημιουργήσει ένα διαιρεμένο μοντέλο του δικτύου εντοπίζοντας «blocks» (κοινότητες) στο εσωτερικό της δομής που συμπυκνώνονται σε μία ενιαία, ομοιογενή, λειτουργική μονάδα. Σε ορισμένες περιπτώσεις σε αυτή την εργασία, αυτή η οικογένεια των προσεγγίσεων παρουσιάζεται συνοπτικά. Για τις περιπτώσεις που δεν καλύπτονται από αυτή την εργασία, ανατρέξτε στο βιβλίο [5].

3.2 Η Επικάλυψη στην Κατηγοριοποίηση

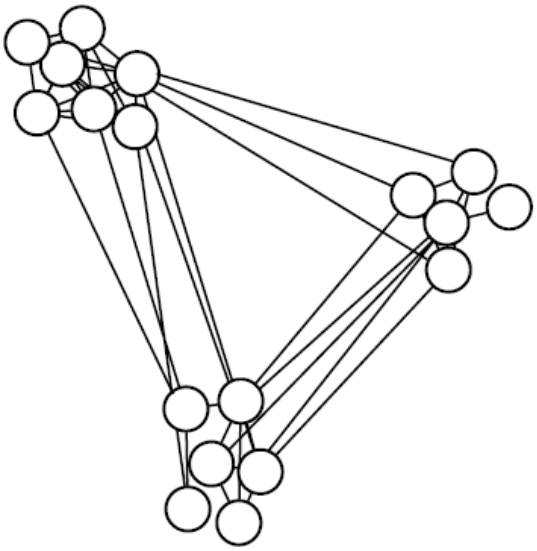
Σε μερικούς από τους ορισμούς της κοινότητας παρατηρείται το φαινόμενο της επικάλυψης. Για παράδειγμα, ένας ορισμός της εσωτερικής πυκνότητας μπορεί να περιλαμβάνει κοινότητες με αραιές εξωτερικές συνδέσεις, δηλαδή, γέφυρες. Στο Κεφάλαιο 5 θα δούμε πως σε αυτόν τον ορισμό η λέξη κλειδί είναι ο σπονδυλωτής (modularity). Ο σπονδυλωτής είναι μία συνάρτηση η οποία εξετάζει τόσο την εσωτερική πυκνότητα μιας κοινότητας, όσο και την απουσία ακμών μεταξύ των κοινοτήτων. Έτσι, οι μέθοδοι που βασίζονται στον σπονδυλωτή θα μπορούσαν να ομαδοποιηθούν σε δύο κατηγορίες. Ωστόσο, ο βαθύτερος ορισμός του σπονδυλωτή

επικεντρώνεται στην εσωτερική πυκνότητα, η οποία είναι και η αιτία για την προτεινόμενη κατηγοριοποίηση. Για να δώσουμε ένα άλλο παράδειγμα, μια προσέγγιση που αφορά τη διάχυση ενδέχεται να εντοπίσει τις ίδιες κοινότητες των οποίων τα μέλη μπορούν να προσεγγίσουν ο ένας τον άλλον μόνο σε λίγα βήματα. Ωστόσο, η προσέγγιση με την διάχυση μπορεί επίσης να βρει κοινότητες που να απαρτίζονται από μέλη που απέχουν μεταξύ τους με μία αυθαίρετη απόσταση.

Πολλές προσεγγίσεις στη βιβλιογραφία δεν ορίζουν επακριβώς τις κοινότητες που θέλουν να εντοπίσουν ή, ακόμη χειρότερα, υποστηρίζουν γενικά ότι σκοπός τους είναι να βρίσκουν τις πυκνές περιοχές του δικτύου. Αυτό δεν αποτελεί πρόβλημα, καθώς ο ακριβής ορισμός της κοινότητας μπορεί να προκύψει από μία ενδελεχή μελέτη της προσέγγισης που περιγράφεται στην παρούσα εργασία. Κανείς δεν έχει απαίτηση από τους ερευνητές να μπορούν να κατηγοριοποιούν τις μεθόδους τους πριν από την αποδοχή μιας καθιερωμένης κατηγοριοποίησης. Ένας από τους στόχους της συγκεκριμένης εργασίας είναι να να υποκινήσει μια συζήτηση σχετικά με το θέμα αυτό. Από τη στιγμή που θα μελετηθεί και θα υπάρχει ακόμη περισσότερη γνώση στο τομέα αυτό, ο κάθε συγγραφέας θα έχει τη δυνατότητα να κατηγοριοποιήσει σωστά την προσέγγισή του.

Προκειμένου να παρατηρηθούν οι εμφανείς διαφορές μεταξύ των προτεινόμενων κατηγοριών, θεωρήστε τα Σχήματα 3.1 και 5.1. Τα σχήματα αυτά απεικονίζουν τις πιο απλές τυπικές κοινότητες που έχουν εντοπιστεί από τους ορισμούς του χαρακτηριστικού απόστασης και της εσωτερικής πυκνότητας. Όπως μπορεί να φανεί, υπάρχουν ορισμένες διαφορές μεταξύ αυτών των δύο παραδειγμάτων.

Η επικάλυψη γενικά συμβαίνει λόγω του γεγονότος ότι αρκετοί αλγόριθμοι λειτουργούν με μερικούς γενικούς ορισμούς της κοινότητας. Οι κατηγορίες που προτείνονται εδώ μπορούν να ομαδοποιηθούν μαζί σε μία ιεραρχία με τις τέσσερις κύριες κατηγορίες που περιγράφονται στην Ενότητα 2.2.



Σχήμα 3.1: Γράφημα που μπορεί να χωριστεί σύμφωνα με το χαρακτηριστικό «απόσταση» μεταξύ των κόμβων του.

Επιπλέον, πολλοί αλγόριθμοι μπορούν να παρουσιάσουν κοινές στρατηγικές στην εξερεύνηση του χώρου αναζήτησης ή στην αξιολόγηση της ποιότητας της διαμέρισης που κάνουν προκειμένου να επιτύχουν βελτίωση πάνω σε αυτό. Θεωρήστε, για παράδειγμα, τις εξής δύο δημοσιεύσεις˙ [6] και [7]. Σε αυτές τις δύο δημοσιεύσεις υπάρχει μια διεξοδική θεωρητική μελέτη σχετικά με το modularity και την πιο γενική μορφή αυτού. Στην πρώτη δημοσίευση, για παράδειγμα, οι συγγραφείς ήταν σε θέση να ορίσουν το modularity με μία τυχαία στρατηγική εξερεύνηση με τα πόδια, αναδεικνύοντας έτσι την επικάλυψή του με έναν από τους αλγορίθμους που είναι ομαδοποιημένοι στην βιβλιογραφία στην κατηγορία «Εγγύτητα» (Closeness) [9].

Η αξιολόγηση της επικάλυψης και των σχέσεων μεταξύ των σημαντικότερων προσεγγίσεων στον εντοπισμό κοινοτήτων δεν είναι υπόθεση απλή, και είναι έξω από το πεδίο δράσης της εν λόγω εργασίας. Εδώ εστιάζουμε στη σχέση μεταξύ ενός αλγορίθμου και σε έναν συγκεκριμένο ορισμό της κοινότητας. Έτσι, δημιουργείται μία χρήσιμη και υψηλού επιπέδου κατηγοριοποίηση για την σύνδεση των αναγκών συγκεκριμένων αναλύσεων (δηλαδή, οι ορισμοί της κοινότητας) με τα εργαλεία που υπάρχουν διαθέσιμα στη βιβλιογραφία.

Βιβλιογραφία

[1] Jon Kleinberg, An Impossibility Theorem for Clustering, Department of Computer Science, Cornell University, Ithaca NY 14853.

[2] B.W. Kernighan, S. Lin, An Efficient Heuristic Procedure for Partioning Graphs.

[3] http://users.eecs.northwestern.edu/~haizhou/357/lec2.pdf.

[4] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, Eric P. Xing, Mixed Membership Stochastic Blockmodels.

[5] P. Doreian, V. Batagelj, and A. Ferligoj, Generalized Blockmodeling.

[6] R. Lambiotte, J.-C. Delvenne and M. Barahona, Laplacian Dynamics and Multiscale Modular Structure in Networks, October 9, 2009.

[7] Jörg Reichardt and Stefan Bornholdt, Statistical mechanics of community detection.

[8] R. Lambiotte, Community detection in complex networks, Department of Mathematics, University of Namur.

[9] Τσιτσιγιάννης Εμμανουήλ, Αλγόριθμοι Εντοπισμού Κοινοτήτων, Φεβρουάριος 2014.

ΚΕΦΑΛΑΙΟ 4 ΤΟ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΤΑΣΗ

4.1 Εισαγωγή

Σε αυτό το Κεφάλαιο παρουσιάζονται οι μέθοδοι εντοπισμού κοινοτήτων που ορίζουν μία κοινότητα σύμφωνα με τον παρακάτω ορισμό:

Ορισμός 4.1 (Κοινότητα με κάποιο χαρακτηριστικό γνώρισμα). Μία κοινότητα που χαρακτηρίζεται από ένα συγκεκριμένο γνώρισμα μέσα σ' ένα πολύπλοκο δίκτυο, αποτελείται από ένα σύνολο οντοτήτων που μοιράζονται ένα συγκεκριμένο σύνολο χαρακτηριστικών (συμπεριλαμβανομένης της ακμής ως ένα χαρακτηριστικό γνώρισμα). Ορίζοντας ως γνώρισμα την *απόσταση* βασισμένοι στις τιμές των χαρακτηριστικών, οι οντότητες που βρίσκονται μέσα σε μια κοινότητα είναι πολύ κοντά η μία στην άλλη, περισσότερο από ό, τι οι οντότητες έξω από αυτήν.

Αυτός ο ορισμός λειτουργεί σύμφωνα με το ακόλουθο πόρισμα:

Πόρισμα 4.1. Δεδομένου ενός συνόλου οντοτήτων και των χαρακτηριστικών τους (τα οποία μπορεί να είναι σχέσεις, ενέργειες ή ιδιότητες), αυτά μπορούν να αναπαρασταθούν με ένα διάνυσμα τιμών. Συνεπώς, πραγματοποιείται μία συσταδοποίηση με βάση τον χώρο (ή το διάνυσμα) με σκοπό να προκύψει η επιθυμητή δομή.

Χρησιμοποιώντας τον παραπάνω ορισμό, θεωρείται ότι το έργο της εύρεσης κοινοτήτων είναι παρόμοιο με το κλασικό πρόβλημα της συσταδοποίησης στην Εξόρυξη Δεδομένων. Στην Εξόρυξη Δεδομένων, η συσταδοποίηση είναι ένας μη εποπτευόμενος γνωστικός τομέας. Μερικές φορές η συσταδοποίηση αναφέρεται και ως *μη εποπτευόμενη κατηγοριοποίηση (unsupervised classification)*. Ο σκοπός ενός αλγορίθμου συσταδοποίησης είναι να αντιστοιχίσει ένα μεγάλο σύνολο από δεδομένα σε ομάδες (συστάδες) έτσι ώστε τα δεδομένα που βρίσκονται στις ίδιες συστάδες να μοιάζουν περισσότερο μεταξύ τους σε σχέση με άλλα δεδομένα που ανήκουν σε οποιαδήποτε άλλη συστάδα. Όσο πιο μεγάλη είναι η ομοιότητα εντός μιας συστάδας και όσο πιο μεγάλη είναι η διαφορά μεταξύ των συστάδων, τόσο πιο καλή (ή αλλιώς διακριτή) είναι η συσταδοποίηση. Η ομοιότητα που αναφέρεται παραπάνω καθορίζεται από το χαρακτηριστικό γνώρισμα της απόστασης, και συνήθως βασίζεται στον

αριθμό των κοινών ιδιοτήτων που έχουν οι οντότητες, ή στις παρόμοιες τιμές αυτών των ιδιοτήτων.

Ένα παράδειγμα των τεχνικών της συσταδοποίησης είναι ο **αλγόριθμος των Κ – μέσων** (K – means algorithm). Ο αλγόριθμος των Κ – μέσων ορίζει ένα πρότυπο σε σχέση με μία τιμή κέντρου βάρους, η οποία είναι συνήθως ο μέσος μιας ομάδας σημείων, και τυπικά εφαρμόζεται σε αντικείμενα εντός ενός συνεχούς n διαστάσεων χώρου. Η τεχνική συσταδοποίησης των Κ – μέσων είναι απλή, και η αρχή της παρουσίασης γίνεται με μία περιγραφή του βασικού αλγορίθμου. Στην αρχή, επιλέγονται Κ αρχικά κέντρα βάρους, όπου το Κ είναι μία παράμετρος ορισμένη από το χρήστη, συγκεκριμένα, το πλήθος των επιθυμητών συστάδων. Κάθε σημείο στη συνέχεια αποδίδεται στο πιο κοντινό κέντρο βάρους, και κάθε σύνολο σημείων που αποδίδεται σε ένα κέντρο βάρους συνιστά μία συστάδα. Στη συνέχεια, το κέντρο βάρους κάθε συστάδας ενημερώνεται με βάση τα σημεία που αποδίδονται στη συστάδα. Τα βήματα της εκχώρησης και της ενημέρωσης επαναλαμβάνονται μέχρι να μην υπάρχει σημείο που να αλλάζει συστάδα, ή ισοδύναμα, μέχρι τα κέντρα βάρους να παραμένουν σταθερά. Ο αλγόριθμος των Κ – μέσων περιγράφεται τυπικά από τον Αλγόριθμο 4.1. Η λειτουργία του αλγορίθμου φαίνεται στο Σχήμα 4.1, το οποίο δείχνει τον τρόπο με τον οποίο, ξεκινώντας από τρία κέντρα βάρους (Κ = 3), οι τελικές συστάδες βρίσκονται σε τέσσερα βήματα εκχώρησης/ενημέρωσης. Σε αυτό, κάθε υπόγραφος δείχνει: 1) τα κέντρα βάρους στην αρχή κάθε επανάληψης και 2) την εκχώρηση των σημείων σε αυτά τα κέντρα βάρους. Τα κέντρα βάρους αναπαρίστανται με το σύμβολο της πρόσθεσης «+». Όλα τα σημεία που ανήκουν στην ίδια συστάδα έχουν το ίδιο σύμβολο-σχήμα.

Αλγόριθμος 4.1 Ο βασικός αλγόριθμος των Κ - μέσων

1: Επίλεξε Κ σημεία ως αρχικά κέντρα βάρους

2: **Επανέλαβε**

3: Σχημάτισε Κ συστάδες αποδίδοντας κάθε σημείο στο πλησιέστερο κέντρο βάρους του

4: Υπολόγισε ξανά το κέντρο βάρους κάθε συστάδας

5: **Μέχρι** να μην αλλάζουν τα κέντρα βάρους

Γενικά, κοιτώντας το δεδομένο γράφημα που θα δίνεται ώστε να πραγματοποιηθεί η συσταδοποίηση, θα υπάρχουν n δεδομένα σημεία x_i, i = 1...n που πρέπει να κατανεμηθούν σε Κ συστάδες. Ο σκοπός είναι να εκχωρηθεί μία συστάδα σε κάθε δεδομένο σημείο. Ο αλγόριθμος των Κ – μέσων είναι μία μέθοδος ομαδοποίησης που έχει ως στόχο να βρίσκει τις θέσεις μ_i, i = 1...Κ των συστάδων που ελαχιστοποιούν την απόσταση από τα δεδομένα σημεία στη συστάδα. Ο αλγόριθμος επιλύει την

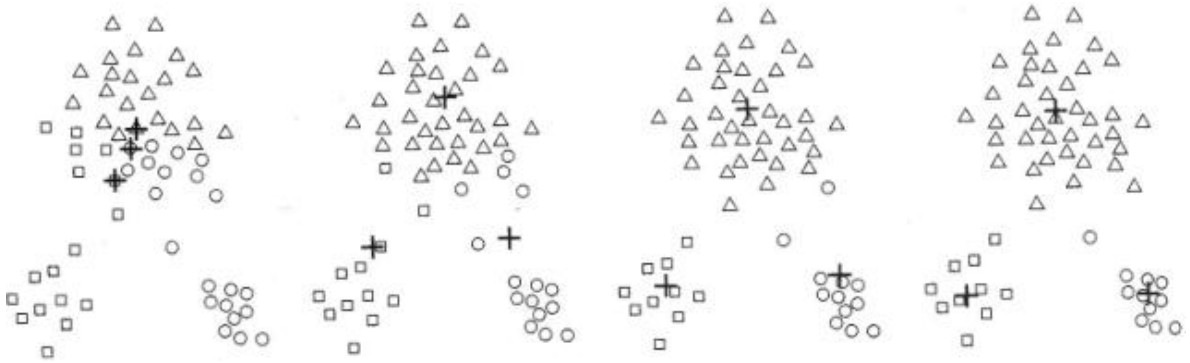
$\arg\min_c \sum_{i=1}^{K} \sum_{x \in c_i} d(x, \mu_i) = \arg\min_c \sum_{i=1}^{K} \sum_{x \in c_i} \|x - \mu_i\|_2^2$, όπου το c_i είναι το σύνολο των σημείων που ανήκουν στο συστάδα i. Ο K - μέσων αλγόριθμος χρησιμοποιεί ως μέτρο την Ευκλείδια απόσταση $d(x, \mu_i) = \|x - \mu_i\|_2^2$. Αυτό το πρόβλημα δεν είναι ασήμαντο (στην πραγματικότητα είναι NP-hard), έτσι ο αλγόριθμος ελπίζει να βρει μόνο το ολικό ελάχιστο, και ενδεχομένως να κολλήσει σε μία διαφορετική λύση.

Αρχικά αποφασίζεται ο αριθμός K των συστάδων. Έπειτα,

1. Αρχικοποιείται το κέντρο των συστάδων, μ_i = κάποια τιμή, i = 1,...,K
2. Αποδίδεται η κοντινότερη συστάδα σε κάθε δεδομένο σημείο,
   $c_i = \{ j : d(x_j, \mu_i) \leq d(x_j, \mu_l), l \neq i, j = 1,...,n \}$
3. Ορίζεται η θέση της κάθε συστάδας στο κέντρο όλων των δεδομένων σημείων που ανήκουν σε αυτή τη συστάδα, $\mu_i = \frac{1}{|c_i|} \sum_{j \in c_i} x_j$, $\forall i$
4. Επαναλαμβάνονται τα βήματα 2 – 3 έως ότου συγκλίνει ο αλγόριθμος.

Ο αλγόριθμος συγκλίνει τελικά σε ένα σημείο, αν και δεν είναι κατ' ανάγκη το ελάχιστο του αθροίσματος των τετραγώνων. Αυτό οφείλεται στο γεγονός ότι το πρόβλημα είναι μη κυρτό και ο αλγόριθμος είναι απλά ένας ευρετικός μηχανισμός, που συγκλίνει σε ένα τοπικό ελάχιστο. Ο αλγόριθμος σταματά όταν οι εκχωρήσεις δεν αλλάζουν από μία επανάληψη στην επόμενη. Σε αυτό το σημείο πρέπει να σημειωθεί ότι ο αριθμός των συστάδων θα πρέπει να ταιριάζει με τα δεδομένα. Μια λανθασμένη επιλογή του αριθμού των συστάδων θα ακυρώσει την όλη διαδικασία. Ένας εμπειρικός τρόπος για να βρεθεί ο καλύτερος αριθμός των συστάδων είναι να δοκιμάσουμε τον αλγόριθμο με διαφορετικό αριθμό συστάδων και να αποτιμήσουμε το άθροισμα των τετραγώνων που προκύπτει.

Στο πρώτο βήμα, το οποίο φαίνεται στο Σχήμα 4.1 α), τα σημεία αποδίδονται στα αρχικά κέντρα βάρους, τα οποία ανήκουν όλα στη μεγαλύτερη ομάδα σημείων. Για αυτό το παράδειγμα, χρησιμοποιείται ο μέσος ως κέντρο βάρους. Μόλις τα σημεία αποδοθούν σε ένα κέντρο βάρους, το κέντρο βάρους ενημερώνεται. Ξανά, το σχήμα για κάθε βήμα δείχνει το κέντρο βάρους στην αρχή του βήματος και την εκχώρηση των σημείων στα κέντρα βάρους. Στο δεύτερο βήμα, τα σημεία αποδίδονται στα πλησιέστερα κέντρα βάρους, τα οποία ενημερώνονται και πάλι. Στα βήματα 2, 3, και 4, τα οποία δίνονται στα Σχήματα 4.1 β), γ) και δ), αντίστοιχα, δύο από τα κέντρα βάρους μετακινούνται στις δύο μικρές ομάδες σημείων στο κάτω μέρος των σχημάτων. Όταν ο αλγόριθμος των K – μέσων, τερματίζει στο Σχήμα 4.1 δ), εξαιτίας του ότι δεν γίνονται άλλες αλλαγές, τα κέντρα βάρους έχουν προσδιορίσει τις φυσικές ομάδες σημείων.



α) 1η επανάληψη β) 2η επανάληψη γ) 3η επανάληψη δ) 4η επανάληψη

Σχήμα 4.1: Χρήση του αλγορίθμου των K – μέσων για την εύρεση τριών συστάδων (K=3).

Μια φυσική προσέγγιση συσταδοποίησης για τον εντοπισμό κοινότητας αποτελούν ορισμένες εξελίξεις του coclustering [21,22] και/ή ορισμένες φασματικές προσεγγίσεις στο πρόβλημα της συσταδοποίησης [23].

Στο άρθρο [4], υπάρχει μία έρευνα σχετικά με αλγόριθμους coclustering, ενώ στο άρθρο [24] υπάρχει μία ενδιαφέρουσα δημοσίευση για τη χωρική συσταδοποίηση. Δεδομένης της πλούσιας βιβλιογραφίας, σε αυτή τη κατηγορία προσεγγίσεων στον εντοπισμό κοινότητας είναι εύκολο να βρεθούν συστάδες που να μην έχουν σχεδόν κανένα χαρακτηριστικό που παρουσιάστηκε μέχρι τώρα. Δεδομένου του γεγονότος ότι κάθε κόμβος και κάθε ακμή αντιπροσωπεύονται από ένα σύνολο ιδιοτήτων, είναι πολύ εύκολο να ληφθούν πολυδιάστατα και πολυτμηματικά αποτελέσματα μέσω μιας απλής συσταδοποίησης σε ένα πολύπλοκο πολυδιάστατο χώρο.

Προκειμένου να καταστούν σαφή τα μειονεκτήματα αυτής της κατηγορίας, θεωρείστε το Σχήμα 3.1, το οποίο απεικονίζει ένα δίκτυο του οποίου οι κόμβοι είναι τοποθετημένοι σύμφωνα με το χαρακτηριστικό απόσταση. Αυτό το χαρακτηριστικό θα μπορούσε να εξετάσει την άμεση σύνδεση της ακμής˙ ωστόσο, αυτό δεν είναι υποχρεωτικό. Οι κόμβοι τότε είναι ομαδοποιημένοι στην ίδια κοινότητα εάν είναι κοντά στο χώρο αυτό (το οποίο μπορεί να εξαρτάται από τον αριθμό των χαρακτηριστικών που μελετούνται). Το Σχήμα 3.1 δείχνει ότι, ανάλογα με τον αριθμό των ιδιοτήτων του/της κάθε κόμβου/ακμής, η βασική δομή του γραφήματος μπορεί να χάσει τη σημασία της. Αυτό μπορεί να οδηγήσει σε αντιδιαισθητικά αποτελέσματα εάν ο αναλυτής προσπαθήσει να εμφανίσει τις ομάδες (συστάδες) κοιτάζοντας μόνο τη δομή του γραφήματος, με αποτέλεσμα την ύπαρξη πολλών ακμών μεταξύ της κοινότητας.

Σε αυτό το Κεφάλαιο εξετάζονται μερικές τεχνικές συσταδοποίησης με ορισμένα αρκετά ενδιαφέροντα χαρακτηριστικά:

- εξελικτική συσταδοποίηση (evolutionary clustering)
- RSN-BD : χρησιμοποιείται για k – χωρισμένα γραφήματα

- MRGC : τεχνική ομαδοποίησης που ασχολείται με τανυστές
- δύο προσεγγίσεις που χρησιμοποιούν σπονδυλωτή για τον εντοπισμό κρυμμένων διαστάσεων για έναν πολυδιάστατο εντοπισμό κοινότητας με έναν κατηγοριοποιητή που μεγιστοποιεί τον αριθμό των κοινών χαρακτηριστικών
- μία Bayesian προσέγγιση για την συσταδοποίηση με βάση την προβλεψιμότητα των χαρακτηριστικών για τους κόμβους που ανήκουν στην ίδια ομάδα
- μία ανάλυση των συνδέσεων που μοιράζονται χαρακτηριστικά σε ένα διμερές γράφημα οντότητα-χαρακτηριστικό.

Υπάρχουν επίσης δημοφιλείς προσεγγίσεις στον τομέα των στατιστικών, όπως είναι η mixture modeling, που εμπίπτουν σε αυτή την κατηγορία, στην οποία οι συγγραφείς συμπεραίνουν με μια log-likelihood προσέγγιση τη θέση των κόμβων σε έναν υψηλών διαστάσεων Ευκλείδιο χώρο με βάση τις τιμές των χαρακτηριστικών τους. Στη συνέχεια προκύπτει στατιστικά και η διαίρεση σε συστάδες.

Μια ενδιαφέρουσα αρχή της συσταδοποίησης είναι το Μήκος Ελάχιστης Περιγραφής (Minimum Description Length - **MDL**) [6,7]. Στην MDL αρχή η βασική ιδέα είναι ότι η οποιαδήποτε κανονικότητα στα δεδομένα (δηλαδή, κοινά χαρακτηριστικά) μπορεί να χρησιμοποιηθεί για να τα συμπιέσει, δηλαδή, για να τα περιγράψει χρησιμοποιώντας λιγότερα σύμβολα από τον αριθμό των συμβόλων που απαιτούνται για να περιγράψουν κατά γράμμα τα δεδομένα [26]. Όσο υπάρχουν περισσότερες κανονικότητες, τόσο τα δεδομένα μπορούν να συμπιεστούν. Δηλαδή, σύμφωνα με την MDL αρχή, όσο περισσότερο μπορούμε να συμπιέσουμε ένα συγκεκριμένο σύνολο δεδομένων, τόσο περισσότερο έχουμε μελετήσει και έχουμε μάθει γι' αυτό. Αυτή είναι μια πολύ ενδιαφέρουσα προσέγγιση, δεδομένου ότι, σε ορισμένες εφαρμογές, επιτρέπεται να πραγματοποιηθεί ο εντοπισμός της κοινότητας χωρίς να οριστεί καμία παράμετρος. Αφού εξεταστούν οι κλασικές προσεγγίσεις συσταδοποίησης, σε αυτή την ενότητα παρουσιάζονται επίσης τρεις βασικοί αλγόριθμοι που εφαρμόζουν μια MDL προσέγγιση στον εντοπισμό κοινοτήτων:

- Autopart : εξ’ όσων γνωρίζουμε, ο πρώτος δημοφιλής εντοπισμός κοινότητας που διατύπωσε τη θεωρία εδάφους για τον MDL εντοπισμό κοινότητας
- Συγκεκριμένου πλαισίου δένδρο συσταδοποίησης (context-specific cluster tree)
- Timefall

4.2 Εξελικτική Συσταδοποίηση (Evolutionary Clustering)

Η Εξελικτική Συσταδοποίηση έγκειται στο πρόβλημα της συσταδοποίησης δεδομένων σε βάθος χρόνου. Η Εξελικτική Συσταδοποίηση θα πρέπει να βελτιστοποιήσει ταυτόχρονα δύο δυνητικά αντικρουόμενα κριτήρια: πρώτον, η συσταδοποίηση ανά πάσα στιγμή θα πρέπει να παραμείνει πιστή όσο το δυνατόν περισσότερο στα τωρινά δεδομένα, και δεύτερον, δεν θα πρέπει να αλλάζει δραματικά από το ένα χρονικό βήμα στο επόμενο.

Εξελικτική Συσταδοποίηση [8] ονομάζεται το πρόβλημα το οποίο επεξεργάζεται χρονοσφραγισμένα δεδομένα με σκοπό να παραχθεί μία ακολουθία από συσταδοποιήσεις, δηλαδή, μία συσταδοποίηση για κάθε χρονικό βήμα του συστήματος. Κάθε συσταδοποίηση στην ακολουθία θα πρέπει να είναι παρόμοια με την συσταδοποίηση που έγινε στο προηγούμενο χρονικό βήμα, και θα πρέπει να αντικατοπτρίζει με ακρίβεια τα δεδομένα που καταφθάνουν κατά τη διάρκεια αυτού του χρονικού βήματος.

Το κύριο πλαίσιο για αυτό το πρόβλημα είναι το ακόλουθο. Κάθε μέρα, νέα δεδομένα έρχονται για την ημέρα, και πρέπει να ενσωματωθούν σε μία συσταδοποίηση. Εάν τα δεδομένα δεν αποκλίνουν από ιστορικές προσδοκίες, η συσταδοποίηση θα πρέπει να είναι «κοντά» σε αυτήν της προηγούμενης ημέρας, παρέχοντας στον χρήστη μία οικεία άποψη των νέων δεδομένων. Ωστόσο, εάν η δομή των δεδομένων αλλάζει σημαντικά, η συσταδοποίηση θα πρέπει να τροποποιηθεί ώστε να αντικατοπτρίζει τη νέα διάρθρωση.

Τα πλεονεκτήματα της Εξελικτικής Συσταδοποίησης σε σύγκριση με την παραδοσιακή συσταδοποίηση εμφανίζονται σε περιπτώσεις στις οποίες η τωρινή (δηλαδή, καθημερινά) συσταδοποίηση δαπανάται τακτικά από ένα χρήστη ή το ίδιο το σύστημα. Σε μία τέτοια τοποθέτηση, η Εξελικτική Συσταδοποίηση είναι χρήσιμη για τους ακόλουθους λόγους:

- *Συνοχή*: ο χρήστης βρίσκει την πιο οικεία συσταδοποίηση της κάθε ημέρας, και έτσι δεν θα χρειάζεται να μάθει ένα εντελώς νέο τρόπο κατηγοριοποίησης των δεδομένων. Ομοίως, οποιαδήποτε γνώση που προέρχεται από μία μελέτη των προηγούμενων συστάδων είναι πιο πιθανό να ισχύει και για μελλοντικές συστάδες.
- *Απομάκρυνση θορύβου*: παρέχοντας μία υψηλής ποιότητας και ιστορικά συνεπή συσταδοποίηση, εξασφαλίζει μεγαλύτερη ανθεκτικότητα κατά του θορύβου λαμβάνοντας σε ισχύ προηγούμενα σημεία δεδομένων.
- *Εξομάλυνση*: εάν οι ακριβείς συστάδες μετατοπίζονται με την πάροδο του χρόνου, η Εξελικτική Συσταδοποίηση παρουσιάζει τον χρήστη με μία ομαλή μετάβαση.
- *Αντιστοιχία Συστάδας*: ως παρενέργεια του συγκεκριμένου πλαισίου, γενικά είναι δυνατό να τοποθετηθούν οι σημερινές συστάδες σε αντιστοιχία με τις χθεσινές. Έτσι, ακόμη και αν η συσταδοποίηση έχει μετατοπιστεί, ο χρήστης θα εξακολουθεί να βρίσκεται μέσα στο ιστορικό πλαίσιο της Εξελικτικής Συσταδοποίησης.

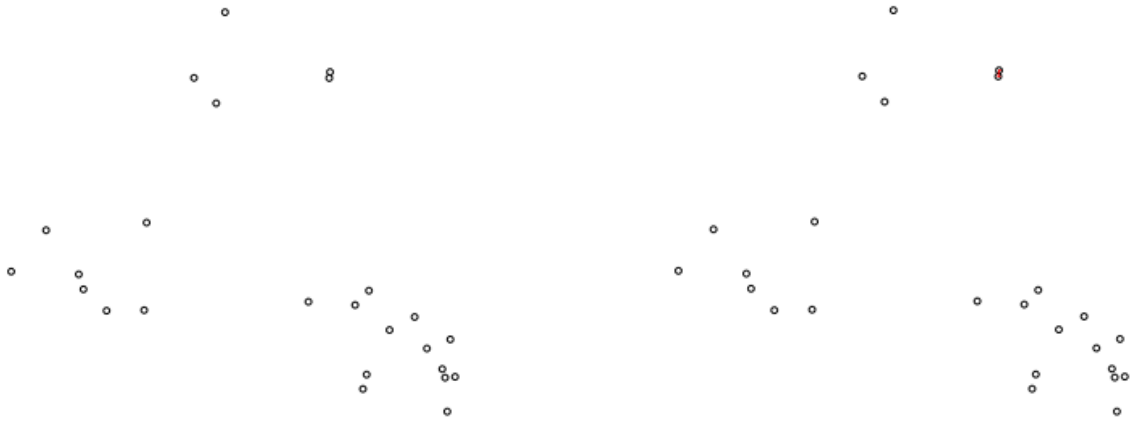
Για να εξεταστούν οι περιορισμοί που προαναφέρονται, ορίζονται οι εξής δύο παράμετροι: *η ποιότητα του στιγμιοτύπου (snapshot quality)* και *το κόστος του ιστορικού (history cost)*. Έστω C_t η συσταδοποίηση που προκύπτει από τον αλγόριθμο με την άφιξη δεδομένων το χρονικό βήμα t. Η ποιότητα του στιγμιοτύπου της C_t μετρά πόσο καλά η συσταδοποίηση C_t αναπαριστά τα δεδομένα στο χρονικό βήμα t. Το κόστος του ιστορικού της συσταδοποίησης είναι ένα μέτρο της απόστασης μεταξύ C_t και C_{t-1}, που η συσταδοποίηση χρησιμοποιεί κατά τη διάρκεια του προηγούμενου χρονικού βήματος. Τυπικά, η ποιότητα του στιγμιοτύπου ορίζεται με βάση τα σημεία των δεδομένων, ενώ το κόστος του ιστορικού είναι μία συνάρτηση των μοντέλων των

συστάδων. Το συνολικό κόστος της ακολουθίας της συσταδοποίησης είναι ένας συνδυασμός της ποιότητας του στιγμιοτύπου και του κόστους του ιστορικού σε κάθε χρονικό βήμα.

Αυτή η τοποθέτηση είναι παρόμοια με τη στοιχειώδη συσταδοποίηση (incremental clustering), αλλά με κάποιες διαφορές. Υπάρχουν δύο βασικές διαφορές. Πρώτον, η έμφαση δίνεται στη βελτιστοποίηση ενός νέου μέτρου ποιότητας που ενσωματώνει μια απόκλιση από το ιστορικό. Δεύτερον, λειτουργεί online (δηλαδή, τα δεδομένα πρέπει να ομαδοποιηθούν κατά τη διάρκεια του χρονικού βήματος t προτού ο αλγόριθμος της Εξελικτικής Συσταδοποίησης ασχοληθεί με οποιαδήποτε δεδομένα για το χρονικό βήμα t + 1˙ αν ο αλγόριθμος έχει πρόσβαση εκ των προτέρων σε όλα τα δεδομένα, τότε είναι offline), όσο άλλα πλαίσια Εξελικτικής Συσταδοποίησης δουλεύουν με ροές δεδομένων (για περαιτέρω μελέτη δείτε το άρθρο [25]).

Στην παγκόσμια βιβλιογραφία, για την υλοποίηση της Εξελικτικής Συσταδοποίησης χρησιμοποιούνται συνήθως δύο κλασικοί αλγόριθμοι της συσταδοποίησης: (1) ο παραδοσιακός αλγόριθμος των K – μέσων που παρέχει μία ομαδοποίηση των σημείων σε ένα διάνυσμα χώρου, και (2) ο αλγόριθμος **Συσσωρευτικής Ιεραρχικής Συσταδοποίησης (Agglomerative Hierarchical Clustering)**, του οποίου το αποτέλεσμα είναι ένα δένδρο συσταδοποίησης. Ο βασικός αλγόριθμος της συσσωρευτικής ιεραρχικής συσταδοποίησης παρουσιάζεται στον Αλγόριθμο 4.2. Εν συντομία, στον συγκεκριμένο αλγόριθμο, 1) τοποθετούμε κάθε δεδομένο σημείο του γραφήματος στην δική του μεμονωμένη ομάδα, 2) συγχωνεύουμε επαναληπτικά τις δύο πιο κοντινές ομάδες, 3) έως ότου όλα τα δεδομένα σημεία να συγχωνευτούν σε μία και μοναδική ομάδα.

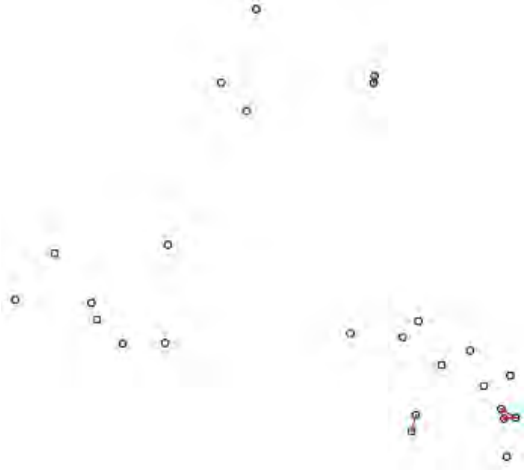
Παρακάτω παρουσιάζεται ένα οπτικό παράδειγμα του αλγορίθμου.



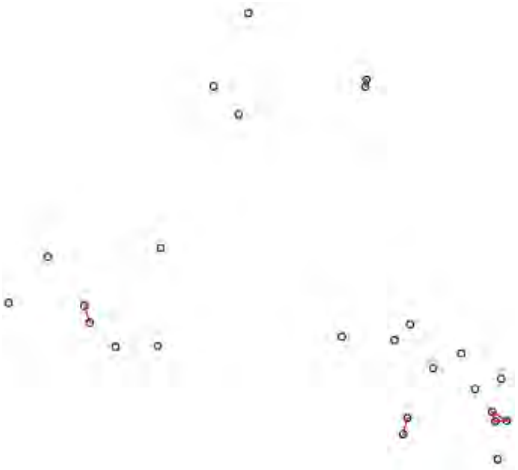
α) μεμονωμένες ομάδων σημείων β) 1η επανάληψη

γ) 2η επανάληψη

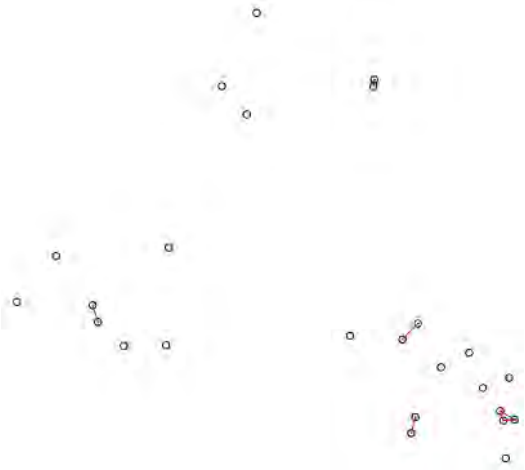
δ) 3η επανάληψη



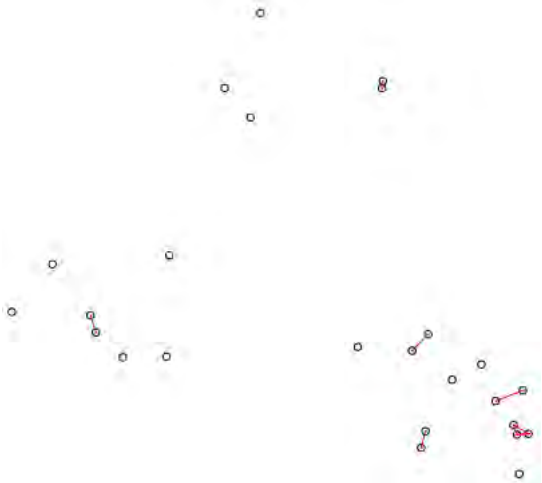
ε) 4η επανάληψη



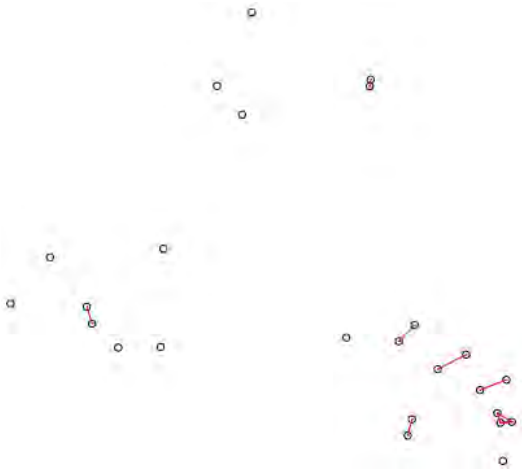
στ) 5η επανάληψη



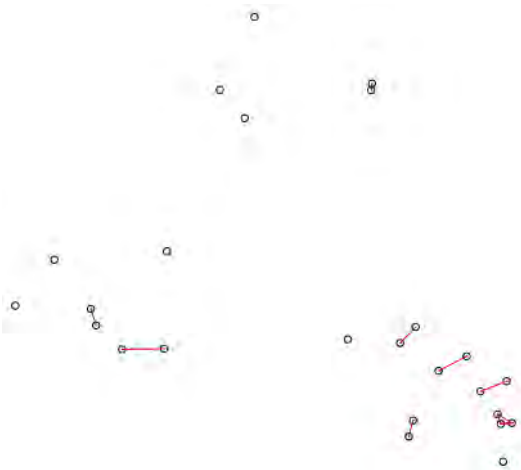
ζ) 6η επανάληψη



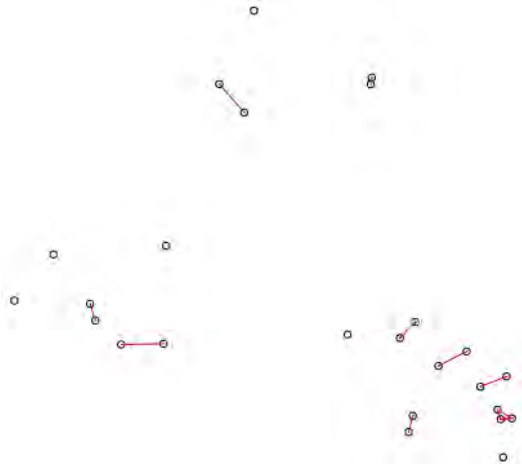
η) 7η επανάληψη



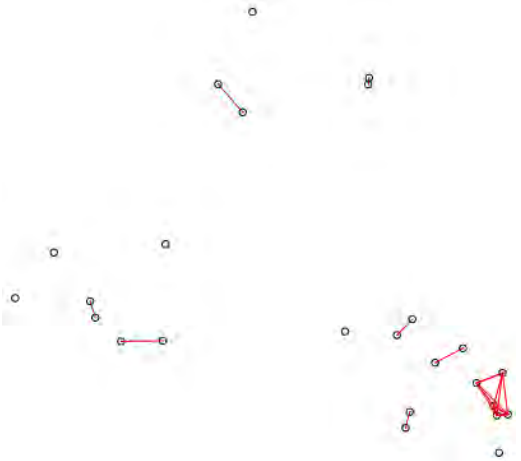
θ) 8η επανάληψη



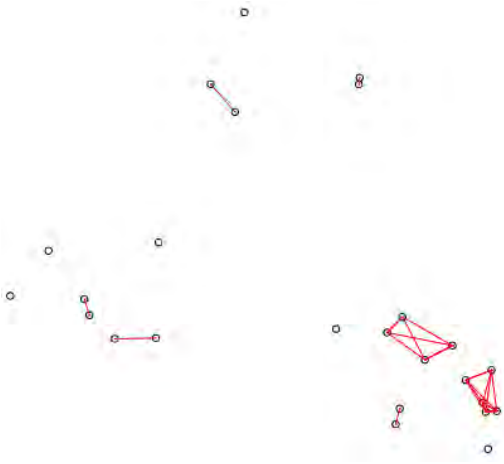
ι) 9η επανάληψη



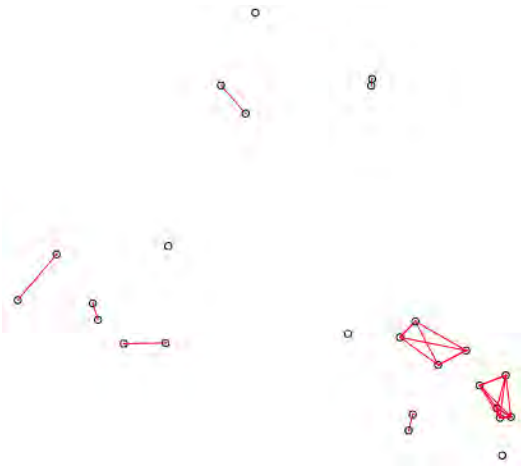
κ) 10η επανάληψη



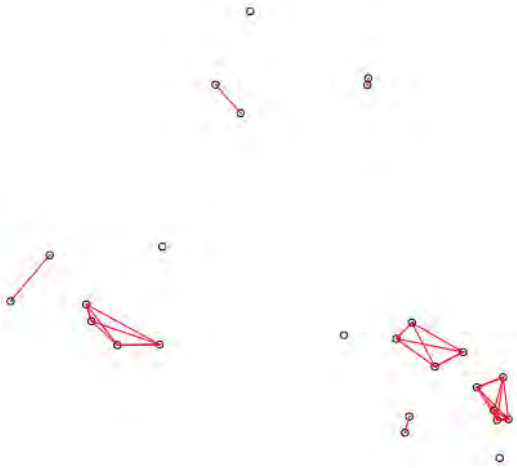
λ) 11η επανάληψη



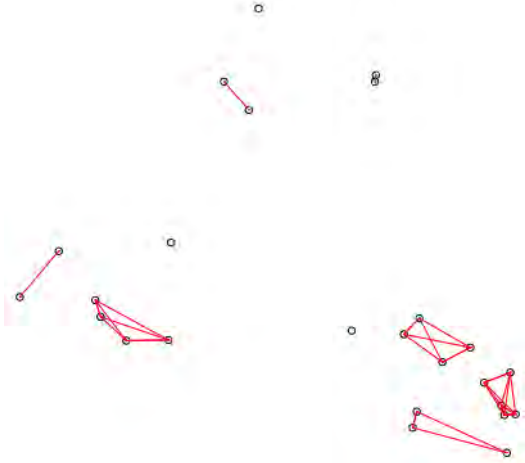
μ) 12η επανάληψη



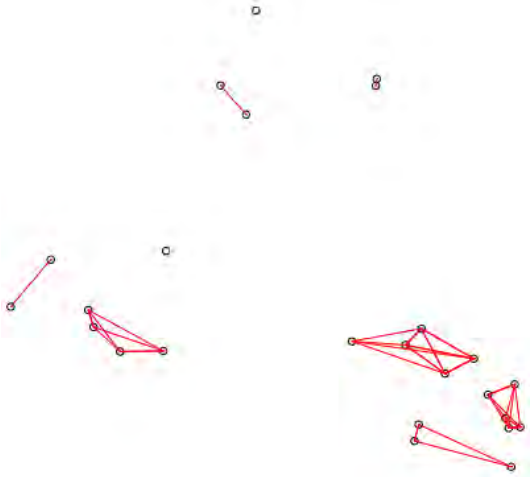
ν) 13η επανάληψη



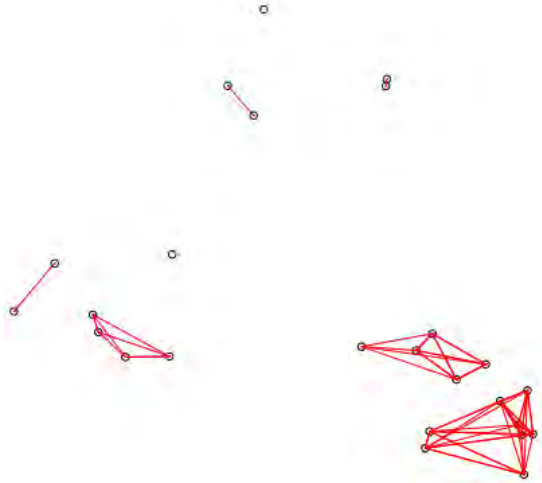
ξ) 14η επανάληψη



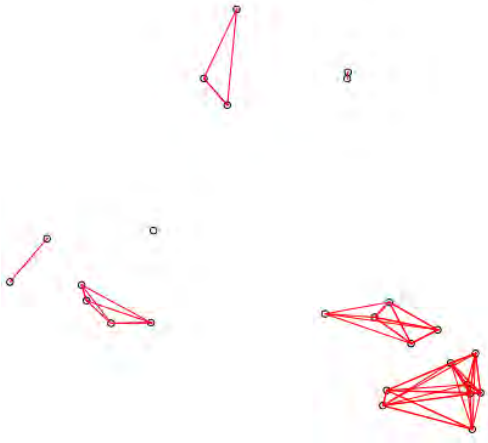
ο) 15η επανάληψη



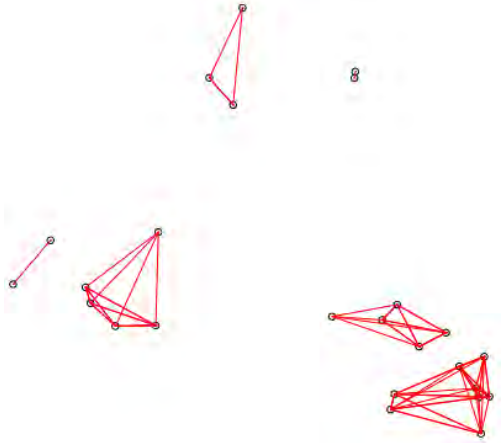
π) 16η επανάληψη



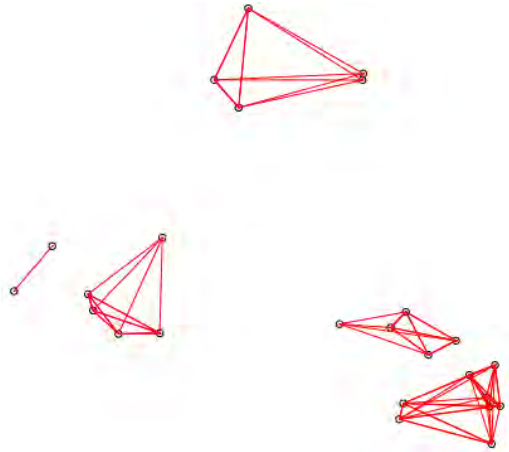
ρ) 17η επανάληψη



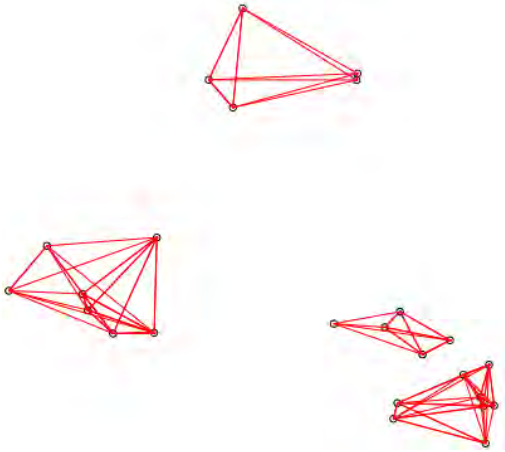
σ) 18η επανάληψη



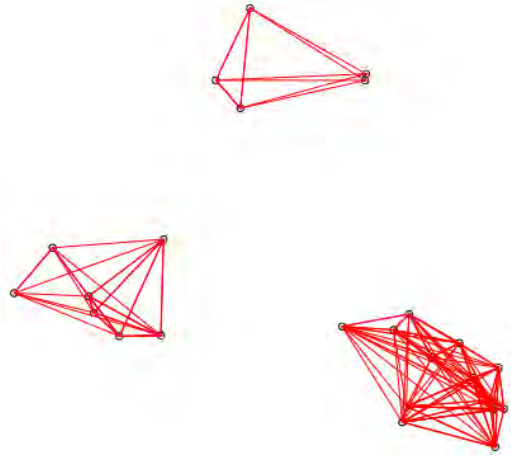
τ) 19η επανάληψη



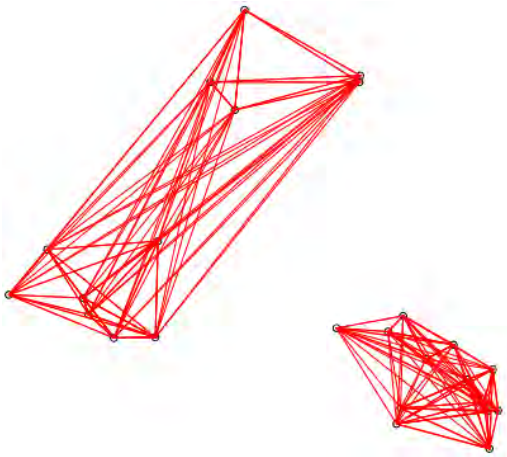
υ) 20η επανάληψη



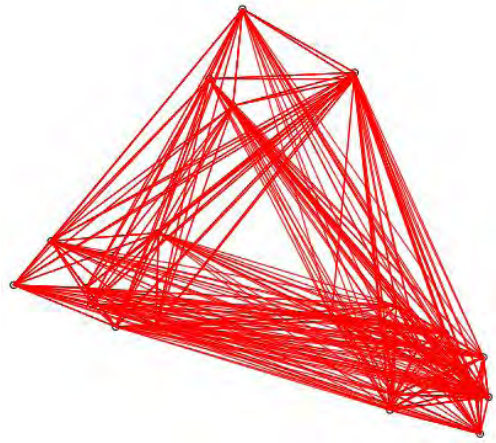
φ) 21η επανάληψη



χ) 22η επανάληψη



ψ) 23η επανάληψη



ω) 24η επανάληψη

Σχήμα 4.2: Στιγμιότυπα τρεξίματος του αλγόριθμου της Συσσωρευτικής Ιεραρχικής Συσταδοποίησης που πραγματοποιείται σε 24 επαναλήψεις.

Το πλαίσιο αυτό μπορεί να προστεθεί σε οποιοδήποτε αλγόριθμο συσταδοποίησης. Η χρονική πολυπλοκότητα θα είναι $O(n^2)$, ιδίως στην συσσωρευτική ιεραρχική συσταδοποίηση. Ωστόσο, το πλαίσιο παρουσιάζεται εδώ επειδή είναι δυνατόν να εφαρμόζει τις αρχές του σε όλους τους άλλους αλγορίθμους εντοπισμού κοινοτήτων που παρουσιάζονται στην παρούσα εργασία.

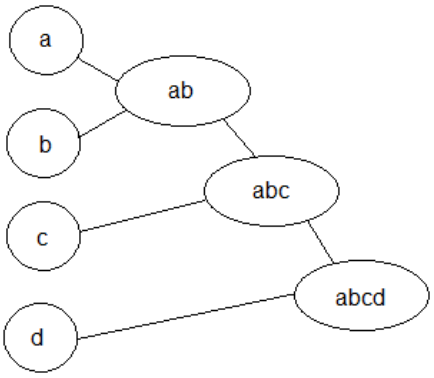
Αλγόριθμος 4.2 Ο βασικός αλγόριθμος Συσσωρευτικής Ιεραρχικής Συσταδοποίησης

1: Υπολόγισε τη μήτρα εγγύτητας, αν είναι απαραίτητο
2: **Επανέλαβε**
3: Συγχώνευσε τις πλησιέστερες δύο συστάδες
4: Ενημέρωσε τη μήτρα εγγύτητας για να απεικονίζει την εγγύτητα μεταξύ της νέας συστάδας και των αρχικών συστάδων
5: **Μέχρι** να απομείνει μία συστάδα

Η εγγύτητα των συστάδων (έστω ότι δύο συστάδες του γραφήματος συμβολίζονται με C, C’, δύο σημεία με x, x’ και d(x,x’) είναι η απόσταση μεταξύ αυτών των σημείων) μπορεί προσδιοριστεί σύμφωνα με ένα από τα παρακάτω κριτήρια:

- $\mathrm{d}(\mathrm{C},\mathrm{C}') = \min_{\mathrm{x}\in\mathrm{C},\mathrm{x}'\in\mathrm{C}'} \mathrm{d}(\mathrm{x},\mathrm{x}')$, τεχνική της ελάχιστης απόστασης (MIN)
- $\mathrm{d}(\mathrm{C},\mathrm{C}') = \max_{\mathrm{x}\in\mathrm{C},\mathrm{x}'\in\mathrm{C}'} \mathrm{d}(\mathrm{x},\mathrm{x}')$, τεχνική της μέγιστης απόστασης (MAX)
- $\mathrm{d}(\mathrm{C},\mathrm{C}') = \frac{\sum_{\mathrm{x}\in\mathrm{C},\mathrm{x}'\in\mathrm{C}'} \mathrm{d}(\mathrm{x},\mathrm{x}')}{|\mathrm{C}||\mathrm{C}'|}$, τεχνική της μέσης απόστασης μεταξύ σημείων

Σε αυτό το σημείο αξίζει να σημειωθούν δύο εφαρμογές αυτού του πλαισίου. Η πρώτη είναι η FacetNet, στην οποία αναπτύσσεται ένα πλαίσιο για την αξιολόγηση της εξέλιξης των κοινοτήτων (για περισσότερα διαβάστε το άρθρο [27]). Η δεύτερη είναι η I–KK, στην οποία παρουσιάζονται οι έννοιες των νανο-κοινοτήτων και της k-κλίκα-με-κλίκα [9]. Οι έννοιες αυτές είναι χρήσιμες για την αξιολόγηση των στιγμιοτύπων και την ιστορική ποιότητα των κοινοτήτων που εντοπίστηκαν σε διάφορα στιγμιότυπα με οποιαδήποτε μέθοδο.



Σχήμα 4.3: Συσσωρευτική Ιεραρχική Συσταδοποίηση. Ξεκίνα με ξεχωριστές συστάδες και, σε κάθε βήμα, συγχώνευσε τα πιο κοντινά ζεύγη συστάδων. Ο προσδιορισμός της εγγύτητας μεταξύ συστάδων μπορεί να γίνει, για παράδειγμα, με την τεχνική της ελάχιστης απόστασης (MIN) ή με αυτήν της μέγιστης απόστασης (MAX).

4.3 Relation Summary Network with Bregman Divergence (RSN-BD)

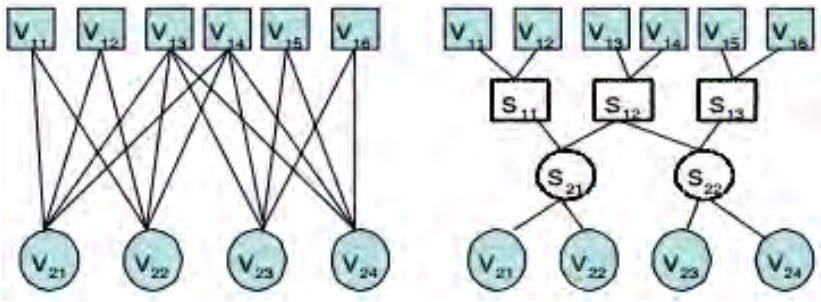
Η μέθοδος Relation Summary Network with Bregman Divergence (RSN-BD) αποτελεί μία προσέγγιση για τον εντοπισμό κοινοτήτων που βασίζεται σε παραδείγματα και δεδομένα του πραγματικού κόσμου που αφορούν πολλαπλούς τύπους αντικειμένων που σχετίζονται μεταξύ τους. Μία φυσική αναπαράσταση αυτού είναι ένας k – χωρισμένος γράφος, που απαρτίζεται από ετερογενείς τύπους κόμβων. Για παράδειγμα, έγγραφα και λέξεις σε ένα σώμα, πελάτες και στοιχεία σε συνεργατικό φιλτράρισμα, συναλλαγές καλαθιού αγοράς (market basket transactions), καθώς επίσης και τα γονίδια και οι συνθήκες σε δεδομένα μικροσυστοιχιών όλα σχηματίζουν ένα 2 – χωρισμένο γράφημα· έγγραφα, λέξεις, και κατηγορίες στον taxonomy mining κλάδο, καθώς επίσης και ιστοσελίδες, ερωτήματα αναζήτησης, και χρήστες του διαδικτύου σε ένα σύστημα διαδικτυακής αναζήτησης όλα σχηματίζουν ένα 3 – χωρισμένο γράφημα· δημοσιεύσεις, λέξεις-κλειδιά, συγγραφείς, και χώροι δημοσίευσης σε επιστημονικό αρχείο σχηματίζουν ένα 4 – χωρισμένο γράφημα.

Μία διαισθητική προσπάθεια ώστε να ανακαλυφθούν οι κρυφές δομές από k – χωρισμένα γραφήματα, είναι να εφαρμοστούν οι υπάρχουσες προσεγγίσεις στην διαμέριση γράφων σε αυτά τα γραφήματα. Αυτή η ιδέα μπορεί να λειτουργήσει σε ορισμένες ειδικές και απλές περιπτώσεις. Ωστόσο, σε γενικές γραμμές, είναι ανέφικτο. Πρώτον, η θεωρία για τη διαμέριση γράφων εστιάζεται στην εύρεση των καλύτερων τομών ενός γραφήματος κάτω από ένα συγκεκριμένο κριτήριο και είναι πολύ δύσκολο να «κοπούν» διαφορετικοί τύποι σχέσεων

(συνδέσεις) ταυτόχρονα για να προσδιοριστούν διαφορετικές κρυμμένες δομές για διαφορετικούς τύπους κόμβων. Δεύτερον, διαμερίζοντας ολόκληρο το k – χωρισμένο γράφημα σε m υπογράφους, κάποιος στην πραγματικότητα υποθέτει ότι όλοι οι διαφορετικοί τύποι κόμβων έχουν τον ίδιο αριθμό συστάδων m, το οποίο σε γενικές γραμμές δεν είναι αλήθεια. Τρίτον, διαμερίζοντας απλώς ολόκληρο το γράφημα σε κομματιασμένα υπογραφήματα, οι κρυφές δομές που προκύπτουν είναι πρόχειρες.

Η βασική ιδέα είναι ότι σε ένα αραιό k – χωρισμένο γράφημα, δύο κόμβοι είναι παρόμοιοι όταν είναι συνδεδεμένοι σε παρόμοιους κόμβους έστω και αν δεν είναι συνδεδεμένοι με τους ίδιους κόμβους. Για να εντοπιστεί αυτή η ομοιότητα, διάφοροι συγγραφείς προτείνουν μία ιδιαίτερη δομή (π.χ προβολή) για να συνδέσουν στενά αυτούς τους δύο κόμβους. Για να γίνει αυτό προστίθεται ένας μικρός αριθμός κρυφών κόμβων. Αυτή η προκύπτουσα δομή ονομάζεται **Relation Summary Network (RSN)**.

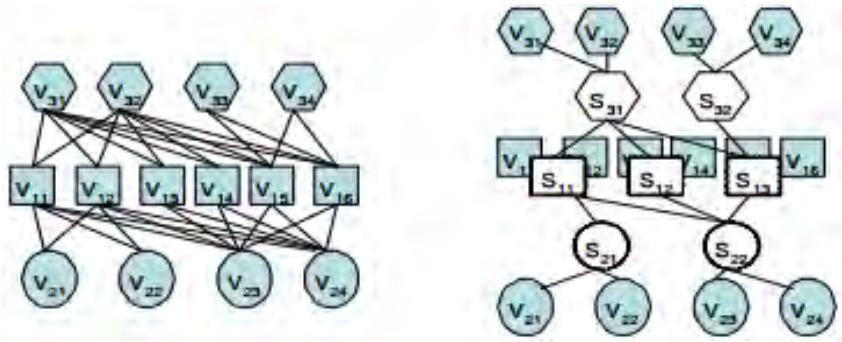
Το Relation Summary Network (RSN) μοντέλο, γενικά, ανακαλύπτει τις κρυμμένες δομές από ένα k – χωρισμένο γράφημα. Η βασική ιδέα του RSN είναι να προσθέτει ένα μικρό αριθμό κρυφών κόμβων στο αρχικό k – χωρισμένο γράφημα έτσι ώστε να καταστήσει σαφείς τις κρυφές δομές του γραφήματος. Για να εξασφαλιστεί ότι το RSN μοντέλο θα εντοπίσει τις επιθυμητές κρυφές δομές του αρχικού γραφήματος, θα πρέπει να έρθει όσο το δυνατόν πιο «κοντά» στο αρχικό γράφημα. Με άλλα λόγια, οι μελετητές στοχεύουν σε ένα βέλτιστο RSN, από το οποίο μπορούν να ανακατασκευάσουν το αρχικό γράφημα όσο το δυνατόν ακριβέστερα. Κάθε κόμβος συνδέεται με έναν και μοναδικό κρυφό κόμβο. Οι αρχικοί κόμβοι συσχετίζονται μεταξύ τους μόνο μέσω των κρυφών κόμβων. Η απόσταση μεταξύ των δύο δομών μπορεί να αξιολογηθεί, συνδέοντας κάθε αρχικό κόμβο με έναν κρυφό και με κάθε ζευγάρι κρυφών κόμβων, εάν και οι δύο κρυφοί κόμβοι συνδέονται με τον ίδιο αρχικό κόμβο. Για να γίνει περισσότερο κατανοητό μπορείτε να κοιτάξετε τα Σχήματα 4.4 και 4.5.



Σχήμα 4.4: Αριστερά βλέπουμε ένα 2 – χωρισμένο γράφημα, και δεξιά το RSN του με τους κόμβους S_{11}, S_{12}, S_{13}, S_{21}, S_{22} να ονομάζονται κρυφοί κόμβοι (hidden nodes) [18].

Στο αριστερό γράφημα του Σχήματος 4.4 φαίνεται ένα 2 – χωρισμένο γράφημα G = (V_1, V_2, E), όπου το V_1 = {v_{11},……,v_{16}} και το V_2 = {v_{21},……,v_{24}} υποδηλώνουν δύο τύπους κόμβων, και το E υποδηλώνει τις ακμές στο G. Ακόμα κι αν αυτό το γράφημα είναι απλό, δεν είναι τετριμμένο ώστε να εντοπιστούν οι κρυμμένες δομές του. Στο δεξί γράφημα του Σχήματος 4.4,

παρουσιάζεται εκ νέου το αρχικό γράφημα και αφού πρώτα έχουν προστεθεί δύο νέα σύνολα κόμβων (αυτοί ονομάζονται κρυφοί κόμβοι), $S_1 = \{s_{11}, s_{12}, s_{13}\}$ και $S_2 = \{s_{21}, s_{22}\}$. Με βάση το νέο γράφημα, οι δομές της συστάδας για κάθε τύπο κόμβων είναι απλή υπόθεση: το σύνολο V_1 έχει τρεις συστάδες, $\{ v_{11}, v_{12} \}$, $\{ v_{13}, v_{14} \}$, $\{ v_{15}, v_{16} \}$, ενώ το V_2 έχει δύο συστάδες $\{ v_{21}, v_{22} \}$, $\{ v_{23}, v_{24} \}$. Αν κοιτάξει κάποιος το υπογράφημα που περιλαμβάνει μόνο τους κρυφούς κόμβους στο Σχήμα 4.4, παρατηρεί ότι αυτό παρέχει έναν ευδιάκριτο σκελετό για την ολική διάρθρωση του γραφήματος, από τον οποίο γίνεται σαφές το πως οι συστάδες των διαφόρων τύπων των κόμβων σχετίζονται μεταξύ τους. Για παράδειγμα, η συστάδα s_{11} συνδέεται με την s_{21}, και η συστάδα s_{12} με αμφότερες τις s_{21} και s_{22}. Με άλλα λόγια, με την εισαγωγή των κρυφών κόμβων στο αρχικό k - χωρισμένο γράφημα γίνονται σαφείς, τόσο οι τοπικές δομές της συστάδας, όσο και οι ολικές δομές της κοινότητας. Επίσης, είναι σημαντικό να τονιστεί ότι το μονοπάτι (v_{13}, s_{12}, s_{21}, v_{22}) στο RSN υποδεικνύει πως υπάρχει ακμή μεταξύ των κόμβων v_{13} και v_{22} στο αρχικό γράφημα τέτοια ώστε το βάρος της ακμής που ενώνει αυτές τις δύο κορυφές είναι ίσο με το βάρος της ακμής που ενώνει τους s_{12} και s_{21}.



(α) (β)

Σχήμα 4.5: Οι δομές της συστάδας του V_2 και του V_3 επηρεάζουν την ομοιότητα μεταξύ του v_{11} και του v_{12} μέσω των κρυφών κόμβων [18].

Το Σχήμα 4.5 δείχνει ένα παράδειγμα το πως οι δομές μιας συστάδας δύο τύπων κόμβων επηρεάζουν την ομοιότητα μεταξύ δύο διαφορετικού τύπου κόμβων. Θεωρείται ότι πρέπει να συσταδοποιηθούν οι κόμβοι που ανήκουν στο σύνολο V_1 (στο Σχήμα 4.5(α) φαίνονται δύο κόμβοι). Οι κλασσικές προσεγγίσεις συσταδοποίησης υπολογίζουν την ομοιότητα μεταξύ των v_{11} και v_{12} βασιζόμενες στα χαρακτηριστικά της σύνδεσής τους, [1,0,1,0] και [0,1,0,1], αντίστοιχα, και ως εκ τούτου, η ομοιότητά τους θεωρείται μηδέν (χαμηλότερο επίπεδο). Πρόκειται για μία χαρακτηριστική περίπτωση σε ένα μεγάλο γράφημα με αραιές συνδέσεις. Τώρα, ας υποτεθεί ότι προκύπτουν κρυφοί κόμβοι για τα V_2 και V_3, όπως φαίνεται στο Σχήμα 4.5(β). Μέσω των κρυφών κόμβων, οι δομές της συστάδας του V_2 αλλάζουν την ομοιότητα μεταξύ του v_{11} και του v_{12} σε 1 (υψηλότερο επίπεδο), δεδομένου ότι τα μειωμένα χαρακτηριστικά της σύνδεσης για αμφότερες τις v_{11} και v_{12} είναι [1,1]. Το παραπάνω είναι ένα πιο λογικό αποτέλεσμα, αφού σε ένα αραιό k – χωρισμένο γράφημα είναι αναμενόμενο δύο κόμβοι να είναι όμοιοι όταν αυτοί συνδέονται με όμοιους κόμβους, ακόμη και αν δεν συνδέονται στους ίδιους κόμβους. Αν

συνεχιστεί αυτό το παράδειγμα, στο επόμενο βήμα, οι v_{11} και v_{12} συνδέονται με τους ίδιους κρυφούς κόμβους στο S_1 (δεν φαίνεται στο Σχήμα 4.5). Έπειτα, αφού προκύπτουν οι κρυφοί κόμβοι για το V_1, οι δομές της συστάδας του V_2 και του V_3 ως αντάλλαγμα μπορεί να επηρεαστούν. Στην πραγματικότητα, αυτή είναι η ιδέα του επαναληπτικού αλγορίθμου για να κατασκευαστεί ένα RSN μοτέλο για ένα k – χωρισμένο γράφημα.

Για την βελτιστοποίηση του προβλήματος, μπορούν να γίνουν πολλές επιλογές για τις συναρτήσεις της απόστασης, οι οποίες με τη σειρά τους συνεπάγονται τις διάφορες υποθέσεις σχετικά με την κατανομή των βαρών των ακμών στο δοσμένο k – χωρισμένο γράφημα. Για παράδειγμα, χρησιμοποιώντας τη συνάρτηση της Ευκλείδειας απόστασης, θεωρείται σιωπηρά η χρήση της κανονικής κατανομής για τα βάρη των ακμών (στο αρχικό γράφημα και στο μετασχηματισμένο γράφημα). Το μέτρο που χρησιμοποιείται σε αυτή την ενότητα είναι μία συνάρτηση απόστασης του k – χωρισμένου γραφήματος με το RSN αυτού, που έχει άμεση σχέση με το άθροισμα της Ευκλείδιας απόστασης μεταξύ του βάρους των ακμών του αρχικού γραφήματος και του βάρους των ακμών στο RSN γράφημα. Έτσι, για την βελτιστοποίηση του προβλήματος μπορεί να χρησιμοποιηθεί οποιαδήποτε από τις αποκλίσεις Bregman. Μία απόκλιση Bregman ορίζει μία τάξη των μέτρων απόστασης για την οποία ούτε η τριγωνική ανισότητα, ούτε η συμμετρία, είναι σεβαστή, και τα μέτρα αυτά καθορίζονται για μήτρες, συναρτήσεις και κατανομές (περισσότερα στην Ενότητα 4.3.1 και στο άρθρο [11]).

Η RSN-BD μέθοδος είναι κατάλληλη για γενικά k – χωρισμένα γραφήματα και όχι μόνο για ειδικές περιπτώσεις όπως συμβαίνει στην δημοσίευση [28]. Στην τελευταία υπάρχει ο περιορισμός ότι οι αριθμοί των συστάδων για διαφορετικούς τύπους κόμβων πρέπει να είναι ίσοι, και ότι οι συστάδες για διαφορετικούς τύπους αντικειμένων πρέπει να έχουν μία-προς-μία αντιστοιχία. Η Relation Summary Network with Bregman Divergence (RSN-BD) μέθοδος συνοψίζεται στον Αλγόριθμο 4.3. Η RSN-BD ενημερώνει επαναληπτικά τις δομές της συστάδας για τους διαφορετικούς τύπους των κόμβων και των σχέσεων μεταξύ των κρυφών κόμβων. Μέσω των κρυφών κόμβων, οι δομές των συστάδων των διαφορετικών τύπων κόμβων αλληλεπιδρούν μεταξύ τους είτε άμεσα είτε έμμεσα. Η μέθοδος RSN-BD είναι εφαρμόσιμη σε ένα ευρύ φάσμα προβλημάτων, δεδομένου ότι δεν έχει περιορισμούς σχετικά με τις δομές του αρχικού k – χωρισμένου γραφήματος. Επιπλέον, τα γραφήματα από διαφορετικές εφαρμογές μπορεί να έχουν διαφορετικές πιθανοτικές κατανομές στις ακμές τους· είναι εύκολο για τον RSN-BD να προσαρμοστεί σε αυτή την κατάσταση χρησιμοποιώντας απλά διαφορετικές αποκλίσεις Bregman, δεδομένου ότι αυτές ανταποκρίνονται σε μια μεγάλη οικογένεια εκθετικών κατανομών, συμπεριλαμβανομένων των πιο γνωστών κατανομών όπως η Κανονική, η Πολυωνυμική και η Poisson. Η συνολική πολυπλοκότητα του αλγορίθμου είναι O (n^2ck).

Αυτός ο κανόνας ενημέρωσης είναι συνεπής με την διαίσθηση για την ακμή μεταξύ δύο κρυφών κόμβων, δηλαδή είναι η «summary relation» για δύο σύνολα τέτοιων κόμβων. Ωστόσο, μια εκπληκτική παρατήρηση είναι ότι η ενημέρωση δεν συνεπάγεται την συνάρτηση απόστασης, δηλαδή, αυτός ο απλός κανόνας ενημέρωσης ισχύει για όλες τις αποκλίσεις Bregman.

Αλγόριθμος 4.3 Relation Summary Network with Bregman Divergences (RSN-BD)

Είσοδος: Ένα k – χωρισμένο γράφημα G = (V_1, …., V_m, E), μία συνάρτηση απόκλισης Bregman D_ϕ, και m θετικοί ακέραιοι k_1, ….., k_m.
'Εξοδος: Ένα RSN γράφημα G^S = (V_1, …., V_m, S_1, ……, S_m, E^S).

1: Αρχικοποίησε το G^S
2: **Επανέλαβε**
3: **για** i = 1 μέχρι m **κάνε**
4: Ενημέρωσε τις ακμές μεταξύ V_i και S_i στο G^S συνδέοντας τον v_{ih} σε κάθε κρυφό κόμβο του συνόλου S_i για να βρεις ποιος κρυφός κόμβος δίνει τις ελάχιστες τιμές για το $D_\phi(G, G^S)$ σύμφωνα με τον τύπο $e^S(v_{ih}, s_{il}) = 1$, για l = arg min $D_\phi(G, G^S_l)$, όπου το G^S_l υποδηλώνει το RSN με το s_{il} να συνδέεται με το v_{ih}
5: **τέλος για**
6: **για** κάθε ζεύγος S_i ~ S_j όπου 1 <= i < j <= m **κάνε**
7: Ενημέρωσε τις ακμές μεταξύ S_i και S_j στο G^S ξαναυπολογίζοντας το βάρος της ακμής μεταξύ ενός ζεύγους κρυφών κόμβων s_{ip} και s_{jq} σύμφωνα με την Εξίσωση 4.1 *.
8: **τέλος για**
9: **Μέχρι** να συγκλίνει

*Εξίσωση 4.1: $e^s(s_{ip}, s_{jq}) = \frac{1}{|U|*|Z|} \sum e(u_{ih}, u_{jl})$, όπου $|U| = \{ u_{ih} : e^s(u_{ih}, s_{ip}) = 1 \}$, δηλαδή οι κόμβοι συνδέονται με τον s_{ip} . $|Z| = \{ u_{jl} : e^s(u_{jl}, s_{jq}) = 1 \}$, δηλαδή οι κόμβοι συνδέονται με τον s_{jq} , με 1 <= p <= k_i , 1<= q <= k_j , 1<= h <= n_i και 1<= l <= n_j.

Περισσότερα μπορείτε να δείτε στο άρθρο [29].

4.3.1 Απόκλιση Bregman

Στην ενότητα αυτή δίνεται μία σύτομη περιγραφή των αποκλίσεων Bregman που είναι μία οικογένεια συναρτήσεων εγγύτητας που μοιράζονται μερικές κοινές ιδιότητες. Ως αποτέλεσμα, είναι δυνατόν να κατασκευαστούν γενικοί αλγόριθμοι εξόρυξης δεδομένων, όπως αλγόριθμοι συσταδοποίησης, οι οποίοι δουλεύουν με κάθε απόκλιση Bregman. Ένα απτό παράδειγμα είναι ο αλγόριθμος συσταδοποίησης των K – μέσων.

Οι αποκλίσεις Bregman είναι συναρτήσεις απώλειας (loss functions) ή συναρτήσεις αποκλίσεων (distortion functions). Για να αντιληφθούμε την ιδέα της συνάρτησης απώλειας, θεωρείστε τα παρακάτω. Έστω x και y δύο σημεία, όπου το y θεωρείται το αρχικό και το x μια απόκλιση ή μία προσέγγιση του y. Για παράδειγμα, το x μπορεί να είναι ένα σημείο το οποίο παρήχθηκε προσθέτοντας τυχαίο θόρυβο στο y. Ο στόχος είναι να μετρηθεί η απόκλιση ή απώλεια που προκύπτει όταν το y προσεγγίζεται από το x. Φυσικά, όσο πιο όμοια είναι τα x και

y, τόσο πιο μικρή είναι η απώλεια ή η απόκλιση. Επομένως, οι αποκλίσεις Bregman μπορούν να χρησιμοποιηθούν ως συναρτήσεις ανομοιότητας.

Πιο επίσημα, υπάρχει ο παρακάτω ορισμός.

Ορισμός 4.2 (Απόκλιση Bregman). Δοθείσης μιας αυστηρά κυρτής συνάρτησης φ (με μερικούς απλούς περιορισμούς, οι οποίοι γενικά ικανοποιούνται), η απόκλιση Bregman (συνάρτηση απώλειας) $D(x,y)$ που παράγεται από αυτή τη συνάστηση φ δίνεται από την ακόλουθη εξίσωση:

$$D(x,y) = \varphi(x) - \varphi(y) - \langle \nabla\varphi(y), (x - y) \rangle$$

όπου $\nabla\varphi(y)$ είναι η κλίση της φ υπολογισμένη στο σημείο y, x – y είναι η διανυσματική διαφορά μεταξύ των x και y, και $\langle \nabla\varphi(y), (x - y) \rangle$ είναι το εσωτερικό γινόμενο μεταξύ των $\nabla\varphi(y)$ και $(x - y)$. Για σημεία στον Ευκλείδιο χώρο, το εσωτερικό γινόμενο είναι απλά το γινόμενο κατά μέλη.

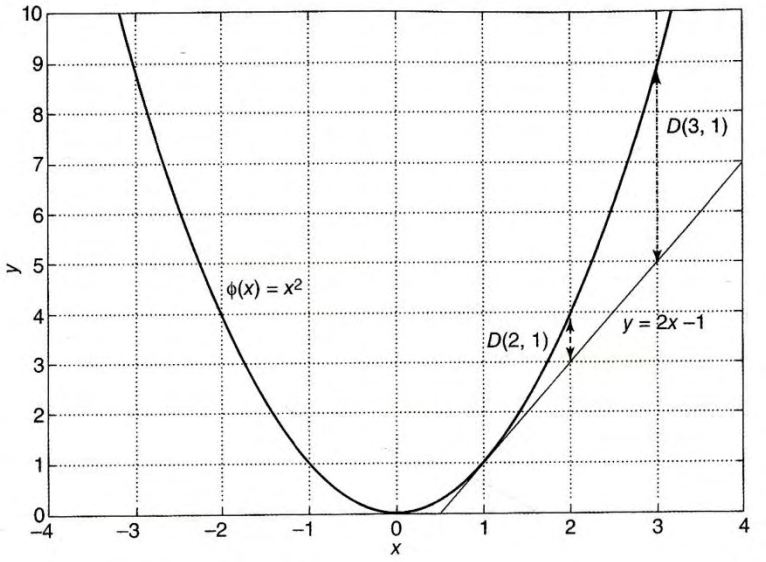
Η $D(x,y)$ μπορεί να γραφεί ως $D(x,y) = \varphi(x) – L(x)$, όπου $L(x) = \varphi(y) + \langle \nabla\varphi(y), (x - y) \rangle$ και αντιπροσωπεύει την εξίσωση ενός επιπέδου εφαπτομένου της φ στο y. Χρησιμοποιώντας ορολογία ανάλυσης, η $L(x)$ είναι η γραμμικοποίηση της φ γύρω από το σημείο y και η απόκλιση Bregman είναι απλώς η διαφορά μεταξύ μιας συνάρτησης και μιας γραμμικής προσέγγισης σε αυτή τη συνάρτηση. Διαφορετικές αποκλίσεις Bregman λαμβάνονται χρησιμοποιώντας διαφορετικές επιλογές της φ.

4.3.1.1 Παράδειγμα

Εδώ παρουσιάζεται ένα συγκεκριμένο παράδειγμα χρησιμοποιώντας την τετραγωνική Ευκλείδια απόσταση αλλά περιορίζεται σε μία διάσταση ώστε να απλοποιηθούν τα μαθηματικά. Έστω x και y πραγματικοί αριθμοί και $\varphi(t)$ η πραγματική συνάρτηση $\varphi(t) = t^2$. Σε αυτήν την περίπτωση, η κλίση ανάγεται στην παράγωγο και το εσωτερικό γινόμενο σε πολλαπλασιασμό. Ειδικότερα, η εξίσωση που δόθηκε στον Ορισμό 4.2 μετατρέπεται στην εξής εξίσωση:

$$D(x,y) = x^2 – y^2 – 2y(x - y) = (x - y)^2 .$$

Το γράφημα για αυτό το παράδειγμα, με y = 1 δίνεται στο Σχήμα 4.6 . Η απόκλιση Bregman δίνεται για δύο τιμές του x: x = 2 και x = 3.



Σχήμα 4.6: Παράδειγμα της απόκλισης Bregman.

4.4 Multi-way Relation Graph Clustering (MRGC)

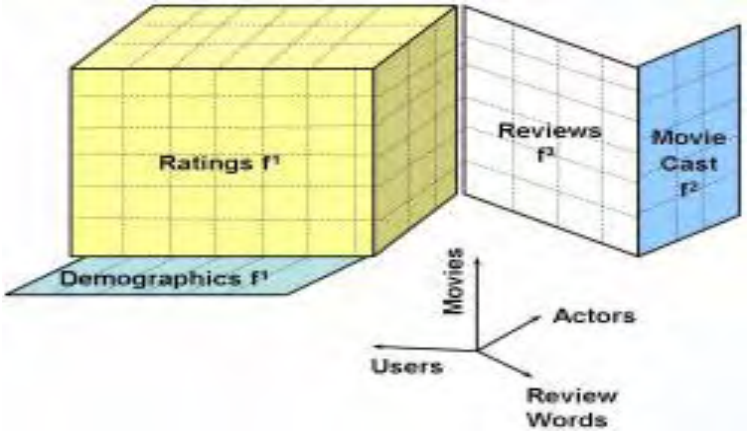
Στον **Multi-way Relation Graph Clustering (MRGC)** αλγόριθμο, κάθε σχέση μεταξύ ενός δεδομένου σύνολου κατηγοριών οντοτήτων αναπαριστάται με έναν πολυδιάστατο τανυστή (ή αλλιώς κύβο από δεδομένα), σύμφωνα με ένα κατάλληλο πεδίο ορισμού, και με τις διαστάσεις να συνδέονται με τις διάφορες κατηγορίες οντοτήτων. Επιπλέον, κάθε κελί στο κύβο κωδικοποιεί τη σχέση μεταξύ ενός συγκεκριμένου συνόλου οντοτήτων και μπορεί είτε να λάβει πραγματικές τιμές, δηλαδή, η σχέση έχει ένα μόνο χαρακτηριστικό γνώρισμα, είτε το ίδιο να αποτελεί ένα διάνυσμα χαρακτηριστικών. Αυτό το γενικό μοντέλο είναι χρήσιμο για εφαρμογές σε πολλούς τομείς που έχουν πολλαπλού τύπου συγγενικά σύνολα δεδομένων. Σε ένα τέτοιο πρόβλημα, η συσταδοποίηση βασίζεται όχι μόνο στις εγγενείς τιμές χαρακτηριστικών των οντοτήτων, αλλά και στις πολλαπλές σχέσεις μεταξύ των οντοτήτων. Επιπλέον, κάθε σχέση μπορεί να περιλαμβάνει πολλαπλά σύνολα οντοτήτων (σε αντίθεση με τα ζεύγη σχέσεων) και οι σχέσεις μπορούν επίσης να έχουν ιδιότητες.

Παράδειγμα 4.4.1. Θεωρήστε τις online εφαρμογές ταινιών όπως το Yahoo! Movies. Οι ιστοσελίδες αυτές έχουν πληροφορίες για την ταινία σε σχέση με την περιγραφή της κάθε ταινίας, οι οποίες μπορούν να αναπαρασταθούν με διάφορους τρόπους/σχέσεις, π.χ,

- f^1 : αξιολογήσεις για (ταινία, θεατής, ηθοποιός) πλειάδες που αντιστοιχούν στις αντιδράσεις των θεατών σχετικά με την απόδοση των ηθοποιών σε διάφορες ταινίες

- f^2 : συνύπαρξη δεικτών για (ταινία, ηθοποιός) ζεύγη που προσδιορίζουν ποιοι ηθοποιοί έπαιξαν σε ποιες ταινίες
- f^3 : γνώμη για (ταινία, λέξεις κριτικής) πλειάδες που κωδικοποιούν τις κριτικές των ταινιών
- f^4 : τιμές (θεατής, δημογραφικά χαρακτηριστικά) που καθορίζουν ορισμένες λεπτομέρειες όπως η ηλικία, το φύλο, κλπ, για διαφορετικούς θεατές.

Το Σχήμα 4.7 δείχνει μια εικόνα αυτού του συνόλου δεδομένων ως ένα σύνολο από τανυστές με μερικές κοινές διαστάσεις, π.χ., η f^1 και η f^4 μοιράζονται τη διάσταση του θεατή, η f^1 και η f^2 μοιράζονται τόσο τη διάσταση του ηθοποιού όσο και αυτήν της ταινίας, ενώ η f^3 έχει και αυτή τη διάσταση της ταινίας.



Σχήμα 4.7: Παράδειγμα πολλαπλού τύπου σχεσιακών δεδομένων στη σύσταση ταινίας.

Η γενική ιδέα είναι ότι κάθε κόμβος και κάθε σχέση είναι μια συλλογή από χαρακτηριστικά/ιδιότητες. Όλα αυτά τα χαρακτηριστικά απεικονίζονται από μία διάσταση του χώρου των σχέσεων. Η Multi-way Relation Graph Clustering (MRGC) μέθοδος προσπαθεί ουσιαστικά κάθε φορά να βρει μία λύση σε μία διάσταση. Βρίσκει τη βέλτιστη συσταδοποίηση ως προς την κάθε διάσταση διατηρώντας κάθε άλλο ενδιάμεσο αποτέλεσμα στις άλλες διαστάσεις σταθερό (η χρονική πολυπλοκότητα είναι O(mD)). Έπειτα αξιολογεί τις λύσεις και κάνει επαναϋπολογισμούς σε όλες τις διαστάσεις, μέχρι να συγκλίνει. Παρά το γεγονός ότι ορίζεται για γραφήματα σχέσεων, το μοντέλο αυτό μπορεί να χρησιμοποιηθεί και για τον εντοπισμό των δομών της κοινότητας σε κοινωνικά δίκτυα. Η MRGC μέθοδος περιλαμβάνει μια επαναληπτική διαδικασία όπου οι αναθέσεις συστάδων της κάθε διάστασης ενημερώνονται ακολουθούμενοι από τον υπολογισμό της Minimum Bregman Information λύσης (περισσότερα στα άρθρα [12,30]). Η μόνη διαφορά είναι ότι οι βέλτιστες αναθέσεις συστάδας και ο MBI υπολογισμός εξαρτώνται από πολλούς τανυστές που σχετίζονται με διαφορετικές σχέσεις.

Ο MRGC αλγόριθμος λειτουργεί σε ένα πλαίσιο πολλαπλών τρόπων συσταδοποίησης, όπου ο στόχος είναι να χαρτογραφηθεί το σύνολο των οντοτήτων που ανήκουν σε ένα (μικρότερο) σύνολο από συστάδες χρησιμοποιώντας μια σειρά από λειτουργίες συσταδοποίησης (δηλαδή,

είναι ένα γενικό πλαίσιο στο οποίο προηγούμενες προσεγγίσεις, όπως αυτή στο άρθρο [22], μπορεί να θεωρηθούν ως ειδικές περιπτώσεις). Ο κρίσιμος μηχανισμός σε αυτό το πρόβλημα είναι το πώς να αξιολογήσει την ποιότητα της «πολλών τρόπων» συσταδοποίησης ώστε να φτάσει τελικά να συγκλίνει. Στην περίπτωση αυτή, διάφοροι συγγραφείς προτείνουν να μετρηθεί από την άποψη του προσεγγιστικού σφάλματος ή την αναμενόμενη Bregman παραμόρφωση [23] μεταξύ του αρχικού τανυστή και του προσεγγιστικού τανυστή που δημιουργείται μετά την εφαρμογή της λειτουργίας της συσταδοποίησης.

4.5 SocDim (Social Dimensions)

Μία βασική (Markov) παραδοχή στον εντοπισμό κοινότητας συχνά είναι ότι η ετικέτα ενός κόμβου εξαρτάται μόνο από τις ετικέτες όλων των γειτόνων του. Ο αλγόριθμος **SocDim** προσπαθεί να προχωρήσει πέρα από αυτή την υπόθεση δημιουργώντας έναν κατηγοριοποιητή ο οποίος όχι μόνο λαμβάνει υπ' όψιν την συνδεσιμότητα ενός κόμβου, αλλά εκχωρεί επιπλέον πληροφορίες για τη σύνδεσή του, δηλαδή, μία περιγραφή της πιθανής σχέσης μεταξύ των κοινωνικών φορέων.

Αυτή η πληροφορία είναι γνωστή και ως *λανθάνουσες κοινωνικές διαστάσεις (latent social dimensions)* και το πλαίσιο που προκύπτει βασίζεται στη *σχεσιακή μάθηση (relational learning)*. Κάθε διάσταση αντιπροσωπεύει έναν πιθανό δεσμό μεταξύ των κοινωνικών φορέων˙ μπορεί να θεωρηθεί ως περιγραφή των πιθανών δεσμών. Με άλλα λόγια, αυτές οι κοινωνικές διαστάσεις περιγράφουν διαφορετικούς δεσμούς των κοινωνικών φορέων που είναι κρυφοί μέσα στο δίκτυο, και η μετέπειτα διακριτική μάθηση μπορεί αυτόματα να καθορίσει ποιες συνεργασίες ανταποκρίνονται καλύτερα με τις ετικέτες των κατηγοριών. Αλλιώς, με αυτές τις κοινωνικές διαστάσεις, μπορεί να χρησιμοποιηθεί η δύναμη της διακριτικής μάθησης, όπως για παράδειγμα ο κατηγοριοποιητής SVM ή η λογιστική παλινδρόμηση, για να επιλεγούν αυτόματα οι σχετικές κοινωνικές διαστάσεις για την κατηγοριοποίηση. Στη φάση της πρόβλεψης, εκτός από τις υπάρχουσες σχεσιακές μεθόδους μάθησης, το συλλογικό συμπέρασμα καθίσταται περιττό όσο οι επιλεγμένες κοινωνικές διαστάσεις έχουν ήδη συμπεριλάβει τις σχετικές πληροφορίες σύνδεσης στο δίκτυο. Εξάλλου, το πλαίσιο που προκύπτει είναι ευέλικτο και επιτρέπει τον συνδυασμό των άλλων χαρακτηριστικών, όπως είναι τα προφίλ των χρηστών ή οι πληροφορίες κοινωνικού περιεχομένου.

4.5.1 Λανθάνουσες κοινωνικές διαστάσεις

Οι κοινωνικές διαστάσεις που προέρχονται από το δίκτυο θα πρέπει να ικανοποιούν τις ακόλουθες ιδιότητες:

- *Ενημερωτικές (Informative)*. Οι κοινωνικές διαστάσεις θα πρέπει να είναι ενδεικτικές των δεσμών μεταξύ των φορέων.
- *Πολυδιάστατες (Plural)*. Ο ίδιος κοινωνικός φορέας μπορεί να αναμιχθεί σε πολλαπλούς δεσμούς, συνεπώς εμφανίζονται σε διαφορετικές κοινωνικές διαστάσεις.
- *Συνεχείς (Continuous)*. Οι φορείς μπορεί να έχουν διαφορετικό βαθμό στις σχέσεις με ένα δεσμό. Ως εκ τούτου, είναι προτιμότερη μία συνεχής τιμή παρά μία διακριτή {0,1}.

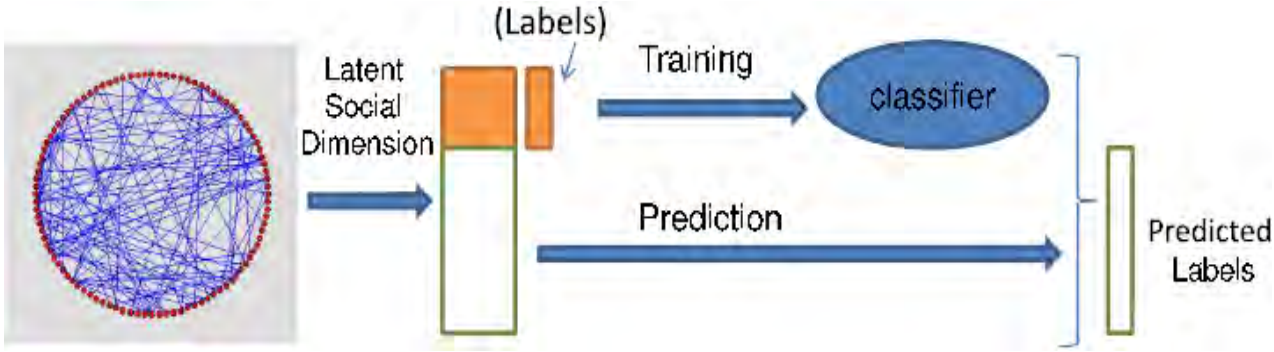
Ο σκοπός είναι να εξαχθούν οι κοινωνικές διαστάσεις που είναι ενδεικτικές των δεσμών μεταξύ των φορέων. Με βάση την ομοφυλία [14], παρόμοιοι φορείς αλληλεπιδρούν σε υψηλότερο ποσοστό από τους ανόμοιους. Έτσι, οι φορείς που μοιράζονται ορισμένες ιδιότητες τείνουν να σχηματίσουν ομάδες με πυκνότερες αλληλεπιδράσεις εντός των ομάδων. Αυτό συνδέεται φυσικά με ένα βασικό πεδίο στην ανάλυση - εντοπισμό κοινότητας των κοινωνικών δικτύων, το οποίο έχει ως στόχο να εντοπίζει κοινότητες που έχουν πυκνότερη αλληλεπίδραση εντός της ομάδας από ό,τι μεταξύ των ομάδων. Ενώ οι περισσότεροι αλγόριθμοι εντοπισμού κοινοτήτων διχοτομούν τους φορείς σε πολλές κομματιασμένες συστάδες, ο συγκεκριμένος αλγόριθμος επιτρέπει στον ίδιο φορέα να συμμετέχει σε διάφορους δεσμούς (συνεργασίες).

Στο σημείο αυτό, θα εξεταστεί εν συντομία η έννοια του σπονδυλωτή (modularity). Περισσότερα παρουσιάζονται στο Κεφάλαιο 5. Θεωρήστε την διαίρεση του πίνακα αλληλεπίδρασης A των n κορυφών και m ακμών, σε k μη επικαλυπτόμενες κοινότητες. Έστω ότι το s_i υποδηλώνει τη συμμετοχή της κορυφής u_i στη κοινότητα, και το d_i το βαθμό της κορυφής i. Ο σπονδυλωτής συμπεριφέρεται σαν ένα στατιστικό τεστ όπου το μηδενικό μοντέλο είναι ένα ομοιόμορφο τυχαίο μοντέλο γράφου, στο οποίο ένας φορέας συνδέεται με άλλους ακολουθώντας την ομοιόμορφη πιθανότητα. Για δύο κόμβους με βαθμούς d_i και d_j, αντίστοιχα, ο αναμενόμενος αριθμός των ακμών μεταξύ των δύο σε ένα ομοιόμορφο τυχαίο μοντέλο γράφου είναι $d_i d_j/2m$. Το modularity μετρά πόσο μακριά αποκλίνει η αλληλεπίδραση από ένα ομοιόμορφο τυχαίο γράφημα με την ίδια κατανομή βαθμού. Ορίζεται ως εξής:

$\mathrm{Q} = \frac{1}{2\mathrm{m}}\sum_{\mathrm{ij}}\left[\mathrm{A}_{\mathrm{ij}} - \frac{d_{\mathrm{i}}d_{\mathrm{j}}}{2\mathrm{m}}\right]\delta(s_{\mathrm{i}}, s_{\mathrm{j}})$, όπου $\delta(s_{\mathrm{i}}, s_{\mathrm{j}}) = 1$ εάν $s_{\mathrm{i}} = s_j$, και 0 σε κάθε άλλη περίπτωση. Σημειώστε ότι το Q θα μπορούσε να είναι αρνητικό, αν οι κορυφές χωρίζονταν σε κακές συστάδες. Όταν το Q είναι θετικό σημαίνει ότι η συσταδοποίηση αποτυπώνει κάποιο βαθμό της δομής της κοινότητας. Σε γενικές γραμμές, ο στόχος είναι να βρεθεί η δομή μιας κοινότητας έτσι ώστε το Q να μεγιστοποιείται.

Έστω ότι $\mathrm{d} \in \mathrm{Z}_{+}^{\mathrm{n}}$ είναι ο βαθμός του κάθε κόμβου, $\mathrm{S} \in \{0,1\}^{\mathrm{n}\times\mathrm{k}}$ ένας πίνακας της κοινότητας που ορίζεται ως εξής: $\mathrm{S} = \begin{cases} 1 & \text{εάν η κορυφή i ανήκει στη κοινότητα j} \\ 0 & \text{αλλιώς} \end{cases}$, και B ο πίνακας του σπονδυλωτή που ορίζεται ως εξής: $\mathrm{B} = \mathrm{A} - \frac{\mathrm{dd}^{\mathrm{T}}}{2\mathrm{m}}$. Τότε, ο σπονδυλωτής μπορεί να μετασχηματιστεί ως εξής: $\mathrm{Q} = \frac{1}{2\mathrm{m}}\mathrm{Tr}(\mathrm{S}^{\mathrm{T}}\mathrm{BS})$. Δεδομένου ότι το S είναι συνεχές, μπορεί να αποδειχθεί ότι το βέλτιστο S είναι τα top-k ιδιοδιανύσματα του modularity πίνακα. Μολονότι ο πίνακας αλληλεπιδράσεων A είναι συνήθως πολύ αραιός, ο modularity πίνακας B είναι πυκνός

και δεν μπορεί να υπολογιστεί και να διατηρηθεί στη μνήμη εάν το n είναι μεγάλο (το οποίο τυπικά ισχύει για πραγματικά κοινωνικά δίκτυα).



Σχήμα 4.8: Σχεσιακή μάθηση μέσω λανθάνουσων κοινωνικών διαστάσεων [42].

4.5.2 Περιγραφή SocDim αλγορίθμου

Σε αυτή την ενότητα, παρουσιάζεται η SocDim μέθοδος που χρησιμοποιείται για τη διακριτική σχεσιακή μάθηση. Η συνολική διαδικασία, βλέπε Σχήμα 4.8, αποτελείται από δύο βήματα:

1. *Εξαγωγή των λανθάνουσων κοινωνικών διαστάσεων που βασίζονται στη συνδεσιμότητα του δικτύου*. Εδώ, επικεντρωνόμαστε στον σπονδυλωτή (modularity – βλέπε Κεφάλαιο 5). Οι διαστάσεις μπορούν να εξαχθούν μέσω των top ιδιοδιανυσμάτων του modularity πίνακα B όπως ορίστηκε παραπάνω. Άλλες προσεγγίσεις στην συσταδοποίηση μπορούν, επίσης, να διερευνηθούν, όπως αναφέρθηκε προηγουμένως. Σημειώστε ότι ένα δίκτυο πραγματικού κόσμου είναι πολύ θορυβώδες συνεπώς κρατάμε εξ' αυτών μόνο τους top εκπροσώπους. Αυτό μειώνει επίσης το υπολογιστικό κόστος των μεγάλης κλίμακας υπολογισμών ενός ιδιοδιανύσματος. Από την στιγμή που οι επισημασμένοι και οι μη επισημασμένοι κόμβοι εμπλέκονται αμφότεροι στον υπολογισμό, οι λανθάνουσες κοινωνικές διαστάσεις μετά τον υπολογισμό είναι διαθέσιμες για όλους τους κόμβους. Με άλλα λόγια, σε αυτό το βήμα χρησιμοποιείται ο σπονδυλωτής προκειμένου να βρεθούν στη δομή του δικτύου οι διαστάσεις στις οποίες τοποθετούνται οι κόμβοι (ακολουθώντας τη θεωρία της ομοφυλίας, η οποία αναφέρει ότι οι φορείς μοιράζονται ορισμένες ιδιότητες, τείνει να σχηματίσει ομάδες). Αυτό συνήθως μπορεί να γίνει σε $O(n^2 \log n)$ βήματα. Αυτό το βήμα μπορεί να παραληφθεί αν ήδη υπάρχει γνώση των κοινωνικών διαστάσεων.

2. *Κατασκευή διακριτικού κατηγοριοποιητή*. Αφού εξαχθούν οι κοινωνικές διαστάσεις, μπορούν να θεωρηθούν ως συνήθη χαρακτηριστικά (συμπεριλαμβανομένων άλλων πιθανών πηγών) και έπειτα να διεξαχθεί επιβλεπόμενη μάθηση. Μπορεί να χρησιμοποιηθεί ένας οποιοσδήποτε κατηγοριοποιητής, όπως είναι ο SVM ή η λογιστική παλινδρόμηση. Αν κάποια άλλα χαρακτηριστικά είναι διαθέσιμα, όπως το προφίλ του χρήστη ή οι πληροφορίες περιεχομένου του blog, μπορούν επίσης να συμπεριληφθούν κατά τη διάρκεια της διακριτικής μάθησης. Αυτό το βήμα είναι ζωτικής σημασίας καθώς ο κατηγοριοποιητής θα καθορίσει ποιες διαστάσεις είναι σχετικές με μία ετικέτα κλάσης της ομάδας. Στην περίπτωση αυτή επιλέγεται ο one-vs-rest γραμμικός SVM λόγω της απλότητας και της επεκτασιμότητάς του [31]. Μπορούν επίσης να εφαρμοστούν πιο ισχυροί μέθοδοι, όπως οι διαρθρωτικοί SVM [32]. Είναι λοιπόν δυνατόν να χρησιμοποιηθούν οι προβλεπόμενες ετικέτες του κατηγοριοποιητή ώστε να ανακατασκευαστεί η οργάνωση της κοινότητας των οντοτήτων. Αυτό το βήμα είναι σημαντικό καθώς ο κατηγοριοποιητής θα καθορίσει ποιες διαστάσεις είναι σχετικές με μία ετικέτα κατηγορίας. Γενικά, η πρόβλεψη είναι εύκολη από τη στιγμή που ο κατηγοριοποιητής είναι έτοιμος, δεδομένου ότι οι λανθάνουσες κοινωνικές διαστάσεις έχουν υπολογιστεί για μη επισημασμένα δεδομένα στο βήμα 1. Σημειώστε ότι το συλλογικό συμπέρασμα δεν είναι απαραίτητο για την πρόβλεψη.

Η εργασία αυτή αποτελεί τη βάση μιας περαιτέρω εξέλιξης η οποία έχει μία ακμο-κεντρική θεώρηση των κοινοτήτων. Ο αλγόριθμος SocDim περιγράφεται περιληπτικά παρακάτω (σε συνδυασμό με το Σχήμα 4.8):

1. Εκπαίδευση:
   - ❖ Εξαγωγή κοινωνικών διαστάσεων για την εκπροσώπηση πιθανών διασυνδέσεων των φορέων
     - Εφαρμοστέες ενότητες: μεγιστοποίηση σπονδυλωτή, Laplacian γράφημα, κλπ.
   - ❖ Κατασκευή ενός κατηγοριοποιητή για να επιλεγούν αυτές οι διακριτικές διαστάσεις
     - Εφαρμοστέες ενότητες: SVM, λογιστική παλινδρόμηση, κλπ.

2. Πρόβλεψη:
   - ❖ Πρόβλεψη των ετικετών με βάση τις λανθάνουσες κοινωνικές διαστάσεις ενός φορέα.

4.6 PMM (Principal Modularity Maximization)

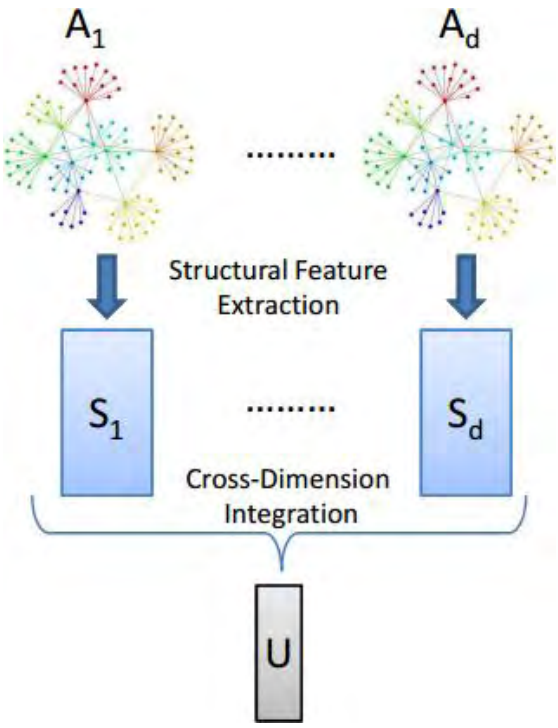
Εδώ παρουσιάζεται μία παραλλαγή της προσέγγισης του σπονδυλωτή (modularity) σε ένα πολυδιάστατο δίκτυο, όπου κάθε διάσταση αντιπροσωπεύει και έναν τύπο της αλληλεπίδρασης μεταξύ των οντοτήτων. Για παράδειγμα, στα διάσημα sites όπου μοιράζονται φωτογραφίες και

βίντεο (όπως τα Flickr και YouTube), ένας χρήστης μπορεί να συνδεθεί με τους φίλους του μέσω πρόσκλησης με e-mail ή με την λειτουργία «add as contacts», οι χρήστες μπορούν επίσης να προσθέσουν ετικέτες/σχόλια στα κοινωνικά περιεχόμενα όπως είναι οι φωτογραφίες και τα βίντεο, ένας χρήστης στο YouTube μπορεί να ανεβάσει ένα βίντεο για να απαντήσει σε ένα ποστάρισμα ενός βίντεο από έναν άλλο χρήστη, και ένας χρήστης μπορεί επίσης να γίνει φαν ενός άλλου χρήστη με εγγραφή στον «περίγυρο» του χρήστη των κοινωνικών περιεχομένων. Ένα δίκτυο αλληλεπίδρασης μπορεί να κατασκευαστεί με βάση κάθε μορφής δραστηριότητα, αναπαριστώντας μία πτυχή των ανθρώπινων αλληλεπιδράσεων. Σε αυτήν την ενότητα γίνεται αναφορά στην μεγιστοποίηση του σπονδυλωτή (modularity maximization), ένα μέτρο που αναπτύχθηκε πρόσφατα για την ποσοτικοποίηση των διαμερίσεων μιας κοινότητας στα κοινωνικά δίκτυα.

Σε αυτό το σημείο θα παρουσιαστεί η **PMM (Principal Modularity Maximization)** μέθοδος, η οποία παρακάμπτει το πρόβλημα της συγκρισιμότητας των διαφορετικών διαστάσεων. Ο στόχος του PMM αλγορίθμου είναι: δεδομένων πολλών διαφορετικών διαστάσεων, βρίσκει αρχικά μία συνοπτική αναπαράστασή τους (στη βιβλιογραφία αυτό το βήμα αναφέρεται ως «Εξαγωγή Δομικού Χαρακτηριστικού - Structural Feature Extraction», υπολογίζοντας τον σπονδυλωτή με την μέθοδο Lanczos. Ο τελευταίος είναι ένας αλγόριθμος όπου βρίσκει τις ιδιοτιμές και τα ιδιοδιανύσματα ενός τετραγωνικού πίνακα, με πολυπλοκότητα $O(mn^2)$), και έπειτα εντοπίζει τις συσχετίσεις μεταξύ αυτών των αναπαραστάσεων (στο «Ολοκλήρωση Διάσχισης των Διαστάσεων - Cross-Dimension Integration» βήμα, χρησιμοποιώντας μια γενικευμένη κανονική ανάλυση συσχέτισης, βλέπε άρθρο [33]).

Όπως προαναφέραμε και φαίνεται και στο Σχήμα 4.9, ο PMM αλγόριθμος αποτελείται από δύο βήματα:

1. *Structural Feature Extraction*. Δομικά χαρακτηριστικά ονομάζονται οι εξαγώγιμες διαστάσεις του δικτύου που είναι ενδεικτικές για τη δομή της κοινότητας. Για τη μεγιστοποίηση του σπονδυλωτή λογαριάζεται μία χαμηλών διαστάσεων ενσωμάτωση χρησιμοποιώντας τα ανώτατα ιδιοδιανύσματα του modularity πίνακα. Με άλλα λόγια, αυτά τα επιλεγμένα ιδιοδιανύσματα αποτελούν τις πιθανές διαμερίσεις της κοινότητας. Έτσι, τα ιδιοδιανύσματα μπορούν να αντιμετωπίζονται ως τα σημαντικά δομικά χαρακτηριστικά που προέρχονται από το δίκτυο.



Σχήμα 4.9: Επισκόπηση του PMM αλγορίθμου [35].

Σε αυτό το σημείο πρέπει να τονιστεί ότι πρώτα απ' όλα πρέπει να εξεταστεί η απομάκρυνση του θορύβου σε κάθε διάσταση του δικτύου. Δεδομένου ότι τα ιδιοδιανύσματα με αρνητικές ή μικρές ιδιοτιμές συμβάλλουν ελάχιστα στον σπονδυλωτή και είναι πολύ πιθανό να είναι θόρυβος, θα πρέπει να παραλείπονται. Σε ένα πολυδιάστατο δίκτυο, μπορούν να εξαχθούν κοινωνικά χαρακτηριστικά από κάθε διάσταση του δικτύου. Θα πρέπει να κρατούνται μόνο τα ιδιοδιανύσματα που έχουν μία θετική ιδιοτιμή. Για τη μείωση του θορύβου μπορούν επίσης να διατηρηθούν μερικοί κορυφαίοι δείκτες της κοινότητας.

2. *Cross-Dimension Integration*. Υποθέτοντας ότι μία λανθάνουσα δομή της κοινότητας μοιράζεται μεταξύ των διαφόρων διαστάσεων σε ένα πολυδιάστατο δίκτυο, αναμένεται ότι τα δομικά χαρακτηριστικά που εξάγονται θα πρέπει να είναι παρόμοια. Ωστόσο, τα χαρακτηριστικά με βάση τη μεγιστοποίηση του σπονδυλωτή δεν είναι μοναδικά. Ανόμοια δομικά χαρακτηριστικά δεν υποδηλώνουν ότι οι αντίστοιχες δομές της κοινότητας είναι δραστικά διαφορετικές.

Μετά το δεύτερο βήμα, σύμφωνα με την παγκόσμια βιβλιογραφία, γίνεται μία χαμηλότερων διαστάσεων ενσωμάτωση, η οποία ακολουθεί το κύριο μοτίβο σε όλες τις διαστάσεις του δικτύου. Τότε, σε αυτή την ενσωμάτωση μπορούν να υλοποιηθούν οι K – μέσοι ώστε να ανακαλυφθεί η εκχώρηση μιας διακριτής κοινότητας.

Αλγόριθμος 4.4 Principal Modularity Maximization αλγόριθμος (PMM)

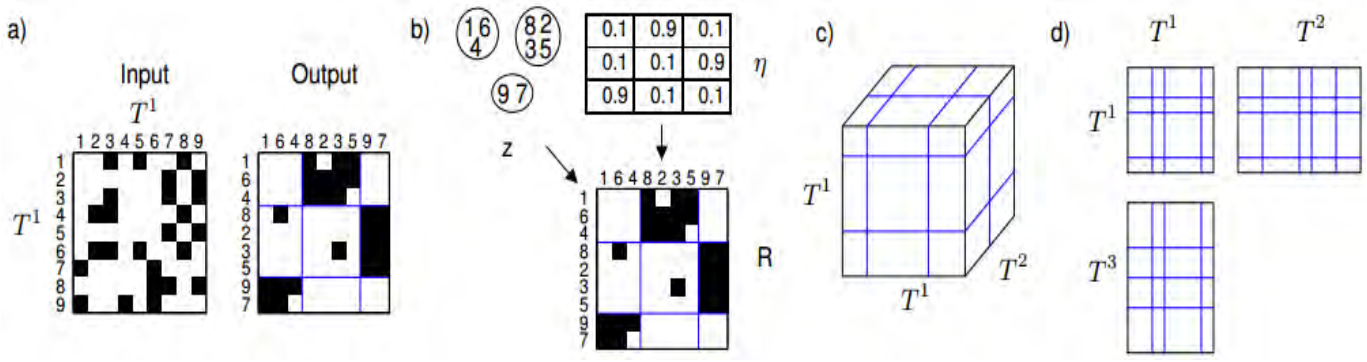
Είσοδος: Ένα δίκτυο A = {A_1, ..., A_d}, k ο αριθμός των κοινοτήτων, l ο αριθμός των δομικών χαρακτηριστικών προς εξαγωγή.
'Εξοδος: Εκχώρηση κοινότητας idx.

1: Υπολόγισε τα κορυφαία l ιδιοδιανύσματα του modularity matrix $\mathrm{B} = \mathrm{A} - \frac{\mathrm{dd^T}}{2\mathrm{m}}$ για κάθε A_i μέσω της Lanczos μεθόδου [15,16]
2: Επέλεξε τα διανύσματα με θετικές ιδιοτιμές ως S_i
3: Υπολόγισε την SVD [34] του X = [S_1, S_2, …., S_d] = UDV^T
4: Κάνε ενσωμάτωση χαμηλότερων διαστάσεων
5: Κανονικοποίησε τις γραμμές του πίνακα που προκύπτει από το προηγούμενο βήμα
6: Υπολόγισε τη συστάδα idx με την μέθοδο των K – μέσων στο αποτέλεσμα του προηγούμενου βήματος

Μία σύνοψη του αλγορίθμου υπάρχει στο άρθρο [35].

4.7 Μοντέλο Άπειρων Σχέσεων (Infinite Relational Model - IRM)

Ας υποτεθεί ότι δίνονται μία ή περισσότερες σχέσεις (δηλαδή, ακμές) που αφορούν έναν ή περισσότερους τύπους (δηλαδή, κόμβους). Ο στόχος του **Μοντέλου Άπειρων Σχέσεων (Infinite Relational Model - IRM)** είναι να διαμοιράσει κάθε κόμβο σε συστάδες (κοινότητες), όπου ένα καλό σύνολο των κατατμήσεων επιτρέπει τις σχέσεις μεταξύ των οντοτήτων να προβλέπονται από τις αναθέσεις της συστάδας. Για παράδειγμα, μπορεί να υπάρχει ένας μονός τύπος *people* και μία μονή σχέση *likes(i, j)*, γεγονός που δείχνει κατά πόσο στο άτομο (person) i αρέσει το άτομο j (person). Γενικά, σκοπός είναι να οργανωθούν οι οντότητες σε συστάδες που σχετίζονται μεταξύ τους με προβλέψιμους τρόπους (Σχήμα 4.10(a)), ομαδοποιώντας ταυτόχρονα τις σχέσεις και τις οντότητες. Επίσης, επιτρέπονται τύποι κατηγορήματα: αν υπάρχουν πολλαπλές σχέσεις που ορίζονται πάνω στον ίδιο τομέα, αυτές ομαδοποιούνται σε έναν τύπο και αναφέρονται ως κατηγορήματα. Για παράδειγμα, μπορεί να υπάρχουν αρκετά κοινωνικά κατηγορήματα που ορίζονται σύμφωνα με το πλαίσιο *people x people : likes(·, ·), admires(·, ·), respects(·, ·), και hates(·, ·).* Έτσι, μπορεί να εισαχθεί ένας τύπος για αυτά τα κοινωνικά κατηγορήματα, και να οριστεί μία τριμερής σχέση *applies(i, j, p)* η οποία είναι αληθής εάν το κατηγόρημα *p* ισχύει για το ζεύγος *(i, j)*. Στόχος είναι να ομαδοποιούνται τα άτομα και τα κατηγορήματα ταυτόχρονα (Σχήμα 4.10(c)). Ο IRM μπορεί να χειριστεί αυθαίρετα πολύπλοκα συστήματα χαρακτηριστικών, οντοτήτων, και σχέσεων: άμα συμπεριλαμβάνονται, για παράδειγμα, δημογραφικά χαρακτηριστικά για τα άτομα, τότε μπορούν να συσταδοποιηθούν τα άτομα, τα κοινωνικά κατηγορήματα, και τα δημογραφικά χαρακτηριστικά ταυτόχρονα.



Σχήμα 4.10: (a) Είσοδος και έξοδος όταν ο IRM εφαρμόζεται σε μία δυαδική σχέση R : T^1 x T^1 -> {0,1}. Ο IRM ανακαλύπτει μία διαμέριση των οντοτήτων, και ο πίνακας εισόδου παίρνει μια σχετικά ευδιάκριτη δομή μπλοκ όταν ταξινομείται σύμφωνα με αυτή τη διαμέριση. (b) Ο IRM υποθέτει ότι η σχέση *R* παράγεται από δύο λανθάνουσες δομές: μία διαμέριση *z* και έναν παραμετροποιημένο πίνακα *η*. Η σχέση *R(i, j)* παράγεται ρίχνοντας ένα μεροληπτικό κέρμα με τιμές $\eta(z_i, z_j)$, όπου τα z_i και z_j είναι οι αναθέσεις των οντοτήτων i και j σε συστάδα. Ο IRM αντιστρέφει αυτό το παραγωγικό μοντέλο για να ανακαλύψει την z και τον η που περιγράφουν καλύτερα τη σχέση R. (c) Συσταδοποίηση σχέσης τριών διαστάσεων R : T^1 x T^1 x T^2 -> {0,1}. Το T^1 μπορεί να είναι ένα σύνολο ατόμων, το T^2 ένα σύνολο κοινωνικών κατηγορημάτων, και η R μπορεί να διευκρινίζει εάν κάθε κατηγόρημα ισχύει για κάθε ζεύγος των ατόμων. Ο IRM ψάχνει για λύσεις, όπου κάθε τρισδιάστατο υπο-μπλοκ περιλαμβάνει ως επί το πλείστον είτε 1s είτε 0s. (d) Ταυτόχρονη συσταδοποίηση τριών σχέσεων. Το T^1 μπορεί να είναι ένα σύνολο ατόμων, το T^2 ένα σύνολο δημογραφικών χαρακτηριστικών, και το T^3 ένα σύνολο ερωτημάτων σχετικά με ένα τεστ προσωπικότητας. Σημειώστε ότι η διαμέριση για το T^1 είναι η ίδια όπου και αν αυτός ο τύπος εμφανίζεται [43].

Επισήμως, ας υποτεθεί ότι τα παρατηρούμενα δεδομένα είναι m σχέσεις που αφορούν n τύπους. Έστω, R^i η i-οστή σχέση, T^j ο j-οστός τύπος, και z^j ένα διάνυσμα των αναθέσεων στις συστάδες για τον T^j. Αυτό που πρέπει να γίνει είναι να συναχθούν οι αναθέσεις στις συστάδες· τελικά το όλο ενδιαφέρον επικεντρώνεται στην μεταγενέστερη κατανομή $P(z^1, \ldots, z^n \mid R^1, \ldots, R^m)$. Αυτή η κατανομή προσδιορίζεται ορίζοντας ένα παραγωγικό μοντέλο για τις σχέσεις και τις αναθέσεις στις συστάδες:

$$P(R^1, \ldots, R^m, z^1, \ldots, z^n) = \prod_{i=1}^{m} P(R^i | z^1, \ldots, z^n) \prod_{j=1}^{n} P(z^j)$$

όπου θεωρείται ότι οι σχέσεις είναι υπό όρους ανεξάρτητες, δεδομένων των αναθέσεων στις συστάδες. Επίσης, για κάθε τύπο οι αναθέσεις στις συστάδες είναι ανεξάρτητες. Για να ολοκληρωθεί το παραγωγικό μοντέλο, πρώτα περιγράφονται τα προγενέστερα των αναθέσεων στις συστάδες διανύσματα, $P(z^j)$, και έπειτα επιδεικνύεται, δεδομένου ενός συνόλου αυτών των διανυσμάτων, πως παράγονται οι σχέσεις.

4.7.1 Παραγωγή συστάδων

Για να επιτρέψει ο IRM τη δυνατότητα να ανακαλυφθεί ο αριθμός των συστάδων στον τύπο T, χρησιμοποιείται προγενέστερη γνώση που αναθέτει κάποια πιθανοτική μάζα σε όλες τις πιθανές διαμερίσεις του τύπου. Μια προγενέστερη λογική θα πρέπει να ενθαρρύνει το μοντέλο για να εισαγάγει μόνο όσες συστάδες δικαιολογούνται από τα δεδομένα. Ακολουθώντας προηγούμενες εργασίες για μη παραμετρικά Bayesian μοντέλα, χρησιμοποιείται μία κατανομή γύρω από τις διαμερίσεις που προκαλείται από μια κινέζικη διαδικασία που ονομάζεται Chinese Restaurant Process (CRP) [36].

Φανταστείτε την «οικοδόμηση ενός διαμερίσματος» ξεκινώντας από το μηδέν: αρχίζοντας με μία μόνο συστάδα που περιέχει ένα μόνο αντικείμενο, και προσθέτοντας αντικείμενα έως ότου όλα τα αντικείμενα να ανήκουν σε συστάδες. Σύμφωνα με την CRP, κάθε συστάδα προσελκύει νέα μέλη ανάλογα με το μέγεθός της. Η κατανομή σε συστάδες για το αντικείμενο i, που προετοιμάζονται για τις αναθέσεις στις συστάδες των αντικειμένων 1, ….. , i - 1 είναι:

$$\mathrm{P}(\mathrm{z_i} = \mathrm{a}|\mathrm{z_1}, \dots, \mathrm{z_{i-1}}) = \begin{cases} \dfrac{\mathrm{n_a}}{\mathrm{i} - 1 + \gamma}, & \mathrm{n_a} > 0, \mathrm{a\ είναι} \\ \dfrac{\gamma}{\mathrm{i} - 1 + \gamma} & \mathrm{μία\ νέα\ συστάδα} \end{cases}$$

Όπου n_a είναι ο αριθμός των αντικειμένων που έχουν ήδη εκχωρηθεί σε μία συστάδα α, και γ είναι μία παράμετρος. Η κατανομή στο z που προκαλείται από το CRP είναι ανταλλάξιμη: η σειρά με την οποία τα αντικείμενα έχουν ανατεθεί σε συστάδες μπορεί να μετατεθεί χωρίς να αλλάξει την πιθανότητα της προκύπτουσας διαμέρισης. Η P(z) μπορεί συνεπώς να υπολογιστεί επιλέγοντας μία αυθαίρετη τάξη και πολλαπλασιάζοντας τις υπό συνθήκη πιθανότητες, όπως ορίζεται παραπάνω. Δεδομένου ότι νέα αντικείμενα μπορούν πάντα να ανατεθούν σε νέες συστάδες, ο IRM έχει αποτελεσματικά πρόσβαση σε μία άπειρη συλλογή συστάδων, από την οποία προκύπτει και το πρώτο συνθετικό του ονόματός του. Η CRP είναι μαθηματικά εύκολη, και σύμφωνη με την διαίσθηση ότι θα πρέπει να ευνοούνται οι διαμερίσεις με μικρό αριθμό συστάδων. Ωστόσο, αυτό δεν αποτελεί μία καθολική λύση για το πρόβλημα της επιλογής του σωστού αριθμού των συστάδων. Ορισμένες φορές μπορεί να υπάρχει προγενέστερη γνώση που δεν συλλαμβάνεται από την CRP: για παράδειγμα, μπορεί να περιμένει κάποιος ότι οι συστάδες θα είναι σε μέγεθος περίπου ίσες. Ακόμα κι έτσι, η CRP αποτελεί ένα χρήσιμο εργαλείο εκκίνησης για τον εντοπισμό των δομών σε διάφορα τέτοια προβλήματα.

4.7.1.1 Παραγωγή σχέσεων από συστάδες

Θεωρείται ότι οι σχέσεις είναι δυαδικές συναρτήσεις, αν και οι επεκτάσεις στα δεδομένα συχνότητας και στα συνεχή δεδομένα είναι απλή υπόθεση. Θα εξεταστεί πρώτα το πρόβλημα με

έναν απλό τύπο T, και μία μόνο διπλότυπη σχέση R : T × T → {0, 1}. Ο τύπος T, για παράδειγμα, θα μπορούσε να είναι μία συλλογή ατόμων, και η σχέση R(i, j) θα μπορούσε να αναφέρει εάν στο άτομο i αρέσει το άτομο j. Το πλήρες παραγωγικό μοντέλο για αυτό το πρόβλημα είναι:

$z|\gamma \sim CRP(\gamma)$

$\eta(a,b)|\beta \sim Beta(\beta,\beta)$ (4.7.1.1.1)

$R(i,j)|z,\eta \sim Bernoulli(\eta(z_i,z_j))$

, όπου a, b ∈ N. Το μοντέλο αναπαρίσταται γραφικά στο Σχήμα 4.10(b).

Εδώ, θεωρείται ότι η τάση μιας οντότητας να συμμετέχει στις σχέσεις καθορίζεται εξ' ολοκλήρου από την ανάθεσή της σε μία συστάδα. Η παράμετρος η(a, b) προσδιορίζει την πιθανότητα ότι υπάρχει σύνδεση μεταξύ κάθε δεδομένου ζεύγους (i, j) όπου το i ανήκει στην συστάδα a και το j ανήκει στην συστάδα b. Σε κάθε είσοδο στον πίνακα η τοποθετούνται συμμετρικοί συζευγμένοι priors (με υπερ-παράμετρο β).

Για να καθοριστεί μία πιο γενική έκδοση του IRM, επεκτείνεται η Σχέση 4.7.1.1.1. Θεωρήστε μία m διαστάσεων σχέση R που περιλαμβάνει n διαφορετικούς τύπους. Έστω ότι d_k είναι η ετικέτα του τύπου που καταλαμβάνει την k διάσταση: για παράδειγμα, η τριών τύπων σχέση R : T^1 x T^1 x T^2 -> {0,1} έχει $d_1 = d_2 = 1$, και $d_3 = 2$. Όπως και πριν, η πιθανότητα ότι ισχύει η σχέση μεταξύ μιας ομάδας οντοτήτων, εξαρτάται μόνο από τις συστάδες των εν λόγω οντοτήτων:

$R(i_1,\ldots, i_m)|z^1,\ldots\ldots, z^n,\eta \sim Bernoulli(\eta(z_{i_1}^{d_1},\ldots\ldots, z_{i_m}^{d_m}))$.

Σε προβλήματα με πολλαπλές σχέσεις, για κάθε σχέση R^i εισάγεται ως παράμετρος ένας πίνακας η^i. Μπορούν να εξαχθούν συμπεράσματα με τη χρήση των Markov Chain Monte Carlo (MCMC) μεθόδων για την πραγματοποίηση δειγματοληψιών από τις μεταγενέστερες αναθέσεις σε συστάδες. Η IRM μέθοδος έχει πολύ υψηλή χρονική πολυπλοκότητα ($O(n^{2c} D)$).

4.8 Εύρεση Φυλών (Find Tribes - FT)

4.8.1 Η βασική ιδέα του αλγορίθμου Εύρεσης Φυλών

Η **Εύρεση Φυλών (FT)** δεν αναπτύχθηκε ρητά για τους σκοπούς του εντοπισμού μιας κοινότητας. Ωστόσο, η τεχνική αυτή μπορεί ακόμη να χρησιμοποιηθεί και για τον προσδιορισμό κάποιου είδους της κοινότητας σε ένα δίκτυο. Είναι πολύ κοντά στον ορισμό που δόθηκε για την κοινότητα με βάση την ενέργεια (action-based): οι οντότητες σε μία ομάδα τείνουν να συμπεριφέρονται με τον ίδιο τρόπο.

Στα σύνολα δεδομένων σε σχεσιακά και κοινωνικά δίκτυα, η κοινωνική δομή μεταξύ των ατόμων προσφέρει μία ζωτικής σημασίας επεξηγηματική δύναμη για διάφορες εργασίες πρόβλεψης. Ρίχνοντας μία πιο λεπτομερή ματιά στις συνδέσεις μεταξύ των οντοτήτων, ιδιαίτερα στους δυναμικούς διαχρονικούς τομείς, αυτό βοηθάει σημαντικά στις αναλύσεις των δεδομένων. Αυτή η ενότητα προσπαθεί μέσω ενός παραδείγματος να συμπεράνει τις στενές σχέσεις μεταξύ ορισμένων συναδέλφων, δεδομένης μιας βάσης δεδομένων από το ιστορικό των συνεργασιών που αναπτύχθηκαν. Συγκεκριμένα, το ζητούμενο είναι η εύρεση κάποιων ομάδων ατόμων, που ονομάζονται **φυλές**, οι οποίες έχουν αφύσικα παρόμοιες αλληλουχίες θέσεων εργασίας στο πλαίσιο μιας μεγάλης βιομηχανίας. Αυτό που πρέπει να γίνει είναι να εντοπιστούν οι εργαζόμενοι που ήταν συν-εργαζομένοι σε πολλαπλές θέσεις εργασίας, και να γίνει διάκριση μεταξύ αυτών που συνεργάστηκαν σκοπίμως, από εκείνους που απλά μοιράζονται συχνά εμφανιζόμενες μορφές απασχόλησης στη βιομηχανία.

Οι οντότητες θα πρέπει να συνδεθούν με αρκετά χαρακτηριστικά. Στόχος του αλγορίθμου είναι να επιστρέψει αυτές τις ομάδες-φυλές που μοιράζονται «ασυνήθιστους» συνδυασμούς των χαρακτηριστικών. Ο περιορισμός αυτός μπορεί εύκολα να γενικευτεί προκειμένου να ληφθούν ακόμη και οι «συνήθεις» ομάδες ως έξοδοι. Ο αλγόριθμος προχωρά με τον εντοπισμό όλων των σημαντικών ζευγών, δηλαδή πόσο σημαντική ή ασυνήθιστη είναι η αλληλουχία των κοινών θέσεων εργασίας. Δεδομένης μιας ακολουθίας θέσεων εργασίας, πρέπει να αποφασιστεί εάν είναι ασυνήθιστο για ένα ζευγάρι συναδέλφων να έχουν εργαστεί από κοινού σε όλες αυτές τις θέσεις εργασίας.

Γενικά, όσον αφορά τη βασική διαδικασία της εύρεσης φυλών, επισήμως δίνεται ένα διμερές γράφημα $G = (R \cup A, E)$ των οντοτήτων (ή αλλιώς αντιπροσώπων) $R = \{r_1, r_2, ..., r_n\}$ και των χαρακτηριστικών (ή αλλιώς οργανισμών) $A = \{a_1, a_2, ..., a_m\}$. Κάθε ακμή $e \in E$ επισημαίνεται από ένα χρονικό διάστημα: $e = (r_i, a_j, tstart_{ij}, tend_{ij})$. Η διαδικασία ανακάλυψης φυλής ξεκινάει με την εύρεση όλων των ζευγών $f_{ij} = (r_i, r_j)$ των ατόμων που έχουν εργαστεί ποτέ μαζί (αυτή θα μπορούσε να είναι μία τεράστια λίστα). Για κάθε ζευγάρι συνοψίζονται οι σχέσεις συναδελφικότητας, παρακολουθώντας τις σχέσεις εργασίας όπου αυτές συμπίπτουν. Επίσης, μπορεί να καταγραφεί μία πρόσθετη πληροφορία, όπως είναι η ημερομηνία που οι αντιπρόσωποι εργάστηκαν για πρώτη φορά σε κάθε δουλειά, καθώς και ο συνολικός χρόνος που ξόδεψαν σε επικαλυπτόμενες δουλειές. Ο αλγόριθμος αποθηκεύει τα ζεύγη σε ένα νέο γράφημα $H = (R, F)$, όπου το $F = \{ f_{ij} \}$, και κάθε ακμή σημειώνεται με:

q_{ij} = { αλληλουχία των θέσεων εργασίας $\{a_x, a_y, ...\}$ που μοιράζονται από τις r_i και r_j ∪ πρόσθετη πληροφορία }.

Για λόγους απλότητας, θα διατηρηθούν μόνο τα ζεύγη οντοτήτων που έχουν τουλάχιστον τρεις κοινές θέσεις εργασίας. Έτσι προκύπτει το γράφημα $H'(R, F')$.

Η στρατηγική γι' αυτό που γίνεται περιστρέφεται γύρω από την ανάπτυξη ενός καλού ορισμού του «ασυνήθιστου». Για μία ομάδα οντοτήτων που θεωρείται ανώμαλη, τα κοινά

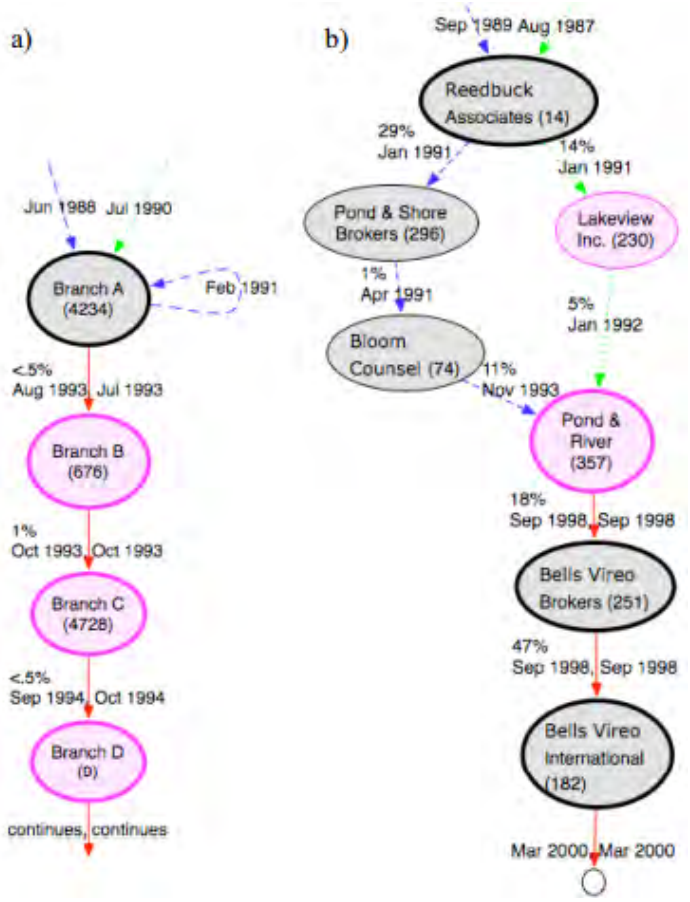
χαρακτηριστικά από μόνα τους δεν χρειάζεται να είναι ασυνήθιστα, αλλά η συγκεκριμένη διάταξή τους πρέπει να είναι.

Ο αλγόριθμος προχωρά με τον εντοπισμό όλων των σημαντικών ζευγών. Για κάθε ακμή στο F', υπολογίζεται ένα αποτέλεσμα/βαθμολογία $c_{ij}(q_{ij})$ (ο αριθμός των γνωρισμάτων στην κοινή ακολουθία, ο αριθμός των χρονικών βημάτων της επικάλυψης, η πιθανοτική αλυσίδα Markov των γνωρισμάτων και ούτω καθ' εξής), το οποίο μετρά πόσο σημαντική ή ασυνήθιστη είναι η ακολουθία των κοινών θέσεων εργασίας. Όταν υπολογίζονται τα σημαντικά αποτελέσματα, τότε επιλέγεται ένα κατώφλι d για αυτά, και διατηρούνται μόνο οι f_{ij} ακμές για τις οποίες ισχύει $c_{ij} >$ d. Στη συνέχεια, υπολογίζονται τα συνδεδεμένα συστατικά του H', τα οποία καθορίζουν και τις φυλές. Η έξοδος του αλγορίθμου περιλαμβάνει μια λίστα φυλών. Η συνολική πολυπλοκότητα του αλγορίθμου είναι $O(mn K^2)$.

Σχήμα 4.11: Παράδειγμα (υποθετικό) του κλάδου-υποκαταστήματος μοτίβων μετάβασης. Η πιο αριστερή ακμή σημαίνει ότι το 80% των εργαζομένων που έχουν ποτέ εργαστεί στον κλάδο του I&D Insurance, αργότερα εργάστηκαν σε αυτόν του Cumulative Sentences. Εμφανίζονται μόνο οι ακμές με τα υψηλά ποσοστά [44].

4.8.2 Ένα πιθανοτικό μοντέλο

Στην ανάπτυξη ενός απλοποιημένου μοντέλου για τα δεδομένα του ιστορικού των θέσεων εργασίας, υπάρχει ένα δίλημμα για το πόσο συγκεκριμένο πρέπει να είναι. Το ζητούμενο είναι ένα ευέλικτο μοντέλο που να μιμείται τα χαρακτηριστικά του κάθε κλάδου, χωρίς να αναπαράγει ακριβώς τα αρχικά δεδομένα. Επιπλέον, η διαδικασία θα πρέπει να είναι προσιτή σε ένα μεγάλο σύνολο δεδομένων. Προσπαθώντας να βρεθεί η σωστή ισορροπία, μοντελοποιείται η κίνηση των οντοτήτων εντός των κλάδων ως μία τροποποίηση μιας αλυσίδας Markov πάνω από τους οργανισμούς, αγνοώντας το χρόνο και τη διάρκεια.



Σχήμα 4.12: Ακολουθίες θέσεων εργασίας για βαθμολόγηση. Οι κόμβοι δείχνουν τους κλάδους και τα μεγέθη τους. Τα βέλη που οδηγούν σε έναν κόμβο δείχνουν τις ημερομηνίες που η νέα δουλειά έχει ξεκινήσει, και τις πιθανότητες μετάβασης. Οι συμπαγείς κόκκινες γραμμές είναι κινήσεις που εκτελούνται και από τις δύο οντότητες στο ζεύγος. Οι διακεκομμένες μπλε γραμμές είναι οι κινήσεις από ένα μέλος μόνο, ενώ οι πράσινες διακεκομμένες από το άλλο. Τα εταιρικά ονόματα είναι κατασκευασμένα για να προτείνουν αντιστοιχίες που είναι ορατές στα πραγματικά δεδομένα. Αυτά τα διαγράμματα έχουν τροποποιηθεί από ζευγάρια που βαθμολογήθηκαν ως σημαντικά. Ως εκ τούτου, οι ημερομηνίες έναρξης των οντοτήτων ταιριάζουν στενά, αν και οι πληροφορίες για τον χρόνο δεν χρησιμοποιήθηκαν στη βαθμολόγηση [44].

Εάν η κάθε οντότητα κατείχε μία θέση εργασίας σε μια χρονική στιγμή και την άλλαζε σε κάθε χρονικό βήμα, αυτή η κίνηση θα μπορούσε να μοντελοποιηθεί χρησιμοποιώντας μια συνηθισμένη αλυσίδα Markov, ως εξής. Κάθε οντότητα επιλέγει έναν κλάδο έναρξης τυχαία. Στη συνέχεια, σε κάθε βήμα, ο επόμενος κλάδος της οντότητας αποφασίζεται πιθανοτικά και βασίζεται μόνο στον τρέχοντα κλάδο. Αγνοείται ο πραγματικός χρόνος που αφιερώνεται σε κάθε εργασία. Σε κάθε βήμα της διαδικασίας Markov, η οντότητα είτε μετακινείται σε έναν νέο κλάδο, είτε εγκαταλείπει το χώρο εργασίας. Αν αυτό ήταν το μοντέλο, τότε θα δινόταν βάση στις εξής ποσότητες: p_i = P(εκκίνηση από τον κλάδο i), και t_{ij} = P(μετάβαση από τον κλάδο i στον κλάδο j | προσωρινά στον κλάδο i). Στη συνέχεια, θα μπορούσε να υπολογιστεί η πιθανότητα το να έχει μία οντότητα μία οποιαδήποτε ακολουθία θέσεων εργασίας ως εξής:

x = P(κλάδος A → κλάδος B → κλάδος C → κλάδος D) = $p_A \cdot t_{AB} \cdot t_{BC} \cdot t_{CD}$.

Οι πιθανότητες είναι ευθέως υπολογισμένες χρησιμοποιώντας:

p_i = # οντοτήτων που ήταν ποτέ στον κλάδο i / # οντοτήτων στη βάση δεδομένων

t_{ij} = # οντοτήτων που εγκαταλείπουν τον κλάδο i και στη συνέχεια πηγαίνουν στον κλάδο j / # οντοτήτων που ήταν ποτέ στον κλάδο i.

Χρησιμοποιώντας τη συνήθη αλυσίδα Markov και την αρχική υπόθεση της ανεξάρτητης κίνησης, θα μπορούσε να βαθμολογηθείε η ακολουθία του Σχήματος 4.12 ως εξής,

1. P(οντότητα 1 κατέχει αυτή την ακολουθία των θέσεων εργασίας) = x.
2. P(οντότητα 1 και 2 έκαστος κατέχει αυτή την ακολουθία των θέσεων εργασίας) = x^2.
3. P(κάποιες δύο οντότητες της βάσης δεδομένων κατέχουν αυτή την ακολουθία των θέσεων εργασίας) ακολουθεί μια διωνυμική κατανομή, με το n = # οντοτήτων στη βάση δεδομένων, και p = x^2.

Περαιτέρω, δεν είναι απαραίτητο να υπολογιστεί ο παρονομαστής του p_i. Για παράδειγμα στο Σχήμα 4.12a), το αποτέλεσμα θα ήταν της μορφής $p_A' \cdot t_{AB} \cdot t_{BC} \cdot t_{CD}$ = (4234)(.005)(.01)(.005). Για τις περιπτώσεις όπου οι οντότητες αρχίζουν ή τελειώνουν σε ξεχωριστές εργασίες, βαθμολογείται η ακολουθία την οποία μοιράζονται.

Αν οι ακολουθίες των θέσεων εργασίας στη βάση δεδομένων ήταν τόσο απλές όπως φαίνεται στο Σχήμα 4.12a), τότε το μοντέλο αυτό θα είναι επαρκές. Όμως, το Σχήμα 4.12b) αποτελεί ένα πιο χαρακτηριστικό παράδειγμα των δεδομένων. Οι οντότητες σε αυτό το παράδειγμα ξεκινούν από τον ίδιο κλάδο, διαχωρίζονται για μερικά χρόνια, επιστρέφουν μαζί, και στη συνέχεια αμφότερες αρχίζουν δύο δουλειές σε σχετικές εταιρίες ταυτόχρονα. Η σημαντικότερη τροποποίηση είναι να επιτραπεί στις οντότητες να έχουν διαφορετικά μονοπάτια μεταξύ των κοινών θέσεων εργασίας, όπως φαίνεται κοντά στην κορυφή του Σχήματος 4.12b). Για να γίνει αυτό, αλλάζει η ποσότητα t_{ij}, η οποία περιγράφει την πιθανότητα μία οντότητα να μετακινείται στον κλάδο j αμέσως μετά τον κλάδο i, σε μία νέα ποσότητα v_{ij}, η οποία με τη σειρά της περιγράφει τη πιθανότητα μία οντότητα να μετακινείται στον κλάδο j σε *οποιοδήποτε* σημείο αφού τελειώσει την εργασία στον κλάδο i. Τώρα, κάθε $v_{ij} > t_{ij}$, και οι πιθανότητες μετάβασης απομάκρυνσης από ένα κλάδο δεν αθροίζονται πλέον στην τιμή 1 ($\sum_i t_{ij} = 1$, αλλά $\sum_i v_{ij} \geq 1$). Για το Σχήμα 4.12b), μπορεί επομένως να υπολογιστεί η P(Reedbuck → Pond & River)(το ποσοστό δεν απεικονίζεται στο σχήμα), χωρίς να λαμβάνονται υπ' όψιν οι ενδιάμεσοι κλάδοι. Αυτή η τροποποίηση είναι πιο ξεκάθαρη από μια εναλλακτική προσέγγιση που θα μπορούσε να επιχειρήσει να υπολογίσει τις πιθανότητες των άμεσων μεταβάσεων κατά μήκος όλων των πιθανών μονοπατιών.

Η άλλη τροποποίηση είναι να επιτραπούν οι ταυτόχρονες θέσεις εργασίας. Αντιμετωπίζονται οι ακολουθίες των κοινών εργασιών σαν να είναι σε μία συγκεκριμένη σειρά. Για παράδειγμα, η οντότητα 1 μπορεί να ξεκινήσει από τον κλάδο A, έπειτα να εισχωρήσει στον B, όσο η οντότητα

2 ξεκινάει από τον κλάδο Β και αργότερα εισχωρεί στον κλάδο Α. Έτσι, οι οντότητες επικαλύπτονται στον κλάδο Β προτού «διπλαρώσουν» στον κλάδο Α, παρόλο που η οντότηα 1 δεν άφησε ποτέ τον κλάδο Β για τον κλάδο Α. Ή, όπως φαίνεται και στο Σχήμα 4.12b), οι οντότητες μπορούν να βρίσκονται αμφότερες ταυτόχρονα στο Bells Vireo, χωρίς η μία να έπεται της άλλης. Για να επεκταθεί το μοντέλο να χειρίζεται τέτοιες καταστάσεις, αντικαθιστάται η ποσότητα v_{ij}, δηλαδή η πιθανότητα μία οντότητα να μετακινείται στον κλάδο j σε οποιοδήποτε σημείο αφότου δουλέψει στον κλάδο i, με μία νέα ποσότητα w_{ij}, δηλαδή την πιθανότητα μία οντότητα να εργάζεται στον κλάδο j σε οποιοδήποτε σημείο ταυτόχρονα με τη δουλειά στον κλάδο i ή και μετά το πέρας αυτής. Οι πιθανότητες μετάβασης που φαίνονται στο Σχήμα 4.12, είναι στη ουσία w_{ij} τιμές, έτσι το παράδειγμα υπολογισμού για το Σχήμα 4.12a) υπολογίζεται όπως συζητήθηκε προηγουμένως, αλλά η σημασία των πιθανοτήτων είναι διαφορετική.

4.8.3 Μία παραλλαγή του μοντέλου

Το πιθανοτικό μοντέλο βαθμολόγησης που περιγράφεται παραπάνω αντιμετωπίζει τις θέσεις εργασίας σε μία ακολουθία που έχει διαταχθεί σύμφωνα με το χρόνο, αλλά δεν λαμβάνει υπόψη πότε συμβαίνουν οι μεταβάσεις. Η μετάβαση θεωρείται εξίσου πιθανή οποτεδήποτε και αν λαμβάνει χώρα. Έτσι, δημιουργείται μία παραλλαγή του μοντέλου, αλλάζοντας την αντιμετώπιση της έννοιας του χρόνου.

Πρώτα απ’ όλα, το μοντέλο ενδιαφέρεται για διαφορετικές πιθανότητες μετάβασης. Θεωρείται ότι η βαθμολόγηση θα είναι πιο ακριβής αν μπορέσουν να αναπαρασταθούν οι κινήσεις των μεμονωμένων συμβάντων, καθώς αυτές αλλάζουν σε βάθος χρόνου στα πρότυπα της βιομηχανίας. Για παράδειγμα, θεωρείστε την περίπτωση όπου το 30% των οντοτήτων στον κλάδο Α τελικά μετακινούνται στον κλάδο Β, αλλά το 99% των οντοτήτων του κλάδου Α το 1997 παρατηρήθηκαν αργότερα στον κλάδο Β. Έτσι, αντί να βαθμολογείται μία μετάβαση που βασίζεται στην πιθανότητα μία οντότητα να μετακινείται από τον κλάδο Α στον κλάδο Β, περιγράφεται ένα πιο συγκεκριμένο γεγονός. Τώρα, μία οντότητα μετακινείται από τον κλάδο Α τη χρονική στιγμή Χ, στον κλάδο Β τη χρονική στιγμή Υ (ειδικότερα, η οντότητα παρατηρείται για πρώτη φορά στον κλάδο Α τη χρονική στιγμή Χ, και έπειτα παρατηρείται για πρώτη φορά στον κλάδο Β τη χρονική στιγμή Υ, που είναι ίση ή μεταγενέστερη της Χ). Ο χρόνος διαιρείται σε κομμάτια, με καθένα από αυτά να αντιπροσωπεύει ένα ή περισσότερα έτη. Κάθε κλάδος έχει τις δικές του τέτοιες διαιρέσεις, που εξαρτώνται από τον αριθμό των εργαζομένων στον κλάδο σε διαφορετικά έτη.

Οι παράμετροι που χρειάζονται γι’ αυτό το νέο μοντέλο απαιτούν την αλλαγή του p_i και (ξανά) του w_{ij}. Σε αυτό το σημείο υπολογίζονται:

p_{iX} = # οντοτήτων που ήταν ποτέ στον κλάδο i κατά τη διάρκεια του χρόνου X / # οντοτήτων στη βάση δεδομένων

y_{iXjY} = # οντοτήτων που ήταν ποτέ στον κλάδο i κατά τη διάρκεια του χρόνου X και στον κλάδο j κατά τη διάρκεια του χρόνου Y (όπου το $Y \geq X$) / # οντοτήτων που ήταν ποτέ στον κλάδο i κατά τη διάρκεια του χρόνου X.

Παρακάτω, θα περιγραφεί εν συντομία μία ακόμη (απλούστερη) παραλλαγή του μοντέλου, που αποδεικνύει πόσο σημαντικός είναι ο χρόνος. Σε αυτή την παραλλαγή αγνοείται ακόμη και η σειρά των κινήσεων των θέσεων εργασίας. Γι' αυτό το μοντέλο, χρησιμοποιείται η αρχική p'_i (και πάλι, δεν χρειάζεται να υπολογιστεί ο παρονομαστής), και μία (τελική) ποσότητα μετάβασης z_{ij}, που αναπαριστά τον αριθμό των οντοτήτων που βρίσκονται ταυτόχρονα στους κλάδους i και j, καθ' όλη τη διάρκεια της καριέρας τους. Υπάρχει μία ασάφεια σε όλο αυτό, στο ότι τώρα θα πρέπει να είναι σε θέση να βαθμολογηθεί ένα σύνολο κοινών κλάδων, ανεξάρτητα από τη σειρά που παρουσιάστηκαν. Όμως, η πιθανότητα μετάβασης από τον i στον j = (z_{ij} / p'_i) ≠ (z_{ij} / p'_j) = πιθανότητα μετάβασης από τον j στον i. Για την ώρα, χρησιμοποιείται η ίδια, χρονική διάταξη των κλάδων όπως χρησιμοποιήθηκε στις άλλες μεθόδους.

4.9 Autopart

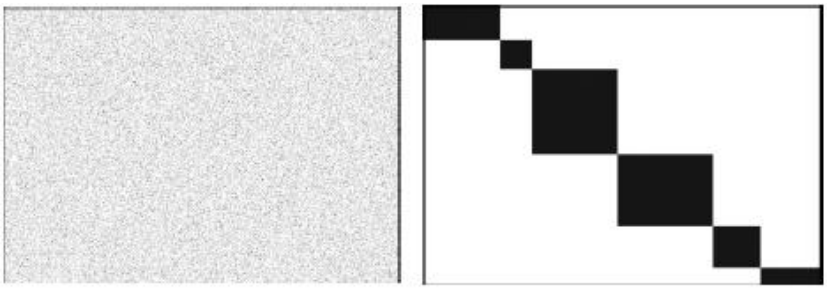
Αρκετές από τις δημοφιλείς μεθόδους για την ανάλυση γραφημάτων, όπως για παράδειγμα ο αλγόριθμος των K – μέσων, απαιτούν από το χρήστη να προσδιορίσει διάφορες παραμέτρους, όπως είναι ο αριθμός των συστάδων, ο αριθμός των διαμερίσεων και ο αριθμός των κύριων συστατικών. Εδώ, παρουσιάζεται μία μέθοδος ομαδοποίησης των κόμβων χρησιμοποιώντας αρχές της θεωρίας της πληροφορίας, για να επιλεγούν τόσο ο αριθμός των ομάδων όσο και η αντιστοίχιση κόμβων-ομάδων. Οι αλγόριθμοι που θα παρουσιαστούν είναι εντελώς απαλλαγμένοι από παραμέτρους (parameter-free), και επίσης πρακτικά «κλιμακώνονται» γραμμικά με το μέγεθος του προβλήματος. Οι αλγόριθμοι αυτοί χρησιμοποιούνται σε προβλήματα που έχουν τους εξής στόχους:

- Συστάδες: «παρόμοιοι» κόμβοι πρέπει να ομαδοποιηθούν σε «φυσικές» συστάδες.
- Όρια: οι αποκλίνουσες ακμές από τη συνολική δομή πρέπει να επισημαίνονται ως όρια (ή ακραίες τιμές).
- Αποστάσεις συστάδων: για κάθε ζεύγος συστάδων, θα πρέπει να ορίζεται ένα μέτρο της «απόστασης» μεταξύ αυτών.

Επιπλέον, οι αλγόριθμοι θα πρέπει να έχουν τις ακόλουθες βασικές ιδιότητες:

- Αυτόματοι: προτείνεται μία διαισθητική, και βάσει αρχών, διατύπωση του προβλήματος, έτσι ώστε ο χρήστης να μην χρειάζεται να ορίσει κάποια παράμετρο.
- Δυνατότητα αναβάθμισης: θα πρέπει να αναβαθμίζονται για μεγάλα γραφήματα
- Στοιχειώδεις: θα πρέπει να επιτρέπουν online επαναϋπολογισμούς των αποτελεσμάτων όταν προστίθενται νέοι κόμβοι και ακμές˙ αυτό θα επιτρέψει τη μέθοδο να προσαρμοστεί στα νέα εισερχόμενα δεδομένα.

Στόχος μας είναι να βρεθούν μοτίβα σε ένα μεγάλο γράφημα, χωρίς παρέμβαση του χρήστη, όπως φαίνεται στο Σχήμα 4.13. Η μέθοδος **Autopart** αποτελεί βασικό συνθετικό της *MDL (Minimum Description Language)* προσέγγισης [40]. Η βασική ιδέα είναι η εξής: ο δυαδικός n × n πίνακας D = $[d_{i,j}]$ αναπαριστά ενώσεις μεταξύ των n κόμβων του γραφήματος (και τα χαρακτηριστικά τους - που αντιστοιχεί σε γραμμές και στήλες στον πίνακα γειτνίασης). Ένα παράδειγμα ενός πιθανού πίνακα γειτνίασης φαίνεται στο Σχήμα 4.13(α). Αν εξορύξουμε σωστά αυτές τις πληροφορίες, θα μπορούσαμε να αναδιατάξουμε τον πίνακα γειτνίασης έτσι ώστε οι «παρόμοιοι» κόμβοι να ομαδοποιούνται μεταξύ τους. Τότε, ο πίνακας γειτνίασης θα αποτελείται από ομογενή ορθογώνια/τετράγωνα blocks υψηλής (χαμηλής) πυκνότητας, αναπαριστώντας το γεγονός ότι ορισμένες ομάδες κόμβων συνδέονται περισσότερο (λιγότερο) με άλλες ομάδες (βλέπε δεξιά πλευρά του Σχήματος 4.13(β)). Για να συμπιεστεί ο πίνακας, θα ήταν προτιμότερο να υπήρχαν μερικά μόνο blocks, το καθένα από τα οποία να είναι πολύ ομογενές. Ωστόσο, έχοντας περισσότερες ομάδες, επιτρέπεται να δημιουργηθούν περισσότερα ομογενή blocks (σε ακραίες περιπτώσεις, έχοντας n ομάδες προκύπτουν n^2 απολύτως ομοιογενή blocks μεγέθους 1 × 1). Έτσι, το καλύτερο σύστημα συμπίεσης θα πρέπει να επιτυγχάνει μία ανταλλαγή μεταξύ αυτών των δύο παραγόντων, με αποτέλεσμα, αυτό το σημείο ανταλλαγής να υποδεικνύει και τον καλύτερο δυνατό αριθμό k των ομάδων των κόμβων. Αυτό επιτυγχάνεται μέσω μιας ειδικής εφαρμογής του γενικού MDL αλγορίθμου, όπου το κόστος συμπίεσης βασίζεται στον αριθμό των bits που απαιτούνται για τη μετάδοση τόσο της «περίληψης» των ομάδων των κόμβων, όσο και του κάθε block δεδομένων των ομάδων. Έτσι, ο χρήστης δεν χρειάζεται να ορίσει κάποια παράμετρο˙ ο αλγόριθμος τις επιλέγει έτσι ώστε να ελαχιστοποιηθούν αυτά τα κόστη. Ο σκοπός του αλγορίθμου είναι να βρεθεί η καλύτερη ομάδα που ελαχιστοποιεί τη συνάρτηση κόστους (συμπίεσης). Για περισσότερες πληροφορίες ελέγξτε το άρθρο [37].



(α) Ο αρχικός πίνακας (β) Ο αναδιατεταγμένος πίνακας

Σχήμα 4.13: Ένα παράδειγμα της MDL αρχής για πίνακες: ο πίνακας στα αριστερά είναι ακριβώς ο ίδιος πίνακας με εκείνον στα δεξιά, αλλά αναδιατεταγμένος, προκειμένου να την περιγράψει απλά.

4.9.1 Παρουσίαση αλγορίθμων

Προηγουμένως, καθορίστηκε ο στόχος του αλγορίθμου: ανάμεσα σε όλες τις πιθανές τιμές για το k, και όλες τις πιθανές ομάδες κόμβων G, επιλέξτε μία διάταξη που μειώνει όσο το δυνατόν περισσότερο το συνολικό κόστος συμπίεσης, όπως υποδεικνύει το μοντέλο MDL (μοντέλο συν τα δεδομένα). Το πρόβλημα θα λυθεί μέσω μιας επαναληπτικής διαδικασίας δύο βημάτων:

1. *InnerLoop*. Δεδομένου ενός αριθμού k των ομάδων των κόμβων, να βρεθεί μία καλή ομαδοποιήση των κόμβων G.
2. *OuterLoop*. Να αναζητηθεί αποτελεσματικά το καλύτερο k (k = 1, 2, . . .).

Αλγόριθμος 4.5 Αλγόριθμος InnerLoop

1: Έστω ότι το t χαρακτηρίζει το δείκτη επανάληψης. Αρχικά, θέστε t = 0. Αν δεν παρέχεται μία G(0), ξεκινήστε με μια αυθαίρετη G(0) αντιστοιχίζοντας κόμβους σε k ομάδες κόμβων. Για αυτή την αρχική διαμέριση, υπολογίστε τους υποπίνακες $D_{i,j}(t)$, και τις αντίστοιχες κατανομές $P_{i,j}(t)$.

2: Για κάθε κόμβο x, συνδέστε την αντίστοιχη γραμμή του πίνακα σε k τμήματα $x_{row,1}, \ldots, x_{row,k}$ σύμφωνα με το G(t) (δηλαδή, $x_{row,\ 1} = \{d_{x,\ u} \mid G_u\ (t) = 1\}$ και ούτω καθ' εξής). Ομοίως, συνδέστε τη στήλη σε k τμήματα $x_{col,1}, \ldots, x_{col,k}$. Υπολογίστε τον αριθμό των άσσων («1» - λέγεται βάρος και συμβολίζεται με το γράμμα w) $w(x_{row,j})$ and $w(x_{col,j})$ (j = 1 . . . k) για όλα αυτά τα τμήματα. Τώρα, εκχωρήσετε τον κόμβο x στην ομάδα κόμβων $G_x(t+1)$ έτσι ώστε

$$G_x(t+1) = arg_{1\le i\le k}\, min\ \Big\{\sum_{j=1}^{k} -[w(x_{row,j})\, log\, P_{i,j}(t) + (n(x_{row,j}) - w(x_{row,j}))log(1 - P_{i,j}(t))\# + w(x_{col,j})\, log\, P_{j,i}(t) + (n(x_{col,j}) - w(x_{col,j}))log(1 - P_{j,i}(t))]\# + d_{x,x}[\, logP_{i,G_x(t)}(t) + logP_{G_x(t),i}(t) - logP_{i,i}(t)]\# + (1 - d_{x,x})[log(1 - P_{i,G_x(t)}(t)) + log(1 - P_{G_x(t),i}(t)) - log(1 - P_{i,i}(t))]\Big\}$$ Εξίσωση (4.9.1.1)

όπου μέχρι και την δεύτερη δίεση δηλώνεται το κόστος της μετατόπισης της γραμμής και της στήλης που αντιστοιχεί τον κόμβο x σε μία νέα ομάδα, ενώ στο υπόλοιπο κομμάτι της

παραπάνω εξίσωσης λογαριάζονται οι διπλοί υπολογισμοί του κελιού $d_{x,x}$ στον πίνακα γειτνίασης.

3: Όσον αφορά το G(t + 1), ξαναϋπολογίστε τους πίνακες $D_{i,j}^{t+1}$, και τις αντίστοιχες κατανομές $P_{i,j}^{t+1}$.

4: Αν δεν υπάρχει περαιτέρω μείωση του συνολικού κόστους, σταματήστε˙ αλλιώς θέστε t = t+1, πηγαίνετε στο βήμα 2 και επαναλάβετε.

Ο InnerLoop αλγόριθμος επαναλαμβάνεται για την ομαδοποίηση G για τον ίδιο αριθμό k των ομάδων των κόμβων. Κάθε επανάληψη βελτιώνει (ή διατηρεί) το κόστος όπως αναφέρεται στο παρακάτω θεώρημα.

Θεώρημα 4.9.1. Μετά από κάθε επανάληψη του InnerLoop, η συνάρτηση κόστους ($\sum_{i=1}^{k}\sum_{j=1}^{k} C(D_{i,j})$) μειώνεται ή διατηρείται η ίδια.

Ο βρόχος τελειώνει όταν το συνολικό κόστος σταματά να βελτιώνεται. Σημειώστε ότι ορισμένες ομάδες είναι δυνατόν να είναι κενές, αλλά αυτό δεν αποτελεί πρόβλημα. Η πολυπλοκότητα του InnerLoop είναι O(w(D) · k · I), όπου I είναι ο αριθμός των επαναλήψεων.

Αλγόριθμος 4.6 Αλγόριθμος OuterLoop

1: Έστω ότι το T χαρακτηρίζει το δείκτη επανάληψης της αναζήτησης. Ξεκινήστε με T = 0 και k(0) = 1.

2: Στην επανάληψη T, δοκιμάστε να αυξήσετε το k: k(T +1) = k(T) +1. Χωρίστε την ομάδα κόμβων r με τη μέγιστη εντροπία σε κάθε κόμβο, δηλαδή $r = \text{argmax}_{1\le i\le k}\sum_{1\le j\le k}\frac{n(D_{i,j})H(P_{i,j})+n(D_{j,i})H(P_{j,i})}{a_i}$

Κατασκευάστε έναν αρχικό επισημασμένο χάρτη αντιστοίχισης G(T + 1) ως εξής: Για κάθε κόμβο x που ανήκει στην ομάδα r (δηλαδή, για κάθε $1 \le x \le n$ τέτοιο ώστε $G_x(T) = r$) τοποθετήστε το στη νέα ομάδα k(T + 1) (δηλαδή, ορίστε $G_x(T + 1) = k(T + 1)$) αν και μόνο αν μειώνει την ανά κόμβο εντροπία της ομάδας r, δηλαδή, αν και μόνο αν $\sum_{1\le j\le k}\frac{n(D'_{r,j})H(P'_{r,j})+n(D'_{j,r})H(P'_{j,r})}{a_r-1} < \sum_{1\le j\le k}\frac{n(D_{r,j})H(P_{r,j})+n(D_{j,r})H(P_{j,r})}{a_r}$, όπου $D'_{r,j}$ είναι το $D_{r,j}$ χωρίς τον κόμβο x. Διαφορετικά, θέστε $G_x(T + 1) = r = G_x(T)$. Αν μετακινηθεί ο κόμβος x στη νέα ομάδα, τότε ενημερώνεται και ο $D_{r,j}$ και ο $D_{j,r}$ αντίστοιχα (για όλα τα $1 \le j \le k$).

3: Εκτελέστε τον αλγόριθμο InnerLoop με το αρχικό G = G(T + 1) και k = k(T + 1) για να βρείτε μία νέα αντιστοίχιση κόμβων G(T + 1) και το αντίστοιχο συνολικό κόστος.

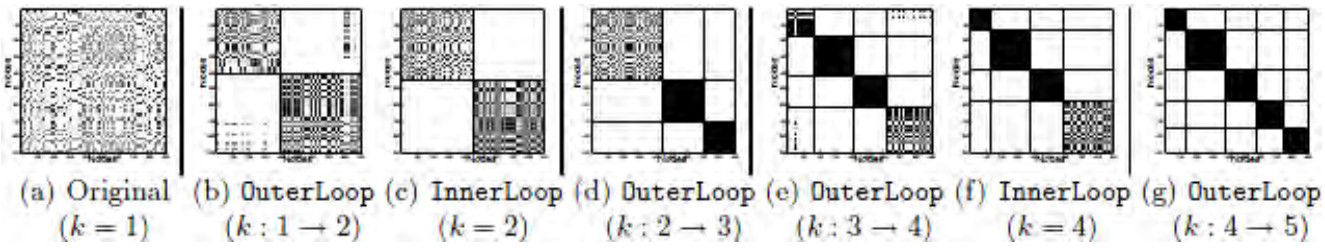
4: Εάν δεν υπάρχει μείωση του συνολικού κόστους, σταματήστε και επιστρέψτε $k^* = k(T)$ και $G^* = G(T)$. Σε αντίθετη περίπτωση, θέστε T = T + 1 και συνεχίστε.

Ο αλγόριθμος OuterLoop προσπαθεί να βρει καλές τιμές του k. Επιλέγει την ομάδα κόμβων με τη μέγιστη εντροπία ανά κόμβο, και τη χωρίζει σε δύο ομάδες. Οι κόμβοι που τοποθετήθηκαν στη νέα ομάδα είναι ακριβώς αυτοί των οποίων η απομάκρυνση μειώνει την εντροπία ανά

κόμβο στην αρχική ομάδα. Όπως αναφέρεται παρακάτω στο Θεώρημα 4.9.2, αυτή η διάσπαση δεν μειώνει ποτέ το κόστος.

Θεώρημα 4.9.2. Στη διάσπαση οποιασδήποτε ομάδας κόμβων, το κόστος είτε μειώνεται είτε παραμένει σταθερό.

Από το Θεώρημα 4.9.1, το ίδιο ακριβώς ισχύει και για τον InnerLoop. Συνεπώς, ο συνολικός αλγόριθμος επίσης μειώνει το κόστος. Ωστόσο, η πολυπλοκότητα της περιγραφής προφανώς αυξάνει με το k. Έχει βρεθεί στην πράξη ότι αυτή η στρατηγική αναζήτησης αποδίδει πολύ καλά. Ο αλγόριθμος OuterLoop εκτελείται k^* φορές, έτσι ώστε η συνολική πολυπλοκότητα της αναζήτησης να είναι $O(w(D)(k^*)^2 I$ – ή αλλιώς $O(mk^2)$). Στην πράξη, το $I \leq 20$ είναι πάντα αρκετό.



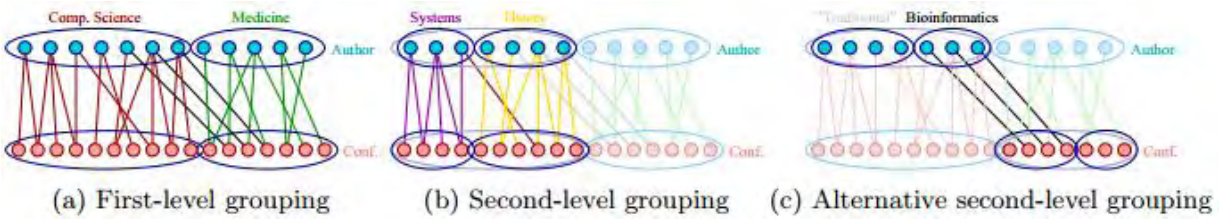
Σχήμα 4.14: Στιγμιότυπα εκτέλεσης αλγορίθμου. Ξεκινώντας με έναν τυχαίο πίνακα «caveman» μετάθεσης (a), ο αλγόριθμος εφαρμόζει OuterLoop και InnerLoop μέχρι να εμφανιστεί η τελική δομή (g). Θα παραλείπονται τα αποτελέσματα του InnerLoop όταν δεν εμφανίζουν καμία βελτίωση. Οι επαναλήψεις του OuterLoop διαχωρίζονται από μαύρες κάθετες γραμμές για λόγους σαφήνειας.

Το Σχήμα 4.14 δείχνει ένα στιγμιότυπο εκτέλεσης του πλήρους αλγορίθμου σε έναν τυχαίο πίνακα «caveman» μετάθεσης (δηλαδή έναν διαγώνιο μπλοκ πίνακα [41]) με Zipfian cave μεγέθη. Ο OuterLoop αυξάνει τον αριθμό των ομάδων των κόμβων, ενώ ο InnerLoop αναδιατάσσει τους κόμβους μεταξύ των ομάδων. Δεν εμφανίζονται διαγράμματα όταν ο InnerLoop δεν μειώνει το συνολικό κόστος. Το σωστό τελικό αποτέλεσμα φαίνεται στο διάγραμμα (g).

4.10 Δένδρο συσταδοποίησης ειδικού περιβάλλοντος (Context-specific Cluster Tree – CCT)

Δεδομένου ενός μεγάλου διμερούς γραφήματος, πώς μπορεί κάποιος να βρει σημαντικές κοινότητες, γρήγορα, και αυτόματα; Προτείνεται να αναζητούνται ιεραρχίες κοινοτήτων, με

κοινότητες-μέσα σε-κοινότητες. Η μέθοδος που παρουσιάζεται σε αυτήν την ενότητα, **το δένδρο συσταδοποίησης ειδικού περιβάλλοντος (Context-specific Cluster Tree – CCT)** βρίσκει αυτές τις κοινότητες σε πολλαπλά επίπεδα, χωρίς παρέμβαση του χρήστη, με βάση τις αρχές της θεωρίας της πληροφορίας (MDL). Πιο συγκεκριμένα, χωρίζει το γράφημα σε σταδιακά όλο και πιο εκλεπτυσμένους υπογράφους, επιτρέποντας στους χρήστες να κινούται γρήγορα από την γενική, τραχύ δομή ενός γραφήματος σε πιο προσανατολισμένα και τοπικά μοτίβα.



Σχήμα 4.15: Ιεραρχία και περιβάλλον [45].

Ως πρόσθετο πλεονέκτημα, καθώς επίσης και ως μία επιπλέον ένδειξη της ποιότητάς του, επιτυγχάνει καλύτερη συμπίεση από τις τυπικές, μη-ιεραρχικές μεθόδους. Το δένδρο CCT που προκύπτει στο τέλος μπορεί να εντοπίσει κατάλληλες συστάδες ειδικού περιβάλλοντος. Παρέχει επίσης ένα αποτελεσματικό σύστημα περίληψης δεδομένων και διευκολύνει την οπτικοποίηση των μεγάλων γραφημάτων, το οποίο από μόνο του είναι ένα δύσκολο και ανοικτό πρόβλημα. Διαισθητικά, ως περιβάλλον καθορίζεται ένα υπογράφημα το οποίο εμμέσως ορίζεται από ένα ζεύγος ομάδων κόμβων πηγής και προορισμού (και, έτσι, περιλαμβάνει ακριβώς εκείνες τις ακμές που συνδέουν τους κόμβους των ομάδων αυτών). Ολόκληρο το γράφημα και μία μονή ακμή είναι τα δύο ακραία περιβάλλοντα, στο γενικό και τοπικό επίπεδο, αντίστοιχα.

Η προσέγγισή αυτή επιτρέπει στους χρήστες να ξεκινήσουν από τις ομάδες των κόμβων και των ακμών που βρίσκονται στο γενικό επίπεδο και γρήγορα να επικεντρωθούν στο κατάλληλο περιβάλλον ώστε να ανακαλύψουν πιο εστιασμένα μοτίβα. Θα απεικονιστεί η αντίληψη και η διαίσθηση πίσω από το προτεινόμενο πλαίσιο με ένα παράδειγμα.

Θεωρήστε ένα σύνολο απο συγγραφείς (μπλε κόμβοι, στην κορυφή του Σχήματος 4.15) και ένα σύνολο διασκέψεων (κόκκινοι κόμβοι, στον πυθμένα του Σχήματος 4.15), με ακμές που υποδεικνύουν ότι ο συγγραφέας εμφανίζεται σε αυτή την διάσκεψη. Εκ πρώτης όψεως, θα μπορούσε κανείς να ανακαλύψει μια φυσική κατάτμηση του γραφήματος σε γενικό επίπεδο ως εξής:

- Ομάδες κόμβων: ας υποτεθεί ότι υπάρχουν δύο ομάδες συγγραφέων, οι επιστήμονες της πληροφορικής και οι ιατρικοί ερευνητές. Επίσης, υποτίθεται ότι υπάρχουν δύο αντίστοιχες ομάδες διασκέψεων. Σε μορφή πίνακα, απεικονίζονται στο Σχήμα 4.16(a) με δύο γραμμών και στηλών διαμερίσεις αντίστοιχα.

- Περιβάλλοντα: Η παραπάνω ομαδοποίηση κόμβων οδηγεί σε τέσσερα περιβάλλοντα (δηλαδή, ομάδες ακμών, ή υπογράφους), ένα για κάθε πιθανό συνδυασμό των δύο ομάδων κόμβων του κάθε τύπου (συγγραφείς και διασκέψεις). Σε μορφή πίνακα, αντιστοιχούν στους τέσσερις υποπίνακες του Σχήματος 4.15(b). Τα κυρίαρχα περιβάλλοντα είναι οι δύο υποπίνακες στη διαγώνιο του Σχήματος 4.16(a) που αντιστοιχούν στην επιστήμη των υπολογιστών (τομή των επιστημόνων της πληροφορικής και των διασκέψεων) και την ιατρική (τομή των γιατρών και ιατρικών διασκέψεων), αντίστοιχα.

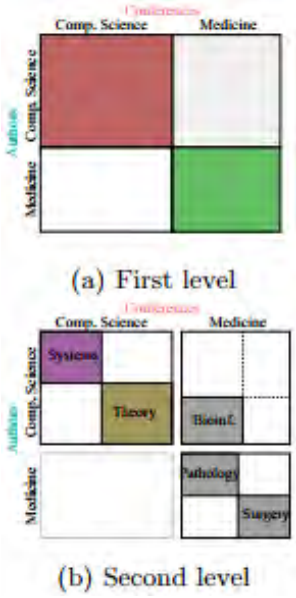
Παρακάτω, θα αναφερθούν λίγα πράγματα για δύο πολύ σημαντικές έννοιες όταν πρέπει να αντιμετωπιστούν τέτοιου είδους προβλήματα: ιεραρχία και περιβάλλον.

Ιεραρχία. Τα γραφήματα συχνά παρουσιάζουν τη δομή κοινότητες-μέσα σε-κοινότητες, οδηγώντας σε μια φυσική αναδρομική διάσπαση της δομής τους, η οποία αποτελεί μία ιεραρχία. Πώς μπορεί να βρεθεί ο κατάλληλος αριθμός των επιπέδων, καθώς και οι ομάδες κόμβων σε κάθε επίπεδο; Αυτά τα δύο ερωτήματα θέτουν πρόσθετες προκλήσεις για το σχεδιασμό του αλγορίθμου.

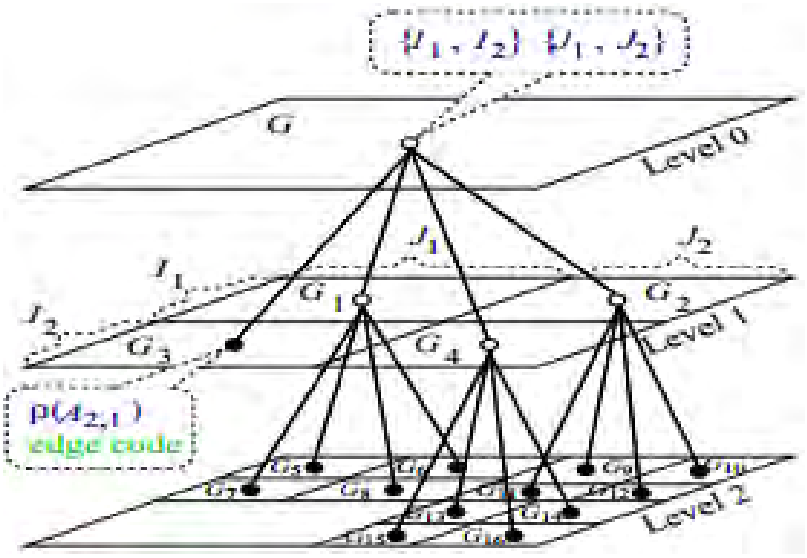
Για παράδειγμα, ας εξεταστεί το περιβάλλον που προκαλείται από τον «επιστήμη των υπολογιστών» συγγραφέα και την «επιστήμη των υπολογιστών» διάσκεψη (βλέπε Σχήμα 4.15(b), ή το επάνω αριστερό μέρος του Σχήματος 4.16(b)). Εκτελώντας μια παρόμοια ανάλυση όπως και πριν, μπορεί να ανακαλυφθεί μία επιπλέον δομή όσον αφορά τις ομάδες κόμβων και τα περιβάλλοντα στο δεύτερο επίπεδο. Το πεδίο της επιστήμης των υπολογιστών μπορεί να υποδιαιρεθεί περαιτέρω σε συστήματα και θεωρίες συγγραφέων, με αντίστοιχη διαίρεση στις διασκέψεις της επιστήμης των υπολογιστών.

Περιβάλλον. Στο παράδειγμα αυτό, τα κυρίαρχα περιβάλλοντα είναι εκείνα της «επιστήμης των υπολογιστών» και η «ιατρική», όπως εξηγήθηκε παραπάνω. Ωστόσο, δεν υπάρχει τίποτα το ιδιαίτερο γι' αυτά τα «διαγώνια» περιβάλλοντα. Στην πραγματικότητα, κάποιος πρέπει επίσης να εξετάσει και τα «εκτός της διαγωνίου» περιβάλλοντα. Για παράδειγμα, το περιβάλλον που ορίζεται από την τομή των «επιστήμης των υπολογιστών» συγγραφέων και των «ιατρική» διασκέψεων (βλέπε Σχήμα 4.15(c)) μπορεί επίσης να διαχωριστεί περαιτέρω σε πολλαπλά χαμηλότερου επιπέδου περιβάλλοντα, με ένα από αυτά να αντιστοιχεί στην «βιοπληροφορική».

Σε γενικές γραμμές, μία συγκεκριμένη επιλογή υπογράφου κατά τη διάρκεια της επαναληπτικής διάσπασης αποτελείται από ένα ζεύγος των ομάδων κόμβων και του περιβάλλοντος που παρέχονται από τις ακμές που τους συνδέουν. Διαφορετικά περιβάλλοντα μπορεί να αποκαλύψουν διάφορες πτυχές των δεδομένων. Λαμβάνοντας αυτή την ιδέα στη λογική της κατάληξη, το συνολικό αποτέλεσμα είναι μία πλούσια ιεραρχία, CCT, που αποτυπώνει τη δομή του γραφήματος σε πολλαπλά επίπεδα.



Σχήμα 4.16: Πίνακας γειτνίασης (βλ. Σχήμα 4.15)[45].



Σχήμα 4.17: CCT που αντιστοιχεί στο Σχήμα 4.16.Οι κενοί κόμβοι δεν αντιστοιχούν σε φύλλα (υπογραφήματα για τα οποία το μοντέλο διάσπασης ήταν το καλύτερο) ενώ οι γεμάτοι κόμβοι αντιστοιχούν σε φύλλα (υπογραφήματα για τα οποία το μοντέλο τυχαίου γραφήματος ήταν το καλύτερο). Τα δύο αναδυόμενα παράθυρα δείχνουν παραδείγματα πληροφορίας σε κάθε διαφορετικού τύπου κόμβου˙ με σκούρο μπλε αναπαρίστανται τμήματα του μοντέλου, ενώ με ανοιχτό πράσινο τμήματα του κώδικα, δεδομένου του μοντέλου [45].

4.10.1 Παρουσίαση του προβλήματος

Μόνο για την συγκεκριμένη ενότητα θα θεωρήσουμε την ύπαρξη των παρακάτω συμβόλων:

Symbol	Definition	Symbol	Definition
A	Binary adjacency matrix.	m_p, n_q	Dimensions of $A_{p,q}$.
m, n	Dimensions of A.	$\|A\|$	Number of elements $\|A\| := mn$.
k, ℓ	No. of source and dest. partitions.	$\rho(A)$	Edge density in $\rho(A) = e(A)/\|A\|$.
$A_{p,q}$	Submatrix for intersection of p-th source and q-th dest. partitions.	$H(\cdot)$	Shannon entropy function.
		$C(A)$	Codelength for A.

Πίνακας 2: Σύμβολα και ορισμοί.

Σε αυτή την παραλλαγή της MDL προσέγγισης, ένας δυαδικός πίνακας γειτνίασης m × n αντιπροσωπεύει ένα διμερές γράφημα με m να συμβολίζονται οι κόμβοι πηγής και με n οι κόμβοι προορισμού. Όπως προαναφέραμε, ο στόχος είναι να κατασκευαστεί αυτόματα μια αναδρομική δομή κοινότητας ενός μεγάλου διμερούς γραφήματος σε πολλαπλά επίπεδα, δηλαδή, ένα δένδρο συσταδοποίησης ειδικού περιβάλλοντος (CCT). Η βασική ιδέα είναι να υποδιαιρεθεί ο πίνακας γειτνίασης σε πλακίδια (ή «περιβάλλοντα»), με ενδεχόμενη αναδιάταξη

των γραμμών και των στηλών, και έπειτα αυτά να συμπιεστούν, είτε ως έχει (αν είναι αρκετά ομοιογενή) είτε με περαιτέρω υποδιαίρεση.

Έστω I := {1, 2, . . . , m} το σύνολο των m κόμβων πηγής, και J := {1,2, . . . , n} το σύνολο των n κόμβων προορισμού. Κάθε ζεύγος κόμβων (i, j), για 1 ≤ i ≤ m και 1 ≤ j ≤ n, μπορεί να συνδέεται με μία ακμή. Έστω ότι ο πίνακας A = [a(i, j)] χαρακτηρίζει τον αντίστοιχο m × n (m, n ≥ 1) δυαδικό πίνακα γειτνίασης.

Ορισμός 4.10.1 (Διμερές γράφημα και υπογράφημα). Το διμερές γράφημα G αποτελείται από την εξής τριάδα: G ≡ (I, J, A). Ένα υπογράφημα αυτού του γραφήματος αποτελείται από την εξής τριάδα: G’ ≡ (I’,J’, A’), όπου I’ ⊆ I, J’ ⊆ J και A’:= [a(i’, j’)] για όλα τα i’ ∈ I’ and j’ ∈ J ’.

Στόχος είναι να ανακαλυφθούν οι ομάδες των ακμών που συνδέουν στενά τις ομάδες των κόμβων πηγής και προορισμού.

Ορισμός 4.10.2 (Διάσπαση υπογράφου). Δεδομένου ενός γράφου G ≡ (I, J, A), αυτό θα διασπαστεί σε ένα σύνολο υπογράφων $\{G_1, G_2, \ldots, G_T\}$ τέτοιοι ώστε η ένωσή τους να ισούται με το αρχικό γράφημα G.

Πιο συγκεκριμένα, επιδιώκεται να διασπαστεί ο αρχικός γράφος σε ένα σύνολο από υπογράφους, οι οποίοι θα πρέπει να έχουν τις ακόλουθες ιδιότητες:

- Συνεκτικότητα: Καθένας από τους υπογράφους θα πρέπει ιδανικά να είναι είτε πλήρως συνδεδεμένος είτε πλήρως αποσυνδεδεμένος, δηλαδή, θα πρέπει να είναι όσο το δυνατόν ομοιογενής.
- Ευελιξία: Η δομή της διάσπασης σε υπογράφους θα πρέπει να είναι αρκετά πλούσια, χωρίς την επιβολή υπερβολικά πολλών περιορισμών.
- Προοδευτικότητα: Η διάσπαση θα πρέπει να επιτρέπει στους χρήστες να περιηγούνται από την γενική δομή σε πιο εστιασμένα και τοπικά μοτίβα, με τη μορφή του σταδιακά όλο και πιο πυκνούς υπογράφους.

Επιπλέον, γίνεται προσπάθεια ώστε να βρεθεί αυτόματα μία τέτοια διάσπαση, χωρίς να απαιτείται καμία παράμετρος από το χρήστη. Για το σκοπό αυτό, χρησιμοποιείται η MDL προσέγγιση σε μια κωδικοποίηση του διμερούς πίνακα γειτνίασης. Η κωδικοποίηση που επιλέγεται είναι ιεραρχική, έτσι ώστε να ικανοποιούνται οι τελευταίες δύο ιδιότητες.

Ορισμός 4.10.3 (Μοντέλο τυχαίου γραφήματος). Σε αυτή την περίπτωση, μπορεί να κωδικοποιηθεί ολόκληρος ο πίνακας χρησιμοποιώντας $C_0(A) := \lceil \log(|A| + 1) \rceil + \lceil |A| H(\rho(A)) \rceil$ bits.

Πιο συγκεκριμένα, χρησιμοποιούνται $\lceil \log(|A| + 1) \rceil$ bits για τη μετάδοση ρ(A) και τελικά $\lceil |A| H(\rho(A)) \rceil$ bits για τη μετάδοση των επιμέρους ακμών. Αυτό προϋποθέτει ότι γνωρίζουμε ήδη το μέγεθος του γραφήματος (δηλαδή, τα m και n). Αυτό μπορεί να υποτεθεί με ασφάλεια για το αρχικό γράφημα G. Για τους υπογράφους του η πληροφορία αυτή παρέχεται από το μοντέλο.

Ορισμός 4.10.4 (Μοντέλο διασπασμένου γραφήματος). Το κόστος της κωδικοποίησης του διασπασμένου μοντέλου είναι $C_1(A):=\lceil \log m \rceil + \lceil \log n \rceil + \left\lceil \log \binom{m}{m_1 \ldots m_k} \right\rceil + \left\lceil \log \binom{n}{n_1 \ldots n_l} \right\rceil + \sum_{p=1}^{k} \sum_{q=1}^{l} C(A_{p,q})$.

Χρειάζονται $\lceil \log m \rceil$ bits για τη μετάδοση k και $\lceil \log n \rceil$ bits για τη μετάδοση ℓ. Επιπλέον, αν υποτεθεί ότι κάθε αντιστοίχιση m κόμβων πηγής σε k κόμβους προορισμού είναι εξίσου πιθανή, τότε χρειάζονται $\left\lceil \log \binom{m}{m_1 \ldots m_k} \right\rceil$ bits για τη μετάδοση της διάσπασης της πηγής $\{I_1, \ldots, I_k\}$, και ομοίως για τη διάσπαση του προορισμού.

Ορισμός 4.10.5 (Συνολικό ιεραρχικό μήκος). Δεδομένης μιας ιεραρχικής διάσπασης, το συνολικό κόστος μήκους για την μετάδοση του γραφήματος (I, J, A) είναι $C(A) := 1 + \min\{C_0(A), C_1(A)\}$.

Ορισμός 4.10.6 (Context-specific Cluster Tree). Το σύνολο όλων των υπογράφων στην προοδευτική, ιεραρχική διάσπαση αποτελεί το δένδρο συσταδοποίησης ειδικού περιβάλλοντος (CCT). Οι κόμβοι φύλλα αντιστοιχούν σε υπογράφους για τους οποίους η καλύτερη επιλογή είναι το μοντέλο τυχαίου γραφήματος. Αυτοί οι υπογράφοι περιλαμβάνουν τη διάσπαση του τελευταίου επιπέδου που περιλαμβάνει τους κόμβους-φύλλα. Λαμβάνοντας υπ' όψιν το μοντέλο, ο κώδικας για τα δεδομένα αποτελείται από τις πληροφορίες για μεμονωμένες ακμές εντός των υπογράφων μόνο στο επίπεδο του φύλλου.

Για παράδειγμα, στο Σχήμα 4.17, ο κόμβος ρίζα θα κωδικοποιούσε την διάσπαση $\{I_1,I_2\}$ και $\{J_1,J_2\}$· αυτό είναι τμήμα του μοντέλου. Ο κόμβος που αντιστοιχεί στο $G_3 \equiv (I_2, J_1, A_{2,1})$ θα κωδικοποιούσε την πυκνότητα $\rho(A_{2,1})$--η οποία αποτελεί επίσης τμήμα του μοντέλου-- και στη συνέχεια, τις επιμέρους ακμές του G_3 χρησιμοποιώντας την κωδικοποίηση της εντροπίας--που αποτελεί μέρος του κώδικα του δεδομένου μοντέλου. Εκτός από την G ρίζα, το CCT αποτελείται από όλους τους 16 κόμβους που αντιστοιχούν στους υπογράφους G_1 έως G_{16}. Η διαμέριση του φύλλου-επιπέδου αποτελείται από 13 γραφήματα $\{G_3,G_5,G_6, \ldots, G_{16}\}$, τα οποία αντιπροσωπεύονται από γεμάτους/χρωματισμένους κόμβους.

Ορισμός 4.10.7 (Περιβάλλον). Λαμβάνοντας υπ' όψιν ως είσοδο ένα ζευγάρι από κόμβο πηγής και προορισμού (I_i,J_i), ένα περιβάλλον του (I_i,J_i) είναι κάθε ζευγάρι (I_c, J_c) έτσι ώστε $I_i \subseteq I_c$ και $J_i \subseteq J_c$.

Με άλλα λόγια, ένα περιβάλλον για κάθε ζεύγος (I_i,J_i) είναι κάθε υπογράφος του αρχικού γράφου που περιλαμβάνει πλήρως τα I_i και J_i. Συνήθως θα περιορίζεται το ζεύγος (I_c, J_c) να αποτελείται μόνο από αυτά τα ζεύγη που εμφανίζονται σε κάποιο κόμβο της ιεραρχικής διάσπασης.

4.10.2 Βρίσκοντας το CCT

Για να φτιαχτεί ένας κλιμακούμενος και πρακτικός αλγόριθμος, επιλέγεται να χρησιμοποιηθεί μία top-down στρατηγική για την κατασκευή της ιεραρχίας, παρά μία bottom-up προσέγγιση. Ξεκινώντας με το αρχικό γράφημα, γίνεται προσπάθεια να βρεθεί μία καλή «πλακόστρωτη σκακιέρα» για το πρώτο επίπεδο της διάσπασης. Στη συνέχεια, επαναπροσδιορίζονται αυτά τα «πλακάκια» και αναδρομικά επιχειρείται η ίδια διαδικασία για κάθε ένα από αυτά.

Ωστόσο, υπάρχουν δύο προβλήματα που πρέπει να αντιμετωπιστούν. Πρώτον, ο αναδρομικός ορισμός της εξίσωσης του μοντέλου του διάσπασμένου γραφήματος είναι αρκετά δαπανηρός ώστε να αξιολογηθεί για κάθε πιθανή ανάθεση των κόμβων σε κατατμήσεις, έτσι αντ' αυτού χρησιμοποιείται η ακόλουθη εξίσωση, $C'_1(A):=\lceil \log m \rceil+\lceil \log n \rceil + \left\lceil \log \binom{m}{m_1 \ldots . m_k} \right\rceil + \left\lceil \log \binom{n}{n_1 \ldots . n_l} \right\rceil + \sum_{p=1}^{k} \sum_{q=1}^{l} C_0(A_{p,q})$, όπου έχει αντικατασταθεί το C με το C_0 στο τελευταίο όρο του τελευταίου αθροίσματος.

Ακόμη και με αυτή την απλοποίηση, η εξεύρεση της βέλτιστης πλακόστρωτης σκακιέρας (δηλαδή, η ανάθεση των κόμβων σε κατατμήσεις) είναι NP-hard [38], ακόμη και αν είναι γνωστός ο αριθμός των πλακιδίων (ή, ισοδύναμα, διασπάσεις κόμβων πηγής και προορισμού). Επιπλέον, αναζητείται και ο αριθμός των πλακιδίων. Ως εκ τούτου, θα χρησιμοποιηθεί μία εναλλακτική ελαχιστοποίηση [38] που συγκλίνει προς ένα τοπικό ελάχιστο.

Αναζητείται αναδρομικά η καλύτερη διάσπαση στην σκακιέρα και σταματάει όταν το μοντέλο του διασπασμένου γραφήματος είναι χειρότερο από το μοντέλο τυχαίου γραφήματος, το οποίο δείχνει ότι το υπογράφημα είναι επαρκώς ομοιογενές. Ο αλγόριθμος αναζήτησης κατόπιν προχωρά σε δύο στάδια: (i) ένα εξωτερικό βήμα, *SPLIT*, που προσπαθεί να αυξήσει σταδιακά τον αριθμό των διασπάσεων της πηγής και του προορισμού, και (ii) ένα εσωτερικό βήμα, *SHUFFLE*, το οποίο, δεδομένου ενός σταθερού αριθμού διασπάσεων, προσπαθεί να βρει την καλύτερη ανάθεση των κόμβων σε κατατμήσεις. Ο ψευδοκώδικας στους Αλγορίθμους 4.7, 4.8, και 4.9 δείχνει τα βήματα της συνολικής διαδικασίας με λεπτομέρεια.

Αλγόριθμος 4.7 Αλγόριθμος SHUFFLE

Ξεκινήστε με μία αυθαίρετη διαμέριση του πίνακα A σε k κατατμήσεις πηγής $I_p^{(0)}$ και ℓ κατατμήσεις στήλης $J_q^{(0)}$. Στη συνέχεια, σε κάθε επανάληψη t εκτελέστε τα παρακάτω βήματα:

1: Γι' αυτό το βήμα, θα κρατηθούν οι κατατμήσεις προορισμού σταθερές, δηλαδή, $J_q^{(t)}$ για όλα τα $1 \le q \le \ell$. Ξεκινήστε με $I_p^{(t+1)} := I_p^{(t)}$ για όλα τα $1 \le p \le k$. Στη συνέχεια, εξετάστε κάθε κόμβο πηγής i, $1 \le i \le n$ και μετακινήστε τον στην p^*-οστή διαμέριση $I_{p^*}^{(t+1)}$ ούτως ώστε η επιλογή να μεγιστοποιεί το κόστος $C'_1(A)$.

2: Παρόμοιο με το βήμα 1, αλλά εδώ εναλλάσσονται οι κόμβοι προορισμού, προκειμένου να βρεθούν νέες διαμερίσεις $J_q^{(t+2)}$ για $1 \le q \le \ell$.
3: Εάν δεν υπάρχει μείωση στο κόστος $C'_1(A)$, σταματήστε. Σε αντίθετη περίπτωση, ορίστε $t \leftarrow t + 2$, πηγαίνετε στο βήμα 1, και επαναλάβετε.

Αλγόριθμος 4.8 Αλγόριθμος SPLIT

Ξεκινήστε με $k^0 = \ell^0 = 1$ και σε κάθε επανάληψη τ:
1: Προσπαθήστε να αυξήσετε τον αριθμό των κατατμήσεων της πηγής, κρατώντας τον αριθμό των κατατμήσεων προορισμού σταθερό. Επιλέξτε να διαχωρίσετε την διάσπαση πηγής p^* με τη μέγιστη ανά-κόμβο εντροπία, δηλαδή, $p^* := \arg\max_{1\le p\le k} \sum_{1\le q\le l} |A_{p,q}| H(p(A_{p,q}))/m_p$. Αυξήστε τον αριθμό των κατατμήσεων γραμμής, $k^{\tau+1} = k^\tau + 1$ και κατασκευάστε μία διαμέριση $\{I_1^{(\tau+1)}, \ldots, I_{k^{T+1}}^{(\tau+1)}\}$ μετακινώντας κάθε κόμβο i της διαμέρισης $I_{p^*}^{(\tau)}$ που θα διαχωριστεί στη νέα διαμέριση πηγής $I_{k^{T+1}}^{(\tau+1)}$, αν και μόνο αν αυτή μειώνει την ανά-κόμβο εντροπία της p^*-οστής διαμέρισης.
2: Εφαρμόστε τον αλγόριθμο SHUFFLE με την αρχική κατάσταση $\{ I_p^{(\tau+1)} \mid 1 \le p \le k^{\tau+1} \}$ και $\{ J_p^{(\tau)} \mid 1 \le p \le \ell^\tau \}$, για να βρείτε καλύτερες αναθέσεις κόμβων στις κατατμήσεις.
3: Εάν δεν υπάρχει μείωση του συνολικού κόστους, σταματήστε και επιστρέψτε $(k, \ell) = (k^\tau, \ell^\tau)$ με τις αντίστοιχες διαμερίσεις. Σε αντίθετη περίπτωση, θέστε $\tau \leftarrow \tau + 1$ και συνεχίστε.
4-6: Παρόμοια με τα βήματα 1-3, αλλά προσπαθήστε να αυξήσετε αντ' αυτού τις διαμερίσεις προορισμού.

Αλγόριθμος 4.9 Ο Ιεραρχικός Αλγόριθμος

1: Δοκιμάστε τον αλγόριθμο SPLIT για να βρείτε το καλύτερο μοντέλο διασπασμένου γραφήματος.
2: Συγκρίνετε το $C'_1(A)$ με εκείνο του μοντέλου τυχαίου γραφήματος, $C_0(A)$.
3: Εάν το μοντέλο διασπασμένου γραφήματος είναι καλύτερο, τότε για κάθε υπογράφημα $(I_p, J_q, A_{p,q})$, για όλα τα $1 \le p \le k$ και $1 \le q \le \ell$, εφαρμόστε αναδρομικά τον Ιεραρχικό αλγόριθμο.

Ο αλγόριθμος SHUFFLE είναι γραμμικός στον αριθμό των ακμών και στον αριθμό των επαναλήψεων. Ο αλγόριθμος SPLIT επικαλείται τον SHUFFLE για κάθε διαχωρισμό, για την χειρότερη συνολικά περίπτωση των $2(k + \ell + 1)$ διασπάσεων. Για κάθε επίπεδο της αναδρομής στον Ιεραρχικό, ο συνολικός αριθμός των ακμών μεταξύ όλων των διασπάσεων ενός επιπέδου είναι το πολύ ίσος με τον αριθμό των ακμών στον αρχικό γράφο. Έτσι, ο συνολικός χρόνος είναι

ανάλογος με τον συνολικό αριθμό των ακμών, όπως επίσης και με το μέσο βάθος των φύλλων και τον αριθμό των διαμερίσεων.

4.11 Timefall

Η μέθοδος **Timefall** χρησιμοποιείται σε κοινωνικά δίκτυα/γραφήματα που εξελίσσονται με την πάροδο του χρόνου. Η κύρια καινοτομία αυτής της προσέγγισης είναι ότι δεν χρειάζεται παραμέτρους που να είναι ορισμένοι από το χρήστη, στηριζόμενη αντ' αυτού στην αρχή του Μήκους Ελάχιστης Περιγραφής (MDL), ώστε να εξαχθούν οι κοινότητες, και για να βρεθούν καλά σημεία αποκοπής στο χρόνο όταν οι κοινότητες αλλάζουν απότομα: ένα σημείο αποκοπής είναι καλό, εάν αυτό οδηγεί σε μικρότερη περιγραφή των δεδομένων.

Γραφήματα που εξελίσσονται με την πάροδο του χρόνου εμφανίζονται σε ένα ευρύ φάσμα επιστημονικών κλάδων και τομέων εφαρμογής. Αυτά τα σύνολα δεδομένων από διάφορους τομείς έχουν το κοινό χαρακτηριστικό ότι μπορούν να αναπαρασταθούν σαν ένα δίκτυο, το οποίο αλλάζει με την πάροδο του χρόνου. Παραδείγματα τέτοιων χρονο-εξελίξιμων γραφημάτων είναι οι βάσεις δεδομένων των δημοσιεύσεων (π.χ. PubMed MEDLINE) και τα δυναμικά on-line κοινωνικά δίκτυα (π.χ. Orkut και Facebook).

Η μέθοδος Timefall («time waterfall»), όπως προαναφέρθηκε, αποτελεί μία προσέγγιση με στόχο την ανάλυση της εξέλιξης ενός δικτύου. Ένα παράδειγμα της απεικόνισης του Timefall εμφανίζεται στο Σχήμα 4.18. Στο Σχήμα 4.18 κάθε κουτί αντιπροσωπεύει μία κοινότητα λέξεων – ένα θέμα προφίλ χρήστη. Συνεπώς, κάθε θέμα μπορεί να περιγραφεί από ένα σύνολο λέξεων-κλειδιών που είναι χαρακτηριστικές για το αντίστοιχο θέμα. Κάθε γραμμή (οριζόντια ομάδα θεμάτων) αντιπροσωπεύει ένα χρονικό βήμα, και τα βέλη μεταξύ των θεμάτων από τα γειτονικά χρονικά βήματα αντιπροσωπεύουν την εξέλιξη των θεμάτων. Παρατηρήστε τις διασπάσεις και τις συγχωνεύσεις των θεμάτων με τη πάροδο του χρόνου.

4.11.1 Περιγραφή προβλήματος

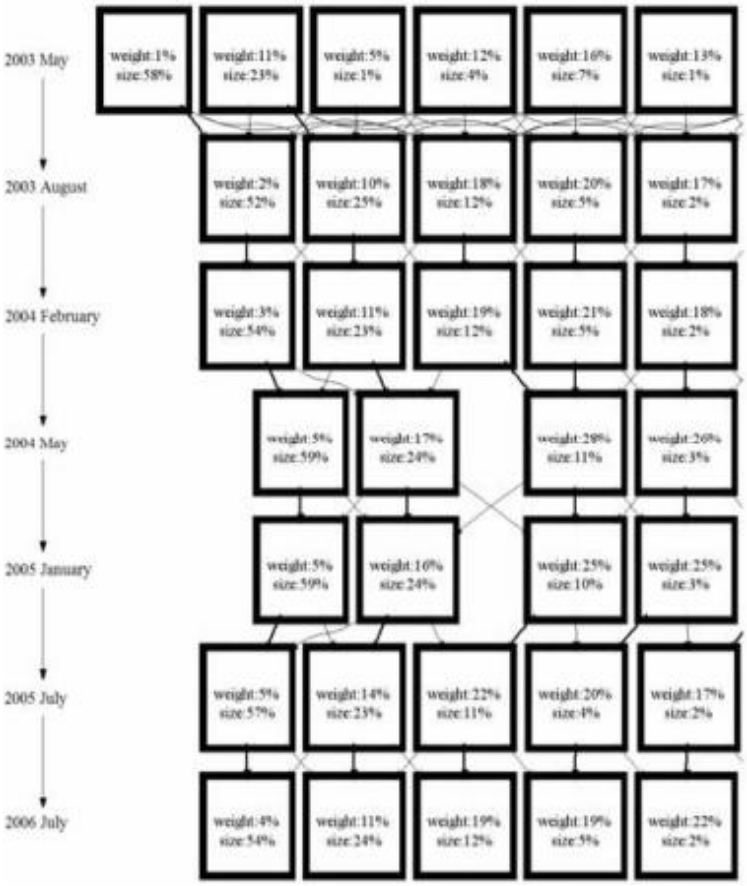
Το πρόβλημα μπορεί συνοπτικά να περιγραφεί ως εξής:

Πρόβλημα: Παρακολούθηση της εξέλιξης ενός απαλλαγμένου από παραμέτρους (parameter-free) δικτύου.

Δεδομένα: n χρονοσφραγισμένα γεγονότα (όπως, π.χ., εργασίες που δημοσιεύθηκαν), καθένα από τα οποία σχετίζεται με αρκετά από m στοιχεία (όπως, τίτλος-λέξεις, και/ή συγγραφέας-ονόματα, και/ή τα ονόματα εκδοτών).

Σκοπός: Η Timefall μέθοδος βρίσκει ταυτόχρονα (α) τις κοινότητες, οι οποίες είναι, ομάδες στοιχείων (π.χ., ερευνητικά θέματα και/ή ερευνητικές κοινότητες), (β) περιγράφει πώς οι κοινότητες εξελίσσονται με την πάροδο του χρόνου (π.χ., εμφανίζονται, εξαφανίζονται, διαχωρίζονται, συγχωνεύονται) και (γ) επιλέγει τα κατάλληλα σημεία αποκοπής στο χρόνο όταν η υπάρχουσα δομή της κοινότητας αλλάζει απότομα.

Χωρίς: καμία ορισμένη από το χρήστη παράμετρο.



Σχήμα 4.18: Timefall απεικόνιση της εξέλιξης των περιγραφών του προφίλ των χρηστών μέσα σε ένα μεγάλο on-line κοινωνικό δίκτυο. Αναλύεται και απεικονίζεται η εξέλιξη του θέματος των 7,5 εκατ. προφίλ κειμένου κατά τη διάρκεια 3 ετών. Η ανάλυση αποκαλύπτει τα θέματα προφίλ (λέξεις-κλειδιά παραλείπονται για λόγους εμπιστευτικότητας) και τα σημεία αποκοπής στο χρόνο , όταν η εξέλιξη είναι ιδιαίτερα απότομη.

4.11.2 Περιγραφή της Timefall μεθόδου

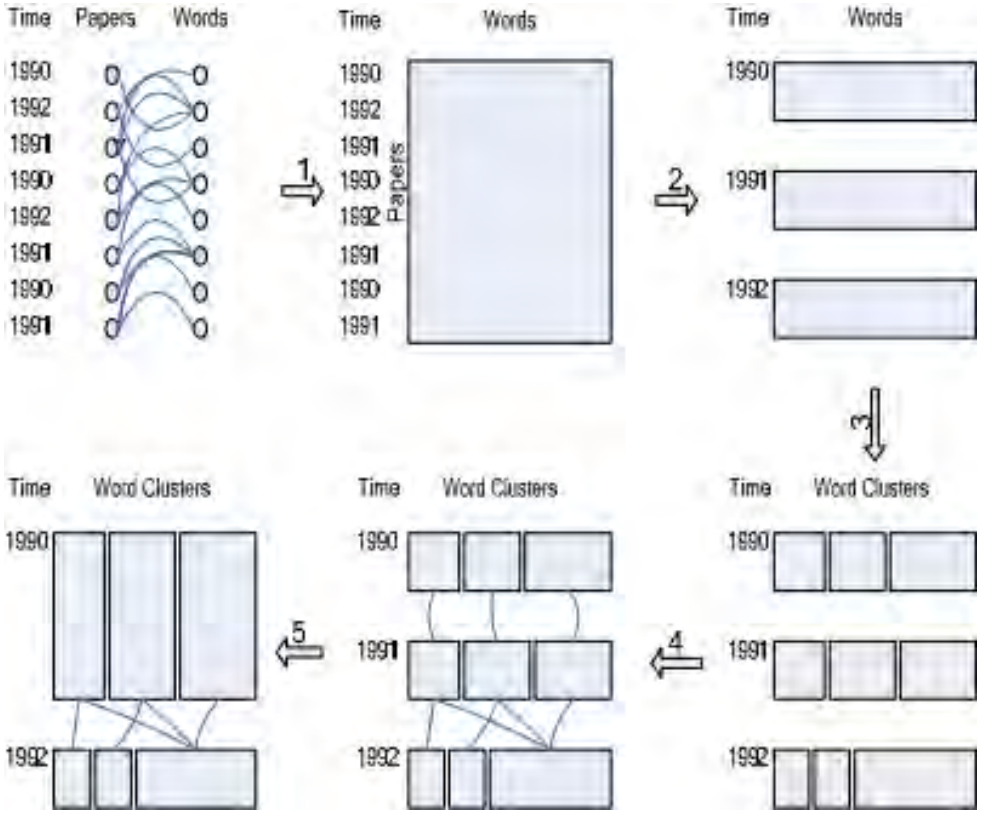
Για λόγους απλότητας θα θεωρηθεί ένα διμερές γράφημα με χρονοσφραγίδες, και θα παρουσιαστεί μία απεικόνιση της εξέλιξης των συστάδων του δικτύου (θέματα, κοινότητες) με την πάροδο του χρόνου. Ο Timefall μπορεί να επεκταθεί για να χειριστεί πολυμερή γραφήματα.

Στόχος είναι να λειτουργεί σε ένα τέτοιο χρονοσφραγισμένο γράφημα, να εντοπίζει αυτόματα τις κοινότητες, την εξέλιξή τους και τα σημεία αποκοπής μεταξύ των εποχών της σταθερής εξέλιξης της κοινότητας. Η διαίσθηση πίσω από αυτή την προσέγγισή είναι να οργανωθεί ο χρονοσφραγισμένος πίνακας γειτνίασης με τέτοιο τρόπο ώστε να είναι εύκολο να συμπιεστεί. Αυτό συνάδει ακριβώς με την MDL προσέγγιση χρησιμοποιώντας κάποια γλώσσα περιγραφής δεδομένων για την παραγωγή όσο το δυνατόν μιας πιο σύντομης περιγραφής των δεδομένων. Σε γενικές γραμμές, η ιδέα είναι να αντιμετωπιστεί το πρόβλημα ως πρόβλημα συμπίεσης, διότι θα καθοδηγήσει να βρεθούν μοτίβα (φυσικά σημεία αποκοπής, φυσικές κοινότητες), εξαλείφοντας την ανάγκη ύπαρξης παραμέτρων που να έχουν οριστεί από τον χρήστη.

Σχεδιάστηκε η γλώσσα περιγραφής του Timefall, η οποία διευκολύνει μία σύντομη χρονικά περιγραφή του γραφήματος παρέχοντας αποτελεσματικά μέσα επικοινωνίας, τα πιο πιθανά χρονικά πρότυπα που παρατηρήθηκαν σε ένα γράφημα, όπως η συγχώνευση, η διάσπαση και η εξέλιξη των κοινοτήτων. Η γλώσσα παρέχει μέσα για να περιγράψει αποτελεσματικά τη διαφορά μεταξύ της συσταδοποίησης δύο πινάκων (γραφήματα). Όταν οι πίνακες περιγράφουν το δίκτυο σε δύο διαδοχικά χρονικά σημεία, τότε η διαφορά στην συσταδοποίηση αντιπροσωπεύει την εξέλιξη της κοινότητας της δικτύου. Αποδεικνύεται ότι οι διαφορές στην συσταδοποίηση δύο πινάκων μπορεί να περιγραφεί αποτελεσματικά με τη χρήση αμοιβαίας πληροφόρησης μεταξύ των πινάκων, δηλαδή, χρησιμοποιούνται οι συστάδες (κοινότητες) τη χρονική στιγμή t για να περιγράψουν αποτελεσματικά τις κοινότητες τη χρονική στιγμή t + 1. Με αυτό τον τρόπο μπορούν να εντοπιστούν μοτίβα όπως ο διαχωρισμός και η συγχώνευση των κοινοτήτων, καθώς επίσης να δημιουργηθούν καινούργιες και να μειωθούν οι παλιές κοινότητες. Χρησιμοποιείται ένας απλός αλγόριθμος αναρρίχησης λόφου (hill climbing algorithm) για την αναζήτηση αυτών των μοτίβων και την αντίστοιχη περιγραφή του μικρού πίνακα (γραφήματος).

Η σχεδιασμένη γλώσσα περιγραφής του δικτύου και ο σύντομης περιγραφής αλγόριθμος αναζήτησης επεκτείνουν την υπάρχουσα Διασταυρωμένη Σχέση (Cross Association) της γλώσσας περιγραφής μέλους κοινότητας και τον αλγόριθμο αναζήτησης˙ ο Cross Association αλγόριθμος αποτελεί τη βάση των MDL αλγορίθμων εντοπισμού κοινοτήτων [39]. Η επέκταση επιτρέπει την αποδοτική μοντελοποίηση των χρονικών μεταβολών στο δίκτυο, και συγχρόνως διατηρώντας την αποτελεσματικότητα και την απουσία των παραμέτρων του αρχικού αλγορίθμου. Επιπλέον, τόσο ο αρχικός Cross Association αλγόριθμος όσο και η ανεπτυγμένη Timefall μέθοδος μπορούν αποτελεσματικά να παραλληλοποιηθούν για να επιτευχθεί γραμμική χρονοβελτίωση. Μια

επισκόπηση της Timefall προσέγγισης για την ανίχνευση της εξέλιξης του δικτύου μπορεί να παρουσιαστεί σε ένα παράδειγμα παρακολούθησης της ανάπτυξης θέματος σε ερευνητικές δημοσιεύσεις, όπως απεικονίζεται στο Σχήμα 4.19.



Σχήμα 4.19: Η Timefall προσέγγιση απεικονίζεται σε ένα διμερές χρονικό γράφημα που αντιπροσωπεύει εργασίες που δημοσιεύθηκαν με την πάροδο του χρόνου και λέξεις από το περιεχόμενό τους. Η προσέγγιση στη αρχή αναπαριστά το διμερές γράφημα με τη μορφή ενός πίνακα γειτνίασης (βήμα 1). Στη συνέχεια χωρίζει τις γραμμές του σύμφωνα με τις χρονικές σφραγίδες τους (βήμα 2) και χρησιμοποιεί τον Cross Association αλγόριθμο για να συσταδοποιήσει τις στήλες των πινάκων σύνδεσης (βήμα 3). Τότε, χρησιμοποιεί την MDL αρχή για να συνδέσει τις συστάδες στηλών των πινάκων (βήμα 4) και να μειώσει τα ασήμαντα χρονικά σημεία στην ιστορία της εξέλιξης του γραφήματος (βήμα 5).

Βιβλιογραφία

[1] http://en.wikipedia.org/wiki/K-means_clustering.

[2] Notes from: Tan, Steinbach, Kumar + Ghosh, The k-means algorithm.

[3] Bo Long, Zhongfei (Mark) Zhang, Xiaoyun Wu, Philip S. Yu, Spectral Clustering for Multi-type Relational Data.

[4] Sara C. Madeira and Arlindo L. Oliveira, Biclustering Algorithms for Biological Data Analysis: A Survey.

[5] http://en.wikipedia.org/wiki/Minimum_description_length.

[6] J. Rissanen, An Introduction to the MDL Principle.

[7] Steven de Rooij, Minimum Description Length Model Selection, Problems and Extensions.

[8] Deepayan Chakrabarti, Ravi Kumar, Andrew Tomkins, Evolutionary Clustering.

[9] Min-Soo Kim, Jiawei Han, A Particle-and-Density Based Evolutionary Clustering Method for Dynamic Networks.

[10] Bo Long, Xiaoyun Wu, Zhongfei (Mark) Zhang, Philip S. Yu, Unsupervised Learning on K-partite Graphs.

[11] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, Joydeep Ghosh, Clustering with Bregman Divergences.

[12] Joydeep Ghosh, Clustering with Bregman Divergences, University of Texas at Austin.

[13] http://www.public.asu.edu/~huanliu/papers/kdd09p.pdf.

[14] Miller McPherson, Lynn Smith-Lovin, James M Cook, BIRDS OF A FEATHER: Homophily in Social Networks.

[15] http://en.wikipedia.org/wiki/Lanczos_algorithm.

[16] J. Almeida, Lanczos Algorithm: Theory and Aplications, York (United Kingdom), April 2012.

[17] http://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo.

[18] Bo Long, Zhongfei Zhang, Philip S. Yu, Practical Relational Community Generation, ICDM 2007.

[19] Nam P. Nguyen, Thang N. Dinh, Sindhura Tokala, My T. Thai Department of Computer and Information Science and Engineering, University of Florida, USA, Overlapping Communities in Dynamic Networks: Their Detection and Mobile Applications.

[20] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, An efficient algorithm for large scale detection of protein families, 2002.

[21] Inderjit S. Dhillon, Dept. of Computer Sciences University of Texas, Austin, Subramanyam Mallela, Dept. of Computer Sciences University of Texas, Austin, Dharmendra S. Modha, IBM Almaden Research Center San Jose, CA, Information-Theoretic Co-clustering.

[22] Hyuk Cho, Inderjit S. Dhillon, Yuqiang Guan, Suvrit Sra, Minimum Sum-Squared Residue Co-clustering of Gene Expression Data.

[23] Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu, Dharmendra S. Modha, A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation.

[24] Jon Kleinberg, Department of Computer Science, Cornell University, Ithaca NY 14853, An Impossibility Theorem for Clustering.

[25] Charu C. Aggarwal, T. J. Watson Resch. Ctr., Jiawei Han, Jianyong Wang UIUC, Philip S. Yu, T. J. Watson Resch. Ctr., A Framework for Clustering Evolving Data Streams.

[26] Thomas M. Cover, Joy A. Thomas, Elements of Information Theory, 1991.

[27] Yu-Ru Lin, Yun Chi, Shenghuo Zhu, Hari Sundaram, Belle L. Tseng, Facetnet: a framework for analyzing communities and their evolutions in dynamic networks.

[28] Bin Gao, Tie-Yan Liu, Xin Zheng, Qian-Sheng Cheng, and Wei-Ying Ma, Consistent Bipartite Graph Co-Partitioning for Star-Structured High-Order Heterogeneous Data Co-Clustering.

[29] Bo Long Zhongfei Zhang, Philip S. Yu, Relational Data Clustering: Models, Algorithms, and Applications.

[30] Srujana Merugu, Co-clustering with Bregman Loss Functions.

[31] Lei Tang, Suju Rajan, Vijay K. Narayanan, Large Scale Multi-Label Classification via MetaLabeler.

[32] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, Yasemin Altun, Support Vector Machine Learning for Interdependent and Structured Output Spaces.

[33] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.3872&rep=rep1&type=pdf

[34] http://en.wikipedia.org/wiki/Singular_value_decomposition.

[35] Lei Tang, Xufei Wang and Huan Liu, Uncovering Groups via Heterogeneous Interaction Analysis.

[36] http://en.wikipedia.org/wiki/Chinese_restaurant_process.

[37] Spiros Papadimitriou, Aristides Gionis, Panayiotis Tsaparas, Risto A. Vaisanen, Heikki Mannila, Christos Faloutsos, Parameter-Free Spatial Data Mining Using MDL.

[38] http://en.wikipedia.org/wiki/NP-hard.

[39] Deepayan Chakrabarti, Spiros Papadimitriou, Dharmendra S. Modha, Christos Faloutsos, Fully Automatic Cross-Associations.

[40] Rissanen, J.: Modeling by shortest data description. Automatica 14 (1978) 465–471.

[41] Watts, D.J.: Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton Univ. Press (1999).

[42] Lei Tang and Huan Liu, Scalable Learning of Collective Behavior based on Sparse Social Dimensions, Data Mining and Machine Learning Laboratory, Computer Science & Engineering, Arizona State University.

[43] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada and Naonori Ueda, Learning Systems of Concepts with an Infinite Relational Model.

[44] Lisa Friedland, David Jensen, Finding Tribes: Identifying Close-Knit Individuals from Employment Patterns.

[45] Spiros Papadimitriou, Jimeng Su, Christos Faloutsos, Philip S. Yu, Hierarchical, Parameter-Free Community Discovery.

[46] Arindam Banerjee Sugato Basuy Srujana Merugu

ΚΕΦΑΛΑΙΟ 5 ΕΣΩΤΕΡΙΚΗ ΠΥΚΝΟΤΗΤΑ

5.1 Εισαγωγή

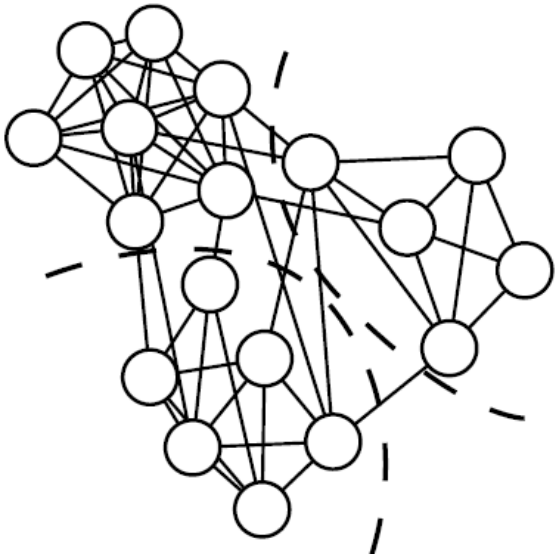
Στο Κεφάλαιο αυτό εξετάζονται ορισμένα προβλήματα που ορίζουν τον εντοπισμό κοινότητας ως μία διαδικασία που καθοδηγείται από τον άμεσο εντοπισμό των πυκνότερων περιοχών του δικτύου. Παρουσιάζονται οι μέθοδοι εντοπισμού κοινοτήτων που ορίζουν μία κοινότητα σύμφωνα με τον παρακάτω ορισμό:

Ορισμός 5.1 (Πυκνή Κοινότητα). Πυκνή κοινότητα σε ένα πολύπλοκο δίκτυο ονομάζεται ένα σύνολο από οντότητες που συνδέονται πυκνά. Για να είναι αυτές συνδεδεμένες πυκνά, μία ομάδα κορυφών πρέπει να έχει έναν αριθμό ακμών σημαντικά υψηλότερο από τον αναμενόμενο αριθμό των ακμών σε ένα τυχαίο γράφημα (το οποίο δεν έχει τη δομή μιας κοινότητας).

Γενικά, το ακόλουθο πόρισμα μοιράζεται από τους αλγορίθμους σε αυτή την κατηγορία:

Πόρισμα 5.1. Δεδομένου ενός γραφήματος, προσπαθήστε να αναπτύξετε ή να συμπτύξετε τις διαμερίσεις κόμβων προκειμένου να βελτιστοποιήσετε μία δεδομένη συνάρτηση πυκνότητας, σταματώντας όταν δεν είναι δυνατή αυτή η προσαύξηση.

Το Σχήμα 5.1 δείχνει ένα δίκτυο στο οποίο οι προσδιορισμένες κοινότητες είναι σημαντικά πυκνότερες από ένα τυχαίο γράφημα με την ίδια κατανομή βαθμού. Μία βασική έννοια για την ικανοποίηση του παραπάνω ορισμού είναι ο **σπονδυλωτής (modularity)**· ένα μέτρο της ποιότητας μιας συγκεκριμένης διαίρεσης ενός δικτύου σε ομάδες. Θεωρήστε τη διαίρεση του γραφήματος σε c μη επικαλυπτόμενες κοινότητες. Έστω το c_i να χαρακτηρίζει την συμμετοχή της κορυφής v_i στην κοινότητα, και το k_i να παριστά το βαθμό της κορυφής i. Ο σπονδυλωτής είναι ένα μέτρο της απόκλισης από μία αναμενόμενη τιμή, δηλαδή, η δομή της κοινότητας ενός ενιαίου μοντέλου τυχαίου γραφήματος με την ίδια αναμενόμενη σειρά βαθμού του αρχικού δικτύου. Σε αυτό το μοντέλο μία οντότητα συνδέεται με άλλες με ομοιόμορφη πιθανότητα. Για δύο κόμβους με βαθμό k_i και k_j αντίστοιχα, ο αναμενόμενος αριθμός των ακμών μεταξύ των δύο σε ένα ομοιόμορφο μοντέλο τυχαίου γραφήματος είναι $\frac{k_i k_j}{2m}$, όπου m είναι ο αριθμός των ακμών του γραφήματος.



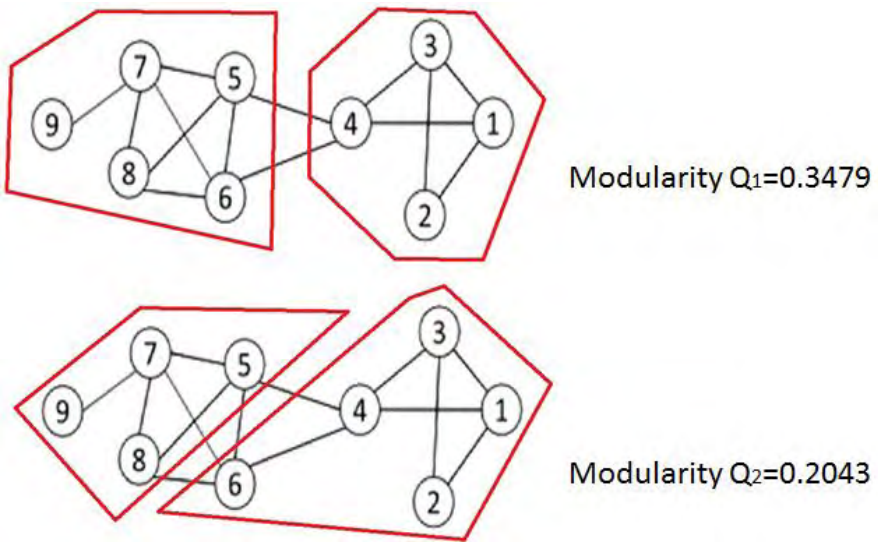
Σχήμα 5.1: Παράδειγμα ενός γραφήματος το οποίο μπορεί να κατανέμεται με την έννοια της εσωτερικής πυκνότητας μεταξύ των κόμβων του.

Το modularity μετράει πόσο μακριά αποκλίνει η αλληλεπίδραση από ένα ενιαίο τυχαίο γράφημα με την ίδια κατανομή βαθμού. Ορίζεται ως:

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)\ ,$$

όπου $\delta(c_i, c_j) = 1$ εάν $c_i = c_j$ (δηλαδή, οι δύο κόμβοι είναι στην ίδια κοινότητα), και 0 αλλιώς, και A_{ij} είναι ο αριθμός των ακμών μεταξύ των κόμβων i και j. Για λόγους συντομίας υποθέτουμε ότι $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$. Μία μεγαλύτερη τιμή του σπονδυλωτή δείχνει μια πυκνότερη αλληλεπίδραση στο εσωτερικό των ομάδων. Βλέπε Σχήμα 5.2. Σημειώστε ότι το Q θα μπορούσε να είναι αρνητικό εάν οι κορυφές διαχωρίζονται σε κακές συστάδες, ή απλά εάν το δίκτυο έχει μη ταξινομημένη ή πολυμερής δομή. $Q > 0$ σημαίνει ότι η συσταδοποίηση αποτυπώνει κάποιο βαθμό από τη δομή της κοινότητας. Ουσιαστικά, ο στόχος είναι να βρεθεί μία δομή της κοινότητας, έτσι ώστε το Q να μεγιστοποιείται. Τιμές που πλησιάζουν το $Q = 1$, που είναι η μέγιστη, δείχνουν ισχυρή δομή της κοινότητας. Στην πράξη, οι τιμές για τέτοια δίκτυα τυπικά εμπίπτουν στο εύρος από περίπου 0,3 έως 0,7. Υψηλότερες τιμές είναι σπάνιες.

Ο σπονδυλωτής εμπλέκεται στο πρόβλημα του εντοπισμού της κοινότητας σε δύο επίπεδα. Κατ' αρχάς, μπορεί να ποσοτικοποιήσει το πόσο καλή είναι μία συγκεκριμένη διαμέριση του δικτύου. Δίνει ένα αποτέλεσμα της ποιότητας της διαμέρισης ακόμη και χωρίς οποιαδήποτε γνώση των πραγματικών κοινοτήτων του δικτύου. Αυτό προσφέρεται ιδιαίτερα για πολύ μεγάλα δίκτυα. Από την άλλη πλευρά, ο σπονδυλωτής δεν είναι η ιδανικότερη λύση για την αξιολόγηση μιας προτεινόμενης διαμέρισης της κοινότητας. Πάσχει από γνωστά προβλήματα, ιδίως στην επίλυση συγκεκριμένων προβλημάτων. Ο σπονδυλωτής αποτυγχάνει να εντοπίσει κοινότητες



Σχήμα 5.2: Παράδειγμα του σπονδυλωτή.

που είναι μικρότερες από μία κλίμακα που εξαρτάται από το συνολικό μέγεθος του δικτύου και από το βαθμό της αλληλεξάρτησης των κοινοτήτων, ακόμη και σε περιπτώσεις όπου οι ενότητες είναι σαφώς καθορισμένες. Επιπλέον, με τον σπονδυλωτή μπορούν να αξιολογηθούν μόνο οι κοινότητες που εξάγονται σύμφωνα με τον ορισμό που προτείνεται στην ενότητα αυτή. Κάθε άλλο είδος ορισμού των κοινοτήτων θα έχει ως αποτέλεσμα μία όχι και τόσο ουσιαστική αξιολόγηση εφαρμόζοντας τον σπονδυλωτή. Για μία εκτενή ανασκόπηση των γνωστών προβλημάτων του σπονδυλωτή βλέπε τα άρθρα [5,6].

Το δεύτερο επίπεδο της χρήσης του σπονδυλωτή στη διαδικασία της διαμέρισης του γραφήματος παρουσιάζεται από αλγορίθμους εντοπισμού κοινοτήτων που βασίζονται στην μεγιστοποίηση του σπονδυλωτή. Αυτοί οι αλγόριθμοι πάσχουν από τα προαναφερθέντα προβλήματα της χρήσης του σπονδυλωτή ως ποιοτικά μέτρα. Ωστόσο, η μεγιστοποίηση του σπονδυλωτή είναι ένα πολύ παραγωγικό πεδίο της έρευνας, και υπάρχουν πολλοί αλγόριθμοι που βασίζονται σε ευρετικές και στρατηγικές για την εξεύρεση των καλύτερων διαμερίσεων του δικτύου.

Θα παρουσιαστεί το βασικό παράδειγμα μιας προσέγγισης που βασίζεται στον σπονδυλωτή, παρέχοντας αναφορές για ήσσονος σημασίας αλγορίθμους μεγιστοποίησης σπονδυλωτή. Μια καλή ανασκόπηση του modularity ιδιοδιανύσματος φαίνεται στο [7].

Ο σπονδυλωτής δεν είναι η μόνη συνάρτηση κόστους η οποία είναι σε θέση να υπολογίσει εάν ένα σύνολο οντοτήτων είναι περισσότερο σχετικό από το αναμενόμενο και επομένως μπορεί να θεωρηθεί ως μία κοινότητα. Οι άλλες μέθοδοι που βασίζονται σε διαφορετικές τεχνικές, αλλά μοιράζονται τον ίδιο ορισμό της κοινότητας που προτείνεται στην παρούσα ενότητα, είναι οι εξής:

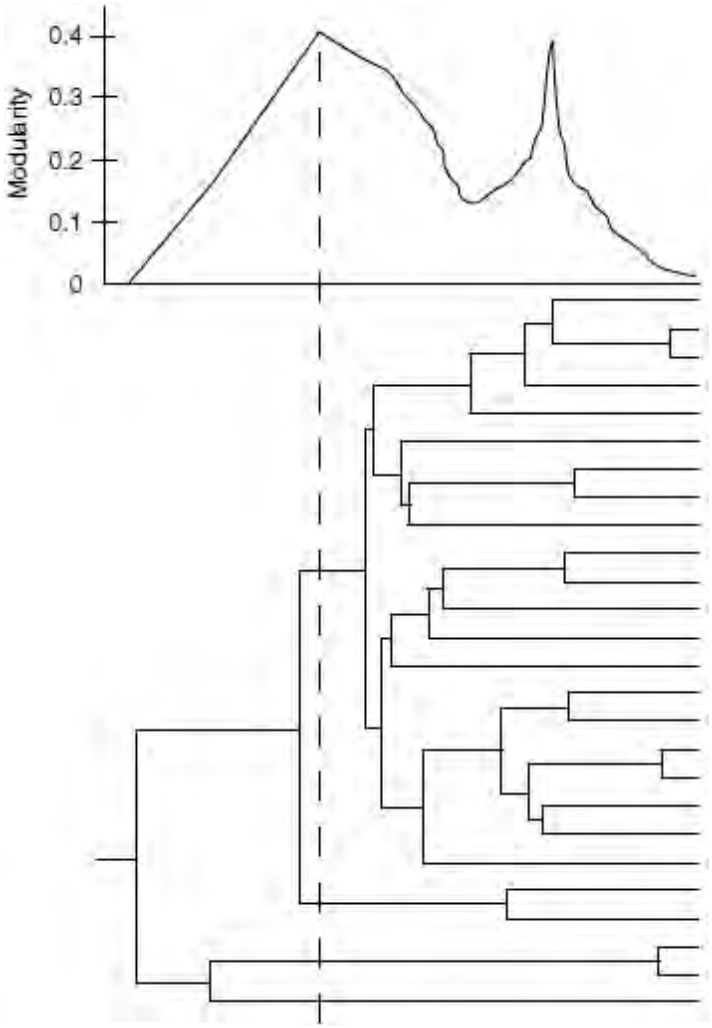
- *MetaFac*, μία τεχνική παραγοντοποίησης υπεργράφου
- έναν φυσικο-χημικό αλγόριθμο χρησιμοποιώντας μία Bayesian προσέγγιση
- μία τοπική προσέγγιση βασισμένη στην πυκνότητα που ονομάζεται LA → IS^2
- και μία άλλη προτεινόμενη συνάρτηση που χρησιμοποιείται για να μετρήσει την εσωτερική τοπική πυκνότητα μιας συστάδας.

Η βελτιστοποίηση μιας συνάρτησης πυκνότητας είναι κατάλληλη για πολλές αναπαραστάσεις γραφημάτων, όπως είναι οι κατευθυνόμενοι (directed) και οι σταθμισμένοι (weighted) γράφοι. Ωστόσο, εκτός από τα προβλήματα του σπονδυλωτή, υπάρχουν και άλλα αδύναμα σημεία. Για παράδειγμα, οι πιο πολύπλοκες δομές δεν είναι ευάγωγες σε αυτή την προσέγγιση, όπως είναι τα πολυδιάστατα δίκτυα. Αν πολλαπλές διαφορετικές ποιοτικές σχέσεις είναι παρούσες σε ένα δίκτυο, πώς πρέπει να υπολογίζεται μία συνεπής τιμή της «πολυσχεσιακής πυκνότητας»; Αυτό είναι σε θέση να το πράξει μία παραλλαγή του modularity για πολυτομικά δίκτυα (για περισσότερα δείτε το [8]). Ωστόσο, πρόσφατες έρευνες θέτουν κάποια ερωτήματα σχετικά με την ασάφεια της πυκνότητας σε πολυδιάστατα δίκτυα, και οι ανησυχίες αυτές δεν εξετάζονται στο παραπάνω paper. Για περισσότερες πληροφορίες ανατρέξτε στα [9,10]. Συνεπώς, δεδομένης της τρέχουσας κατάστασης δεν είναι δυνατόν να υπάρχει μία καθαρή λύση όσον αφορά την πολυδιάστατη πυκνότητα, και οι προσεγγίσεις που παρουσιάζονται σε αυτή τη κατηγορία πάσχουν σε αυτό το σενάριο.

5.2 Σπονδυλωτής (Modularity)

Ο σπονδυλωτής, γενικά, είναι ένα μέτρο της δομής των δικτύων ή των γραφημάτων. Σχεδιάστηκε για να μετρά την ισχύ (δύναμη) της διαίρεσης ενός δικτύου σε ενότητες (που ονομάζονται επίσης ομάδες, συστάδες ή κοινότητες). Δίκτυα με υψηλό modularity έχουν *πυκνές* συνδέσεις μεταξύ των κόμβων *μέσα* στις ενότητες αλλά *αραιές* συνδέσεις μεταξύ των κόμβων σε *διαφορετικές* ενότητες. Ο σπονδυλωτής χρησιμοποιείται συχνά σε μεθόδους βελτιστοποίησης για τον εντοπισμό της δομής της κοινότητας μέσα σε δίκτυα. Ωστόσο, έχει αποδειχθεί ότι πάσχει από ένα όριο ανάλυσης και, επομένως, δεν είναι σε θέση να ανιχνεύσει μικρές κοινότητες.

Το modularity αποτελεί μία ιδιότητα ενός δικτύου και μία συγκεκριμένη προτεινόμενη κατανομή του εν λόγω δικτύου σε κοινότητες. Αυτός μετρά, όταν η διαμέριση είναι καλή, με την έννοια ότι υπάρχουν πολλές ακμές εντός των κοινοτήτων και μόνο μερικές μεταξύ αυτών. Το να βρεθεί μια διαμέριση που να παρέχει τη μέγιστη τιμή του modularity, αυτό αποτελεί ένα NP-complete πρόβλημα.



Σχήμα 5.3: Ένα δενδρόγραμμα για τον αλγόριθμο μεγιστοποίησης του modularity, με ένα διάγραμμα των τιμών του modularity που προκύπτουν λόγω της διαμέρισης. Καθώς προχωράμε αριστερά στο δέντρο οι κορυφές ενώνονται για να σχηματίσουν όλο και μεγαλύτερες κοινότητες, όπως υποδεικνύεται από τις γραμμές, μέχρι να φτάσουμε τέρμα αριστερά, όπου όλες ενώνονται σε μια ενιαία κοινότητα. Μία εγκάρσια τομή του δένδρου σε οποιοδήποτε επίπεδο, όπως υποδεικνύεται από τη διακεκομμένη γραμμή, θα δώσει τις κοινότητες σε αυτό το επίπεδο.

Ως εκ τούτου, έχουν προταθεί πολλοί άπληστοι ευρετικοί μηχανισμοί. Μετά από μία πρωτοποριακή εργασία που είχε προτείνει για τον σπονδυλωτή [11], ο Newman παρουσίασε μία αποτελεσματική στρατηγική για τη μεγιστοποίηση του σπονδυλωτή και συγκεκριμένα τη κατ' επανάληψη συγχώνευση των δύο κοινοτήτων, των οποίων η συνένωση παράγει τη μεγαλύτερη αύξηση στο Q. Για ένα δίκτυο n κορυφών, μετά από n - 1 τέτοιες ενώσεις έχουμε μείνει με μία ενιαία κοινότητα και ο αλγόριθμος σταματά. Η όλη διαδικασία μπορεί να αναπαρασταθεί ως ένα δέντρο του οποίου τα φύλλα είναι οι κορυφές του αρχικού δικτύου και του οποίου οι εσωτερικοί κόμβοι αντιστοιχούν στις ενώσεις. Αυτό το δενδρόγραμμα αντιπροσωπεύει μία ιεραρχική διάσπαση του δικτύου σε κοινότητες σε όλα τα επίπεδα, το οποίο πρέπει να κοπεί στην αιχμή (peak) του σπονδυλωτή, προκειμένου να βρεθούν οι

κοινότητες, όπως απεικονίζεται στο Σχήμα 5.3. Η λειτουργία του αλγορίθμου προϋποθέτει την εξεύρεση των αλλαγών στο Q που θα προέκυπταν από την συνένωση του κάθε ζεύγους των κοινοτήτων, επιλέγοντας το μεγαλύτερο από αυτά, και εκτελώντας την αντίστοιχη συνένωση. Ένας τρόπος για να προβλεφθεί (και να εφαρμοστεί) αυτή η διαδικασία είναι να θεωρηθεί ένα δίκτυο ως ένα πολυγράφημα, στο οποίο μία ολόκληρη κοινότητα αντιπροσωπεύεται από μια κορυφή, δεσμίδες ακμών συνδέουν μία κορυφή στην άλλη, και οι ακμές στο εσωτερικό των κοινοτήτων αντιπροσωπεύονται από αυτόνομες ακμές. Στον αλγόριθμο του άρθρου [11], αυτή η λειτουργία γίνεται ρητά σε ολόκληρο τον πίνακα, αλλά εάν ο πίνακας γειτνίασης είναι αραιός (που αναμένεται στα πρώτα στάδια της διαδικασίας) η λειτουργία μπορεί να εκτελεστεί αποτελεσματικότερα χρησιμοποιώντας δομές δεδομένων για αραιούς πίνακες. Δυστυχώς, ο υπολογισμός του ΔQ_{ij} και η εξεύρεση του ζεύγους i, j με το μεγαλύτερο ΔQ_{ij} απαιτεί πολύ χρόνο.

Το Σχήμα 5.3 δείχνει άλλη μία ακόμη ιδιότητα. Έχει ανακαλυφθεί ότι ο σπονδυλωτής δεν έχει μία μόνο κορυφή (ένα peak), δεδομένων όλων των πιθανών κατατμήσεων, αλλά υπάρχουν πολλά τοπικά βέλτιστα. Επιπλέον, τα πραγματικά δίκτυα έχουν πολλά σχεδόν ολικά βέλτιστα σε διάφορα σημεία [6] (η δεξιότερη κορυφή στο Σχήμα 5.3) και δεν μπορεί κάποιος εξ' αρχής να γνωρίζει το που ο αλγόριθμος εντοπίζει τη λύση του.

Στον αλγόριθμο αυτό, παρά τη διατήρηση του πίνακα γειτνίασης και τον υπολογισμό του ΔQ_{ij}, διατηρείται και ενημερώνεται ο πίνακας τιμών του ΔQ_{ij}. Από τη στιγμή που όταν δεν συνενώνονται δύο κοινότητες με καμία ακμή μεταξύ τους δεν μπορεί ποτέ να παραχθεί μία αύξηση στο Q, χρειάζεται μόνο να αποθηκευτεί το ΔQ_{ij} για αυτά τα ζεύγη i, j που είναι ενωμένα με μία ή περισσότερες ακμές. Δεδομένου ότι αυτός ο πίνακας έχει την ίδια στήριξη με τον πίνακα γειτνίασης, θα είναι παρομοίως αραιός, ώστε να μπορεί να αναπαρασταθεί και πάλι με αποτελεσματικές δομές δεδομένων. Επιπλέον, γίνεται χρήση μιας αποτελεσματικής δομής δεδομένων για την παρακολούθηση του μεγαλύτερου ΔQ_{ij}.

Η βελτιστοποίηση που προτείνεται είναι να εφοδιαστεί ένας πίνακας που να περιέχει μόνο τις τιμές των κοινοτήτων, δηλαδή τις αλλαγές στον σπονδυλωτή Q όταν ενώνονται οι κοινότητες i και j. Ο αλγόριθμος μπορεί τώρα να ορισθεί ως εξής.

1. Υπολογίστε τις αρχικές τιμές του $\Delta Q_{i,j}$ και παρακολουθείστε το μεγαλύτερο στοιχείο της κάθε γραμμής του πίνακα ΔQ,
2. Επιλέξτε το μεγαλύτερο $\Delta Q_{i,j}$ μεταξύ αυτών των μεγαλύτερων στοιχείων.
3. Ενώστε τις αντίστοιχες κοινότητες,
4. Ενημερώστε τον πίνακα ΔQ και τη συλλογή των μεγαλύτερων στοιχείων.
5. Αυξήστε τον Q από τον $\Delta Q_{i,j}$. Επαναλάβετε αυτό το τελευταίο βήμα μέχρι το δενδρόγραμμα να είναι πλήρες.

Στο άρθρο [12] η προσέγγιση της μεγιστοποίησης του σπονδυλωτή είναι προσαρμοσμένη στην περίπτωση ενός κατευθυνόμενου δικτύου. Συνεπώς, δίνεται μία αναπαράσταση του γραφήματος με πίνακα, αλλά αυτός ο πίνακας δεν είναι συμμετρικός. Ο αλγόριθμος βασίζεται στο άρθρο [13]. Πιο πρόσφατες εργασίες δείχνουν, επίσης, την εφαρμογή της προσέγγισης του

σπονδυλωτή για τις επικαλυπτόμενες κοινότητες [14]. Έχει επίσης προταθεί μια τοπική αξιολόγηση του σπονδυλωτή, διαιρώντας το γράφημα σε γνωστά, συνοριακά και ανεξερεύνητα σύνολα.

Μία άλλη βελτιστοποίηση των προσεγγίσεων που βασίζονται στον σπονδυλωτή παρουσιάζεται στο άρθρο [15]. Αυτός είναι βασικά ένας διαιρετικός αλγόριθμος που βελτιστοποιεί το modularity Q χρησιμοποιώντας μια ευρετική αναζήτηση. Η αναζήτηση αυτή βασίζεται σε ένα μέτρο λ, που εξαρτάται από το βαθμό του κόμβου, και η κανονικοποίησή του περιλαμβάνει όλες τις συνδέσεις στο δίκτυο μετά την άθροιση. Ο κόμβος που επιλέγεται, σε έναν πρωτότυπο εξωτερικό αλγόριθμο βελτιστοποίησης [16] είναι πάντα ο κόμβος με τη χειρότερη λ_i τιμή. Έχουν προταθεί μια σειρά από άλλες στρατηγικές βελτιστοποίησης (μείωση του μεγέθους [17], προσομοιωμένη ανόπτηση [18]).

Τελικά, παρουσιάζεται η τελευταία άπληστη προσέγγιση που είναι σύμφωνη με τον κλασικό ορισμό του σπονδυλωτή [19]. Το προηγούμενο μεγαλύτερο γράφημα που χρησιμοποιήθηκε για τη δοκιμή του σπονδυλωτή ήταν 5,5 εκατομμύρια κόμβοι [20], με τη βελτίωση αυτή είναι δυνατόν να αναβαθμιστεί σε 100 εκατομμύρια κόμβους. Ο αλγόριθμος χωρίζεται σε δύο φάσεις που επαναλαμβάνονται επαναληπτικά. Για κάθε κόμβο i θεωρήστε τους γείτονες J του i και αξιολογήστε το κέρδος στον σπονδυλωτή που θα λάβει χώρα με την αφαίρεση του i από την κοινότητά του και τοποθετώντας το στην κοινότητα του J. Ο κόμβος i στη συνέχεια τοποθετείται στην κοινότητα για την οποία το κέρδος είναι το μέγιστο μέχρι καμία ατομική κίνηση να μπορεί να βελτιώσει το modularity. Η δεύτερη φάση περιλαμβάνει τη δημιουργία ενός νέου δικτύου, του οποίου οι κόμβοι είναι τώρα οι κοινότητες που βρέθηκαν κατά τη διάρκεια της πρώτης φάσης. Είναι τότε δυνατόν να εφαρμοστεί εκ νέου το πρώτο στάδιο στο προκύπτον σταθμισμένο δίκτυο και να επαναλαληφθεί η ίδια διαδικασία. Αυτή η μέθοδος έχει δοκιμαστεί στο WebGraph UKUnion [21], σε συνεργατικά δίκτυα [22], καθώς και σε δίκτυα κινητής τηλεφωνίας.

Ένα ιδιαίτερα ενδιαφέρον πλαίσιο εργασίας στον σπονδυλωτή είναι ο πολυτομικός σπονδυλωτής (Multislice Modularity)[8]. Επεκτείνεται το απλό μοντέλο του σπονδυλωτή (το τυχαίο γράφημα) σε ένα νέο πολύπλοκο (πολυδιάστατο). Χρησιμοποιούνται διάφορες γενικεύσεις, δηλαδή μία πρόσθετη παράμετρος που ελέγχει την σύζευξη μεταξύ των διαστάσεων, βασίζοντας τη λειτουργία τους στην ισοδυναμία μεταξύ του σπονδυλωτή, όπως είναι οι συναρτήσεις ποιότητας, και την Laplacian δυναμική των πληθυσμών των τυχαίων περιπατητών [23]. Βασικά, το παραπάνω επεκτείνεται επιτρέποντας πολυδιάστατα μονοπάτια για τον τυχαίο περιπατητή [24], εξετάζοντας τους διαφορετικούς τύπους σύνδεσης με διαφορετικά βάρη [25], και μια διαφορετική εξάπλωση αυτών των βαρών μεταξύ των διαστάσεων [26].

Προκειμένου να αναπαρασταθούν και τα στιγμιότυπα και οι διαστάσεις του δικτύου, χρησιμοποιείται η μέθοδος του τεμαχισμού. Κάθε κομμάτι s ενός δικτύου αντιπροσωπεύεται από ένα στοιχείο A_{ijs} του πίνακα γειτνίασης του δικτύου μεταξύ των κόμβων i και j, το οποίο

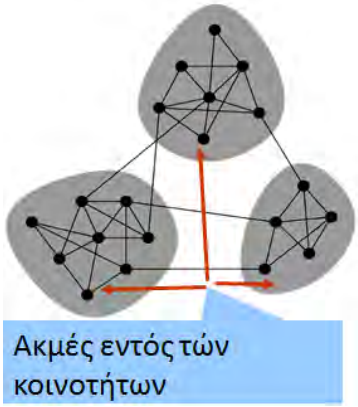
κανονικά θα είναι 0 ή 1 αν και μεγαλύτερες τιμές είναι πιθανές σε δίκτυα όπου επιτρέπονται οι πολλαπλές ακμές· δηλαδή, γενικά

$$A_{ij} = \begin{cases} 1, & \text{εάν οι κορυφές i , j ενώνονται} \\ 0, & \text{σε οποιαδήποτε άλλη περίπτωση} \end{cases} .$$

Επίσης, καθορίζονται σύνδεσμοι C_{jrs} μεταξύ των κομματιών που συνδέουν τον κόμβο j στο κομμάτι r με τον εαυτό του στο κομμάτι s. Επισημαίνονται τα πλεονεκτήματα του κάθε κόμβου χωριστά σε κάθε κομμάτι, έτσι ώστε $k_{js} = \sum_i A_{ijs}$ και $c_{js} = \sum_r C_{jsr}$, και συνολικά $\kappa_{js} = k_{js} + c_{js}$. Στη συνέχεια καθορίζεται ένα σχετικό πολυτομικό μοντέλο. Ο προκύπτων πολυτομικός γενικευμένος ορισμός του σπονδυλωτή είναι ο ακόλουθος:

$$Q = \frac{1}{2\mu}\sum_{ijsr}\left\{\left(A_{ijs} - \gamma_s \frac{k_{is}k_{js}}{2m_s}\delta_{sr}\right) + \delta_{ij}C_{jsr}\right\}\delta(c_{is}, c_{jr}) .$$

Σε αυτή τη γενίκευση, το γ_sείναι η παράμετρος ανάλυσης, που μπορεί (ή δεν μπορεί) να είναι διαφορετική για κάθε κομμάτι. Εάν γ_s = 1 για κάθε s, τότε αυτός ο τύπος εκφυλίζεται στη συνήθη ερμηνεία του σπονδυλωτή ως μία καταμέτρηση του συνολικού βάρους των ακμών μεταξύ των κομματιών μείον το βάρος που αναμένεται τυχαία. Σε αντίθετη περίπτωση, εξετάζονται οι σύνδεσμοι C_{jsr} μεταξύ των κομματιών. Ο C_{jsr} παίρνει τιμές από το 0 έως το ∞. Αν C_{jsr} = 0 τότε εκφυλίζεται και πάλι με τον συνήθη ορισμό του σπονδυλωτή. Διαφορετικά, οι ποιοτικο-βελτιστοποιημένες κατατμήσεις αναγκάζουν την ανάθεση της κοινότητας ενός κόμβου να παραμείνει η ίδια σε όλα τα κομμάτια στα οποία εμφανίζεται ο συγκεκριμένος κόμβος. Επιπλέον, η πολυτομική ποιότητα μειώνεται σε σχέση με εκείνη του πίνακα γειτνίασης αθροιστικά για τις συνεισφορές από τα μεμονωμένα κομμάτια με ένα μοντέλο που σέβεται το βαθμό των κατανομών των ατομικών συνεισφορών. Η γενικότητα αυτού του πλαισίου εργασίας επιτρέπει επίσης διαφορετικά βάρη να συμπεριληφθούν σε όλους τους C_{jsr} συνδέσμους. Μετά τον καθορισμό της νέας συνάρτησης της ποιότητας, ο αλγόριθμος που απαιτείται για την εξαγωγή των κοινοτήτων μπορεί να είναι ένας από τους πολλούς αλγορίθμους που βασίζονται στην έννοια του modularity.



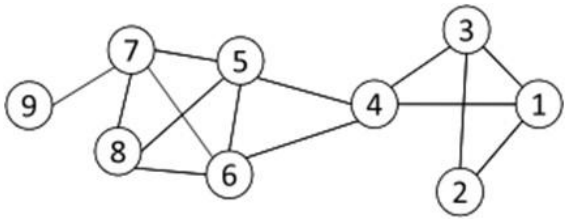
Σχήμα 5.4: Εσωτερικές ακμές [13].

Ένας περιοριστικός παράγοντας είναι ότι, ανάλογα με την εφαρμογή, δεν μπορούν να επιστραφούν όλα τα χαρακτηριστικά (π.χ., μόνο η πολυτομική εφαρμογή είναι σε θέση να εξετάσει πολλαπλές διαστάσεις).

Συνοπτικά, όσον αφορά τον σπονδυλωτή θα πρέπει να γνωρίζουμε τα εξής βασικά:
➢ **Q** = (πλήθος ακμών εντός των κοινοτήτων) – (αναμενόμενο πλήθος εσωτερικών ακμών σε ένα τυχαίο γράφημα με τον ίδιο βαθμό κόμβων)
➢ $\mathbf{k_i}$: βαθμός του κόμβου i
➢ **m** : πλήθος ακμών γραφήματος
➢ $\mathbf{A_{ij}}$= 1 αν (i,j) ∈ στο σύνολο E των ακμών του γραφήματος, 0 αλλιώς
➢ $\frac{\mathbf{k_i k_j}}{\mathbf{2m}}$: αναμενόμενος αριθμός ακμών μεταξύ i και j σε ένα τυχαίο γράφημα με τον ίδιο βαθμό κόμβων
➢ **δ(c_i, c_j)** = 1 εάν c_i = c_j (δηλαδή, οι δύο κόμβοι είναι στην ίδια κοινότητα), και 0 αλλιώς
➢ $\mathbf{Q} = \frac{\mathbf{1}}{\mathbf{2m}} \sum_{\mathbf{ij}} \left[\mathbf{A_{ij}} - \frac{\mathbf{k_i k_j}}{\mathbf{2m}}\right] \boldsymbol{\delta}(\mathbf{c_i}, \mathbf{c_j})$
➢ $\mathbf{B_{ij}} = \mathbf{A_{ij}} - \frac{\mathbf{k_i k_j}}{\mathbf{2m}}$

Ενδεικτικά, παρακάτω δίνεται ένα στιγμιότυπο ενός παραδείγματος σπονδυλωτή, όπου υπολογίζεται μέχρι και ο πίνακας $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$.

Έστω ότι το παρακάτω δίκτυο αποτελείται από 9 κορυφές και 14 ακμές (δηλαδή m = 14), και έχει την παρακάτω μορφή:



Ο πίνακας γειτνίασης αποτελείται από 9 γραμμές και 9 στήλες, όπου κάθε γραμμή/στήλη αντιπροσωπεύει και έναν κόμβο σε αύξοντα αριθμό. Έτσι ο πίνακας αυτός με τις ανάλογες συνδέσεις των ακμών του παραπάνω δικτύου έχει την εξής μορφή:

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

Επιπλέον, από το συγκεκριμένο δίκτυο αντλείται και η εξής πληροφορία:

k_1=3 , k_2=2 , k_3=3 , k_4=4 , k_5=4 , k_6=4 , k_7=4 , k_8=3 , k_9=1

Έπειτα, υπολογίζοντας για κάθε ζεύγος κορυφών τον όρο B_{ij} σύμφωνα με την εξίσωση $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$, π.χ. $B_{11} = A_{11} - \frac{k_1 k_1}{2m} = 0 - \frac{9}{28} = -0.3214$,

$$B_{12} = A_{12} - \frac{k_1 k_2}{2m} = 1 - \frac{6}{28} = 0.7857, \text{ κ.ο.κ}$$

Προκύπτει ο εξής πίνακας B:

$$B = \begin{pmatrix} -0.32 & 0.79 & 0.68 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.79 & -0.14 & 0.79 & -0.29 & -0.29 & -0.29 & -0.29 & -0.21 & -0.07 \\ 0.68 & 0.79 & -0.32 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.57 & -0.29 & 0.57 & -0.57 & 0.43 & 0.43 & -0.57 & -0.43 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & -0.57 & 0.43 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & 0.43 & -0.57 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & -0.57 & 0.43 & 0.43 & -0.57 & 0.57 & 0.86 \\ -0.32 & -0.21 & -0.32 & -0.43 & 0.57 & 0.57 & 0.57 & -0.32 & -0.11 \\ -0.11 & -0.07 & -0.11 & -0.14 & -0.14 & -0.14 & 0.86 & -0.11 & -0.04 \end{pmatrix}$$

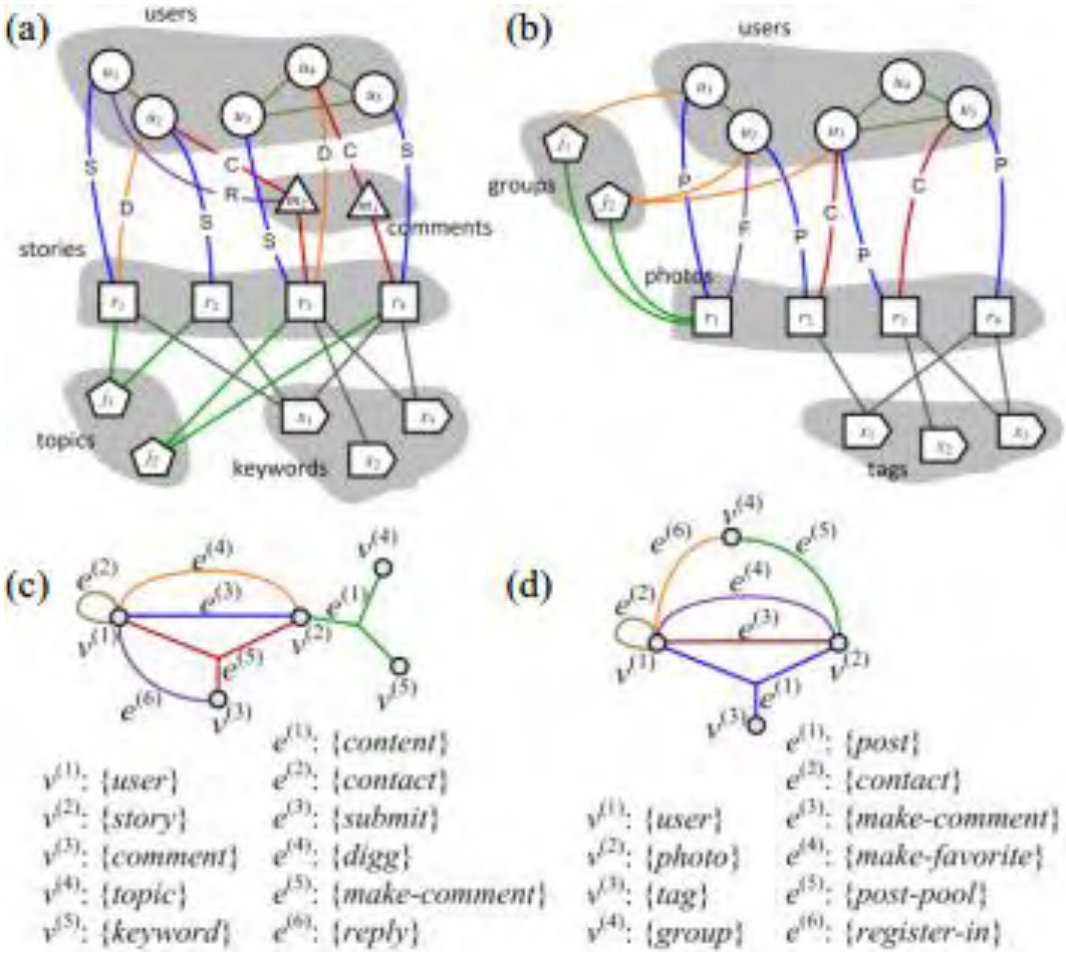
Ο αλγόριθμος συνεχίζεται όπως ακριβώς περιγράφεται παραπάνω στην παρούσα ενότητα. Ένα παράδειγμα μπορείτε να δείτε και στον σύνδεσμο [27].

5.3 MetaFac (MetaGraph Factorization)

Σε αυτή την ενότητα παρουσιάζεται ο αλγόριθμος του **MetaFac(MetaGraph Factorization)**, ένα πλαίσιο εργασίας που εξάγει δομές κοινοτήτων από διάφορα κοινωνικά περιβάλλοντα και αλληλεπιδράσεις. Η παρουσίασή του έχει τρεις βασικές συνεισφορές: (1) μεταγράφημα, ένα σχεσιακό υπεργράφημα που αναπαριστά πολυσχεσιακά και πολυδιάστατα κοινωνικά δεδομένα, (2) μία αποτελεσματική μέθοδος παραγοντοποίησης για την εξαγωγή κοινοτήτων σε ένα δεδομένο μεταγράφημα, (3) μία on-line μέθοδος για να χειρίζεται χρονικά μεταβαλλόμενες σχέσεις μέσω της στοιχειώδους παραγοντοποίησης του μεταγραφήματος. Στην πράξη, υπάρχουν οντότητες οι οποίες συνδέονται με διαφορετικά είδη αντικειμένων και με διάφορους τρόπους (π.χ. σε μέσα κοινωνικής δικτύωσης μέσω tagging, σχολιάζοντας, ή δημοσιεύοντας μια φωτογραφία, βίντεο ή κείμενο). Ο στόχος είναι να ανακαλυφθεί μια λανθάνουσα δομή της κοινότητας στο μεταγράφημα, για παράδειγμα, το κοινό περιβάλλον των ενεργειών των χρηστών σε μέσα κοινωνικής δικτύωσης. Με άλλα λόγια, ο αλγόριθμος ασχολείται με συστάδες που αποτελούνται από οντότητες που αλληλεπιδρούν μεταξύ τους με συνεκτικό τρόπο. Σε αυτό το μοντέλο, ένα σύνολο οντοτήτων του ίδιου τύπου ονομάζεται *έδρα/πλευρά* (facet). Επιπλέον, μία αλληλεπίδραση μεταξύ δύο ή περισσότερων εδρών ονομάζεται *σχέση* (relation).

Πιο συγκεκριμένα, σε αυτή την ενότητα η μελέτη επικεντρώνεται κυρίως στον εντοπισμό δομών κοινοτήτων σε πλούσια μέσα κοινωνικής δικτύωσης, μέσω της ανάλυσης των χρονικά μεταβαλλόμενων πολυσχεσιακών δεδομένων από ιστοσελίδες κοινωνικής δικτύωσης. Οι διαδικτυακοί τόποι κοινωνικής δικτύωσης, όπως το Flickr, το Digg και το Facebook, επιτρέπουν ένα ευρύ φάσμα ενεργειών για τα αντικείμενα των πολυμέσων - π.χ. ανέβασμα φωτογραφιών, υποβολή και σχολιασμό ειδήσεων, bookmarking και τοποθέτηση ετικετών (tags), σημείωση εγγράφων, δημιουργία web-συνδέσεων, καθώς και ενέργειες σε σχέση με τους άλλους χρήστες (π.χ. κοινή χρήση πολυμέσων και συνδέσμων με ένα φίλο). Το κλειδί σε τέτοιου είδους πληροφορίες κοινωνικών πολυμέσων, όπως η σύσταση των πολυμέσων, βασίζεται στην κατανόηση του περιβάλλοντος αυτών των ενεργειών - πώς σχετίζονται με άλλες ενέργειες, οι χρήστες και τα αντικείμενα των πολυμέσων. Για παράδειγμα, σε έναν χρήστη μπορεί να δοθούν κίνητρα για να αναζητήσει μια ιστορία, μετά την προβολή των σελιδοδεικτών του φίλου του.

Το πρόβλημα έχει δύο προκλήσεις: (1) σε μέσα κοινωνικής δικτύωσης, το περιβάλλον των ενεργειών των χρηστών αλλάζει συνεχώς και συν-εξελίσσεται, π.χ. σε σχέση με άλλες ενέργειες χρηστών, αναδυόμενες έννοιες και ιστορικές προτιμήσεις χρηστών. Ως εκ τούτου, το κοινωνικό περιβάλλον περιέχει τις χρονοεξελισσόμενες πολυδιάστατες σχέσεις, (2) το κοινωνικό περιβάλλον καθορίζεται από τα διαθέσιμα χαρακτηριστικά του συστήματος που επιτρέπουν αλληλεπιδράσεις μεταξύ των αντικειμένων των πολυμέσων και των ατόμων. Ως εκ τούτου, το κοινωνικό περιβάλλον είναι μοναδικό σε κάθε ιστοσελίδα κοινωνικής δικτύωσης. Για παράδειγμα, το Σχήμα 5.5 δείχνει τις κυριότερες ενέργειες που διατίθενται στο Digg και Flickr, καθώς και τα συναφή αντικείμενα των πολυμέσων. Στο Digg, οι χρήστες μπορούν να υποβάλουν / ψηφίσουν (digg) / σχολιάσουν μια ιστορία, να απαντήσουν σε σχόλιο, να απαντήσουν στην απάντηση, κλπ. Οι χρήστες του Flickr μπορούν να δημοσιεύουν, να επισημαίνουν και να σχολιάζουν πάνω σε μία φωτογραφία, να κάνουν αιτήσεις για νέους φίλους, να επισημαίνουν μια φωτογραφία ως αγαπημένη, να συμμετέχουν σε μια ομάδα ανταλλαγής φωτογραφιών, κλπ.



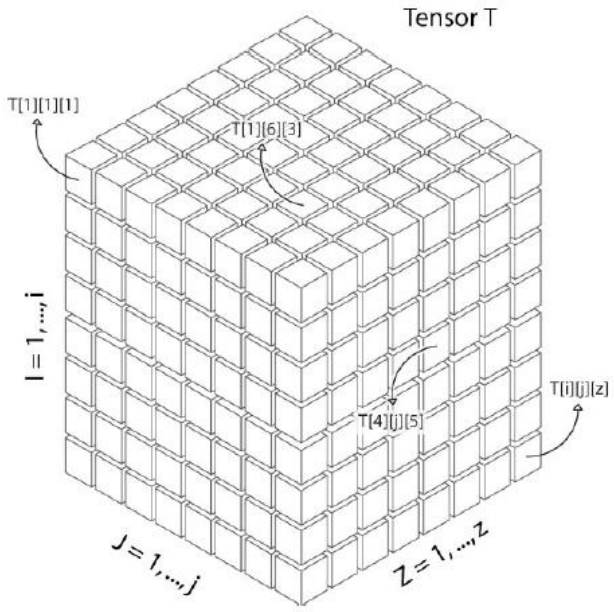
Σχήμα 5.5: Το κοινωνικό περιβάλλον των ενεργειών των χρηστών ποικίλλει μεταξύ των ιστοσελίδων κοινωνικής δικτύωσης - προτείνεται μία αναπαράσταση μεταγραφήματος για να μοντελοποιηθούν διάφορα κοινωνικά περιβάλλοντα. (a) οι κύριες ενέργειες και τα συσχετιζόμενα αντικείμενα στο Digg, (b) οι κύριες ενέργειες και τα συσχετιζόμενα αντικείμενα στο Flickr, (c) μία αναπαράσταση μεταγραφήματος για το Digg, (d) ένα μεταγράφημα για το Flickr.

5.3.1 Προκαταρκτικά στους τανυστές

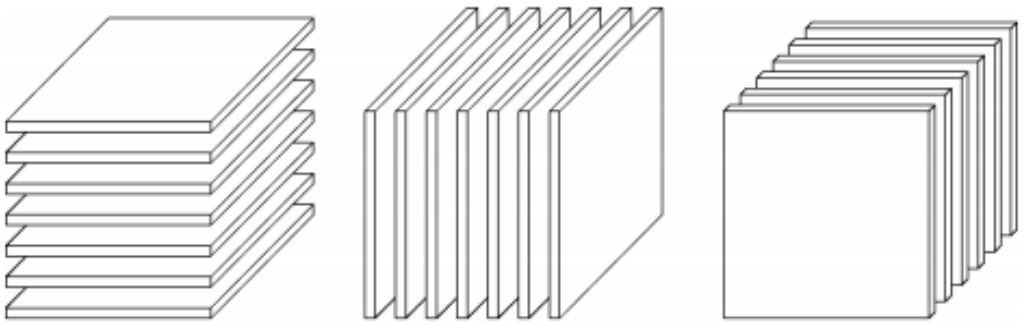
Αυτή η ενότητα παρέχει μερικές σημειώσεις και ένα ελάχιστο υπόβαθρο γύρω από τους **τανυστές (tensors).** Προτείνεται στους αναγνώστες να απευθυνθούν για περαιτέρω μελέτη στο άρθρο [28].

Ένας τανυστής αποτελεί μία μαθηματική αναπαράσταση ενός πολλών-τρόπων πίνακα. Δηλαδή, μπορεί να αναπαρασταθεί ως ένας πολυδιάστατος πίνακας αριθμητικών τιμών. Η *τάξη* (ή αλλιώς βαθμός) ενός τανυστή είναι ο αριθμός των διαστάσεων του πίνακα που απαιτούνται για την αναπαράστασή του, ή ισοδύναμα, ο αριθμός των δεικτών που απαιτούνται για να επισημανθεί ένα στοιχείο του εν λόγω πίνακα. Ένα διάνυσμα αποτελεί έναν πρώτης τάξης τανυστή, ένας πίνακας έναν δεύτερης τάξης, ενώ ένας ανώτερης τάξης τανυστής έχει τρεις ή περισσότερες διαστάσεις. Χρησιμοποιείται το **x** ως ένα διάνυσμα, το **X** σαν ένα πίνακας, και το $\mathcal{X}$

ως ένας τανυστή. Η *διαστατικότητα* ενός τρόπου είναι ο αριθμός των στοιχείων σε αυτόν τον τρόπο. Χρησιμοποιείται το I_q για να υποδηλωθεί η διάσταση του τρόπου q. Π.χ., ο τανυστής $\mathcal{X} \in R_+^{I_1 \times I_2 \times I_3}$ διαθέτει τρεις τρόπους με διαστατικότητες I_1, I_2, και I_3 αντίστοιχα.

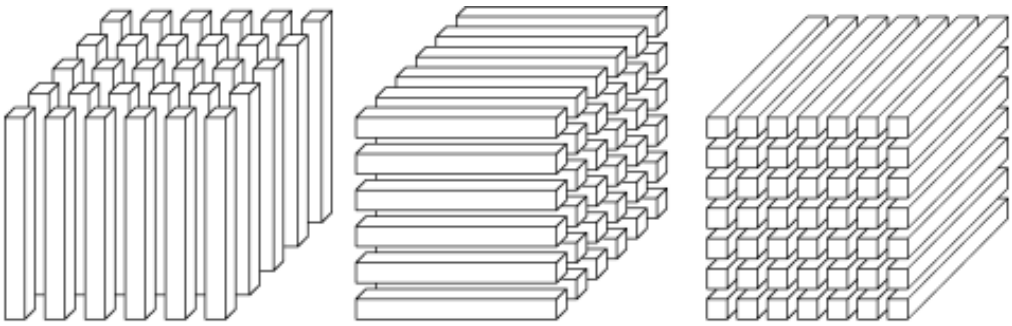


Σχήμα 5.6: Ένας τρίτης τάξης τανυστής T [28].



(α) (β) (γ)

Σχήμα 5.7: Τα κομμάτια ενός τρίτης τάξης τανυστή, (α) οριζόντια T(i , . , .), (β) πλευρικά T(. , j , .), (γ) μετωπικά T(. , . , z) [28].



(α) (β) (γ)

Σχήμα 5.8: Οι ίνες ενός τρίτης τάξης τανυστή, (α) στήλες T(. , j , z), (β) γραμμές T(i , . , z), (γ) σωλήνες (i , j , .) [28].

Το R_+ δείχνει ότι όλα τα στοιχεία του τανυστή $\mathcal{X}$ έχουν μη αρνητικές τιμές, το οποίο συνήθως ισχύει για ένα τανυστή δεδομένων. Το (i_1, i_2, i_3)-στοιχείο ενός τρίτης τάξης τανυστή συμβολίζεται με $x_{i_1 i_2 i_3}$. Οι δείκτες συνήθως κυμαίνονται από 1 έως την εκδοχή όπου συναντάται το κεφαλαίο γράμμα του, π.χ. $i_1 = 1, ..., I_1$.

Symbol	Description
$\mathbf{x}$	a vector (boldface lower-case letter)
$\mathbf{X}$	a matrix (boldface capital letter)
$\mathcal{X}$	a tensor (boldface Euler script letter)
$I_1,\ldots,I_M$	the dimensionality of mode 1, …, M
$v^{(q)}$	a vertex $v^{(q)} \in V$ represents the facet $v^{(q)}$
$e^{(r)}$	a hyperedge $e^{(r)} \subseteq V$ represents the relation $e^{(r)}$
V	the set of all facets $V=\{v^{(q)}\}$, or the set of all vertex indices
E	the set of all relations $E=\{e^{(r)}\}$ or all hyperedge indices
G	a metagraph $G=(V,E)$, where V is a set of facets/vertices and E is a set of relations/hyperedges
K, L	constants

Πίνακας 3: Περιγραφή συμβόλων

5.3.2 Διαμόρφωση του προβλήματος

Στην ενότητα αυτή ορίζεται το πρόβλημα του εντοπισμού της λανθάνουσας δομής της κοινότητας που αντιπροσωπεύει το περιβάλλον των ενεργειών των χρηστών σε κοινωνικά δίκτυα. Το πρόβλημα διαθέτει τρία μέρη: (1) πώς να αναπαρασταθούν πολυσχεσιακά κοινωνικά δεδομένα (ενότητα 5.3.2.1), (2) πώς να αποκαλυφθούν με συνέπεια οι λανθάνουσες κοινότητες δια μέσου πολλών σχέσεων, και (3) πώς μπορούν να παρακολουθούνται οι κοινότητες με την πάροδο του χρόνου (ενότητα 5.3.2.2).

5.3.2.1 Αναπαράσταση με μεταγράφημα

Γενικά, εισάγεται το *μεταγράφημα*, ένα σχεσιακό υπεργράφημα για την αναπαράσταση πολλαπλών σχεσιακών και πολυδιάστατων κοινωνικών δεδομένων. Χρησιμοποιείται ένα μεταγράφημα για να ρυθμιστεί το σχεσιακό περιβάλλον ειδικά για τα χαρακτηριστικά του συστήματος - αυτό είναι το κλειδί για να γίνει η ανάλυση της κοινότητας, προσαρμοσμένη στα διάφορα περιβάλλοντα κοινωνικών πολυμέσων, π.χ. Digg και Flickr (Σχήμα 5.5). Θα χρησιμοποιηθεί το παράδειγμα του Digg για να τονιστούν τρεις έννοιες: *έδρα* (facet), *σχέση* (relation), και *σχεσιακό υπεργράφημα* (relational hypergraph).

Όπως φαίνεται στο Σχήμα 5.5(a), το Digg επιτρέπει διάφορες ενέργειες για την ανταλλαγή ειδήσεων – οι χρήστες μπορούν να υποβάλουν (που υποδεικνύεται από την γραμμή με την ένδειξη «S») μία είδηση που σχετίζεται με ένα συγκεκριμένο θέμα. Θα μπορούσαν να ψηφίσουν (ή αλλιώς digg, γραμμή «D») ή να σχολιάσουν (γραμμή «C») πάνω στην υποβληθείσα ιστορία, να απαντήσουν (γραμμή «R») σε ένα σχόλιο που δημιουργήθηκε από άλλους χρήστες, ή ακόμα και να απαντήσουν σε μία απάντηση (δεν φαίνεται στο σχήμα), κλπ. Για να περιγραφεί το περιβάλλον των ενεργειών, ονομάζεται *έδρα* ένα σύνολο αντικειμένων ή οντοτήτων του ίδιου τύπου˙ μία έδρα χρήστης αποτελείται από ένα σύνολο χρηστών, μία έδρα ιστορίας αποτελείται από ένα σύνολο ιστοριών, κλπ. Τις αλληλεπιδράσεις μεταξύ των εδρών τις ονομάζουν *σχέση*. Μία σχέση μπορεί να περιλαμβάνει δύο (δηλαδή δυαδική σχέση) ή περισσότερες έδρες, π.χ. η σχέση «digg» περιλαμβάνει δύο έδρες (χρήστης, ιστορία), και η «make-comment» είναι μία 3-εδρών σχέση (χρήστης, ιστορία, σχόλιο). Μία έδρα μπορεί να είναι σιωπηρή, ανάλογα με το αν οι οντότητες των εδρών αλληλεπιδρούν με άλλες έδρες, π.χ. το σύνολο των digg αντικειμένων μπορεί να παραλειφθεί λόγω της μη αλληλεπίδρασης με άλλες έδρες.

Επισήμως, ορίζεται η q-οστή έδρα ως $v^{(q)}$ και το σύνολο όλων των εδρών ως V. Ένα σύνολο στιγμιοτύπων μιας M-τρόπων σχέσης e στις έδρες $v^{(1)}$, $v^{(2)}$,…, $v^{(M)}$ είναι ένα υποσύνολο του Καρτεσιανού γινομένου $v^{(1)} \times v^{(2)} \times \ldots v^{(M)}$. Μία ιδιαίτερη σχέση σημειώνεται με $e^{(r)}$ όπου το r είναι ο δείκτης της σχέσης. Οι παρατηρήσεις μιας M-τρόπων σχέσης $e^{(r)}$ παρουσιάζονται ως ένας M-τρόπων τανυστής δεδομένων $\mathcal{X}^{(r)}$.

Τώρα εισαγάγεται ένα πολυσχεσιακό υπεργράφημα (συμβολίζεται ως μεταγράφημα στην παρούσα εργασία) για να περιγράψει τον συνδυασμό των σχέσεων και των πλευρών σε ένα περιβάλλον κοινωνικών πολυμέσων. Ένα υπεργράφημα είναι ένα γράφημα όπου οι ακμές, που ονομάζονται υπερακμές, συνδέονται με ένα οποιοδήποτε αριθμό κορυφών. Η ιδέα είναι να χρησιμοποιηθεί μία M-υπερακμή για να αναπαρασταθούν οι αλληλεπιδράσεις των M εδρών: κάθε έδρα ως μία κορυφή και κάθε σχέση ως μία υπερακμή σε ένα υπεργράφημα. Ένα μεταγράφημα ορίζει μία συγκεκριμένη δομή των αλληλεπιδράσεων μεταξύ των εδρών, και όχι μεταξύ των στοιχείων των εδρών.

Επισήμως, για ένα σύνολο των εδρών V = { $v^{(q)}$ } και ένα σύνολο σχέσεων E = { $e^{(r)}$ }, κατασκευάζεται ένα μεταγράφημα της μορφής G = (V, E). Για να μειωθεί η πολυπλοκότητα των

συμβόλων, τα V και E αντιπροσωπεύουν επίσης το σύνολο όλων των δεικτών των κορυφών και των ακμών αντίστοιχα. Μία υπερακμή/σχέση $e^{(r)}$ θεωρείται ότι αποτελεί ένα *γεγονός* μιας έδρας/κορυφής $v^{(q)}$ εάν $v^{(q)} \in e^{(r)}$, το οποίο αναπαριστάται από το $v^{(q)} \sim e^{(r)}$ ή $e^{(r)} \sim v^{(q)}$. Π.χ. στο Σχήμα 5.5(c) το $v^{(1)}$ αντιπροσωπεύει την έδρα χρήστης, και το $e^{(5)} = \{v^{(1)}, v^{(2)}, v^{(3)}\}$ αντιπροσωπεύει τη σχέση «make-comment». Αυτά τα σύμβολα συνοψίζονται στον παραπάνω Πίνακα 4.

5.3.2.2 Εντοπισμός κοινότητας σε μεταγράφημα

Θα επισημοποιηθεί το πρόβλημα του εντοπισμού κοινότητας ως μία λανθάνουσα εξαγωγή χώρου από πολυσχεσιακά κοινωνικά δεδομένα που αντιπροσωπεύονται από ένα μεταγράφημα. Στόχος είναι να ανακαλυφθούν οι λανθάνουσες δομές της κοινότητας που αντιπροσωπεύουν το περιβάλλον των ενεργειών των χρηστών σε δίκτυα κοινωνικής δικτύωσης. Το ενδιαφέρον στρέφεται σε συστάδες ατόμων που αλληλεπιδρούν μεταξύ τους κατά τρόπο συνεκτικό. Μερικές από τις αλληλεπιδράσεις μπορεί να είναι έμμεσες, π.χ δύο χρήστες μπορούν να σχολιάσουν τις ίδιες ιστορίες, και οι αλληλεπιδράσεις μπορούν να ενισχυθούν περαιτέρω από άλλες αλληλεπιδράσεις. Ως εκ τούτου, μία κοινότητα θεωρείται ως ένας λανθάνων/κρυμμένος χώρος σταθερών αλληλεπιδράσεων ή σχέσεων μεταξύ των χρηστών και των αντικειμένων.

Με την παραδοχή συνεπών/σταθερών αλληλεπιδράσεων σε μία κοινότητα, η αλληλεπίδραση μεταξύ δύο οποιοδήποτε οντοτήτων (χρήστες ή αντικείμενα πολυμέσων) i και j σε μία κοινότητα k, που γράφεται ως x_{ij}, μπορεί να θεωρηθεί ως μία συνάρτηση των σχέσεων μεταξύ της κοινότητας k με την οντότητα i και της k με την j. Έστω [z] ένας υπερδιαγώνιος τανυστής, όπου το σύμβολο [·] μετασχηματίζει ένα διάνυσμα z σε ένα υπερδιαγώνιο τανυστή θέτοντας τα στοιχεία τανυστή $z_{k\ldots k} = z_k$ και τα άλλα στοιχεία ως 0. Επίσης, έστω $U_{(q)}$ ένας $I_q \times K$ πίνακας, όπου το I_q είναι το μέγεθος του $v^{(q)}$. Ο τανυστής πυρήνας [z] και οι παράγοντες $\{ U_{(q)} \}$ είναι, γενικά, υποχρεωμένοι να περιέχουν μη αρνητικές τιμές πιθανότητας (p_k είναι η πιθανότητα αλληλεπίδρασης στην k-οστή κοινότητα). Η μη αρνητική διάσπαση/παραγοντοποίηση τανυστή μπορεί να θεωρηθεί ως ανακάλυψη κοινότητας σε μία μόνο σχέση. Οι αλληλεπιδράσεις σε μέσα κοινωνικής δικτύωσης είναι πιο περίπλοκες - συνήθως περιλαμβάνουν πολλαπλές δύο- ή πολλών- τρόπων σχέσεις. Με τη χρήση διαφόρων μεταγραφημάτων, αντιπροσωπεύεται ένα ευρύ σύνολο από σχεσιακά περιβάλλοντα με την ίδια μορφή και ορίζεται το πρόβλημα εντοπισμού κοινότητας σε ένα μεταγράφημα, με τα ακόλουθα δύο τεχνικά ζητήματα.

Το πρώτο ζήτημα είναι το πώς να εξαχθεί η δομή της κοινότητας ως συνεκτικούς, αλληλεπιδραστικούς, κρυμμένους χώρους από τα παρατηρούμενα κοινωνικά δεδομένα που ορίζονται σε ένα μεταγράφημα. Αυτό ορίζεται ως εξής:

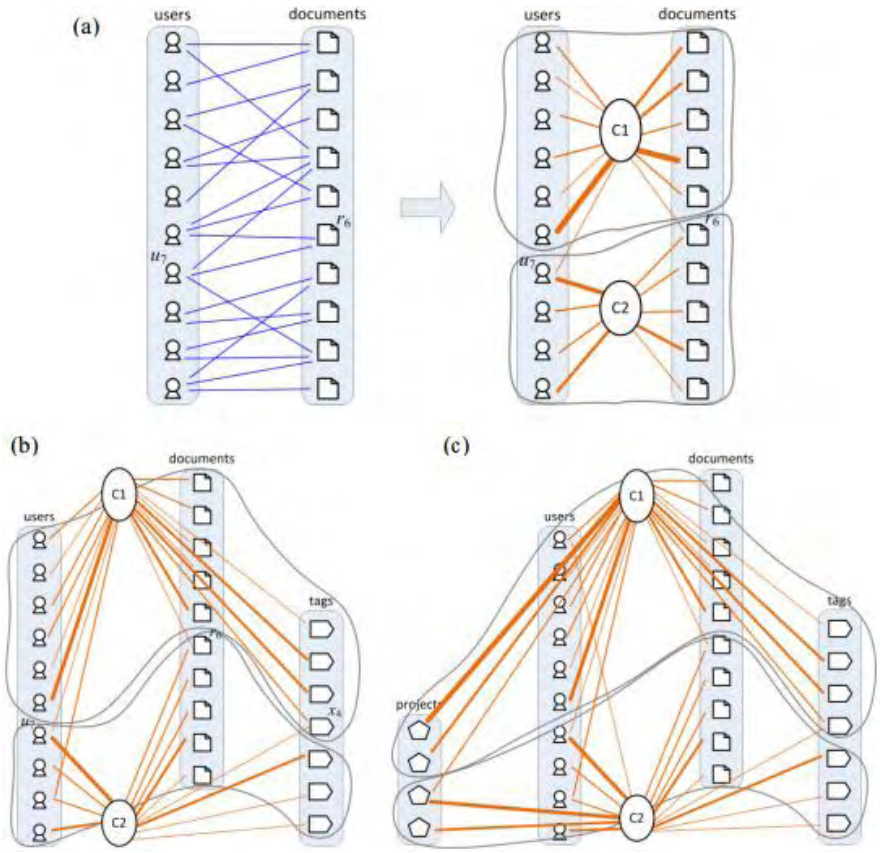
Πρόβλημα (Metagraph Factorization, ή MF): Δεδομένου ενός μεταγραφήματος G=(V, E) και ενός συνόλου παρατηρούμενων τανυστών δεδομένων $\{ \mathcal{X}^{(r)} \}$, r ∈ E, που ορίζεται στο G, να

βρεθεί ένας μη αρνητικός τανυστής πυρήνα [z], και παράγοντες { $U^{(q)}$ }, q ∈ V, για τις αντίστοιχες έδρες V = { $v^{(q)}$ } (από τη στιγμή που το E αντιπροσωπεύει επίσης το σύνολο όλων των δεικτών των ακμών, οι συμβολισμοί r ∈ E και $e^{(r)}$ ∈ E είναι ανταλλάξιμοι. Ομοίως, και οι q ∈ V και $v^{(q)}$ ∈ V είναι ανταλλάξιμοι).

Το δεύτερο ζήτημα αφορά τη δυναμική φύση των ανθρωπίνων δραστηριοτήτων - αυτές οι αλληλεπιδράσεις μπορεί να είναι συνεπείς κατά τη διάρκεια μιας σύντομης χρονικής περιόδου, αλλά είναι απίθανο να είναι συνεπείς όλη την ώρα. Το πρόβλημα, το πώς να εξαχθεί η δομή της κοινότητας ως κρυμμένους χώρους συνεκτικής αλληλεπίδρασης από τα χρονικά εξελισσόμενα δεδομένα, δεδομένου ενός μεταγραφήματος, ορίζεται ως εξής:

Πρόβλημα (MF για χρονικά εξελισσόμενα δεδομένα, ή MFT): Δεδομένου ενός μεταγραφήματος G=(V, E) και ενός διαδοχικού συνόλου παρατηρούμενων τανυστών δεδομένων { $\mathcal{X}_t^{(r)}$ }, r ∈ E, που ορίζεται στο G τη χρονική στιγμή t=1,2,…, βρείτε έναν μη αρνητικό τανυστή πυρήνα [z_t] και παράγοντες { $U_t^{(q)}$ }, q ∈ V, για τις αντίστοιχες έδρες V = { $v^{(q)}$ } για κάθε χρονική στιγμή t.

Λόγω του γεγονότος ότι ο MetaFac αλγόριθμος περιλαμβάνει πολλές και πολύπλοκες μαθηματικές εκφράσεις, για την περαιτέρω ανάγνωση του αλγορίθμου σας παραπέμπουμε στα παρακάτω άρθρα: [39,40].



Σχήμα 5.9: (a) Μία εικόνα του πώς δύο κοινότητες έχουν συνεπείς/σταθερές αλληλεπιδράσεις σε ένα δίκτυο χρήστη (user)-εγγράφου (document), π.χ. η αλληλεπίδραση μεταξύ του χρήστη u_7 και του

εγγράφου r_6 φαίνεται και από τη σχέση τους με τις κοινότητες C_1 και C_2. (b) Ένα παράδειγμα για το πώς δύο κοινότητες έχουν μία τριών-τρόπων αλληλεπίδραση μεταξύ των χρηστών (users), των εγγράφων (documents), και των ετικετών (tags). (c) Εντοπισμός δύο κοινοτήτων για μία τεσσάρων-τρόπων αλληλεπίδραση μεταξύ των χρηστών (users), των εγγράφων (documents), των ετικετών (tags), και των εργασιών (projects) [39].

Η χρονική πολυπλοκότητα του αλγορίθμου είναι O(mnD).

5.4 Παραλλαγμένος Bayes (Variational Bayes)

Μεγάλης κλίμακας δίκτυα που περιγράφουν πολύπλοκες αλληλεπιδράσεις ανάμεσα σε μία πληθώρα αντικειμένων, έχουν βρει εφαρμογή σε ένα ευρύ φάσμα τομέων, από τη βιολογία μέχρι την τεχνολογία της πληροφορίας. Σε αυτές τις εφαρμογές κάποιος συχνά επιθυμεί να μοντελοποιεί δίκτυα, καταστέλλοντας την πολυπλοκότητα της πλήρους περιγραφής, διατηρώντας σχετικές πληροφορίες σχετικά με τη δομή των αλληλεπιδράσεων. Ένα τέτοιο μοντέλο δικτύου ομαδοποιεί τους κόμβους σε ενότητες, ή «κοινότητες», με διαφορετικές πυκνότητες ενδο-και δια-συνδεσιμότητας για τους κόμβους στις ίδιες ή σε διαφορετικές ενότητες. Στην ενότητα αυτή παρουσιάζεται ένα υπολογιστικά αποδοτικό Bayesian πλαίσιο εργασίας για την συναγωγή του αριθμού των ενοτήτων, των παραμέτρων του μοντέλου, και των αναθέσεων των ενοτήτων για ένα τέτοιο μοντέλο.

Το πρόβλημα της εύρεσης ενοτήτων σε δίκτυα (ή «εντοπισμού κοινότητας») έχει λάβει μεγάλη προσοχή σε προβλήματα της φυσικής, όπου πολλές προσεγγίσεις [29] εστιάζονται στη βελτιστοποίηση μιας συνάρτησης κόστους που βασίζεται στην ενέργεια με σταθερές παραμέτρους στις πιθανές αναθέσεις των κόμβων σε κοινότητες. Οι συγκεκριμένες συναρτήσεις κόστους ποικίλουν, αλλά οι περισσότερες συγκρίνουν μία συγκεκριμένη διαμέριση κόμβων με ένα απεριόριστο μοντέλο˙ τα δύο πιο δημοφιλή είναι το μοντέλο διαμόρφωσης και μια περιορισμένη έκδοση του στοχαστικού μπλοκ μοντέλου (SBM)[30]. Αν και πολύ προσπάθεια έχει γίνει για το πώς να βελτιστοποιηθούν αυτές οι λειτουργίες κόστους, έχει δοθεί λιγότερη προσοχή στο τι πρέπει να βελτιστοποιηθεί. Επεκτείνονται πρόσφατες πιθανοτικές επεξεργασίες των συνδυαζόμενων δικτύων [31] για να αναπτυχθεί μία λύση στο πρόβλημα που στηρίζεται στη συναγωγή κατανομών επί των παραμέτρων του μοντέλου, σε αντίθεση με τις a priori υποστηριζόμενες τιμές των παραμέτρων, για τον προσδιορισμό της αρθρωτής δομής ενός δεδομένου δικτύου. Οι ανεπτυγμένες τεχνικές είναι, βάσει αρχών, ερμηνεύσιμες, υπολογιστικά αποδοτικές, και αποδεδειγμένα γενικεύουν αρκετές προηγούμενες μελέτες στον εντοπισμό κοινότητας.

Καθορίζεται ένα N-κόμβων δίκτυο με έναν πίνακα γειτνίασης A, όπου A_{ij} = 1 αν υπάρχει ακμή μεταξύ των κόμβων i και j και A_{ij} = 0 αλλιώς, και ορίζεται το $\sigma_i \in \{1, \dots, K\}$ να αποτελεί την αόρατη συμμετοχή του i-οστού κόμβου στην κοινότητα. Χρησιμοποιείται ένα περιορισμένο

SBM, το οποίο αποτελείται από μία multinomial κατανομή πάνω από αναθέσεις σε ενότητα με βάρη $\pi_\mu \equiv p(\sigma_i = \mu \mid \underset{\pi}{\rightarrow})$ και Bernoulli κατανομές πάνω από ακμές που περιέχονται εντός και μεταξύ των ενοτήτων με βάρη $\theta_c \equiv p\left(A_{ij} = 1 \mid \sigma_i = \sigma_j, \underset{\theta}{\rightarrow}\right)$ και $\theta_d \equiv p\left(A_{ij} = 1 \mid \sigma_i \neq \sigma_j, \underset{\theta}{\rightarrow}\right)$, αντιστοίχως. Εν ολίγοις, για να παραχθεί ένα τυχαίο μη-κατευθυνόμενο γράφημα στο πλαίσιο αυτού του μοντέλου «κυλιέται» ένας K-πλευρών κύβος (ωθείται από $\underset{\pi}{\rightarrow}$) N φορές για τον προσδιορισμό των αναθέσεων στις ενότητες για κάθε έναν από τους N κόμβους· τότε αναστρέφεται το ένα από τα δύο μεροληπτικά κέρματα (είτε για ένδο-ή δια-σύνδεση ενότητας, που ωθείται από τα θ_c, θ_d, αντίστοιχα) για καθένα από τα N(N - 1)/2 ζεύγη κόμβων για να προσδιοριστεί εάν το ζεύγος είναι συνδεδεμένο. Η επέκταση σε κατευθυνόμενα γραφήματα είναι απλή.

Χρησιμοποιώντας αυτό το μοντέλο, οι συγγραφείς γράφουν την από κοινού πιθανότητα $p\left(A, \underset{\sigma}{\rightarrow} \middle| \underset{\pi}{\rightarrow}, \underset{\theta}{\rightarrow}, K\right) = p(A| \underset{\sigma}{\rightarrow}, \underset{\theta}{\rightarrow}) p(\underset{\sigma}{\rightarrow} | \underset{\pi}{\rightarrow})$ ως

$$p\left(A, \underset{\sigma}{\rightarrow} \middle| \underset{\pi}{\rightarrow}, \underset{\theta}{\rightarrow}\right) = \theta_c^{c_+}(1-\theta_c)^{c_-}\theta_d^{d_+}(1-\theta_d)^{d_-}\prod_{\mu=1}^{K}\pi_\mu^{n_\mu} \quad (1)$$

όπου $c_+ \equiv \sum_{i>j} A_{ij}\delta_{\sigma_i,\sigma_j}$ είναι ο αριθμός των ακμών που περιέχονται στις κοινότητες,

$c_- \equiv \sum_{i>j}(1-A_{ij})\delta_{\sigma_i,\sigma_j}$ είναι ο αριθμός των ακμών που δεν περιέχονται στις κοινότητες,

$d_+ \equiv \sum_{i>j} A_{ij}(1-\delta_{\sigma_i,\sigma_j})$ είναι ο αριθμός των ακμών μεταξύ διαφορετικών κοινοτήτων,

$d_- \equiv \sum_{i>j}(1-A_{ij})(1-\delta_{\sigma_i,\sigma_j})$ είναι ο αριθμός των ακμών που δεν υπάρχουν μεταξύ διαφορετικών κοινοτήτων, και

$n_\mu \equiv \sum_{i=1}^{N}\delta_{\sigma_i,\mu}$ είναι ο αριθμός «κατοχής» της μ$^{\text{th}}$ ενότητας. Ορίζοντας $H \equiv -\ln p\left(A, \underset{\sigma}{\rightarrow} \middle| \underset{\pi}{\rightarrow}, \underset{\theta}{\rightarrow}\right)$ και ομαδοποιώντας ξανά τους όρους, ανακτείται (μέχρι και πρόσθετες σταθερές) μια γενικευμένη εκδοχή:

$H = -\sum_{i>j}\left(J_L A_{ij} - J_G\right)\delta_{\sigma_i,\sigma_j} + \sum_{\mu=1}^{K} h_\mu \sum_{i=1}^{N}\delta_{\sigma_i,\mu}$ (2) , όπου $J_G = \ln(1-\theta_d)/(1-\theta_c)$ και $J_L = \ln\frac{\theta_c}{\theta_d} + J_G$,και $h_\mu = -\ln\pi_\mu$. Η συγκεκριμένη προσέγγιση χρησιμοποιεί έναν μέσης-διαταραχής υπολογισμό ώστε να βγει ένα συμπέρασμα για τις κατανομές πάνω από αυτές τις παραμέτρους. Για να γίνει αυτό, παίρνονται οι Beta (B) και Dirichlet (D) κατανομές πάνω από τα $\underset{\theta}{\rightarrow}$ και $\underset{\pi}{\rightarrow}$ αντίστοιχα:

$$p\left(\underset{\theta}{\rightarrow}\right)\left(\underset{\pi}{\rightarrow}\right) \equiv B(\theta_c; c_{+0}, c_{-0})B(\theta_d; d_{+0}, d_{-0})D(\underset{\pi}{\rightarrow}; \underset{n_0}{\rightarrow}). \quad (3)$$

Αυτές οι εκ των προτέρων κατανομές σύζευξης (conjugate prior), ορίζονται στο πλήρες φάσμα των $\underset{\theta}{\rightarrow}$ και $\underset{\pi}{\rightarrow}$, αντίστοιχα, και οι λειτουργικές μορφές τους διατηρούνται όταν ενσωματώνονται ενάντια στο μοντέλο για τη λήψη ανανεωμένων παραμέτρων κατανομών. Οι υπερπαράμετροί

τους $\{c_{+0}, c_{-0}, d_{+0}, d_{-0}, \underset{n_0}{\rightarrow}\}$ λειτουργούν ως μία ψευδο-αρίθμηση που αυξάνει τις αριθμήσεις των παρατηρούμενων ακμών και των αριθμών κατοχής.

Σε αυτό το πλαίσιο εργασίας, το πρόβλημα του εντοπισμού ενότητας (κοινότητας) μπορεί να διατυπωθεί ως εξής:

Δεδομένου ενός πίνακα γειτνίασης A, προσδιορίστε τον πιθανότερο αριθμό των ενοτήτων (δηλαδή τις κατεχόμενες spin καταστάσεις) $\mathrm{K}^* = \mathrm{argmax}_{\mathrm{K}}\, \mathrm{p}(\mathrm{K}|\mathrm{A})$ και συμπεράνετε μεταγενέστερες κατανομές κατά τη διάρκεια των παραμέτρων του μοντέλου (δηλαδή σταθερές σύζευξης και χημικά δυναμικά) $p(\underset{\pi}{\rightarrow}, \underset{\theta}{\rightarrow}|A)$, καθώς επίσης και τις λανθάνουσες αναθέσεις ενότητας (δηλαδή τις spin καταστάσεις) $p(\underset{\sigma}{\rightarrow}, |A)$. Ελλείψει μιας εκ των προτέρων πεποίθησης για τον αριθμό των ενοτήτων, απαιτείται από την $p(K)$ να είναι επαρκώς αδύνατη έτσι ώστε η μεγιστοποίηση της $p(K) \propto p(A|K)p(K)$ να είναι ισοδύναμη με την μεγιστοποίηση της $p(A|K)$.

Μια πιο φυσική διαισθητική ερμηνεία των αποδείξεων είναι η μέσης-διαταραχής συνάρτηση διαμέρισης ενός spin-glass, η οποία υπολογίζεται περιθωριοποιώντας πάνω από τις πιθανές αποσβεσθέντες τιμές των παραμέτρων $\underset{\theta}{\rightarrow}$ και $\underset{\pi}{\rightarrow}$ καθώς και τις spin ρυθμίσεις $\underset{\sigma}{\rightarrow}$:

$$Z = p(A|K) = \sum_{\vec{\sigma}} \int d\underset{\theta}{\rightarrow} \int d\underset{\pi}{\rightarrow} p\left(A, \underset{\sigma}{\rightarrow} \middle| \underset{\pi}{\rightarrow}, \underset{\theta}{\rightarrow}\right) p\left(\underset{\theta}{\rightarrow}\right) p\left(\underset{\pi}{\rightarrow}\right) =$$
$$\sum_{\vec{\sigma}} \int d\underset{\theta}{\rightarrow} \int d\underset{\pi}{\rightarrow} e^{-H} p\left(\underset{\theta}{\rightarrow}\right) p\left(\underset{\pi}{\rightarrow}\right) \text{(4), (5)}$$

Για να διευκολυνθούν μεγάλης κλίμακας δίκτυα χρησιμοποιείται μία παραλλαγμένη προσέγγιση που είναι γνωστή στη στατιστική κοινότητα της φυσικής [32] και έχει βρει πρόσφατα εφαρμογή στη στατιστική και μηχανική μάθηση˙ κοινώς ονομάζεται **Variational Bayes (VB)** [33,42,43]. Οι συγγραφείς συνεχίζουν την προσέγγιση παίρνοντας τον αρνητικό λογάριθμο του Z, και χρησιμοποιώντας την ανισότητα του Gibbs προκύπτει το εξής:

$$-\ln Z = -\ln \sum_{\vec{\sigma}} \int d\underset{\theta}{\rightarrow} \int d\underset{\pi}{\rightarrow} q(\underset{\sigma}{\rightarrow}, \underset{\pi}{\rightarrow}, \underset{\theta}{\rightarrow}) \frac{p(A, \vec{\sigma}, \vec{\pi}, \vec{\theta}|K)}{q(\vec{\sigma}, \vec{\pi}, \vec{\theta})} \text{ (6)}$$

$$\leq -\sum_{\vec{\sigma}} \int d\underset{\theta}{\rightarrow} \int d\underset{\pi}{\rightarrow} q\left(\underset{\sigma}{\rightarrow}, \underset{\pi}{\rightarrow}, \underset{\theta}{\rightarrow}\right) \ln \frac{p(A, \vec{\sigma}, \vec{\pi}, \vec{\theta}|K)}{q(\vec{\sigma}, \vec{\pi}, \vec{\theta})} \text{ (7).}$$

Δηλαδή, πρέπει πρώτα να πολλαπλασιάσουν και να διαιρέσουν με μία αυθαίρετη προσεγγιστική κατανομή $q(\underset{\sigma}{\rightarrow}, \underset{\pi}{\rightarrow}, \underset{\theta}{\rightarrow})$ και έπειτα να βάλουν άνω όρια στον expectation log. Ορίζουν την ποσότητα να ελαχιστοποιείται, η έκφραση στην ανισότητα 7, όπως την μεταβολική ενέργεια $F\{q; A\}$, ένα λειτουργικό της $q(\underset{\sigma}{\rightarrow}, \underset{\pi}{\rightarrow}, \underset{\theta}{\rightarrow})$.

Έπειτα, επιλέγεται μία παραγοντοποιημένη προσεγγιστική κατανομή $q\left(\underset{\sigma}{\rightarrow}, \underset{\pi}{\rightarrow}, \underset{\theta}{\rightarrow}\right) = q_{\vec{\sigma}}(\underset{\sigma}{\rightarrow}) q_{\vec{\pi}}(\underset{\pi}{\rightarrow}) q_{\vec{\theta}}(\underset{\theta}{\rightarrow})$ με $q_{\vec{\pi}}\left(\underset{\pi}{\rightarrow}\right) = D(\underset{\pi}{\rightarrow}; \underset{n}{\rightarrow})$ και $q_{\vec{\theta}}\left(\underset{\theta}{\rightarrow}\right) = q_c(\theta_c) q_d(\theta_d) = B(\theta_c; c_+, c_-) B(\theta_d; d_+, d_-)$˙ όπως στο πεδίο της μέσης τιμής

παραγοντοποιείται η $q_{\vec{\sigma}}(\vec{\sigma})$ ως $q(\sigma_i = \mu) = Q_{i\mu}$, ένας N- με- K πίνακας ο οποίος δίνει την πιθανότητα ότι ο i-οστός κόμβος ανήκει στην μ-οστή ενότητα. Αξιολογώντας την $F\{q; A\}$ με αυτή τη λειτουργική μορφή για την $q(\vec{\sigma}, \vec{\pi}, \vec{\theta})$, δίνει μια συνάρτηση των μεταβολικών παραμέτρων $\{c_+, c_-, d_+, d_-, \vec{n}\}$ και τα στοιχεία του πίνακα $Q_{i\mu}$, τα οποία μπορούν ακολούθως να ελαχιστοποιούνται με τη λήψη των κατάλληλων παραγώγων.

Συνοψίζεται ο προκύπτον *επαναληπτικός αλγόριθμος*, ο οποίος αποδεδειγμένα συγκλίνει σε ένα τοπικό ελάχιστο της $F\{q; A\}$ και παρέχει ελεγχόμενες προσεγγίσεις στην απόδειξη $p(A|K)$ καθώς και στις μεταγενέστερες $p(\vec{\pi}, \vec{\theta}|A)$ και $p(\vec{\sigma}, |A)$:

Αρχικοποίηση. Αρχικοποιήστε τον N- με- K πίνακα Q = Q_0 και ορίστε $c_+ = c_{+0}$, $c_- = c_{-0}$, $d_+ = d_{+0}$, $d_- = d_{-0}$ και $n_\mu = n_{\mu 0}$.

Κύριος βρόχος. Μέχρι τη σύγκλιση στην $F\{q; A\}$:

(1) Ενημερώστε την αναμενόμενη τιμή των σταθερών σύζευξης και χημικών δυναμικών
$J_L = \psi(c_+) - \psi(c_-) - \psi(d_+) + \psi(d_-)$ (8)
$J_G = \psi(d_-) - \psi(d_+ + d_-) - \psi(c_-) + \psi(c_+ + c_-)$ (9)
$h_\mu = \psi\left(\sum_\mu n_\mu\right) - \psi(n_\mu)$, (10)
όπου *ψ(x)* είναι η συνάρτηση δίγαμμα

(2) Ενημερώστε την μεταβολική κατανομή για κάθε spin σ_i
$Q_{i\mu} \propto \exp\{\sum_{j\neq i}[J_L A_{ij} - J_G]\, Q_{i\mu} - h_\mu\}$ (11) , κανονικοποιημένα έτσι ώστε $\sum_\mu Q_{i\mu} = 1$,για όλα τα i

(3) Ενημερώστε την μεταβολική κατανομή πάνω από τις παραμέτρους $n_\mu = n_\mu + n_{\mu 0} = \sum_{i=1} Q_{i\mu} + n_{\mu 0}$ (12) , με το i να πηγαίνει μέχρι το N,
$c_+ = c_+ + c_{+0} = \frac{1}{2} Tr(Q^T A Q) + c_{+0}$ (13)
$c_- = c_- + c_{-0} = \frac{1}{2} Tr(Q^T (un^T - Q)) - c_+ + c_{-0}$ (14)
$d_+ = d_+ + d_{+0} = M - c_+ + d_{+0}$ (15)
$d_- = d_- + d_{-0} = C - M - c_- + d_{-0}$, (16)
όπου $C = N(N-1)/2$, $M = \sum_{i>j} A_{ij}$, και u είναι ένα N- με -1 διάνυσμα των άσσων («1»)

(4) Υπολογίστε την ενημερωμένη βελτιστοποιημένη free ενέργεια $F\{q; A\} = -\ln \frac{Z_c Z_d Z_\pi}{\overrightarrow{Z_c Z_d Z_\pi}} + \sum_{\mu=1}^{K} \sum_{i=1}^{N} Q_{i\mu} \ln Q_{i\mu}$, (17)
όπου $Z_\pi = B(\vec{n})$ είναι η beta συνάρτηση.

Δεδομένου ότι αυτό αποδεδειγμένα συγκλίνει σε ένα τοπικό βέλτιστο, ο αλγόριθμος VB είναι καλύτερα υλοποιήσιμος με πολλαπλές τυχαία επιλεγμένες αρχικοποιήσεις του Q_0 μέχρις ότου να βρεθεί το ολικό ελάχιστο της $F\{q; A\}$.

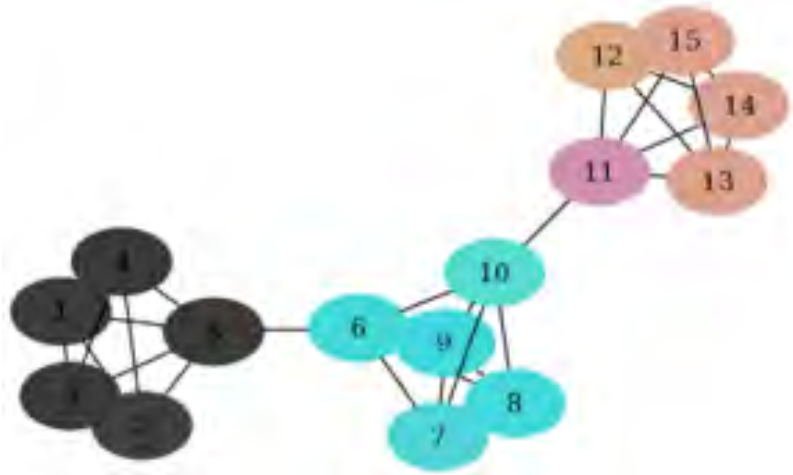
Αυτή η μέθοδος είναι πολύ πιο ακριβής από ό, τι άλλες προσεγγιστικές μέθοδοι, όπως είναι η Bayesian Information Criterion (BIC) [34,41] και η Integrated Classification Likelihood (ICL)[44,45], και υπολογιστικά είναι λιγότερο ακριβή από εμπειρικές μεθόδους όπως η cross-validation (CV)[46,47], στην οποία κάποιος πρέπει να εκτελέσει τη σχετική διαδικασία μετά την τοποθέτηση του μοντέλου για κάθε εξεταζόμενη τιμή του K.

Στο Σχήμα 5.10 παρουσιάζεται ένα γραφικό παράδειγμα εκτέλεσης του αλγορίθμου.

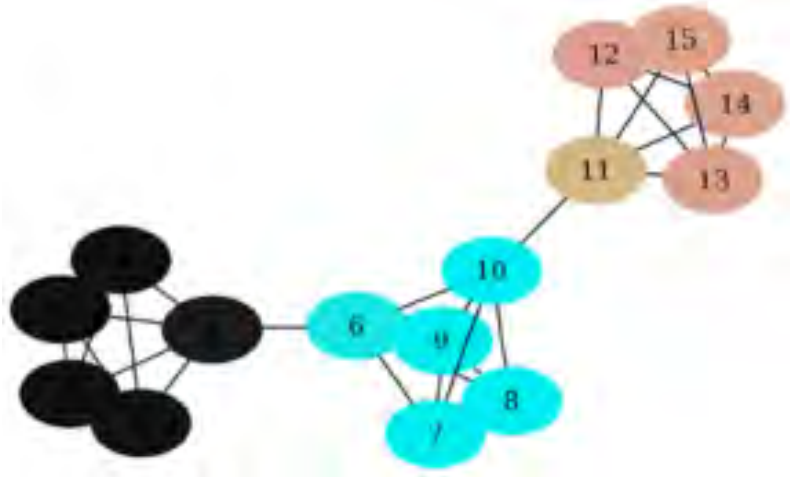
α) 1η επανάληψη

β) 2η επανάληψη

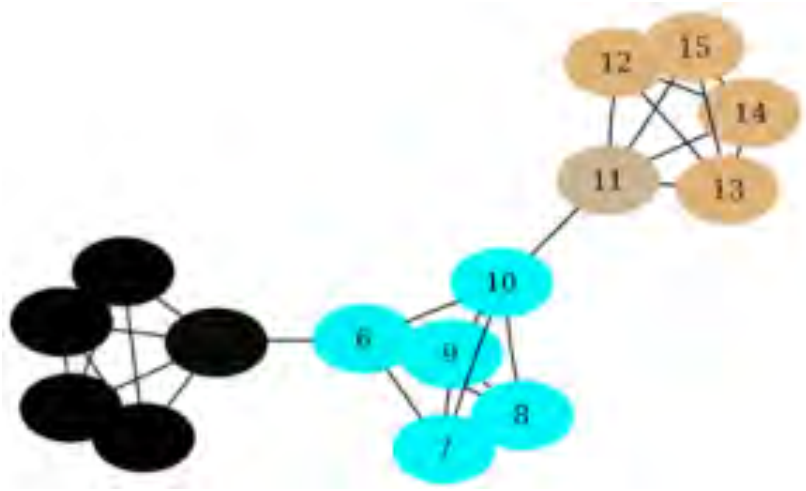
γ) 3η επανάληψη



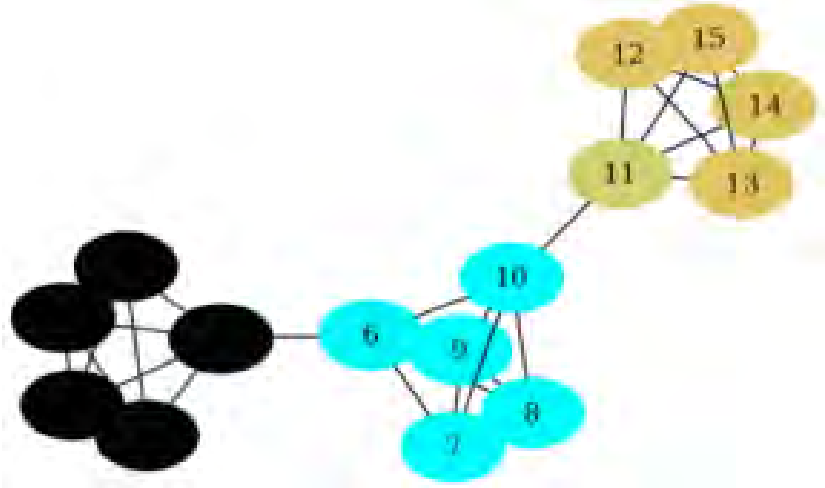
δ) 4η επανάληψη



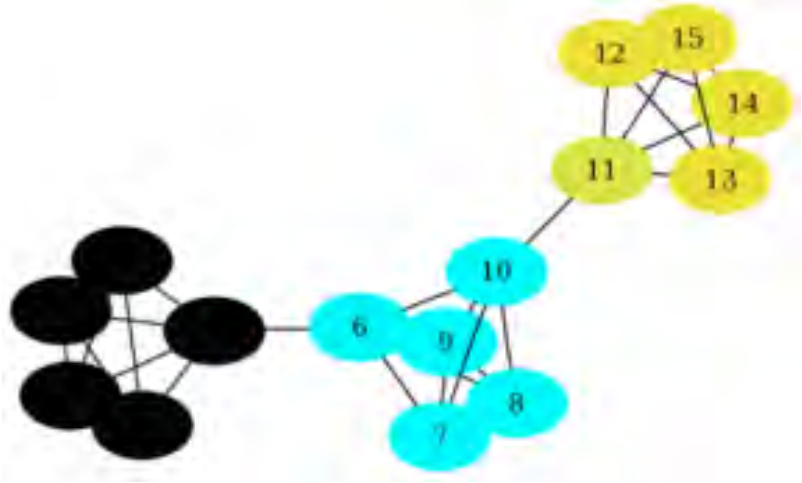
ε) 5η επανάληψη



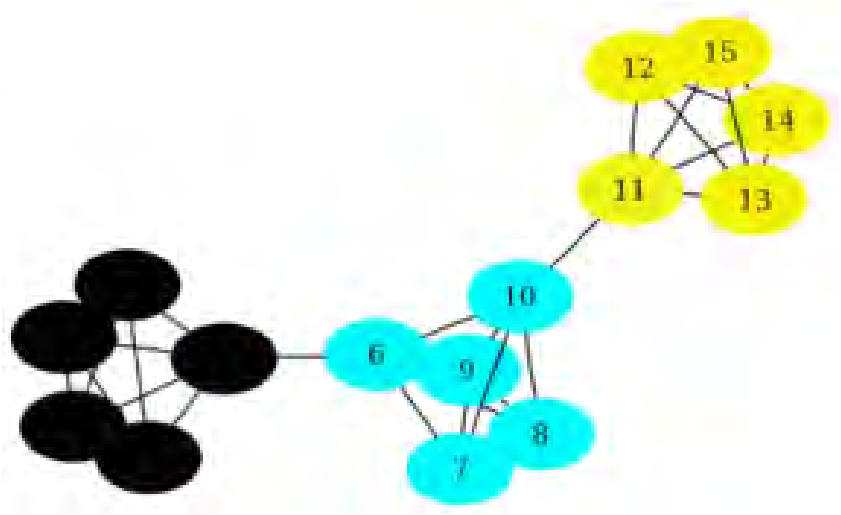
στ) 6η επανάληψη



ζ) 7η επανάληψη



η) 8η επανάληψη



θ) 9η επανάληψη

Σχήμα 5.10: Παράδειγμα της Bayesian προσέγγισης σε 9 επαναλήψεις. Παραγωγικό μοντέλο: die rolling για τις αναθέσεις σε κοινότητες, και coin-flipping για τις ακμές [50].

5.5 LA→ IS²

Τα άτομα (φορείς) σε μία κοινωνική κοινότητα τείνουν να σχηματίζουν ομάδες και ενώσεις που αντικατοπτρίζουν τα ενδιαφέροντά τους. Είναι σύνηθες για τους φορείς να ανήκουν σε διάφορες τέτοιες ομάδες. Οι ομάδες μπορεί να είναι, ή μπορεί να μην είναι, γνωστές στο ευρύ κοινωνικό σύνολο. Μερικές ομάδες είναι στην πραγματικότητα «κρυμμένες», εσκεμμένα ή όχι, στον επικοινωνιακό μικρόκοσμο. Μία ομάδα που προσπαθεί σκόπιμα να κρύψει την επικοινωνιακή της συμπεριφορά στο πλήθος των παρασκηνιακών επικοινωνιών, μπορεί να σχεδιάζει κάποια ανεπιθύμητη ή ενδεχομένως ακόμα και κακόβουλη δραστηριότητα. Είναι σημαντικό να ανακαλύπτονται τέτοιες ομάδες πριν ακόμα επιχειρήσουν την ανεπιθύμητη δραστηριότητά τους.

Στην ενότητα αυτή παρουσιάζεται ένας νέος αποδοτικός αλγόριθμος για την εύρεση επικαλυπτόμενων κοινοτήτων που θα έχουν ως δεδομένα το ποιος επικοινώνησε με ποιον. Πρωταρχικό κίνητρο για την εύρεση τέτοιων κοινωνικών κοινοτήτων είναι να χρησιμοποιηθούν αυτές οι πληροφορίες ώστε να φιλτραριστούν περαιτέρω οι κρυμμένες κακόβουλες ομάδες με βάση κάποια άλλα κριτήρια˙ δείτε το [35]. Ωστόσο, η γνώση όλων των κοινοτήτων μπορεί να είναι ζωτικής σημασίας για την ανάλυση της κοινωνικής συμπεριφοράς και της εξέλιξης της κοινωνίας στο σύνολό της, καθώς επίσης και τα μεμονωμένα μέλη αυτής [36].

Η μαθηματική διατύπωση του προβλήματος του καθορισμού των συστάδων, που μπορεί ενδεχομένως να είναι επικαλυπτόμενες, παρουσιάζεται στο παρακάτω άρθρο [37], η οποία

ορίζει μία συστάδα ως ένα τοπικά βέλτιστο υπογράφημα σε σχέση με έναν δεδομένο κανόνα. Αυτή η έννοια της συστάδας εξετάζεται εν συντομία στην ενότητα 5.5.1. Δεδομένου ότι τοπικά βέλτιστα υπογραφήματα μπορεί να επικαλύπτονται, η διατύπωση αυτή επιτρέπει την ύπαρξη επικαλυπτόμενων συστάδων. Ο αλγόριθμος για την εύρεση τέτοιων τοπικά βέλτιστων υπογράφων αποτελείται από δύο μέρη: την **αρχικοποίηση**, RaRe, η οποία δημιουργεί συστάδες «σπόρου», ή αλλιώς τροφοδοτούμενες συστάδες, και την **βελτίωση**, IS, η οποία σαρώνει επανειλημμένα τις κορυφές προκειμένου να βελτιώσει τις τρέχουσες συστάδες μέχρι κάποια να φθάσει σε μια τοπικά βέλτιστη συλλογή των συστάδων.

Το επίκεντρο αυτής της ενότητας είναι να παράσχει ένα νέο αποδοτικό αλγόριθμο για την αρχικοποίηση των συστάδων σπόρων και την εκτέλεση των επαναληπτικών βελτιώσεων. Συγκεκριμένα, στην παρούσα ενότητα, παρουσιάζεται μία διαδικασία που λέγεται **Link Aggregate (LA)** για την αρχικοποίηση των συστάδων, και η διαδικασία **IS^2** που βελτιώνει επαναληπτικά οποιοδήποτε δεδομένο σύνολο των συστάδων. Και στις δύο διαδικασίες ως είσοδος θεωρείται ένα μέτρο πυκνότητας που συμβολίζεται με W, και για το οποίο γίνεται προσπάθεια ώστε να βελτιστοποιηθεί. Στην ουσία το W δεν είναι τίποτε άλλο παρά μόνο ο *μέσος βαθμός* W_{ad} που ορίζεται για ένα σύνολο κόμβων της συστάδας C ως εξής: $\mathrm{W_{ad}(C)} = \frac{2|\mathrm{E(C)}|}{|\mathrm{C}|}$, όπου το E(C) είναι το σύνολο των ακμών που έχουν και τα δύο ακριανά τους σημεία εντός της συστάδας C. Ο συνδυασμένος αλγόριθμος αναπτύσσει επικαλυπτόμενους υπογράφους σε ένα γενικό γράφημα. Αξίζει να σημειωθεί ότι ο συγκεκριμένος αλγόριθμος μπορεί να εφαρμοστεί σε μεγάλα (~ 10^6) δίκτυα κόμβων.

5.5.1 Συστάδες

Στην παρούσα ενότητα υιοθετείται η ιδέα που διατυπώθηκε στο άρθρο [35] όπου μία ομάδα φορέων C σε ένα κοινωνικό δίκτυο σχηματίζει μία κοινότητα, εάν η συνάρτηση «πυκνότητας» της επικοινωνίας επιτυγχάνει ένα τοπικό μέγιστο στη συλλογή των ομάδων που είναι «κοντά» στην C. Δύο ομάδες θεωρούνται κοντά, εάν γίνουν πανομοιότυπες με την αλλαγή της συμμετοχής ενός μόνο φορέα.

Έτσι, μία ομάδα αποτελεί μία κοινότητα αν η προσθήκη κάθε νέου μέλους στην ομάδα, ή η αφαίρεση οποιουδήποτε τρέχοντος μέλους από αυτήν, μειώνει τον μέσο όρο των ανταλλαγών της επικοινωνίας. Συστάδα ονομάζεται το αντίστοιχο υπογράφημα του γραφήματος που αναπαριστά τις επικοινωνίες στο κοινωνικό δίκτυο.

Η συσταδοποίηση είναι μία σημαντική τεχνική για την ανάλυση των δεδομένων με μία ποικιλία εφαρμογών σε τομείς όπως η εξόρυξη δεδομένων, η βιοπληροφορική, και οι κοινωνικές επιστήμες. Παραδοσιακά, βλέπε για παράδειγμα το άρθρο [38], η συσταδοποίηση γίνεται κατανοητή ως μία κατάτμηση των δεδομένων σε υποσύνολα. Ο περιορισμός αυτός είναι πάρα πολύ σοβαρός και ταυτόχρονα περιττός στην περίπτωση των κοινοτήτων που λειτουργούν

σε ένα κοινωνικό δίκτυο. Ο ορισμός, όπως έχει διατυπωθεί προηγουμένως, επιτρέπει τον ίδιο φορέα να είναι μέλος διάφορων συστάδων. Εξάλλου, ο αλγόριθμος είναι σχεδιασμένος να ανιχνεύει τέτοιες επικαλυπτόμενες κοινότητες.

5.5.2 Οι αλγόριθμοι

5.5.2.1 Ο αλγόριθμος Link Aggregate (LA)

Ο αλγόριθμος IS αποδίδει καλά στην ανακάλυψη κοινοτήτων δεδομένης μιας καλής αρχικής εκτίμησης, για παράδειγμα, όταν οι αρχικές «μαντεψιές» είναι οι έξοδοι ενός άλλου αλγορίθμου ομαδοποίησης όπως ο RaRe, σε αντίθεση με τυχαίες ακμές στο δίκτυο επικοινωνίας. Στην παρούσα ενότητα αναφέρεται ένας διαφορετικός, αποδοτικός αλγόριθμος αρχικοποίησης.

Ο αλγόριθμος Rank Removal (RaRe) ξεκινάει από την κατάταξη όλων των κόμβων σύμφωνα με κάποιο κριτήριο, όπως είναι το Page Rank. Οι υψηλά ταξινομημένοι κόμβοι στη συνέχεια απομακρύνονται σε ομάδες μέχρι να σχηματιστούν μικρές συνδεδεμένες συνιστώσες (ονομάζονται πυρήνες της συστάδας). Στη συνέχεια αυτοί οι πυρήνες διευρύνονται με την προσθήκη κάθε αφαιρεμένου κόμβου σε οποιαδήποτε συστάδα της οποίας η πυκνότητα βελτιώνεται με την προσθήκη του. Ενώ αυτή η προσέγγιση ήταν επιτυχής όσον αφορά την ανακάλυψη συστάδων, το βασικό της μειονέκτημα ήταν η αναποτελεσματικότητα. Αυτό οφείλεται εν μέρει στο γεγονός ότι οι τάξεις και οι συνδεδεμένες συνιστώσες πρέπει να επανυπολογίζονται κάθε φορά που ένα τμήμα των κόμβων απομακρύνεται. Η πολυπλοκότητα του RaRe βελτιώνεται σημαντικά όταν οι τάξεις υπολογίζονται μόνο μία φορά.

Δεδομένου ότι οι συστάδες θα εξεταστούν από τον IS αλγόριθμο, ο αλγόριθμος χρειάζεται να βρει μόνο τις κατά προσέγγιση συστάδες. Ο αλγόριθμος IS θα «καθαρίσει» τις συστάδες. Με αυτό κατά νου, ο νέος αλγόριθμος LA εστιάζει στην αποτελεσματικότητα. Οι κόμβοι διατάσσονται σύμφωνα με κάποιο κριτήριο, για παράδειγμα σύμφωνα με το PageRank σε φθίνουσα σειρά, και στη συνέχεια εξετάζονται ακολουθιακά σύμφωνα με αυτή τη διάταξη. Όπως φαίνεται και στον παρακάτω ψευδοκώδικα, ένας κόμβος προστίθεται σε κάθε συστάδα εάν η προσθήκη βελτιώνει την πυκνότητα της συστάδας. Εάν ο κόμβος δεν προστίθεται σε κάποια συστάδα, τότε δημιουργεί μία νέα συστάδα. Σημειώστε ότι κάθε κόμβος βρίσκεται σε τουλάχιστον μία συστάδα. Συστάδες που είναι πολύ μικρές για να είναι σχετικές με την συγκεκριμένη εφαρμογή μπορεί τώρα να μειωθούν. Η πολυπλοκότητα μπορεί να οριοθετηθεί λαμβάνοντας υπ' όψιν τον αριθμό των συστάδων C της εξόδου. Η πολυπλοκότητα του LA αλγορίθμου είναι O(|C||E| + |V|).

Αλγόριθμος 5.1 Ο Link Aggregate (LA) Αλγόριθμος

διαδικασία LA (G = (V, E), W)
C ← Ø ;
Διέταξε τις κορυφές $u_1, u_2, \ldots, u_{|V|}$ σύμφωνα με το PageRank;
για i = 1 **έως** |V| **κάνε**
  added ← false ;
  για όλα $D_j \in C$ **κάνε**
    εάν $W(D_j \cup u_i) > W(D_j)$ **τότε**
      $D_j \leftarrow D_j \cup u_i$;
      added ← true;
  εάν added = false **τότε**
    $C \leftarrow C \cup \{\{u_i\}\}$;
επέστρεψε C ;

Κάθε ακμή πλησίον της u_i τοποθετείται σε δύο κλάσεις για κάθε συστάδα στο C_i, είτε το άλλο άκρο της ακμής βρίσκεται εντός της συστάδας είτε έξω από αυτή. Η πυκνότητα της συστάδας με την προσθήκη της u_i μπορεί να υπολογιστεί γρήγορα (Ο (1)) και να συγκριθεί με την εκάστοτε πυκνότητα.

5.5.2.2 Ο βελτιωμένος Iterative Scan Αλγόριθμος (IS^2)

Ο αρχικός αλγόριθμος IS δημιουργεί ρητά μία συστάδα που είναι τοπικά μέγιστη λαμβάνοντας υπ' όψιν το μέτρο της πυκνότητας, ξεκινώντας από μία «τροφοδοτούμενη» υποψήφια συστάδα. Έπειτα, την ενημερώνει προσθέτοντας ή αφαιρώντας κάθε φορά έναν κόμβο, όσο το μέτρο βελτιώνεται αυστηρά. Ο αλγόριθμος σταματά όταν δεν μπορεί πλέον να επιτευχθεί περαιτέρω βελτίωση με μία μόνο αλλαγή. Η αρχική διαδικασία περιλαμβάνει διαδοχική επανάληψη ολόκληρης της λίστας των κόμβων έως ότου η πυκνότητα της συστάδας να μην μπορεί να βελτιωθεί.

Ο νέος αλγόριθμος IS^2, που βασίζεται στον IS, δίνεται σε μορφή ψευδοκώδικα παρακάτω. Προκειμένου να μειωθεί η πολυπλοκότητα του IS, γίνεται η ακόλουθη παρατήρηση. Οι μοναδικοί κόμβοι που είναι ικανοί να αυξήσουν την πυκνότητα της συστάδας είναι τα μέλη της ίδιας της συστάδας (τα οποία μπορούν να αφαιρεθούν) ή τα μέλη της αμέσως πιο κοντινής γειτονιάς της συστάδας, δηλαδή εκείνοι οι κόμβοι που βρίσκονται πλησίον ενός κόμβου μέσα στη συστάδα. Έτσι, αντί να επισκέπτεται κάποιος κάθε κόμβο σε κάθε επανάληψη, μπορούν να παρακαμφθούν όλοι οι κόμβοι εκτός από αυτούς που ανήκουν σε μία από αυτές τις δύο ομάδες. Εάν η γειτονιά μιας συστάδας είναι αρκετά μικρότερη από ολόκληρο το γράφημα, αυτό θα βελτίωνε σημαντικά την πολυπλοκότητα του αλγορίθμου.

Αλγόριθμος 5.2 Ο IS^2 Αλγόριθμος

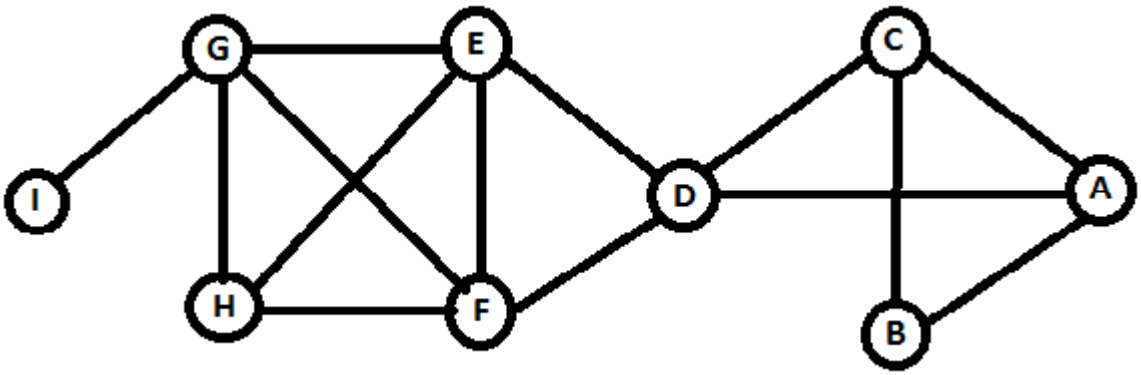
```
διαδικασία IS² (seed, G, W)
C ← seed ;
w ← W(C);
increased ← true;
όσο increased κάνε
    N ← C;
    για ολα u ∈ C κάνε
            N ← N ∪ adj(u);
    για όλα u ∈ N κάνε
            εάν u ∈ C τότε
                    C' ← C \ {u};
            αλλιώς
                    C' ← C ∪ {u};
            εάν W(C') > W(C) τότε
                    C ← C';
    εάν W(C) = w τότε
            increased ← false;
    αλλιώς
            w ← W(C);
επέστρεψε C;
```

Σημειώστε ότι ο συγκεκριμένος αλγόριθμος δεν είναι αυστηρά ίδιος με τον αρχικό IS, από τη στιγμή που ενδεχομένως ένας κόμβος που λείπει από το σύνολο γειτονιάς N μπορεί να γίνει γείτονας της συστάδας όσο εξετάζονται οι κόμβοι. Αυτός ο κόμβος έχει την ευκαιρία να εισχωρήσει στην συστάδα στον αρχικό IS αλγόριθμο, όσο στον IS^2 θα παρακάμπτεται. Αυτό δεν είναι τίποτα από τη στιγμή που ο κόμβος θα έχει τη δυνατότητα να μπει στη συστάδα στην επόμενη επανάληψη του IS^2.

Αυτός ο αλγόριθμος παρέχει τόσο μία πιθανή μείωση, όσο και αύξηση στην πολυπλοκότητα. Η μείωση λαμβάνει χώρα όταν η συστάδα, καθώς επίσης και οι γειτονιές αυτής, είναι μικρές σε σύγκριση με τον αριθμό των κόμβων στο γράφημα. Αυτό είναι το πιθανότερο σενάριο σε ένα αραιό γράφημα. Σε αυτή την περίπτωση, με βάση το σύνολο της γειτονιάς N, χρειάζεται ένα σχετικά σύντομο χρονικό διάστημα σε σχέση με την εξοικονόμηση χρόνου παρακάμπτοντας κόμβους έξω από τη γειτονιά. Η αύξηση της πολυπλοκότητας μπορεί να συμβεί όταν η γειτονιά της συστάδας είναι μεγάλη. Εδώ, η εύρεση της γειτονιάς είναι δαπανηρή, καθώς και η εξοικονόμηση χρόνου θα μπορούσε να είναι μικρή δεδομένου ότι λίγοι κόμβοι απουσιάζουν από το N. Αυτή την ιδιότητα μία μεγάλη συστάδα θα μπορούσε να έχει σε ένα πυκνό γράφημα. Σε αυτή την περίπτωση, είναι προτιμότερος ο αρχικός αλγόριθμος IS.

5.5.3 Παράδειγμα

Σε αυτή την ενότητα θα παρουσιαστεί ένα παράδειγμα ώστε να γίνει περισσότερο κατανοητός ο αλγόριθμος που περιγράφηκε παραπάνω. Έστω ότι το αρχικό δίκτυο είναι αυτό που φαίνεται στο Σχήμα 5.11.

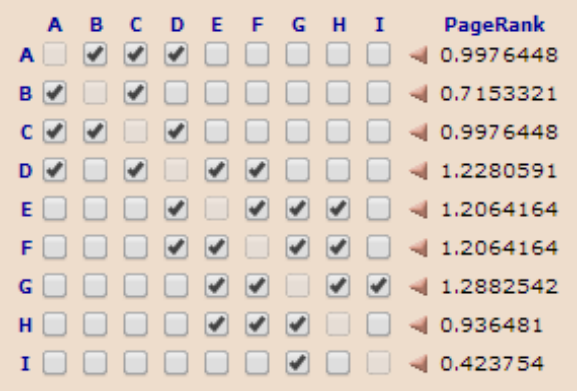


Σχήμα 5.11: Το αρχικό γράφημα στο οποίο θα εφαρμοστεί ο LA→ IS^2 αλγόριθμος.

Το συγκεκριμένο δίκτυο αποτελείται συνολικά από 9 κορυφές και 14 ακμές. Ο αλγόριθμος, όπως προαναφέρθηκε, χωρίζεται σε δύο βήματα. 1) στο LA, και 2) στο IS^2.

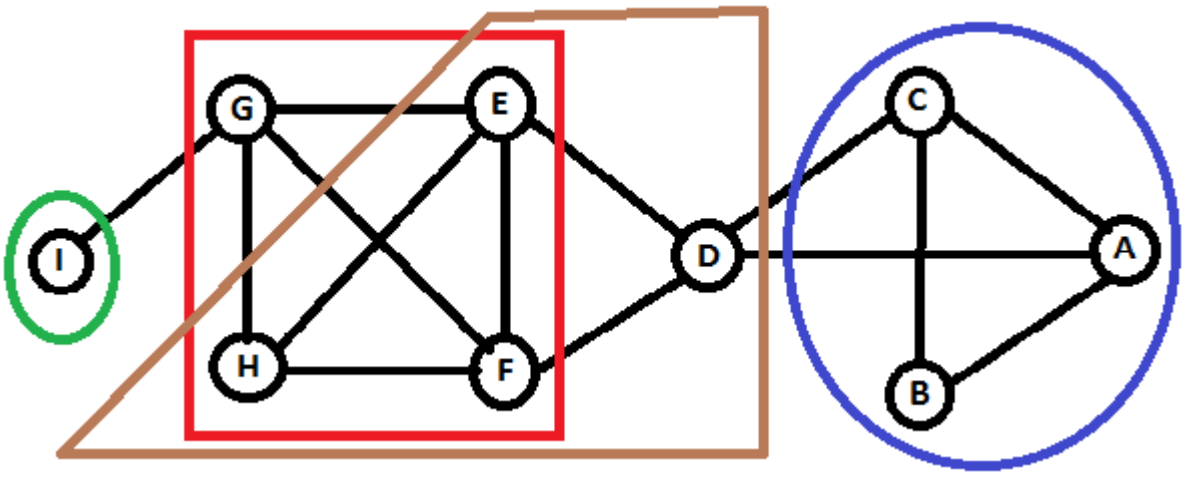
1) LA βήμα:

Αρχικά, διατάσσονται οι κορυφές σύμφωνα με τον κανόνα του PageRank. Για να γίνει αυτό εύκολα και γρήγορα προτείνονται οι παρακάτω δύο σύνδεσμοι [48,49], όπου ο χρήστης το μόνο που έχει να κάνει είναι να σημειώσει τις κορυφές και τις συνδέσεις οποιουδήποτε γραφήματος και αν τον ενδιαφέρει, και έπειτα η ιστοσελίδα αυτόματα του βγάζει τα αποτελέσματα του PageRank των κορυφών του γραφήματος. Στο συγκεκριμένο παράδειγμα οι κορυφές διατάσσονται με τη φθίνουσα σειρά [G, D, E, F, A, C, H, B, I] σύμφωνα με το στιγμιότυπο που φαίνεται στο Σχήμα 5.12.



Σχήμα 5.12: Οι κορυφές διατάχθηκαν σύμφωνα με το PageRank [48].

Αρχικά, σύμφωνα με την παραπάνω διάταξη, εξετάζεται ο κόμβος G. Αυτός δημιουργεί την πρώτη συστάδα $C_1 = \{ G \}$ με $W_{C1} = 0$. Έπειτα, ο κόμβος D δημιουργεί τη συστάδα $C_2 = \{ D \}$ με $W_{C2} = 0$. Στη συνέχεια, ο κόμβος E συνεισφέρει και στην C_1 και στην C_2, με αποτέλεσμα οι συγκεκριμένες συστάδες να γίνονται $C_1 = \{ G,E \}$ και $C_2 = \{ D,E \}$ με $W_{C1} = 1$ και $W_{C2} = 1$, αντίστοιχα. Εξετάζοντας την κορυφή F προκύπτει $C_1 = \{ G,E,F \}$ και $C_2 = \{ D,E,F \}$ με $W_{C1} = 2$ και $W_{C2} = 2$, αντίστοιχα. Στη συνέχεια η κορυφή A δε συνεισφέρει σε καμία από τις δύο συστάδες και δημιουργεί μια τρίτη $C_3 = \{ A \}$ με $W_{C3} = 0$. Η κορυφή C συνεισφέρει στην συστάδα $C_3 = \{ A,C \}$ με $W_{C3} = 1$. Η κορυφή H συνεισφέρει στις $C_1 = \{ G,E,F,H \}$ και $C_2 = \{ D,E,F,H \}$ με $W_{C1} = 3$ και $W_{C2} = 2.5$, αντίστοιχα. Εξετάζοντας τη κορυφή B προκύπτει $C_3 = \{ A,C,B \}$ με $W_{C3} = 2$. Τέλος, η κορυφή I δημιουργεί μία νέα συστάδα $C_4 = \{ I \}$ με $W_{C4} = 0$. Με το τέλος του LA βήματος, το αρχικό γράφημα χωρίζεται στις τροφοδοτούμενες συστάδες, που φαίνεται στο Σχήμα 5.13, οι οποίες θα χρησιμοποιηθούν στο επόμενο IS^2 βήμα.



Σχήμα 5.13: Οι τέσσερις τροφοδοτούμενες συστάδες που προκύπτουν μετά το πέρας του LA βήματος.

2) IS^2 βήμα:

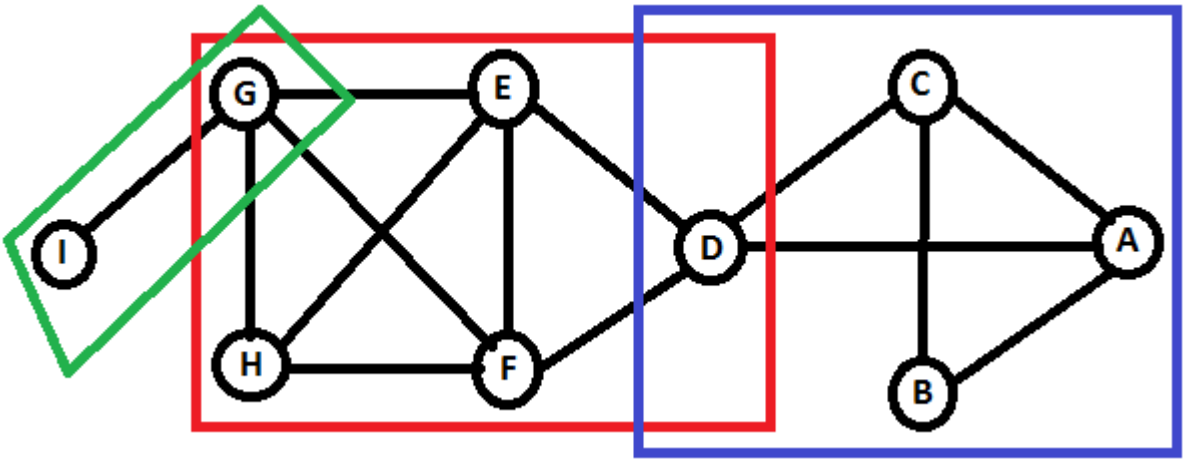
Σε αυτό το βήμα εξετάζονται επαναληπτικά μία-μία οι συστάδες που προέκυψαν προηγουμένως. Θα παραλειφθεί η αναλυτική παρουσίαση των αλλαγών που γίνονται σε κάθε επανάληψη, αλλά θα παρουσιαστούν σχηματικά τα αποτελέσματα αυτών.

Μετά το τέλος της πρώτης επανάληψης προκύπτουν οι συστάδες:

C_1 = { G,E,F,H,D } με W_{C1} = 3.2,

C_2 = { A,B,C,D } με W_{C2} = 2.5,

C_3 = { I,G } με W_{C3} = 1.



Σχήμα 5.14: Οι τρεις συστάδες που προκύπτουν μετά το πέρας της πρώτης επανάληψης του IS^2 βήματος.

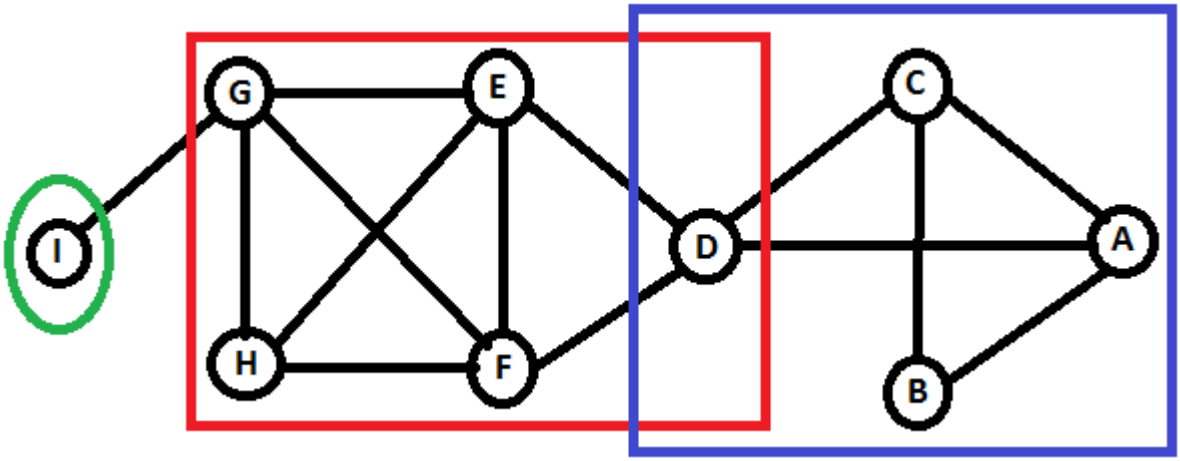
Μετά το τέλος της δεύτερης επανάληψης προκύπτουν οι συστάδες:

C_1 = { G,E,F,H,D } με W_{C1} = 3.2,

C_2 = { A,B,C,D } με W_{C2} = 2.5,

C_3 = { I }.

Σε αυτό το σημείο σταματάει ο αλγόριθμος επειδή δεν γίνεται περαιτέρω αύξηση του $W(C_j)$.



Σχήμα 5.15: Η τελική συσταδοποίηση που προκύπτει εφαρμόζοντας τον LA→ IS^2 αλγόριθμο στο αρχικό γράφημα, όπου φαίνονται ξεκάθαρα δύο επικαλυπτόμενες κοινότητες που έχουν κοινή κορυφή την D.

5.6 Τοπική Πυκνότητα (Local Density)

Σε αυτή την ενότητα παρουσιάζεται μία «κλασσική», θα λέγαμε, προσέγγιση η οποία είναι χαρακτηριστική για την κατηγορία αυτή. Προτείνεται ένας αλγόριθμος συσταδοποίησης που εφαρμόζει τοπικές μεθόδους για την ομαδοποίηση ολόκληρου του γραφήματος. Πολλές από τις υπάρχουσες μελέτες επικεντρώνονται στις τοπικές μεθόδους συσταδοποίησης χρησιμοποιώντας διάφορα ποιοτικά μέτρα. Αρχικά, ορίζεται ένα ποιοτικό μέτρο, αυτό της πυκνότητας το οποίο στη συνέχεια βελτιστοποιείται, έτσι ώστε να συγχωνεύονται συστάδες αναδρομικά, αν και μόνο αν, βέβαια, η κίνηση αυτή προκαλεί μία αύξηση στην ποιοτική συνάρτηση. Δηλαδή, με την συγκεκριμένη προσέγγιση μπορούν και υπολογίζονται οι συστάδες στα γραφήματα, μία κάθε φορά, στηριζόμενοι μόνο στις *γειτονιές* των κορυφών που περιλαμβάνονται στην τρέχουσα υποψήφια συστάδα. Μια «τοπική» προσέγγιση, χωρίς παραμέτρους, για την εύρεση μιας καλής συστάδας που περιέχει μία συγκεκριμένη κορυφή ή ένα σύνολο κορυφών εξετάζοντας έναν μόνο περιορισμένο αριθμό κορυφών κάθε φορά, οδηγεί στην περιοχή γύρω από την υπό εξέταση κορυφή.

5.6.1 Περιγραφή αλγορίθμου

Στον αλγόριθμο που περιγράφεται, όταν αναφέρεται η έννοια συστάδα εννοούνται κορυφές που συνδέονται με *πολλές εσωτερικές* συνδέσεις, και *μόνο λίγες εξωτερικές* συνδέσεις. Η

ποιοτική συνάρτηση, όπως αναφέρθηκε προηγουμένως, είναι ο εσωτερικός βαθμός μιας συστάδας C, δηλαδή, ο αριθμός των ακμών που συνδέουν τις κορυφές της C μεταξύ τους, και συμβολίζεται ως εξής: $deg_{int}(C) = |\{(u, v) \in E | u, v \in C\}|$. Σε αυτό το σημείο θα πρέπει να αναφερθεί και ένα ανάλογο μέτρο, το οποίο συμβολίζεται με $deg_{ext}(C) = |\{(u, v) \in E | u \in C, v \in V \setminus C\}|$, το οποίο στην ουσία περιγράφει τον αριθμό των ακμών που συνδέουν τις κορυφές της C με γειτονικές κορυφές εκτός συστάδας.

Έτσι, είναι δυνατόν να οριστεί η **τοπική πυκνότητα** της συστάδας ως εξής:

$$\delta_l(C) = \frac{deg_{int}(C)}{|C|(|C|-1)}$$, όπου |C| είναι το μέγεθος της συστάδας C, δηλαδή ο αριθμός των κορυφών μέσα σε αυτή.

Βελτιστοποιώντας το $\delta \in [0,1]$ από μόνο του δημιουργούνται μικρές κλίκες που οδηγούν σε μεγαλύτερους αλλά ελαφρώς πιο αραιούς υπογράφους, το οποίο δεν είναι συχνά πρακτικό. Αντ' αυτού, για συστάδες που περιέχουν μόνο λίγες συνδέσεις σε σχέση με το υπόλοιπο γράφημα, προτείνεται η βελτιστοποίηση ενός μέτρου που ονομάζεται **σχετική πυκνότητα**, και το οποίο συμβολίζεται ως εξής:

$$\delta_r(C) = \frac{deg_{int}(C)}{deg_{int}(C) + deg_{ext}(C)}.$$

Το τελικό ποιοτικό μέτρο που χρησιμοποιείται είναι η **σύνθετη καταλληλότητα**, που συμβολίζεται ως εξής: $f(C) = \delta_l(C)\delta_r(C)$.

Ο σκοπός είναι να εξεταστούν στοχαστικά υποσύνολα της V που περιέχουν τις u κορυφές, και να επιλεγεί το υποψήφιο με το **μέγιστο** f ως C(u). Η αρχική συστάδα C'(u) μιας κορυφής u περιέχει την ίδια την u και όλες τις κορυφές πλησίον της u. Κάθε βήμα της αναζήτησης μπορεί είτε να προσθέτει μία νέα κορυφή που βρίσκεται δίπλα σε μία ήδη συμπεριληφθείσα κορυφή, είτε να αφαιρεί μία συμπεριληφθείσα κορυφή.

Εάν ένα δεδομένο γράφημα G έχει ένα k-κορυφών υπογράφημα C για το οποίο $f(C) \geq \gamma$ για κάποιο σταθερό $k \in N$ και $\gamma \in [0, 1]$, τόσο το $\delta_l(C)$ όσο και το $\delta_r(C)$ από μόνα τους παραπέμπουν σε NP-complete προβλήματα απόφασης.

5.6.2 Εφαρμογή του αλγορίθμου

Στον αλγόριθμο που παρουσιάζεται στην ενότητα αυτή, η επιλογή της αρχικής κορυφής από την οποία ξεκινάει το τρέξιμο του αλγορίθμου είναι τυχαία. Συνήθως είναι μία κορυφή με το μεγαλύτερο βάρος. Καθώς προχωράει ο αλγόριθμος, εξετάζονται, κατά προτίμηση, κατά σειρά, από την εκάστοτε γειτονιά, οι κορυφές με το μεγαλύτερο βάρος, και από αυτές αυτή που μεγιστοποιεί τη σύνθετη καταλληλότητα. Η αναζήτηση πρέπει να διατηρεί μόνο α) τη λίστα των

κορυφών που συμπεριλαμβάνονται στην C, β) το $deg_{int}(C)$ και γ) το $deg_{ext}(C)$. Η ενημέρωση για το C '= C ∪ {u} γίνεται θέτοντας

$deg_{int}(C') = deg_{int}(C) + k$ και $deg_{ext}(C') = deg_{ext}(C) - k + l$, όπου

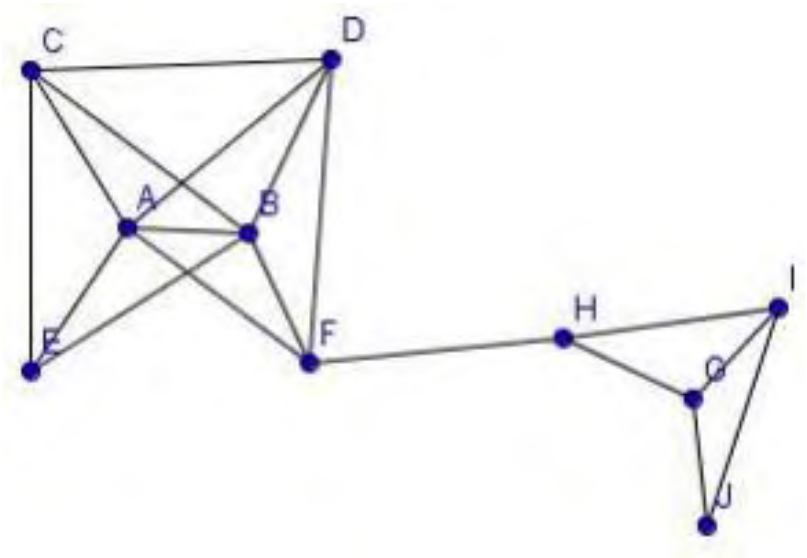
$k = |C \cap \Gamma(u)|$, με το Γ(u) να συμβολίζει το σύνολο των γειτόνων της κορυφής u και $l = \deg(u) - k$.

Για μία νέα κορυφή u, το k είναι ο αριθμός των ακμών μεταξύ του k και της C, όπου το C συμβολίζει την παλιά συστάδα. Το l είναι ο αριθμός των ακμών που δεν είναι παρακείμενες στην C. Στην τελική, διαπιστώνεται ότι το άθροισμα $k + l$ είναι ίσο με τον βαθμό της κορυφής u.

Στο τέλος, αφού υπολογιστούν κατά σειρά τα μεγέθη $\delta_l(C')$, $\delta_r(C')$ και $f(C')$, εάν η σύνθετη καταλληλότητα αυξάνεται, τότε συμπεριλαμβάνεται η υποψήφια κορυφή στην εξεταζόμενη συστάδα, αλλιώς η διαδικασία σταματάει και παραμένει η συστάδα ως έχει.

5.6.2.1 Παράδειγμα

Θα προχωρήσουμε τώρα σε ένα παράδειγμα για να δούμε πως λειτουργεί ο αλγόριθμος. Έστω το παρακάτω γράφημα το οποίο αποτελείται από m = 10 κορυφές. Με λεξικογραφική σειρά ορίζουμε τη λίστα με τα βάρη των κορυφών του γραφήματος:
[5 , 5 , 4 , 4 , 3 , 4 , 3 , 3 , 3 , 2].



Σχήμα 5.16: Το αρχικό γράφημα πριν την εφαρμογή του αλγορίθμου.

Στο πρώτο βήμα, ξεκινάμε διαλέγοντας την κορυφή με το μεγαλύτερο βαθμό. Ανάμεσα από την κορυφή Α και Β ας πάρουμε την Α. Σε αυτό το βήμα, οι τιμές των μέτρων που μας ενδιαφέρουν παίρνουν τις εξής τιμές:

- C’ = {A}
- $|C'|$ = 1
- Γ = {B,C,D,E,F}
- $\deg_{\text{int}}(C')$ = 0
- $\deg_{\text{ext}}(C')$ = 5
- f(C’) = 0

Με το τέλος του πρώτου βήματος εντός της συστάδας C έχουμε την κορυφή Α.

Στο δεύτερο βήμα, για να διαλέξουμε την επόμενη υποψήφια κορυφή για να μπει στη συστάδα, κοιτάμε την παραπάνω λίστα Γ των γειτόνων, και επιλέγουμε την κορυφή Β που έχει μεγαλύτερο βάρος από τις άλλες. Σε αυτό το βήμα, οι τιμές των μέτρων που μας ενδιαφέρουν παίρνουν τις εξής τιμές:

- C’ = {A,B}
- $|C'|$ = 2
- Γ = {C,D,E,F}
- $\deg_{\text{int}}(C')$ = 1
- $\deg_{\text{ext}}(C')$ = 8
- $\delta_l(C')$ = 1/2
- $\delta_r(C')$ = 1/9
- f(C’) = 1/18 = 0.055 > 0

Με το τέλος του δεύτερου βήματος εντός της συστάδας C έχουμε τις κορυφές Α, Β.

Στο τρίτο βήμα, για να διαλέξουμε την επόμενη υποψήφια κορυφή για να μπει στη συστάδα, κοιτάμε την λίστα Γ των γειτόνων του δευτέρου βήματος. Από τις τέσσερις κορυφές οι C,D,F έχουν μεγαλύτερο βάρος. Οποιαδήποτε και να διαλέξουμε εξ’ αυτών προκύπτει ακριβώς το ίδιο αποτέλεσμα. Ας επιλέξουμε ως υποψήφια κορυφή την C. Σε αυτό το βήμα, οι τιμές των μέτρων που μας ενδιαφέρουν παίρνουν τις εξής τιμές:

- C’ = {A,B,C}
- $|C'|$ = 3
- Γ = {D,E,F}
- $\deg_{\text{int}}(C')$ = 3
- $\deg_{\text{ext}}(C')$ = 8
- $\delta_l(C')$ = 1/2
- $\delta_r(C')$ = 3/11
- f(C’) = 0.136 > 0.055

Με το τέλος του τρίτου βήματος εντός της συστάδας C έχουμε τις κορυφές A, B, C.

Στο τέταρτο βήμα επιλέγουμε ως υποψήφια κορυφή την D. Σε αυτό το βήμα, οι τιμές των μέτρων που μας ενδιαφέρουν παίρνουν τις εξής τιμές:

- C’ = {A,B,C,D}
- |C’| = 4
- Γ = {E,F}
- $\deg_{\text{int}}(C')$ = 6
- $\deg_{\text{ext}}(C')$ = 6
- $\delta_l(C')$ = 1/2
- $\delta_r(C')$ = 1/2
- f(C’) = 0.25 > 0.136

Αν δοκιμάζαμε να επιλέξουμε την κορυφή F αντί για την D θα βλέπαμε ότι το f(C’) θα ήταν μικρότερο από αυτό που βρήκαμε παραπάνω. Με το τέλος του τέταρτου βήματος εντός της συστάδας C έχουμε τις κορυφές A, B, C, D.

Στο πέμπτο βήμα επιλέγουμε ως υποψήφια κορυφή την F. Σε αυτό το βήμα, οι τιμές των μέτρων που μας ενδιαφέρουν παίρνουν τις εξής τιμές:

- C’ = {A,B,C,D,F}
- |C’| = 5
- Γ = {E,H}
- $\deg_{\text{int}}(C')$ = 9
- $\deg_{\text{ext}}(C')$ = 4
- $\delta_l(C')$ = 9/20
- $\delta_r(C')$ = 9/13
- f(C’) = 0.31 > 0.25

Σε αυτό το βήμα εάν διαλέγαμε την κορυφή E αντί για την F, το αποτέλεσμα θα ήταν το ίδιο. Με το τέλος του πέμπτου βήματος εντός της συστάδας C έχουμε τις κορυφές A, B, C, D, F.

Στο έκτο βήμα επιλέγουμε ως υποψήφια κορυφή την E. Σε αυτό το βήμα, οι τιμές των μέτρων που μας ενδιαφέρουν παίρνουν τις εξής τιμές:

- C’ = {A,B,C,D,E,F}
- |C’| = 6
- Γ = {H}
- $\deg_{\text{int}}(C')$ = 12
- $\deg_{\text{ext}}(C')$ = 1
- $\delta_l(C')$ = 6/15
- $\delta_r(C')$ = 12/13

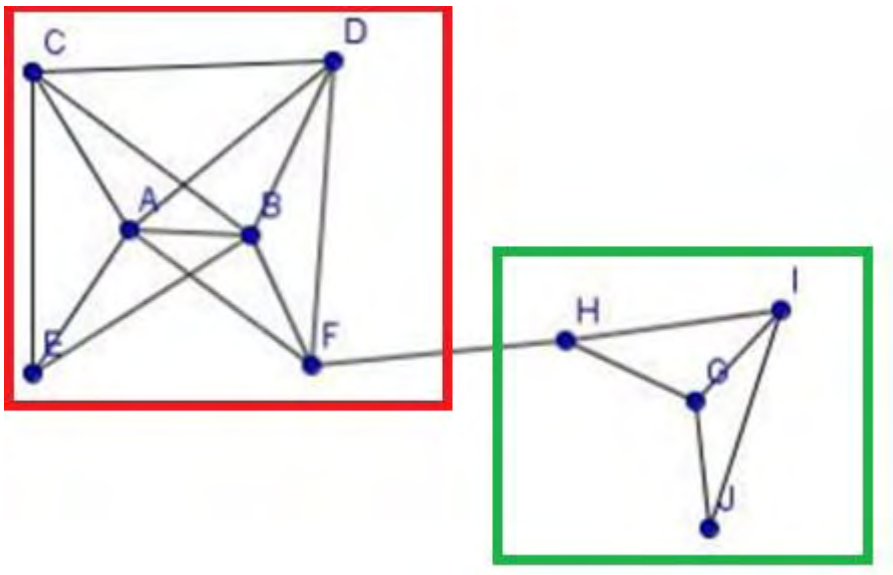
- f(C’) = 0.369 > 0.31

Σε αυτό το βήμα εάν διαλέγαμε την κορυφή Η αντί για την Ε, θα βλέπαμε ότι η τιμή της σύνθετης καταλληλότητας μειώνεται, το οποίο σημαίνει ότι η υποψήφια κορυφή Η δεν θα έμπαινε στη συστάδα. Έτσι, θα καταλήγαμε να δοκιμάσουμε την κορυφή Ε. Με το τέλος του έκτου βήματος εντός της συστάδας C έχουμε τις κορυφές A, B, C, D, E, F.

Στο έβδομο βήμα επιλέγουμε ως υποψήφια κορυφή την Η. Σε αυτό το βήμα, οι τιμές των μέτρων που μας ενδιαφέρουν παίρνουν τις εξής τιμές:

- C’ = {A,B,C,D,E,F}
- |C’| = 7
- Γ = {H}
- $\deg_{\text{int}}(C') = 13$
- $\deg_{\text{ext}}(C') = 2$
- $\delta_l(C') = 13/42$
- $\delta_r(C') = 13/15$
- f(C’) = 0.268 < 0.369

Σε αυτό το βήμα το f(C’) μειώνεται, οπότε συμπληρώθηκε η συστάδα χωρίς να συμπεριλαμβάνεται η κορυφή Η. Με το τέλος του έβδομου βήματος εντός της πρώτης συστάδας C έχουμε τελικά τις κορυφές {A, B, C, D, E, F}. Ομοίως παίρνουμε και τη δεύτερη συστάδα {G, H, I, J}.



Σχήμα 5.17: Η εφαρμογή του αλγορίθμου χώρισε το γράφημα σε δύο συστάδες.

Βιβλιογραφία

[1] Sanjeev Arora, Rong Ge, Sushant Sachdeva, Grant Schoenebeck, Finding Overlapping Communities in Social Networks: Toward a Rigorous Approach.

[2] Jierui Xie, Stephen Kelley, Boleslaw K. Szymanski, Overlapping Community Detection in Networks: the State of the Art and Comparative Study.

[3] Lior Rokach, Department of Industrial Engineering Tel-Aviv University, Oded Maimon Department of Industrial Engineering Tel-Aviv University, CLUSTERING METHODS.

[4] Anil K. Jain, Richard C. Dubes, Algorithms for Clustering Data, Michigan State University.

[5] Santo Fortunato, Community detection in graphs.

[6] Benjamin H. Good, Yves-Alexandre de Montjoye, and Aaron Clauset, The performance of modularity maximization in practical contexts.

[7] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices.

[8] Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnela, Community Structure in Time-Dependent, Multiscale and Multiplex Networks.

[9] Michele Berlingerio, Michele Coscia, Fosca Giannotti, Finding and Characterizing Communities in Multidimensional Networks.

[10] Michele Berlingerio, Michele Coscia, Fosca Giannotti, Finding Redundant and Complementary Communities in Multidimensional Networks.

[11] M. E. J. Newman, Fast algorithm for detecting community structure in networks.

[12] E. A. Leicht and M. E. J. Newman, Community structure in directed networks.

[13] M. E. J. Newman, Modularity and community structure in networks.

[14] V. Nicosia, G. Mangioni, V. Carchiolo, M. Malgeri, Extending the definition of modularity to directed graphs with overlapping communities.

[15] Jordi Duch and Alex Arenas, Community detection in complex networks using Extremal Optimization.

[16] Per Bak, Kim Sneppen, Punctuated Equilibrium and Criticality in a Simple Model of Evolution.

[17] A Arenas, J Duch, A Fernandez and S Gomez, Size reduction of complex networks preserving modularity.

[18] Roger Guimerà and Luís A Nunes Amaral, Cartography of complex networks.

[19] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre, Fast unfolding of communities in large networks.

[20] Ken Wakita, Toshiyuki Tsurumi, Finding Community Structure in Mega-scale Social Networks [Extended Abstract].

[21] Paolo Boldi, Massimo Santini, Sebastiano Vigna, A Large Time-Aware Web Graph.

[22] Matthew L. Wallace and Yves Gingras, A new approach for detecting scientific specialties from raw cocitation networks.

[23] R. Lambiotte, J.-C. Delvenne and M. Barahona, Laplacian Dynamics and Multiscale Modular Structure in Networks.

[24] Sergio Gomez, Pablo Jensen and Alex Arenas, Analysis of community structure in networks of correlated data.

[25] Michael J. Barber, Modularity and community detection in bipartite networks.

[26] V.A. Traag and Jeroen Bruggeman, Community detection in networks with positive and negative links.

[27] http://en.wikipedia.org/wiki/Modularity_(networks).

[28] Brett W. Bader and Tamara G. Kolda, Algorithm 862: MATLAB Tensor Classes for Fast Algorithm Prototyping.

[29] J. Reichardt and S. Bornholdt, Phys. Rev. E 74, 016110 (2006) και M. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004).

[30] P. Holland and S. Leinhardt, Sociological Methodology 7, 1 (1976) και F. McSherry, in IEEE Symposium on Foundations of Computer Science (2001), pp. 529{537.

[31] M. B. Hastings, Phys. Rev. E 74, 035102(R) (2006) και M. E. J. Newman and E. A. Leicht, PNAS 104, 9564 (2007).

[32] R. P. Feynman, Statistical Mechanics, A Set of Lectures(W. A. Benjamin, 1972), ISBN 0805325085.

[33] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul, Machine Learning 37, 183 (1999).

[34] G. Schwarz, The Annals of Statistics 6, 461 (1978).

[35] J. Baumes, M. Goldberg, M. Magdon-Ismail, and W. Wallace, Discovering Hidden Groups in Communication Networks.

[36] M. E. J. Newman, The structure and function of complex networks.

[37] J. Baumes, M. Goldberg, M. Krishnamoorty, M. Magdon-Ismail, and N. Preston. Finding communities by clustering a graph into overlapping subgraphs. Proceedings of IADIS Applied Computing 2005, pages 97–104, February 2005.

[38] U. Brandes, M. Gaertler, and D. Wagner. Experiments on graph clustering algorithms. Lecture Notes in Comuter Science, Di Battista and U. Zwick (Eds.):568–579, 2003.

[39] Lin Y.-R., Sun J., Sundaram H., Kelliher A., Castro P. & Konuru R, Community discovery via metagraph factorization, *ACM Trans. Knowl. Discov. Data* 5, 17:1–17:44 (2011).

[40] Christian Bockermann, Felix Jungermann, Stream-based Community Discovery via Relational Hypergraph Factorization on Evolving Networks.

[41] Matthew J. Beal, Zoubin Ghahramani, Variational Bayesian Learning of Directed Graphical Models with Hidden Variables.

[42] Matthew J. Beal, Variational Algorithms For Approximate Bayesian Inference, M.A., M.Sci., Physics, University of Cambridge, UK (1998).

[43] Masa-aki Sato, Online Model Selection Based on the Variational Bayes.

[44] C. Biernacki, G. Celeux, and G. Govaert, IEEE Trans. Pattern Anal. Mach. Intell. 22, 719 (2000).

[45] C. A. Hugo Zanghi and V. Miele, Fast online graph clustering via Erdos-Reyni mixture (2007), sSB-RR-8.

[46] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, Mixed membership stochastic blockmodels (2007).

[47] M. Stone, J. Royal Stat. Soc. 36, 111 (1974).

[48] http://www.webworkshop.net/pagerank_calculator.php?.

[49] http://williamcotton.com/pagerank-explained-with-javascript.

[50] http://jakehofman.com/talks/nips_20071208_jmh_chw_static.pdf.

ΚΕΦΑΛΑΙΟ 6 ΣΥΝΔΥΑΣΤΙΚΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗ

6.1 Εισαγωγή

Υπάρχει ένας αριθμός πλαισίων για τον εντοπισμό μιας κοινότητας που χρησιμοποιούν έναν μηδαμινό ορισμό της κοινότητας, ή δεν έχουν καθόλου ορισμό. Αυτές οι μέθοδοι συχνά υποθέτουν ότι υπάρχουν ορισμένα επιθυμητά χαρακτηριστικά της κοινότητας που δεν παρέχονται από πολλούς αλγόριθμους. Καθορίζουν λειτουργίες προ-επεξεργασίας ή/και μετα-επεξεργασίας και στη συνέχεια τις εφαρμόζουν σε έναν αριθμό άλλων διαφόρων γνωστών μεθόδων που δεν εξαγάγουν κοινότητες με τα επιθυμητά χαρακτηριστικά. Με τον τρόπο αυτό βελτιώνουν τα αποτελέσματα.

Βασικά, ο ορισμός που διατηρήθηκε είναι:

Ορισμός 6.1 (Κοινότητα). Κοινότητες σε ένα πολύπλοκο δίκτυο αποτελούν σύνολα τα οποία παρουσιάζουν ορισμένα ιδιαίτερα χαρακτηριστικά, ανεξάρτητα από το λόγο για τον οποίο ομαδοποιούνται οι κόμβοι τους.

Φυσικά, οι διαδικασίες και τα χαρακτηριστικά αυτών των προσεγγίσεων εξαρτώνται τόσο από την προ/μετά-επεξεργασία όσο και από την «φιλοξενούμενη» μέθοδο. Αυτό το Κεφάλαιο εξετάζει τη συσταδοποίηση που βασίζεται σε στατιστικά μοντέλα. Είναι συχνά βολικό και αποτελεσματικό να θεωρείται ότι τα δεδομένα έχουν παραχθεί ως αποτέλεσμα μιας στατιστικής διαδικασίας και να περιγράφονται τα δεδομένα βρίσκοντας το στατιστικό μοντέλο που εφαρμόζεται καλύτερα σε αυτά, όπου το στατιστικό μοντέλο περιγράφεται σε σχέση με μία κατανομή και ένα σύνολο παραμέτρων για αυτήν την κατανομή. Σε ένα υψηλό επίπεδο, αυτή η διαδικασία εμπεριέχει την απόφαση επιλογής ενός στατιστικού μοντέλου για τα δεδομένα και εκτίμησης των παραμέτρων του μοντέλου από τα δεδομένα. Αυτό το Κεφάλαιο περιγράφει ένα συγκεκριμένο είδος στατιστικού μοντέλου, τα *συνδυαστικά μοντέλα (mixture models)* [1], τα οποία μοντελοποιούν τα δεδομένα χρησιμοποιώντας ένα πλήθος στατιστικών κατανομών. Κάθε κατανομή αντιστοιχεί σε μία συστάδα και οι παράμετροι κάθε κατανομής δίνουν μία περιγραφή της αντίστοιχης συστάδας, συνήθως σε σχέση με το κέντρο και τη διασπορά της.

Στη συνέχεια του Κεφαλαίου, αφού δοθεί μια περιγραφή των συνδυαστικών μοντέλων, εξετάζεται ο τρόπος με τον οποίο μπορούν να εκτιμηθούν οι παράμετροι για τα μοντέλα των

στατιστικών δεδομένων. Αρχικά περιγράφεται ο τρόπος με τον οποίο μπορεί να χρησιμοποιηθεί η διαδικασία που ονομάζεται **εκτίμηση μέγιστης πιθανοφάνειας (maximum likelihood estimation - MLE)** [2] για την εκτίμηση παραμέτρων σε απλά στατιστικά μοντέλα και έπειτα παρουσιάζεται ο τρόπος με τον οποίο μπορεί να επεκταθεί αυτή η προσέγγιση για να εκτιμηθούν οι παράμετροι των συνδυαστικών μοντέλων. Ειδικότερα, περιγράφεται ο γνωστός αλγόριθμος **προσδοκίας – μεγιστοποίησης (expectation maximization algorithm -EM)** [3,6,7,8], ο οποίος κάνει μία αρχική τυχαία υπόθεση για τις παραμέτρους και στη συνέχεια βελτιώνει επαναληπτικά αυτές τις εκτιμήσεις.

6.2 Συνδυαστικά μοντέλα

Τα συνδυαστικά μοντέλα [9] θεωρούν τα δεδομένα ως ένα σύνολο παρατηρήσεων, από ένα συνδυασμό διαφορετικών κατανομών πιθανοτήτων. Οι κατανομές πιθανοτήτων μπορεί να είναι οποιεσδήποτε, αλλά συχνά θεωρείται ότι είναι πολυμεταβλητές κανονικές κατανομές, δεδομένου ότι αυτός ο τύπος είναι εύκολα κατανοητός, μαθηματικά εύκολος για να εργαστεί κάποιος και έχει αποδειχθεί ότι παράγει καλά αποτελέσματα σε πολλές περιπτώσεις. Αυτοί οι τύποι κατανομών μπορούν να μοντελοποιήσουν ελλειψοειδείς συστάδες.

Εννοιολογικά, τα συνδυαστικά μοντέλα αντιστοιχούν στην ακόλουθη διαδικασία παραγωγής δεδομένων. Δοθέντων διάφορων κατανομών, συνήθως ίδιου τύπου, αλλά με διαφορετικές παραμέτρους, επιλέγεται τυχαία μία από αυτές τις κατανομές και παράγεται ένα αντικείμενο από αυτήν. Η διαδικασία επαναλαμβάνεται m φορές, όπου το m είναι το πλήθος των αντικειμένων.

Πιο τυπικά, ας υποτεθεί ότι υπάρχουν K κατανομές και m αντικείμενα, X = { x_1,…..,x_m }. Έστω ότι η j-οστή κατανομή έχει παραμέτρους θ_j και ότι Θ είναι το σύνολο όλων των παραμέτρων, δηλαδή Θ = { θ_1,…..,θ_k }. Τότε, prob($x_i|\theta_j$) είναι η πιθανότητα του i-οστού αντικειμένου να προέρχεται από την j-οστή κατανομή. Η πιθανότητα ότι έχει επιλεχθεί η j-οστή κατανομή για να παράγει ένα αντικείμενο δίνεται από το βάρος w_j, $1 \leq j \leq K$, όπου αυτά τα βάρη (πιθανότητες) υπόκεινται στον περιορισμό ότι αθροίζονται στη μονάδα, δηλαδή $\sum_{j=1}^{K} w_j = 1$. Τότε η πιθανότητα ενός αντικειμένου x δίνεται από την Εξίσωση 6.1.

$$\text{prob}(x|\Theta) = \sum_{j=1}^{K} w_j \, p_j(x|\theta_j) \qquad (6.1)$$

Αν τα αντικείμενα παράγονται με έναν ανεξάρτητο τρόπο, τότε η πιθανότητα όλου του συνόλου των αντικειμένων, είναι απλά το γινόμενο των πιθανοτήτων κάθε ξεχωριστού x_i.

$$\text{prob}(X|\Theta) = \prod_{i=1}^{m} \text{prob}(x_i|\Theta) = \prod_{i=1}^{m} \sum_{j=1}^{K} w_j p_j(x_i|\Theta_j) \qquad (6.2)$$

Για τα συνδυαστικά μοντέλα, κάθε κατανομή περιγράφει μία διαφορετική ομάδα, δηλαδή μία διαφορετική συστάδα. Χρησιμοποιώντας στατιστικές μεθόδους, μπορούν να εκτιμηθούν οι

παράμετροι αυτών των κατανομών από τα δεδομένα και επομένως να περιγράφουν αυτές τις κατανομές (συστάδες). Επίσης, είναι δυνατόν να προσδιοριστεί ποια αντικείμενα ανήκουν σε ποιες συστάδες. Ωστόσο, η συνδυαστική μοντελοποίηση δεν παράγει μία σαφή εκχώρηση αντικειμένων σε συστάδες, αλλά αντίθετα, δίνει την πιθανότητα με την οποία ένα συγκεκριμένο αντικείμενο ανήκει σε μία συγκεκριμένη συστάδα.

6.3 Εκτίμηση των παραμέτρων του μοντέλου με χρήση της MLE

Δοθέντος ενός στατιστικού μοντέλου για τα δεδομένα, είναι απαραίτητο να εκτιμηθούν οι παράμετροι του μοντέλου. Μία τυπική προσέγγιση που χρησιμοποιείται για αυτή την εργασία, είναι η εκτίμηση μέγιστης πιθανοφάνειας, η οποία εξηγείται παρακάτω.

Αρχικά, ας θεωρηθεί ένα σύνολο από m σημεία, τα οποία παράγονται από μία μονοδιάστατη κατανομή Gauss. Υποθέτοντας ότι τα σημεία παράγονται ανεξάρτητα, η πιθανότητα των σημείων είναι απλά το γινόμενο των ξεχωριστών τους πιθανοτήτων. Χρησιμοποιώντας τη συνάρτηση πυκνότητας πιθανότητας για μία μονοδιάστατη κατανομή Gauss σε ένα σημείο x, η πιθανότητα μπορεί να γραφεί με τον τρόπο που φαίνεται στην Εξίσωση 6.3. Δεδομένου ότι αυτή η πιθανότητα θα είναι ένας πολύ μικρός αριθμός, τυπικά χρησιμοποιείται η λογαριθμική πιθανότητα, όπως φαίνεται στην Εξίσωση 6.4.

$$\text{prob}(X|\Theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \qquad (6.3)$$

$$\log \text{prob}(X|\Theta) = -\sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2} - 0.5m \log 2\pi - m \log \sigma \qquad (6.4)$$

Θα ήταν επιθυμητό να βρεθεί μια διαδικασία για την εκτίμηση των μ και σ αν αυτά είναι άγνωστα. Μια προσέγγιση είναι να επιλεγούν οι τιμές των παραμέτρων για τις οποίες τα δεδομένα είναι πιο πιθανά. Με άλλα λόγια, επιλέγονται τα μ και σ που μεγιστοποιούν την Εξίσωση 6.3. Αυτή η προσέγγιση είναι γνωστή ως η **αρχή της μέγιστης πιθανοφάνειας (maximum likelihood principle)**, και η διαδικασία εφαρμογής αυτής της αρχής για την εκτίμηση των παραμέτρων μιας στατιστικής κατανομής από τα δεδομένα είναι γνωστή ως **εκτίμηση της μέγιστης πιθανοφάνειας (maximum likelihood estimation - MLE)** [4,5].

Η συγκεκριμένη αρχή ονομάζεται έτσι επειδή, δοθέντος ενός συνόλου δεδομένων, η πιθανότητα των δεδομένων, θεωρούμενη ως συνάρτηση των παραμέτρων, ονομάζεται *συνάρτηση πιθανοφάνειας*. Για να γίνουν τα παραπάνω εμφανή, η Εξίσωση 6.3 ξαναγράφεται ως Εξίσωση 6.5 για να δοθεί έμφαση στο ότι οι στατιστικές παράμετροι μ και σ θεωρούνται ως μεταβλητές και επιπλέον ότι τα δεδομένα θεωρούνται σταθερά. Για πρακτικούς λόγους, χρησιμοποιείται συχνά η λογαριθμική πιθανοφάνεια. Η συνάρτηση λογαριθμικής πιθανοφάνειας, η οποία εξάγεται από τη λογαριθμική πιθανότητα της Εξίσωσης 6.4, δίνεται

στην Εξίσωση 6.6. Αξίζει να σημειωθεί ότι οι τιμές των παραμέτρων που μεγιστοποιούν τη λογαριθμική πιθανοφάνεια επίσης μεγιστοποιούν την πιθανοφάνεια δεδομένου ότι η συνάρτηση λογάριθμος είναι μονότονα αύξουσα.

$$\text{πιθανοφάνεια}(\Theta|X) = L(\Theta|X) = \prod_{i=1}^{m} \frac{1}{2\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad (6.5)$$

$$\log \text{πιθανοφάνεια}(\Theta|X) = l(\Theta|X) = -\sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2} - 0.5m \log 2\pi - m \log \sigma \quad (6.6)$$

6.4 Ο ΕΜ αλγόριθμος

Μπορεί επίσης να χρησιμοποιηθεί η προσέγγιση της μέγιστης πιθανοφάνειας για να εκτιμηθούν οι παράμετροι ενός συνδυαστικού μοντέλου. Στην απλούστερη περίπτωση, γίνεται γνωστό ποια αντικείμενα προέρχονται από ποιες κατανομές και η κατάσταση ανάγεται σε μία εκτίμηση των παραμέτρων μιας απλής κατανομής δοθέντων των δεδομένων, από αυτήν την κατανομή. Για τις πιο κοινές κατανομές, οι εκτιμήσεις μέγιστης πιθανοφάνειας των παραμέτρων υπολογίζονται από απλές σχέσεις που εμπεριέχουν τα δεδομένα.

Σε μια πιο γενική (και πιο ρεαλιστική) κατάσταση, δεν είναι γνωστό ποια σημεία έχουν παραχθεί από ποια κατανομή. Επομένως, δεν είναι δυνατός ο άμεσος υπολογισμός της πιθανότητας κάθε σημείου δεδομένων και επομένως, θα φαινόταν ότι δεν είναι δυνατή η χρήση της αρχής της μέγιστης πιθανοφάνειας για την εκτίμηση των παραμέτρων. Η λύση στο πρόβλημα αυτό είναι ο **αλγόριθμος Expectation Maximization-EM** [6], ο οποίος δίνεται στον Αλγόριθμο 6.1. Συνοπτικά, δοθείσας μιας τυχαίας εκτίμησης των τιμών των παραμέτρων, ο αλγόριθμος ΕΜ υπολογίζει την πιθανότητα κάθε σημείου να ανήκει σε κάθε κατανομή, και έπειτα χρησιμοποιεί αυτές τις πιθανότητες για να υπολογίσει μια νέα εκτίμηση των παραμέτρων (αυτές οι παράμετροι είναι εκείνες που μεγιστοποιούν την πιθανοφάνεια). Αυτή η επαναληπτική διαδικασία συνεχίζεται μέχρι οι εκτιμήσεις των παραμέτρων είτε να μην αλλάζουν είτε να αλλάζουν ελάχιστα. Επομένως, πάλι χρησιμοποιείται η εκτίμηση μέγιστης πιθανοφάνειας αλλά μέσω επαναληπτικής αναζήτησης.

Αλγόριθμος 6.1 Ο Expectation Maximization (EM) Αλγόριθμος

---

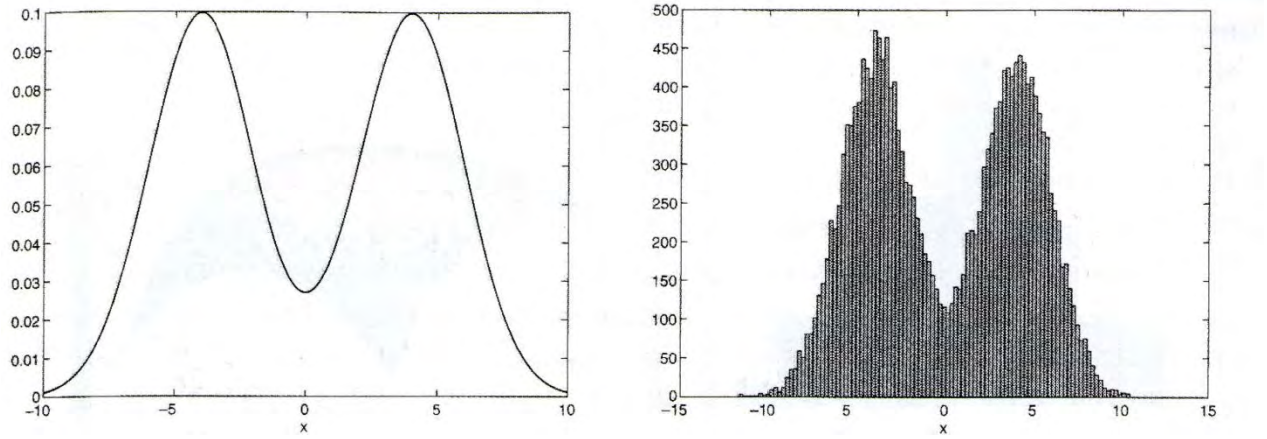
1. Επίλεξε ένα αρχικό σύνολο παραμέτρων του μοντέλου.
   (όπως με τους Κ-μέσους, αυτό μπορεί να γίνει τυχαία, ή με ένα πλήθος τρόπων).
2. **Επανέλαβε**
3. **Βήμα Προσδοκίας**. Για κάθε αντικείμενο, υπολόγισε την πιθανότητα κάθε αντικειμένου να ανήκει σε κάθε κατανομή, δηλαδή υπολόγισε την ποσότητα *prob*(κατανομή j | x_i,Θ).

4. **Βήμα Μεγιστοποίησης**. Δοθέντων των πιθανοτήτων από το βήμα προσδοκίας, βρες τις νέες εκτιμήσεις των παραμέτρων που μεγιστοποιούν την αναμενόμενη πιθανοφάνεια.
5. **Μέχρι** να μην αλλάζουν οι παράμετροι.
   (Εναλλακτικά, σταμάτα αν η αλλαγή των παραμέτρων είναι κάτω από ένα κατώφλι).

Ο αλγόριθμος ΕΜ είναι παρόμοιος με τον αλγόριθμο των Κ – μέσων που παρουσιάστηκε στον Αλγόριθμο 4.1. Πράγματι, ο αλγόριθμος των Κ – μέσων για Ευκλείδια δεδομένα είναι μια ειδική περίπτωση του αλγορίθμου ΕΜ για σφαιρικές κατανομές Gauss με ίσους πίνακες συνδιακύμανσης, αλλά διαφορετικούς μέσους. Το βήμα προσδοκίας αντιστοιχεί στο βήμα εκχώρησης κάθε αντικειμένου σε μια συστάδα του αλγορίθμου των Κ – μέσων. Αντίθετα, κάθε αντικείμενο εκχωρείται σε κάθε συστάδα (κατανομή) με κάποια πιθανότητα. Το βήμα μεγιστοποίησης αντιστοιχεί στον υπολογισμό των κέντρων βάρους της συστάδας. Αντίθετα, όλες οι παράμετροι των κατανομών, καθώς επίσης και οι παράμετροι βαρών, επιλέγονται να μεγιστοποιούν την πιθανοφάνεια. Αυτή η διαδικασία είναι συχνά απλή, καθώς οι παράμετροι υπολογίζονται συνήθως χρησιμοποιώντας τις σχέσεις που εξάγονται από την εκτίμηση της μέγιστης πιθανοφάνειας. Για παράδειγμα, για μία μονή κατανομή Gauss, η εκτίμηση μέγιστης πιθανοφάνειας του μέσου είναι ο μέσος των αντικειμένων στην κατανομή. Στο περιβάλλον των συνδυαστικών μοντέλων και του αλγορίθμου ΕΜ, ο υπολογισμός του μέσου τροποποιείται για να λάβει υπ’ όψιν το γεγονός ότι κάθε αντικείμενο ανήκει σε μία κατανομή με μία συγκεκριμένη πιθανότητα. Αυτό φαίνεται καλύτερα στο ακόλουθο παράδειγμα.

6.4.1 Παράδειγμα αλγορίθμου ΕΜ

Το παράδειγμα αυτό δείχνει τον τρόπο με τον οποίο λειτουργεί ο αλγόριθμος ΕΜ, όταν εφαρμόζεται στα δεδομένα του Σχήματος 6.1. Για να παραμείνει το παράδειγμα όσο το δυνατόν πιο απλό, ας υποτεθεί ότι είναι γνωστό ότι η τυπική απόκλιση των δύο κατανομών είναι ίση με 2.0 και ότι τα σημεία έχουν παραχθεί με ίση πιθανότητα και από τις δύο κατανομές. Θα γίνεται αναφορά στην αριστερή και δεξιά κατανομή ως κατανομή 1 και 2 αντίστοιχα.



(α) Συνάρτηση πυκνότητας πιθανότητας για το συνδυαστικό μοντέλο

(β) 20.000 σημεία που παράγονται από το συνδυαστικό μοντέλο

Σχήμα 6.1: Συνδυαστικό μοντέλο που αποτελείται από δύο κανονικές κατανομές με μέσους -4 και 4, αντίστοιχα. Και οι δύο κατανομές έχουν τυπική απόκλιση ίση με 2.

Ο αλγόριθμος ΕΜ ξεκινάει με μία τυχαία εκχώρηση τιμών για τα μ_1 και μ_2, έστω μ_1 = - 2 και μ_2 = 3. Επομένως, οι αρχικές παράμετροι θ(μ,σ) για τις δύο κατανομές είναι, αντίστοιχα, θ_1 = (-2,2) και θ_2 = (3,2). Το σύνολο των παραμέτρων για ολόκληρο το συνδυαστικό μοντέλο είναι Θ = { θ_1, θ_2 }.

Για το βήμα προσδοκίας του ΕΜ, επιθυμείται να υπολογιστεί η πιθανότητα ενός σημείου να προέρχεται από μια συγκεκριμένη κατανομή, δηλαδή να υπολογιστούν οι πιθανότητες *prob*(κατανομή 1| x_i,Θ) και *prob*(κατανομή 2| x_i,Θ). Αυτές οι τιμές μπορούν να εκφραστούν από την ακόλουθη εξίσωση , η οποία είναι μια άμεση εφαρμογή του γνωστού κανόνα του Bayes:

$$prob(\text{κατανομή } j \mid x_i,\Theta) = \frac{0.5\ \mathrm{prob}(x_i|\theta_j)}{0.5\ \mathrm{prob}(x_i|\theta_1) + 0.5\ \mathrm{prob}(x_i|\theta_2)}\ , \qquad (6.7)$$

όπου 0.5 είναι η πιθανότητα (βάρος) κάθε κατανομής και το j είναι 1 ή 2.

Για παράδειγμα, έστω ότι ένα από τα σημεία είναι 0. Χρησιμοποιώντας την συνάρτηση πυκνότητας του Gauss, υπολογίζεται ότι $\mathrm{prob}(0|\theta_1) = 0.12$ και $\mathrm{prob}(0|\theta_2)$ = 0.06. Χρησιμοποιώντας αυτές τις τιμές και την Εξίσωση 6.7, προκύπτει ότι *prob*(κατανομή 1|0, Θ) = 0.12 / (0.12 + 0.06) = 0.66 και *prob*(κατανομή 2|0, Θ) = 0.06 / (0.12 + 0.06) = 0.33. Αυτό σημαίνει ότι το σημείο 0 είναι δύο φορές πιο πιθανό να ανήκει στην κατανομή 1 σε σχέση με την κατανομή 2 με βάση τις τρέχουσες υποθέσεις για τις τιμές των παραμέτρων.

Αφού υπολογιστούν οι πιθανότητες της ιδιότητας μέλους συστάδας για όλα τα 20000 σημεία, υπολογίζονται οι νέες εκτιμήσεις για τα μ_1 και μ_2 (χρησιμοποιώντας τις Εξισώσεις 6.8 και 6.9) στο βήμα μεγιστοποίησης του αλγορίθμου ΕΜ. Ας παρατηρηθεί ότι η νέα εκτίμηση για

το μέσο μιας κατανομής είναι απλά ένας σταθμισμένος μέσος των σημείων, όπου τα βάρη είναι οι πιθανότητες των σημείων να ανήκουν στην κατανομή, δηλαδή οι τιμές *prob*(κατανομή j | x_i).

$$\mu_1 = \sum_{i=1}^{20000} x_i \frac{\text{prob}(\text{κατανομή } 1|x_i,\Theta)}{\sum_{i=1}^{20000} \text{prob}(\text{κατανομή } 1|x_i,\Theta)} \quad (6.8)$$

$$\mu_2 = \sum_{i=1}^{20000} x_i \frac{\text{prob}(\text{κατανομή } 2|x_i,\Theta)}{\sum_{i=1}^{20000} \text{prob}(\text{κατανομή } 2|x_i,\Theta)} \quad (6.9)$$

Τα ίδια βήματα επαναλαμβάνονται μέχρι οι εκτιμήσεις των μ_1 και μ_2 είτε να μην αλλάζουν είτε να αλλάζουν πολύ λίγο. Ο Πίνακας 4 δίνει τις πρώτες επαναλήψεις του αλγορίθμου ΕΜ όταν εφαρμόζεται στο σύνολο των 20000 σημείων. Γι' αυτά τα δεδομένα, γίνεται γνωστό ποια κατανομή παρήγαγε ποιο σημείο, οπότε μπορεί να υπολογιστεί ο μέσος των σημείων από κάθε κατανομή. Οι μέσοι είναι μ_1 = -3.98 και μ_2 = 4.03.

Επανάληψη	μ_1	μ_2
0	-2.00	3.00
1	-3.74	4.10
2	-3.94	4.07
3	-3.97	4.04
4	-3.98	4.03
5	-3.98	4.03

Πίνακας 4: Πρώτες επαναλήψεις του αλγορίθμου ΕΜ στο σύνολο των 20000 σημείων.

6.4.2 Πλεονεκτήματα και περιορισμοί με τη χρήση του αλγορίθμου ΕΜ

Η εύρεση των συστάδων με μοντελοποίηση των δεδομένων, χρησιμοποιώντας συνδυαστικά μοντέλα και εφαρμόζοντας τον αλγόριθμο ΕΜ για την εκτίμηση των παραμέτρων αυτών των μοντέλων, έχει μία ποικιλία πλεονεκτημάτων και μειονεκτημάτων.

Θεωρώντας την αρνητική πλευρά, ο αλγόριθμος ΕΜ μπορεί να είναι αργός, δεν είναι πρακτικός για μοντέλα με μεγάλο πλήθος συνιστωσών και δε λειτουργεί καλά όταν οι συστάδες περιέχουν μόνο λίγα σημεία ή αν τα σημεία δεδομένων είναι σχεδόν συγγραμικά. Αντιμετωπίζει επίσης πρόβλημα στην εκτίμηση του πλήθους των συστάδων, ή πιο γενικά, στην επιλογή της ακριβής μορφής του μοντέλου που θα χρησιμοποιηθεί. Αυτό το πρόβλημα τυπικά έχει αντιμετωπιστεί εφαρμόζοντας μία προσέγγιση του Bayes, η οποία, γενικά μιλώντας, δίνει τις πιθανότητες ενός μοντέλου σε σχέση με ένα άλλο, με βάση μία εκτίμηση που λαμβάνεται από τα δεδομένα. Τα συνδυαστικά μοντέλα μπορεί επίσης να έχουν δυσκολία με το θόρυβο και τις

ακραίες τιμές, παρά το γεγονός ότι έχει γίνει προσπάθεια για να αντιμετωπιστεί αυτό το πρόβλημα.

Στη θετική πλευρά, τα συνδυαστικά μοντέλα είναι πιο γενικά, για παράδειγμα, από τους K – μέσους [10], επειδή μπορούν να χρησιμοποιήσουν διαφόρων τύπων κατανομές. Ως αποτέλεσμα, τα συνδυαστικά μοντέλα (με βάση τις κατανομές Gauss) μπορούν να βρουν συστάδες διαφορετικών μεγεθών και ελλειπτικών σχημάτων. Επίσης, μία προσέγγιση βάσει του μοντέλου, παρέχει έναν συστηματικό τρόπο εξάλειψης ενός μέρους της πολυπλοκότητας που σχετίζεται με τα δεδομένα. Για να δούμε τα υποδείγματα στα δεδομένα, είναι συχνά απαραίτητο να απλοποιηθούν τα δεδομένα, και η προσαρμογή τους σε ένα μοντέλο είναι ένας καλός τρόπος να γίνει αυτό, αν το μοντέλο ταιριάζει καλά στα δεδομένα. Επιπλέον, είναι εύκολο να χαρακτηριστούν οι συστάδες που παράγονται, δεδομένου ότι μπορούν να περιγραφούν από ένα μικρό πλήθος παραμέτρων. Τέλος, πολλά σύνολα δεδομένων είναι πράγματι το αποτέλεσμα τυχαίων διαδικασιών, και επομένως, θα πρέπει να ικανοποιούν τις στατιστικές υποθέσεις αυτών των μοντέλων.

Βιβλιογραφία

[1] http://en.wikipedia.org/wiki/Mixture_model#Expectation_maximization_.28EM.29.

[2] http://en.wikipedia.org/wiki/Maximum_likelihood.

[3] http://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm.

[4] https://files.nyu.edu/mrg217/public/mle_introduction1.pdf.

[5] http://statweb.stanford.edu/~susan/courses/s200/lectures/lect11.pdf.

[6] M. E. J. Newman and E. A. Leicht, Mixture models and exploratory analysis in networks, Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA.

[7] Chuong B Do & Serafim Batzoglou, What is the expectation maximization algorithm?, 2008.

[8] Dr. Simon J.D. Prince, The Expectation-Maximization (EM) Algorithm, Dept. of Computer Science, University College London.

[9] Douglas Reynolds, Gaussian Mixture Models, MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA.

[10] http://statweb.stanford.edu/~tibs/stat315a/LECTURES/em.pdf.

ΚΕΦΑΛΑΙΟ 7 ΣΥΜΠΕΡΑΣΜΑΤΑ

Ο σκοπός της συγκεκριμένης διπλωματικής ήταν να δημιουργήσει έναν μικρό οδηγό για το πρόβλημα του εντοπισμού κοινότητας σε ένα πολύπλοκο δίκτυο με βάση πάντα το τι θεωρεί ο κάθε αναλυτής ως κοινότητα, και επομένως ποιον αλγόριθμο θα χρησιμοποιήσει ώστε να εντοπίσει τις κοινότητες σε αυτό το δίκτυο.

Πρώτα απ' όλα αντιμετωπίστηκε το πρόβλημα της έλλειψης ενός καθολικά αποδεκτού ορισμού του τι είναι μια κοινότητα. Όπως επισημαίνεται και από τον Fortunato [1], η έλλειψη ενός θεωρητικού πλαισίου έχει μερικές σημαντικές συνέπειες όχι μόνο στην καθ΄εαυτού προσέγγιση του εντοπισμού κοινότητας (δηλαδή, δεδομένης μιας διαφωνίας στον ορισμό της «κοινότητας», πως μπορεί να εξαχθεί μία κοινότητα από το δίκτυο) αλλά και σε άλλες πτυχές του προβλήματος. Μία από αυτές τις πτυχές είναι, για παράδειγμα, η αξιολόγηση ενός αλγορίθμου.

Έχει προταθεί ένας διαφορετικός ορισμός της κοινότητας, και σε αυτή τη βάση δημιουργείται μια νέα κατηγοριοποίηση των μεθόδων του εντοπισμού κοινότητας που βασίζονται στις σχέσεις του κάθε ορισμού της κοινότητας, χρησιμοποιώντας το γενικό ορισμό. Έχουν αξιολογηθεί οι προσεγγίσεις σύμφωνα με τον ορισμό των γενικών κατηγοριών, όπως το χαρακτηριστικό απόσταση, η εσωτερική πυκνότητα, και η συνδυαστική συσταδοποίηση. Η κατηγοριοποίηση αυτή αποτελεί μία προτεινόμενη απάντηση στα προβλήματα που επισημαίνονται από τον Fortunato. Κάθε κύρια μέθοδος στη συνέχεια παρουσιάζεται συνοπτικά, η πολυπλοκότητά της και τα δυνατά και αδύναμα σημεία της κατηγορίας στην οποία ανήκει.

Ένα σημαντικό πρόβλημα που έχει εντοπιστεί είναι η ανάγκη για μια εκτενή μελέτη της επικάλυψης μεταξύ των ορισμών της κοινότητας. Όπως αναφέρεται στην Ενότητα 3.1, υπάρχουν αρκετές πολύπλοκες συνδέσεις μεταξύ των διαφορετικών ορισμών και διαφορετικών αλγορίθμων. Θα ήταν άξιο αναφοράς να δημιουργηθεί μια ακριβής γραφική αναπαράσταση αυτής της επικάλυψης στην οποία οι κόμβοι είναι οι συνδεδεμένοι αλγόριθμοι, εάν μοιράζονται μέρος του ορισμού τους της κοινότητας, μερικών χαρακτηριστικών της εισόδου/εξόδου, ή ορισμένες συναρτήσεις ποιότητας. Αυτό το πολυδιάστατο πολύπλοκο δίκτυο θα μπορούσε να μελετηθεί έτσι ώστε να υπάρχει μια πιο σαφή και λεπτομερή άποψη στο πρόβλημα του εντοπισμού μιας κοινότητας.

Μία άλλη συμβολή της παρούσας εργασίας είναι η ενσωμάτωση των σημαντικά καινοτόμων χαρακτηριστικών ενός αλγορίθμου διαμέρισης ενός γράφου η οποία ίσως δεν έχει εξεταστεί σε άλλες παρόμοιες εργασίες. Ο ορισμός των διαφορετικών χαρακτηριστικών είναι κρίσιμη, διότι προφανώς δεν υπάρχει η «τέλεια μέθοδος». Ωστόσο, οι μέθοδοι που είναι ή δεν είναι σε θέση να εξετάσουν πολλαπλές διαστάσεις, οι αλγόριθμοι που αντιμετωπίζουν ή όχι επικαλυπτόμενες κοινότητες, και ούτω κάθε εξής, μπορούν να κατηγοριοποιηθούν όπως εξετάστηκε παραπάνω. Επιλέχθηκαν να περιλαμβάνονται νέα χαρακτηριστικά, όπως είναι η πολυδιαστατικότητα, καθώς προσθέτουν μία αναλυτική δύναμη που περιγράφει καλύτερα τα διάφορα φαινόμενα του πραγματικού κόσμου.

Ένα άλλο ανοιχτό ζήτημα είναι να οριστεί και να προβλεφθεί ποια θα είναι τα πιο σημαντικά χαρακτηριστικά στο μέλλον. Υπάρχει ένα ενδιαφέρον ιδιαίτερα στην πολυδιαστατικότητα [2,3,4,5,6], που εκλαμβάνεται ως ένα χαρακτηριστικό που είναι μέρος της λύσης και όχι μόνο ως μία υπό επεξεργασία είσοδος. Με άλλα λόγια, να μην θεωρείται η πολυδιαστατικότητα μόνο ως είσοδος, αλλά και να εξαγάγονται πραγματικές πολυδιάστατες κοινότητες. Ένα άλλο ενδιαφέρον χαρακτηριστικό θα μπορούσε να είναι η ταυτόχρονη παρουσία μιας ιεραρχικής και επικαλυπτόμενης οργάνωσης της δομής της κοινότητας, δεδομένου ότι τα δύο αυτά χαρακτηριστικά δεν θεωρούνται πλέον ως αμοιβαία αποκλειόμενα [7].

Βιβλιογραφία

[1] S. Fortunato, Community detection in graphs, Phys Rep 486(3–5) (2010), 175–174.

[2] D. Cai, Z. Shao, X. He, X. Yan, and J. Han, Community mining from multi-relational networks, In Proceedings of the 2005 European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05), (Porto, Portugal), 2005.

[3] L. Tang and H. Liu, Scalable learning of collective behavior based on sparse social dimensions, In CIKM, 2009.

[4] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J. Onnela, Community Structure in Time-Dependent, Multiscale, and Multiplex Networks, Science 328 (2010), 876–878.

[5] M. Berlingerio, M. Coscia, and F. Giannotti, Finding and characterizing communities in multidimensional networks, ASONAM, Kaohsiung, Taiwan, IEEE, 2011.

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

[7] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Link communities reveal multi-scale complexity in networks, Nature 466 (2010), 761–764.