ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάλυση Μεγάλων Δεδομένων Διαδικτύου-των-Αντικειμένων με Αλγόριθμους Ανίχνευσης Κοινοτήτων σε Δίκτυα Υπερβολικών Χώρων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τσιτσεκλής Χ. Κωνσταντίνος

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάλυση Μεγάλων Δεδομένων Διαδικτύου-των-Αντικειμένων με Αλγόριθμους Ανίχνευσης Κοινοτήτων σε Δίκτυα Υπερβολικών Χώρων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τσιτσεκλής Χ. Κωνσταντίνος

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 9η Οκτωβρίου 2017.
Αθήνα, Οκτώβρης 2017

.....	.....	.....
Παπαβασιλείου Συμεών	*Βαρβαρίγου Θεοδώρα*	*Ρουσσάκη Ιωάννα*
Καθηγητής Ε.Μ.Π.	*Καθηγήτρια Ε.Μ.Π.*	*Επίκουρη Καθηγήτρια Ε.Μ.Π.*

(Υπογραφή)

.....

ΚΩΝΣΤΑΝΤΙΝΟΣ ΧΡ. ΤΣΙΤΣΕΚΛΗΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

ΠΕΡΙΛΗΨΗ

Ο σκοπός της παρούσας Διπλωμάτικης Εργασίας είναι η ανάπτυξη μιας αποδοτικής μεθοδολογίας Ανίχνευσης Κοινοτήτων για την ανάλυση Μεγάλων Δεδομένων που προκύπτουν από τοπολογίες Διαδικτύου-των-Αντικειμένων. Για τον σκοπό αυτό γίνεται χρήση δικτύων ενσωματωμένων στον Υπερβολικό Χώρο. Ο χώρος αυτός λόγω των ιδιοτήτων του, επιλέγεται ως ένας «φυσικός» χώρος για την αναπαράσταση Σύνθετων Δικτύων.

Για την ανάπτυξη της μεθοδολογίας γίνεται τροποποίηση του αλγόριθμου Ανίχνευσης Κοινοτήτων των Newman-Girvan. Στον αλγόριθμο των Newman-Girvan γίνεται χρήση της έννοιας της Κεντρικότητας Ενδιαμεσικότητας Ακμών. Στο πλαίσιο της παρούσας εργασίας προτείνεται μια προσέγγιση υπολογισμού της μετρικής αυτής με χρήση υπερβολικής γεωμετρίας. Συγκεκριμένα, προτείνεται ένας προσεγγιστικός αλγόριθμος ο οποίος υπολογίζει τις Κεντρικότητες των Ακμών βασισμένος στις αποστάσεις των κόμβων μετά από ενσωμάτωσή των αντίστοιχων δικτύων στον Υπερβολικό Χώρο. Η νέα αυτή μετρική ονομάζεται Υπερβολική Κεντρικότητα Ενδιαμεσικότητας Ακμής (Υ.Κ.Ε.Α.). Επειδή ο υπολογισμός της Κεντρικότητας Ενδιαμεσικότητας (είτε κορυφής, είτε ακμής) βασίζεται στον υπολογισμό συντομότερων μονοπατιών επιλέγεται για την ενσωμάτωση του δικτύου η Ενσωμάτωση Rigel. Η ενσωμάτωση αυτή τοποθετεί τους κόμβους με τέτοιο τρόπο στον Υπερβολικό Χώρο ώστε η απόσταση δυο κόμβων να προσεγγίζει την απόσταση τους στο δίκτυο. Στη συνέχεια, η μετρική αυτή χρησιμοποιείται στον προτεινόμενο αλγόριθμο Ανίχνευσης Κοινοτήτων Hyperbolic Newman-Girvan (HNG).

Η απόδοση του αλγόριθμου HNG μελετήθηκε τόσο για Γράφους Εγγύτητας, δίκτυα δηλαδή που προκύπτουν από την ένωση σημείων παρατηρήσεων όσο και για Κοινωνικά Δίκτυα τύπου μικρής-κλίμακας, αλλά και για διάφορα μοντέλα κατασκευής Σύνθετων Δικτύων. Ο αλγόριθμος ολοκληρώθηκε σε μικρότερο χρονικό διάστημα από τον αλγόριθμο των Newman-Girvan στην περίπτωση των Γράφων Εγγύτητας που εξετάστηκαν, καθώς επίσης έδωσε ικανοποιητικά αποτελέσματα Αρθρωτότητας για τα υπόλοιπα δίκτυα. Τα αποτελέσματα αυτά παρουσιάζονται αναλυτικά τόσο με πίνακες όσο και με διαγράμματα στο αντίστοιχο κεφάλαιο της Εργασίας (Κεφάλαιο 8).

Λέξεις Κλειδιά: Ανίχνευση Κοινοτήτων, Ομαδοποίηση, Μεγάλα Δεδομένα, Διαδίκτυο-των-Αντικειμένων, Κεντρικότητα Ενδιαμεσικότητας Ακμής, Υπερβολικός Χώρος, Υπερβολική Γεωμετρία, Κοινωνικά Δίκτυα, Ενσωμάτωση Δικτύου.

ABSTRACT

The purpose of this Diploma Thesis is the development of an efficient method for Community Detection to be used in the analysis of Big Data sets that stem from Internet-Of-Things topologies. For this purpose, network graph embedding in the Hyperbolic Space is employed. This space is selected based on its properties as a "natural" space for the representation of Complex Networks.

For the development of the method, the Newman-Girvan Community Detection algorithm is modified. The Newman-Girvan algorithm makes use of the notion of Edge Betweenness Centrality. In this Thesis an approximation for the computation of this network metric based on hyperbolic geometry is proposed. Specifically, an approximation algorithm is proposed, which computes the Edge Betweenness Centrality based on node distances after the corresponding graphs are embedded in Hyperbolic Space. This new metric is named Hyperbolic Edge Betweenness Centrality (HEBC). As the computation of Betweenness Centrality (either of nodes or edges) is based on the computation of shortest paths, Rigel Embedding is chosen for the stage of the network embedding. This embedding algorithm assigns node coordinates in a way that the distance between two nodes in the Hyperbolic Space approximates the length of the shortest path joining them in the original network. The HEBC metric is then used in the proposed Community Detection algorithm called Hyperbolic Newman-Girvan.

The efficiency of the proposed Community Detection algorithm was studied for Proximity Graphs, meaning graphs generated by linking observation points, as well as for small-scale Social Networks but also for artificial Complex Networks generated by various models. The proposed algorithm completed faster than the traditional Newman-Girvan algorithm for the Proximity Graphs studied and achieved good Modularity scores for the rest of the networks. These results are presented thoroughly using tables and diagrams in the respective chapter of the Thesis (Chapter 8).

Keywords: Community Detection, Clustering, Big Data, Internet-Of-Things, Edge Betweenness Centrality, Hyperbolic Space, Hyperbolic Geometry, Online Social Networks, Network Embedding.

Ευχαριστίες

Καταρχήν, θα ήθελα να ευχαριστήσω τον Καθηγητή της σχολής «Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών», κύριο Συμεών Παπαβασιλείου για την επίβλεψη της διπλωματικής μου εργασίας καθως και για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα.

Ευχαριστώ επίσης ιδιαίτερα τον μεταδιδακτορικό ερευνητή Βασίλη Καρυώτη για την πολύτιμη βοήθεια του, το αμείωτο ενδιαφέρον του αλλά και τον χρόνο που διέθεσε για την ολοκλήρωση της εργασίας μου. Ακόμα, ευχαριστώ τον Κωνσταντίνο Σωτηρόπουλο για την βοήθεια και τις προτάσεις του σε διάφορες φάσεις της εργασίας.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ στον πατέρα μου Χρήστο, στην μητέρα μου Ελένη και στην αδερφή μου Ευαγγελία γιατί η ολόπλευρη στήριξή τους έπαιξε καθοριστικό ρόλο στις σπουδές μου.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1

Εισαγωγή

1.1 Ομαδοποίηση Σε Σύνθετα Δίκτυα και Μεγάλα Δεδομένα

Η ραγδαία ανάπτυξη των δικτύων επικοινωνιών, του Διαδικτύου, των Κοινωνικών Δικτύων, των υποδομών δικτύωσης αλλά και η αύξηση του πλήθους των διασυνδεδεμένων συσκευών που οδηγεί στην ολοένα και μεγαλύτερη ανάπτυξη του λεγόμενου Διαδικτύου-των–Αντικειμένων, ΔτΑ, (Internet-of-Things, IoT), είναι μερικοί από τους λόγους που η μελέτη και η ανάλυση των δικτύων καθίσταται αναγκαία για ένα πλήθος εφαρμογών. Τα δίκτυα τα οποία προκύπτουν ως αποτέλεσμα των παραπάνω αλλά και άλλων παραγόντων ορίζονται ως Σύνθετα Δίκτυα (Complex Networks) και η μελέτη τους εμπίπτει στην επιστήμη της «Ανάλυσης Κοινωνικών Δικτύων» ή πιο γενικά της «Ανάλυσης Σύνθετων Δικτύων», για την οποία παρά το «νεαρό» της ηλικίας της υπάρχει ήδη σημαντικός όγκος βιβλιογραφίας. Ενδεικτικά αναφέρονται τα [Kad11], [BEJ13], [KSP13] και άλλα. Τέτοια δίκτυα παρουσιάζουν ενδιαφέρουσες τοπολογικές ιδιότητες οι οποίες δεν απαντώνται σε πολύ μικρά ή σε τετριμμένα δίκτυα. Μια τέτοια ιδιότητα, η οποία έχει απασχολήσει πολλούς επιστήμονες, είναι η ιδιότητα που έχουν οι κόμβοι σε τέτοια δίκτυα να σχηματίζουν ομάδες. Οι ομάδες αυτές αποτελούν σύνολα κόμβων οι οποίοι έχουν περισσότερες κοινές συνδέσεις μεταξύ τους από ότι με κόμβους που δεν ανήκουν στην ομάδα. Οι ομάδες αυτές, συχνά αναφέρονται ως Κοινότητες.

Η αναζήτηση Κοινοτήτων είναι ένας επιστημονικός τομέας που δεν περιορίζεται στα δίκτυα όπου έχουν προταθεί αρκετοί αλγόριθμοι όπως ο αλγόριθμος Ομαδοποίησης με Αλυσίδες Markov [Don00], Φασματικοί αλγόριθμοι [DoM04], μέθοδοι όπως ο αλγόριθμος των Newman-Girvan και μέθοδοι Μεγιστοποίησης της Αρθρωτότητας που αναλύονται στην συνέχεια στην εργασία. Επεκτείνεται και σε σημεία στον χώρο, χρονοσειρές, και άλλα [EKS96], [HaW79], [Lia05]. Είναι γεγονός ότι όλο και περισσότερο αυξάνεται ο ρυθμός παραγωγής δεδομένων. Τα δεδομένα που παράγονται από τις εφαρμογές, από δίκτυα έξυπνων συσκευών ή από ολόκληρες τοπολογίες αισθητήρων παρουσιάζουν και αυτά την τάση να σχηματίζουν ομάδες, με την έννοια ότι παρατηρήσεις με παρεμφερείς τιμές υποδηλώνουν συνάφεια των φυσικών συσκευών που τις μέτρησαν, είτε στον χώρο, είτε στον χρόνο. Παρουσιάζεται λοιπόν και εδώ η ανάγκη για την ανακάλυψη των συναφειών αυτών, η ανάγκη δηλαδή για ομαδοποίηση των παρατηρήσεων.

Το μεγάλο μέγεθος των Σύνθετων Δικτύων αλλά και ο όγκος των δεδομένων που παράγονται από τοπολογίες σαν αυτές που περιγράφηκαν στην προηγούμενη παραγράφο που χαρακτηρίζονται ως Μεγάλα Δεδομένα, απαιτούν τεχνικές ανάλυσης ικανές να διαχειριστούν τον μεγάλο όγκο δεδομένων αλλά και να παράγουν αποτελέσματα σε αποδεκτό χρονικό διάστημα. Στην προσπάθεια για γρήγορη ανάλυση, προτάθηκε μεταξύ άλλων η απονομή συντεταγμένων σε κάθε παρατήρηση ή κόμβο σε έναν Υπερβολικό Χώρο συντεταγμένων. Ο συγκεκριμένος μετρικός χώρος προτάθηκε λόγω της ιδιότητας του να μπορεί να αναπαριστά με καλή ακρίβεια τις σχέσεις μεταξύ των κόμβων ή την εγγύτητα μεταξύ των παρατηρήσεων.

1.2 Αντικείμενο Διπλωματικής και Συνεισφορά

Στην παρούσα εργασία προτείνουμε ένα ενιαίο πλαίσιο τόσο για την ανάλυση τοπολογιών Σύνθετων Δικτύων όσο και για την ανάλυση Μεγάλων Δεδομένων. Για την τελευταία χρησιμοποιούμε μια μέθοδο για να παράγουμε από τα δεδομένα γράφους εγγύτητας. Τέτοιοι γράφοι σχηματίζονται θεωρώντας κάθε σημείο στο χώρο ως κόμβο ο οποίος συνδέεται με άλλους αν ικανοποιείται κάποια συνθήκη. Στην συνέχεια, ενσωματώνουμε το δίκτυο στον Υπερβολικό Χώρο κάνοντας χρήση της Ενσωμάτωσης Rigel. Η επιλογή του Υπερβολικού Χώρου έγινε καθώς από τις μελέτες άλλων επιστημόνων φαίνεται η υπερβολική γεωμετρία είναι εκείνη που εξηγεί καλύτερα την εξέλιξη των Σύνθετων Δικτύων. Η συγκεκριμένη ενσωμάτωση δίνει κατάλληλες συντεταγμένες στους κόμβους του δικτύου έτσι ώστε η μεταξύ τους απόσταση να είναι με μικρό σφάλμα κοντά στο μήκος του συντομότερου μονοπατιού που τους συνδέει. Αυτό μας εξασφαλίζει έναν γρήγορο υπολογισμό του μήκους του συντομότερου μονοπατιού που συνδέει δυο κόμβους σε χρόνο $O(1)$, αντί να χρειαστεί να εκτελούμε κάθε φορά κάποιον από τους παραδοσιακούς αλγόριθμους εύρεσης συντομότερου μονοπατιού ή να διατηρούμε εκτενείς πίνακες στη μνήμη.

Αξιοποιώντας λοιπόν το γεγονός αυτό, προτείνουμε μια μέθοδο για τον προσεγγιστικό υπολογισμό της Κεντρικότητας Ενδιαμεσικότητας Ακμής (Κ.Ε.Α) κάνοντας χρήση της υπερβολικής γεωμετρίας. Τα πειράματα που παρουσιάζονται παρακάτω στην εργασία αποδεικνύουν ότι σε συγκεκριμένους τύπους δικτύων η προσέγγιση είναι αρκετά ακριβής (~80% σε δίκτυα Ελεύθερης Κλίμακας για τις δέκα ακμές με την υψηλότερη τιμή Κεντρικότητας).

Στην συνέχεια, τροποποιούμε με τέτοιο τρόπο τον αλγόριθμο Ομαδοποίησης των Newman-Girvan ώστε να χρησιμοποιεί την Υπερβολική Κεντρικότητα Ενδιαμεσικότητας Ακμής. Ακόμα, σε αντίθεση με τον υπάρχοντα αλγόριθμο που αφαιρεί την ακμή με την υψηλότερη τιμή Κ.Ε.Α. εμείς προτείνουμε την αφαίρεση μιας Δέσμης ακμών σε κάθε βήμα, μέχρι να διαμεριστεί το δίκτυο στον αριθμό των κοινοτήτων που έχουμε ορίσει. Η μεθοδολογία αυτή δίνει αρκετά ικανοποιητικά αποτελέσματα και σε πολλές περιπτώσεις είναι ταχύτερη από τον αλγόριθμο Newman-Girvan.

Για τον έλεγχο της μεθόδου μας χρησιμοποιήθηκαν πραγματικά Κοινωνικά Δίκτυα μικρής κλίμακας, δίκτυα που προέκυψαν ως γράφοι εγγύτητας από παρατηρήσεις στον δισδιάστατο Ευκλείδειο χώρο, αλλά και τεχνητά Σύνθετα Δίκτυα μεγαλύτερου μεγέθους για τα οποία χρησιμοποιήθηκαν μερικά από τα γνωστότερα μοντέλα κατασκευής τους.

1.3 Οργάνωση Κειμένου

Η Διπλωματική Εργασία αποτελείται από εννέα Κεφάλαια. Στα Κεφάλαια 2 έως 6 γίνεται η απαιτούμενη εισαγωγή σε έννοιες που χρησιμοποιούνται και παρουσιάζεται το θεωρητικό υπόβαθρο της εργασίας. Στην συνέχεια στο έβδομο Κεφάλαιο παρουσιάζεται η προτεινόμενη μέθοδος, ενώ στο όγδοο παρουσιάζονται τα αποτελέσματα για διάφορους τύπους δικτύων.

Πιο συγκεκριμένα:

- Στο Κεφάλαιο 2 γίνεται η αναγκαία εισαγωγή σε βασικές έννοιες της Θεωρίας Γραφημάτων, του βασικού μαθηματικού εργαλείου αναπαράστασης και ανάλυσης δικτύων. Επίσης, παρουσιάζεται ο συμβολισμός και η ορολογία που ακολουθείται στην συνέχεια της εργασίας.

- Στο Κεφάλαιο 3 παρουσιάζονται τα μοντέλα κατασκευής Σύνθετων Δικτύων που χρησιμοποιούμε στην εργασία και αναλύονται βασικές έννοιες και μετρικές τους. Συγκεκριμένα, στο Κεφάλαιο αυτό γίνεται η ανάλυση της έννοιας της Κεντρικότητας για στοιχεία ενός δικτύου. Ειδικότερα, εδώ γίνεται η πρώτη αναφορά στην Κεντρικότητα Ενδιαμεσικότητας Ακμής που θα μας απασχολήσει και στο Κεφάλαιο 7.

- Στο Κεφάλαιο 4 γίνεται μια σύντομη αναφορά στα Μεγάλα Δεδομένα και παρουσιάζονται δυο τεχνικές για την παραγωγή γράφου από ένα σύνολο παρατηρήσεων.

- Το Κεφάλαιο 5 περιλαμβάνει την παρουσίαση του σημαντικού αλγόριθμου Ομαδοποίησης Newman-Girvan τον οποίο τροποποιούμε στο πλαίσιο της παρούσας εργασίας. Ακόμα παρουσιάζεται η έννοια της Αρθρωτότητας, μιας Ομαδοποίησης, καθώς και ο αλγόριθμος Ομαδοποίησης που εξασφαλίζει την μέγιστη Αρθρωτότητα.

- Στο Κεφάλαιο 6 δίνονται βασικά στοιχεία για την Υπερβολική Γεωμετρία και παρουσιάζονται δυο μοντέλα αναπαράστασης στον Υπερβολικό Χώρο. Στη συνέχεια, εξηγούνται δυο δημοφιλείς μέθοδοι για την Ενσωμάτωση ενός γράφου δικτύου στον Υπερβολικό Χώρο.

- Στο Κεφάλαιο 7 εξηγείται θεωρητικά ο προτεινόμενος αλγόριθμος Ομαδοποίησης, ο οποίος προκύπτει από τροποποίηση του αλγόριθμου Newman-Girvan.

- Στο Κεφάλαιο 8 δίνονται τα αποτελέσματα των πειραμάτων που εκτελέστηκαν για την παρουσίαση των δυνατών σημείων, αλλά και των αδυναμιών της καινούργιας μεθόδου.

- Τέλος, στο Κεφάλαιο 9 συνοψίζονται τα συμπεράσματα της παρούσας εργασίας και προτείνονται πιθανές οδοί μελλοντικής μελέτης.

2

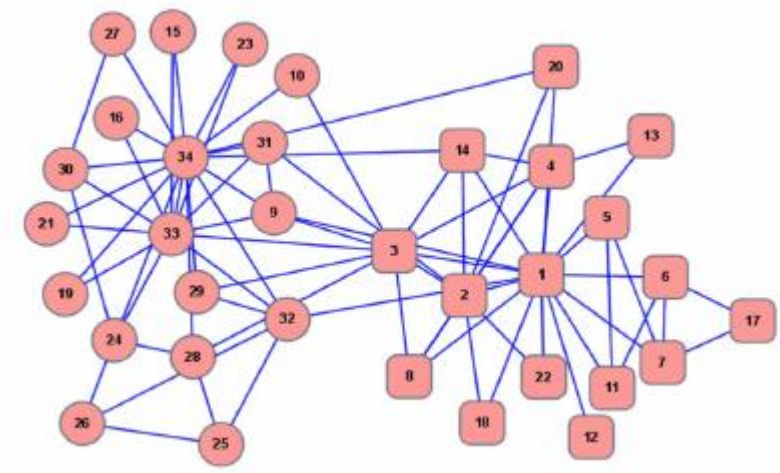
Εισαγωγικές Έννοιες Θεωρίας Γραφημάτων

Η αναπαράσταση σύνθετων συστημάτων όπως κοινωνικών, βιολογικών, ηλεκτρικών δικτύων, δικτύων υπολογιστών και πολλών άλλων χρειάζεται μια κομψή μαθηματική αναπαράσταση που να διευκολύνει την μελέτη τους και να αναπαριστά τις σχέσεις μεταξύ των συστατικών τους μερών με φυσικό τρόπο. Η Θεωρία Γραφημάτων αποτελεί το κατάλληλο μαθηματικό εργαλείο για τον σκοπό αυτό. Στο παρόν κεφάλαιο παρουσιάζονται συνοπτικά μερικές βασικές έννοιες της Θεωρίας Γραφημάτων, οι οποίες χρησιμοποιούνται εκτενώς στο πλαίσιο της παρούσας εργασίας.

Για τον σχεδιασμό γραφημάτων τόσο σε αυτό το Κεφάλαιο όσο και στα υπόλοιπα χρησιμοποιήθηκε η σελίδα http://graphonline.ru/en/.

2.1 Βασικές Έννοιες

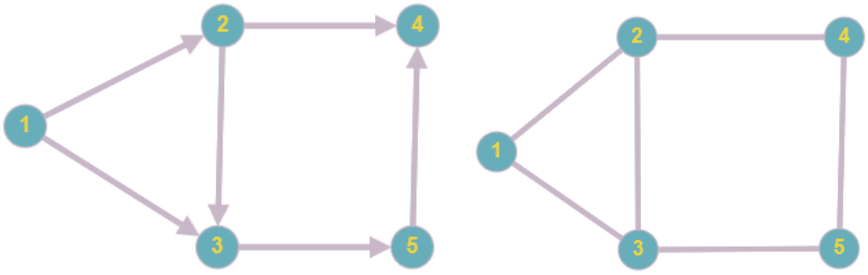
Ένα γράφημα (ή γράφος) ορίζεται ως ένα ζεύγος $G = (V, E)$, όπου $E \subseteq V^2$ και $E \cap V = \emptyset$. Τα στοιχεία του συνόλου V ονομάζονται κορυφές (vertices) (ή κόμβοι) και τα στοιχεία του συνόλου *E,* ακμές (edges ή links). Μια ακμή πρόκειται ουσιαστικά για ένα δισύνολο *(u,v)* με στοιχεία $u, v \in V$, το οποίο υποδηλώνει τη σύνδεση των δυο κορυφών. Μια ακμή προσπίπτει σε μια κορυφή αν η κορυφή αποτελεί το ένα από τα δυο άκρα της.



Σχήμα 1: Παράδειγμα γράφου, πρόκειται για το γράφο του κοινωνικού δικτύου Zachary's Karate Club [Res].

Ένας γράφος $H' = (V', E')$, όπου $V' \subseteq V$ και $E' \subseteq E$, ονομάζεται *υπογράφος* (subgraph) του G. Ένας υπογράφος του οποίου όλοι οι κόμβοι συνδέονται μεταξύ τους ονομάζεται κλίκα. Παράδειγμα απεικόνισης ενός γράφου στο επίπεδο αποτελεί το σχήμα του Σχήματος 1.

Ένα γράφημα ονομάζεται κατευθυνόμενο (directed) αν οι ακμές του έχουν προσανατολισμό. Έτσι η μια κορυφή δηλώνει την αφετηρία και η άλλη το πέρας της ακμής. Αν κάτι τέτοιο δεν ισχύει, τότε το γράφημα είναι *μη-κατευθυνόμενο* (undirected). Στο Σχήμα 2 μπορούμε να δούμε πως σχεδιάζεται ένα κατευθυνόμενο γράφημα. Οι ακμές του σχεδιάζονται σαν βέλη όπου η κορυφή τους υποδηλώνει την κατεύθυνση της ακμής.



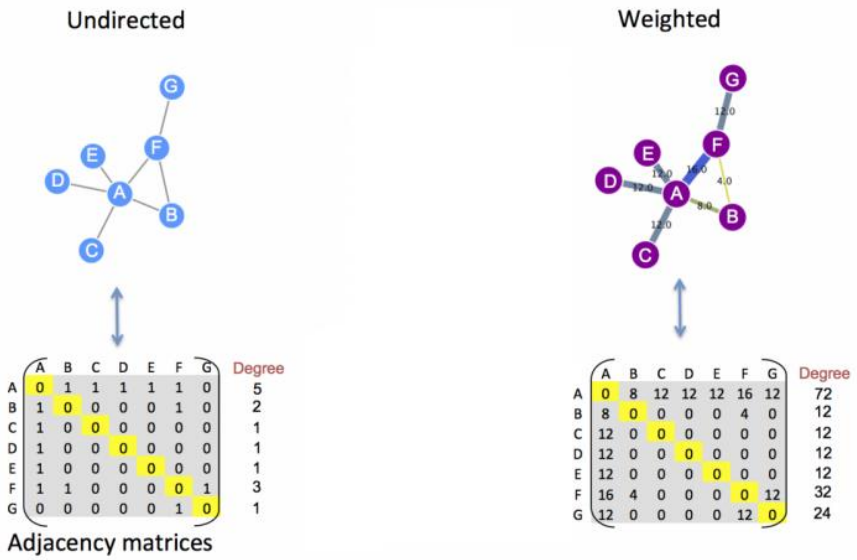
Σχήμα 2: Αριστερά ένας κατευθυνόμενος γράφος με πέντε κορυφές, δεξιά ο αντίστοιχος μη-κατευθυνόμενος γράφος.

Δυο κορυφές *u,v* ονομάζονται γειτονικές αν $(u, v) \in E$, αν υπάρχει δηλαδή ακμή που τις ενώνει απευθείας. Το σύνολο των γειτόνων μιας κορυφής u συμβολίζεται με $N_G(u)$. Ο αριθμός των γειτόνων μιας κορυφής *u,* $|(N_G(u))|$, σε ένα μη-κατευθυνόμενο γράφημα *G,* υποδηλώνει τον βαθμό της, ο οποίος συμβολίζεται $d_G(u)$. Ένας γράφος, για τον οποίο κάθε κόμβος του έχει τον ίδιο βαθμό, d, ονομάζεται d-κανονικός.

Πολλές φορές οι ακμές συμβολίζουν μια φυσική ποσότητα, όπως μήκος-απόσταση, χρόνος διάσχισης, χωρητικότητα, κ.α. Αν στην ανάλυση του γραφήματος αυτή η φυσική ποσότητα πρέπει να ληφθεί υπόψη τότε σε κάθε ακμή αποδίδεται ένας αριθμός που προσδιορίζει την ποσότητα αυτή. Στην περίπτωση αυτή λέμε ότι οι ακμές έχουν βάρος ή κόστος και ο γράφος είναι γράφος με βάρη (weighted graph). Ένας γράφος με βάρη μπορεί να είναι είτε κατευθυνόμενος είτε μη-κατευθυνόμενος.

Μαθηματικά ένας γράφος μπορεί να αναπαρασταθεί χρησιμοποιώντας πίνακα γειτνίασης (adjacency matrix). Ο πίνακας γειτνίασης είναι ένας τετραγωνικός πίνακας A, διάστασης $n \times n$, όπου n το πλήθος των κορυφών του γραφήματος. Κάθε στοιχείο του, A_{ij}, δηλώνει την ύπαρξη ή όχι της ακμής (i, j). Στην περίπτωση γράφου χωρίς βάρη στις ακμές, ο πίνακας έχει «1» στις θέσεις που αντιστοιχούν σε ακμές που ανήκουν στον γράφο και μηδέν στις υπόλοιπες. Αν έχουμε γράφο με βάρη στις ακμές, η τιμή του πίνακα είναι ίση με το βάρος κάθε ακμής και στην περίπτωση αυτή είναι σύνηθες ο πίνακας γειτνίασης να συμβολίζεται με τον γράμμα W (weight matrix). Στο Σχήμα 3 δίνονται παραδείγματα για γράφο χωρίς βάρη αλλά και για γράφο με βάρη στις ακμές. Στην περίπτωση κατευθυνόμενου γράφου χωρίς βάρη, οι ακμές που έχουν την

αφετηρία τους σε κάποιο κόμβο i και το πέρας τους σε κάποιο κόμβο j έχουν τιμή $A_{ij} = 1$ και $A_{ji} = -1$.



Σχήμα 3: Πίνακας γειτνίασης για γράφο χωρίς βάρη (αριστερά) και πίνακας βαρών για γράφο με βάρη στις ακμές. [EBI]

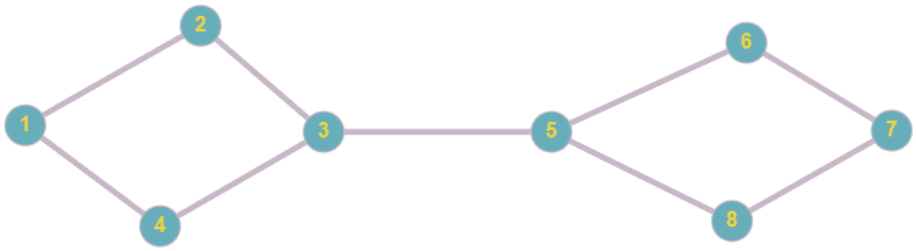
2.2 Περίπατοι – Μονοπάτια

Ένας *περίπατος* π μήκους k στο γράφημα *G,* είναι μια ακολουθία από εναλλασσόμενες ακμές και κορυφές, $\pi = u_o e_1 u_1 e_2 \ldots u_{(k-1)} e_k u_k$, τέτοια ώστε $e_i = (u_{(i-1)}, u_i), \forall 1 \leq i \leq k$.

Ένα *μονοπάτι* πρόκειται για έναν περίπατο χωρίς επαναλαμβανόμενες κορυφές. Αν η αρχική και η τελική κορυφή ταυτίζονται τότε λέμε ότι έχουμε έναν *κύκλο* στο *G.*

Ένα μονοπάτι *P* από έναν κόμβο *u* σε ένα κόμβο *v* του γράφου *G,* λέγεται ελάχιστο ή συντομότερο (shortest path) αν δεν υπάρχει κανένα άλλο μονοπάτι με μήκος μικρότερο του *P* που να συνδέει τους δυο αυτούς κόμβους. Αν για κάθε ζεύγος κόμβων στον G υπάρχει μονοπάτι που τους συνδέει, τότε ο γράφος είναι *συνδεδεμένος (connected)*. Σε διαφορετική περίπτωση ο γράφος δεν είναι συνδεδεμένος και χωρίζεται σε *συνεκτικές συνιστώσες (connected components)*, οι οποίες αποτελούνται από συνδεδεμένους υπογράφους του *G*.

Σε έναν συνδεδεμένο γράφο μια ακμή που η αφαίρεσή της προκαλεί την διαίρεση του γράφου σε περισσότερες από μια συνεκτικές συνιστώσες, ονομάζεται *γέφυρα (bridge*), Σχήμα 4.



Σχήμα 4: Στον παραπάνω γράφο η ακμή (3,5) αποτελεί γέφυρα, η αφαίρεση της οποίας δημιουργεί δυο συνεκτικές συνιστώσες.

Η απόσταση δυο κορυφών $u, v \in V$ του G, $d(u, v)$, ορίζεται ως το μήκος του ελάχιστου μονοπατιού που τις συνδέει. Σε έναν συνδεδεμένο γράφο η απόσταση είναι μια μετρική, άρα ικανοποιεί τις παρακάτω ιδιότητες:

1.$d(u, u) = 0, \forall u \in V$

2.$d(u, v) > 0, \forall u, v \in V με\ u \neq v$

3.$d(u, v) = d(v, u), \forall u, v \in V$

4.$d(u, v) \geq d(u, w) + d(w, v),\ \forall u, v, w \in V$ (τριγωνική ανισότητα)

2.3 Δέντρα

Ως Δέντρο (tree) νοείται ένας γράφος, έστω $T = (V, E)$, για τον οποίο ισχύει μία από τις παρακάτω ισοδύναμες προτάσεις:

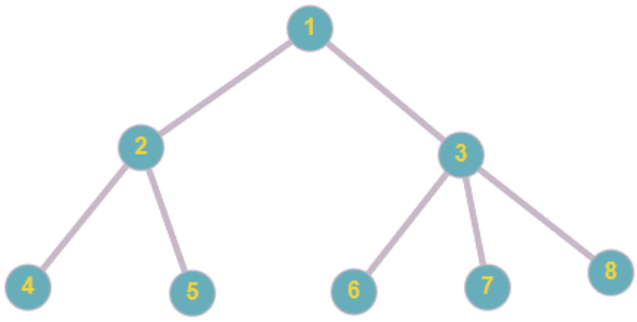
1. Το T είναι συνεκτικό χωρίς κύκλους.

2. Το T είναι συνεκτικό και ισχύει ότι $|E| = |V| - 1$.

3. Το T δεν έχει κύκλους και $|E| = |V| - 1$.

4. Για κάθε ζεύγος κορυφών $u, v \in V$υπάρχει μοναδικό μονοπάτι P που τις συνδέει.

5. Το T είναι συνεκτικό και κάθε ακμή του είναι γέφυρα.



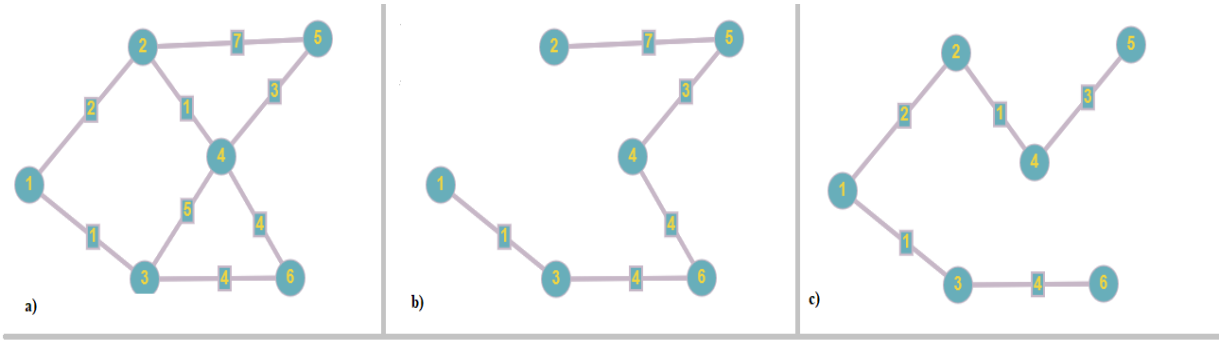
Σχήμα 5: Δέντρο με 8 κορυφές

Ως επίπεδα σε ένα δέντρο νοούνται τα σύνολα κόμβων που απέχουν το ίδιο από την ρίζα. Για παράδειγμα στο δέντρο του Σχήματος 5 οι κόμβοι {2,3} ανήκουν στο ίδιο επίπεδο αφού απέχουν και οι δυο απόσταση ίση με 1 από την ρίζα του δέντρου. Αντίστοιχα, οι κόμβοι {4,5,6,7,8} που απέχουν απόσταση 2 από τον κόμβο 1.

2.3.1 Δέντρα Επικάλυψης Ελάχιστου Κόστους

Ένα δέντρο επικάλυψης (spanning tree) ενός γράφου $G = (V, E)$ είναι ένα υποσύνολο των ακμών του γράφου $E' \subseteq E$ τέτοιο ώστε ο γράφος $G' = (V, E')$ να είναι δέντρο, δηλαδή να έχει τις ιδιότητες που ορίστηκαν παραπάνω. Οι γράφοι των τμημάτων b και c του Σχήματος 6 είναι δέντρα επικάλυψης του γράφου που παρουσιάζεται στο τμήμα a.

Ένα Δέντρο Επικάλυψης Ελάχιστου Κόστους (Minimum Spanning Tree, MST) είναι ένα δέντρο επικάλυψης του οποίου το άθροισμα των βαρών των ακμών του είναι το ελάχιστο από όλα τα πιθανά δέντρα επικάλυψης. Για τον υπολογισμό του MST οι δημοφιλέστεροι αλγόριθμοι είναι αυτοί των Kruskal [Kru56] και Prim [Pri57].



Σχήμα 6: a) Ένας γράφος 6 κορυφών και 8 ακμών, b) Ένα συνεκτικό δέντρο του (a) , c) Το ελάχιστο συνεκτικό δέντρο του (a).

Οι έννοιες που παρουσιάστηκαν παραπάνω αποτελούν θεμέλιο για την κατανόηση βασικών εννοιών που χρησιμοποιούνται στην ανάλυση Σύνθετων Δικτύων, με την οποία ασχολούμαστε εκτενώς στην παρούσα εργασία. Πολλές από τις μετρικές που χρειάζονται για την μελέτη και την κατανόηση των Σύνθετων Δικτύων, όπως οι διάφορες Κεντρικότητες, η Ενσωμάτωση και άλλες βασίζονται στις παραπάνω βασικές έννοιες.

3

Έννοιες Σύνθετων Δικτύων

Η ραγδαία ανάπτυξη των δικτύων επικοινωνιών, υπολογιστών και των πλατφορμών κοινωνικής δικτύωσης οδήγησε στην ανάγκη για περαιτέρω μελέτη των δικτύων αυτών, τα οποία συνοψίζονται με τον όρο Σύνθετα Δίκτυα (complex networks). Τα δίκτυα αυτά παρουσιάζουν συγκεκριμένες ιδιότητες που δεν απαντώνται ή δεν είναι τόσο έντονες σε μικρότερα δίκτυα, όπως ο σχηματισμός κοινοτήτων, ο υψηλός συντελεστής ομαδοποίησης (clustering coefficient) και άλλες. Επίσης, ιδιαίτερη σημασία αποκτά σε τέτοια δίκτυα η έννοια της Κεντρικότητας αφού ενδιαφέρει η εύρεση των πλέον σημαντικών, από διάφορες απόψεις, κόμβων στο δίκτυο.

Εκτός από την ανάλυση πραγματικών Σύνθετων Δικτύων όταν αυτό είναι δυνατό, έχουν προταθεί διάφορα μοντέλα για την μελέτη και την προσομοίωση τους. Τα μοντέλα αυτά μπορούν να χωριστούν σε χωρικά (spatial) και σχεσιακά (relational). Χωρικά λέγονται τα δίκτυα των οποίων οι κόμβοι συνδέονται μεταξύ τους ανάλογα με την θέση τους σε κάποιο γεωμετρικό μετρικό χώρο. Μοντέλο τέτοιων δικτύων αποτελούν οι Τυχαίοι Γεωμετρικοί Γράφοι. Σχεσιακά μοντέλα είναι αυτά τα οποία προκύπτουν όταν οι κόμβοι του δικτύου συνδέονται ανάλογα με τοπολογικές ιδιότητες που έχουν στο δίκτυο, όπως για παράδειγμα βάσει του βαθμού κόμβου. Σε αυτά τα μοντέλα εντάσσονται τα Δίκτυα Ελεύθερης Κλίμακας και τα Δίκτυα Μικρού Κόσμου.

Στο κεφάλαιο αυτό θα γίνει η παρουσίαση των παραπάνω μοντέλων καθώς και των Τυχαίων Γράφων, οι οποίοι χρησιμοποιούνται κυρίως ως μέτρο σύγκρισης για μεθοδολογίες καθώς προσομοιώνουν επακριβώς πολύ λίγα από τα πραγματικά συστήματα. Ακόμα θα παρουσιαστούν μετρικές που είναι απαραίτητες για την μελέτη Σύνθετων Δικτύων, όπως ο Συντελεστής Ομαδοποίησης και οι Κεντρικότητες. Μια πιο αναλυτική μελέτη των αντικειμένων που παρουσιάζονται στο κεφάλαιο αυτό μπορεί να βρεθεί στο [KSP13].

3.1 Μοντέλα Κατασκευής Τεχνητών Σύνθετων Δικτύων

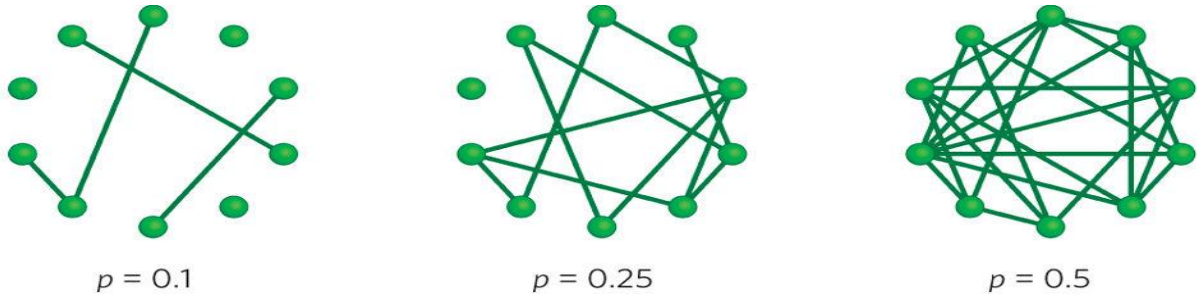
Σε αυτή την ενότητα θα εξεταστούν μοντέλα κατασκευής τεχνητών Σύνθετων Δικτύων. Συγκεκριμένα θα δούμε το μοντέλο των Τυχαίων Γράφων που παράγονται από τα μοντέλα των Gilbert και Erdos-Renyi, τους Τυχαίους Γεωμετρικούς Γράφους, τα Δίκτυα Ελεύθερης Κλίμακας που παράγονται σύμφωνα με το μοντέλο που ορίστηκε από τους Barabasi και Albert και τέλος το μοντέλο του Μικρού Κόσμου με την μεθοδολογία κατασκευής των Watts-Strogatz.

3.1.1 Τυχαίοι Γράφοι

Όταν οι κόμβοι ενός γράφου συνδέονται τυχαία μεταξύ τους τότε παράγεται ένας τυχαίος γράφος. Τα δημοφιλέστερα μοντέλα κατασκευής τους είναι τα μοντέλα του Gilbert και των Erdos-Renyi. Συνοπτικές περιγραφές των μοντέλων δίνονται παρακάτω.

- *Το μοντέλο του Gilbert*

  Στο μοντέλο που πρότεινε ο Gilbert [Gil59] για την κατασκευή ενός τυχαίου γράφου $G(n,p)$, σε ένα σύνολο n απομονωμένων (δηλαδή χωρίς κανένα γείτονα) κόμβων προστίθενται διαδοχικά ακμές μεταξύ τους. Κάθε ακμή έχει πιθανότητα να εμφανιστεί ίση με p, ανεξάρτητα από τις άλλες. Όσο η μεταβλητή τείνει στο 1 τόσο ο γράφος τείνει στον πλήρη γράφο n κόμβων, όπως φαίνεται και από το παράδειγμα του Σχήματος 7. Το μοντέλο αυτό αποτελεί ένα σχεσιακό μοντέλο κατασκευής γράφου καθώς οι συνδέσεις μεταξύ των κόμβων του δεν εξαρτώνται από την θέση των κόμβων στον χώρο.



Σχήμα 7: Τυχαίος Γράφος βάσει του μοντέλου Gilbert για τρεις διαφορετικές τιμές του p [Nat]

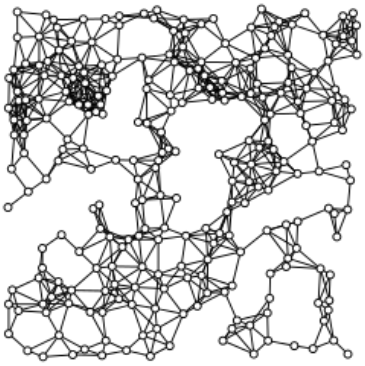
- *Το μοντέλο των Erdos-Renyi*

  Σε αυτό το μοντέλο [ErR59] ένας τυχαίος γράφος $G(n,M)$ προκύπτει επιλέγοντας τον τυχαία, με ομοιόμορφη κατανομή πιθανότητας, από όλους τους πιθανούς γράφους με n κορυφές και M ακμές. Στην περίπτωση που $pn^2 \to \infty$ το μοντέλο του Gilbert γίνεται ισοδύναμο με αυτό των Erdos-Renyi στην περίπτωση που $M = \binom{n}{2}p$.

Οι τυχαίοι γράφοι χρησιμοποιούνται περισσότερο σαν γράφοι αναφοράς για την σύγκριση μεθόδων αλλά μπορούν να μοντελοποιήσουν Νευρωνικά Δίκτυα, δίκτυα ομότιμων χρηστών (peer-to-peer) [LCC02], κ.α.

3.1.2 Τυχαίοι Γεωμετρικοί Γράφοι

Ένας Τυχαίος Γεωμετρικός Γράφος (Random Geometric Graph, RGG) *G(N,r)* είναι ένας γράφος N κόμβων οι οποίοι έχουν συντεταγμένες σε έναν Γεωμετρικό Χώρο και συνδέονται μεταξύ τους μόνο αν η απόστασή τους δεν υπερβαίνει την μεταβλητή r. Τέτοιος γράφος είναι και αυτός που απεικονίζεται στο Σχήμα 8. Οι Τυχαίοι Γεωμετρικοί γράφοι είναι ένα χωρικό μοντέλο κατασκευής γράφων και χρησιμοποιούνται για την προσομοίωση συστημάτων όπου οι σχέσεις μεταξύ των συστατικών του μερών εξαρτώνται από την θέση τους στον χώρο.

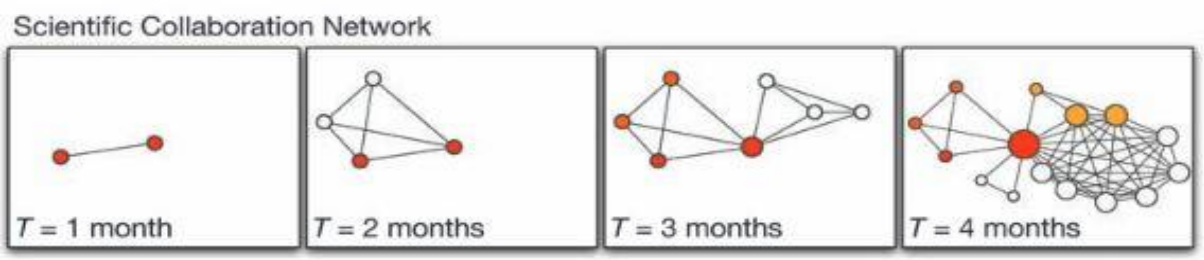


Σχήμα 8: Παράδειγμα ενός τυχαίου γεωμετρικού γράφου [Wik]

Το συγκεκριμένο μοντέλο αποτελεί ένα απλό μαθηματικό μοντέλο κατασκευής γράφου και οδηγεί σε δίκτυα με μεγάλο μέσο μήκος μονοπατιού. Οι Τυχαίοι Γεωμετρικοί Γράφοι αποτελούν επιτυχημένα μοντέλα για την προσομοίωση και την μελέτη ασύρματων ad-hoc δικτύων [MuP10].

3.1.3 Δίκτυα Ελεύθερης Κλίμακας

Η μελέτη μεγάλων τοπολογιών Σύνθετων Δικτύων αποκάλυψε ότι η πιθανότητα $P(k)$ ένας κόμβος να έχει βαθμό ίσο με κ, ακολουθεί ένα νόμο δύναμης (power-law), όπου $P(\kappa) \sim \kappa^{-\gamma}$. Η παράμετρος γ πειραματικά έχει προσδιοριστεί ότι συνήθως κυμαίνεται μεταξύ $2 \leq \gamma \leq 3$, αν και έχουν παρατηρηθεί και Δίκτυα Ελεύθερης-Κλίμακας (Scale-Free) με υψηλότερη τιμή. Το γεγονός ότι υπάρχουν δίκτυα που αναπτύσσονται με αυτό τον τρόπο είναι κάτι που τα προγενέστερα μοντέλα κατασκευής Σύνθετων Δικτύων, τα οποία βασίζονταν στην τυχαιότητα, δεν μπορούν να προβλέψουν. Αυτό συμβαίνει επειδή δεν λαμβάνουν υπόψη τόσο την ανάπτυξη (είσοδος νέων κόμβων στο δίκτυο) όσο και την προτίμηση-στη-σύνδεση (preferential attachment). Η τελευταία έννοια σημαίνει ότι ένας νέος κόμβος που εισέρχεται σε ένα δίκτυο είναι πιθανότερο να συνδεθεί με κόμβους με υψηλό βαθμό. Έτσι οι παλιοί κόμβοι καθίστανται πιο «ισχυροί», από άποψη βαθμού, ενώ οι νεότεροι λιγότερο. Η λογική της προτίμησης-στη-σύνδεση μπορεί να συνοψιστεί στη φράση «ο πλούσιος γίνεται πλουσιότερος». Στο παράδειγμα του Σχήματος 9 μπορούμε να παρατηρήσουμε πως οι αρχικοί κόμβοι αποκτούν όλο και περισσότερες συνδέσεις καθώς νέοι κόμβοι εισέρχονται στο δίκτυο.



Σχήμα 9: Η εξέλιξη του δικτύου των επιστημονικών συνεργασιών σε χρονικό διάστημα τεσσάρων μηνών, [Bar09].

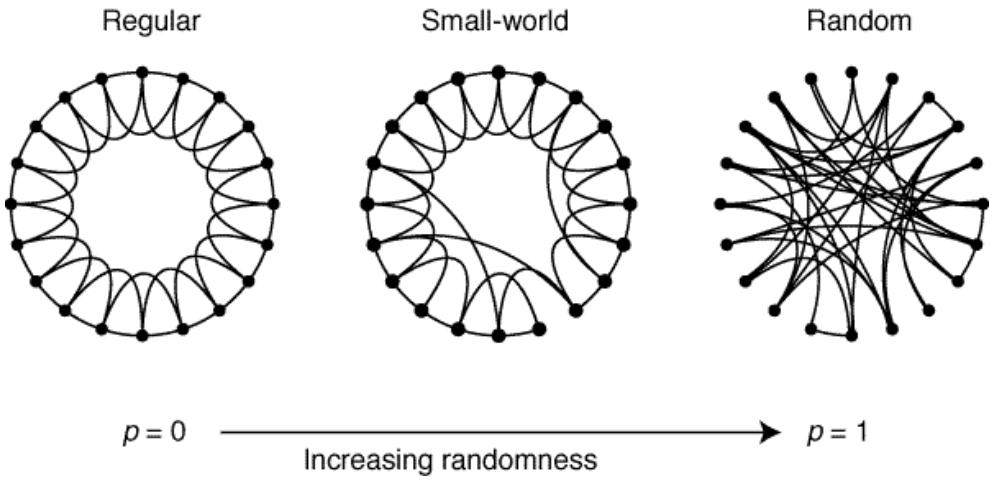
Οι Barabasi και Albert [BaA99] ήταν οι πρώτοι που ασχολήθηκαν με τα Δίκτυα Ελεύθερης Κλίμακας και πρότειναν ένα μαθηματικό μοντέλο για την κατασκευή τους. Σύμφωνα με αυτό, το δίκτυο αποτελείται αρχικά από m_0 κόμβους. Σε κάθε χρονική στιγμή t ένας νέος κόμβος εισέρχεται στο δίκτυο. Ο κόμβος αυτός θα συνδεθεί με $m < m_0$ κόμβους. Η πιθανότητα Π ένας κόμβος i να συνδέεται με τον νέο κόμβο εξαρτάται από τον βαθμό του k_i και είναι ίση με $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$. Μετά από t βήματα το μοντέλο συγκλίνει σε μια κατάσταση ανεξάρτητη από το μέγεθός του όπου η πιθανότητα ένας κόμβος να έχει βαθμό ίσο με k ακολουθεί έναν νόμο-δύναμης με $\gamma = 2.9 \pm 0.1$.

Επειδή εμπειρικά έχει παρατηρηθεί ότι τα Κοινωνικά Δίκτυα, ο Παγκόσμιος Ιστός, Δίκτυα Μεταφορών και άλλα ακολουθούν το λεγόμενο power-law κατανομή βαθμού, το μοντέλο των Barabasi και Albert αποτελεί μια καλή προσέγγιση για την μελέτη τους.

3.1.4 Δίκτυα Μικρού Κόσμου

Στα δίκτυα μικρού-κόσμου (small world networks) αν και οι περισσότεροι κόμβοι δεν είναι γείτονες μεταξύ τους, συνήθως υπάρχει μονοπάτι μικρού μήκους που συνδέει μη γειτονικούς κόμβους. Πολλά δίκτυα παρουσιάζουν αυτό που έγινε γνωστό ως ιδιότητα του μικρού κόσμου. Ο Stanley Milgram σε ένα κοινωνιολογικό πείραμα του [Mil67], ζήτησε από 296 ανθρώπους να προωθήσουν ένα γράμμα προς έναν προορισμό προωθώντας το κάθε φορά σε κάποιον γνωστό τους, τον οποίο θεωρούν την καλύτερη επιλογή ώστε να φτάσει το γράμμα γρηγορότερα. Από τα 64 γράμματα που έφτασαν εν τέλει στον προορισμό τους διαπιστώθηκε ότι κατά μέσο όρο είχαν κάνει 5.5-6 βήματα. Αργότερα σε ένα πείραμα των Leskovec και Horvitz το 2008 [LeH08], στο οποίο αναλύθηκαν 240 εκατομμύρια λογαριασμοί χρηστών στην υπηρεσία Microsoft Instant Messenger, βρέθηκε ότι οι κόμβοι της μεγαλύτερης συνεκτικής συνιστώσας του δικτύου απέχουν κατά μέσο όρο 6.6-7 βήματα.

Το μοντέλο κατασκευής τεχνητών Δικτύων Μικρού-Κόσμου, που θα χρησιμοποιηθεί στα πλαίσια της διπλωματικής εργασίας, είναι το μοντέλο των Watts-Strogatz [WaS98]. Σε αυτό, ξεκινώντας από ένα διατεταγμένο πλέγμα N κόμβων, με πιθανότητα p κόμβοι που δεν ήταν γειτονικοί επανασυνδέονται, ανεξάρτητα από την πρότερη μεταξύ τους απόσταση. Όσο η πιθανότητα p αυξάνει τόσο ο γράφος που παράγεται τείνει στο να γίνει Τυχαίος Γράφος όπως φαίνεται και από το παράδειγμα του Σχήματος 10.



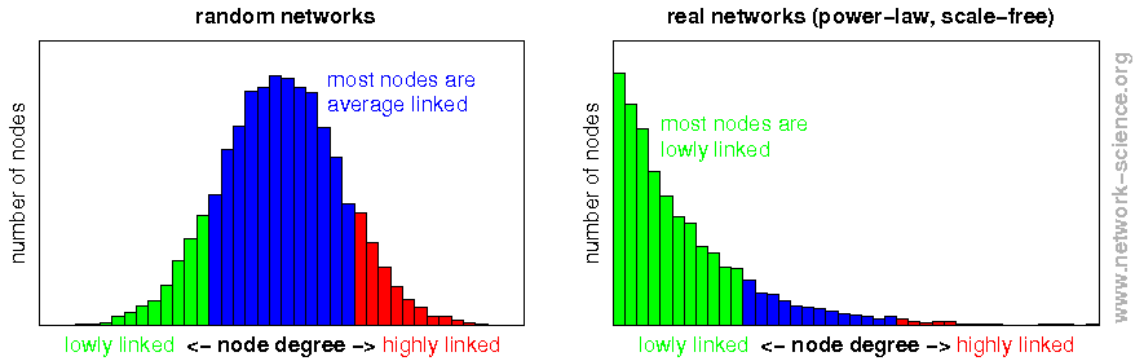
Σχήμα 10: Εξέλιξη ενός γράφου που παράγεται με την μέθοδο των Watts-Strogatz όσο αυξάνει η πιθανότητα p.

3.2 Μετρικές Σύνθετων Δικτύων

Στο κεφάλαιο αυτό παρουσιάζονται εκείνες οι μετρικές των δικτύων που είναι απαραίτητες τόσο για μια βασική ανάλυση των Σύνθετων Δικτύων όσο και για να γίνει κατανοητή η προτεινόμενη μέθοδος της παρούσας εργασίας.

3.2.1 Κατανομή Βαθμού

Η έννοια του βαθμού έχει οριστεί στο Κεφάλαιο 2. Η κατανομή βαθμού (degree distribution) $P(k)$, έχει οριστεί ως το ποσοστό των κόμβων που έχουν βαθμό ίσο με k. Όπως είδαμε τα Scale-Free δίκτυα έχουν μια κατανομή βαθμού $P(k) \sim \kappa^{-\gamma}$, αντιθέτως τα τυχαία δίκτυα παρουσιάζουν μια κανονική (Gaussian) κατανομή βαθμού, η διαφορά αυτή στην κατανομή βαθμού φαίνεται παρακάτω στο Σχήμα 11.



Σχήμα 11: Σύγκριση κατανομής βαθμού σε ένα τυχαίο γράφο (αριστερά) και ένα δίκτυο ελεύθερης κλίμακας. Όπως βλέπουμε στο Scale-Free δίκτυο υπάρχουν λίγοι κόμβοι υψηλού βαθμού σε σχέση με τον τυχαίο γράφο, [Net].

3.2.2 Μέσο Μήκος Μονοπατιού

Το μέσο μήκος μονοπατιού (average path length) είναι ο μέσος όρος του μήκους όλων των συντομότερων μονοπατιών μεταξύ όλων των κόμβων του δικτύου. Ορίζεται ως:

$$l_G = \frac{1}{n * (n-1)} \sum_{i,j} d(i,j) \qquad (1)$$

Όσο πιο πυκνός είναι ένας γράφος τόσο πιο μικρό αναμένεται να είναι το μέσο μήκος μονοπατιού. Επίσης, αναμένεται ένας γράφος του οποίου οι κόμβοι σχηματίζουν κοινότητες να έχει μικρό μέσο μήκος μονοπατιού. Αυτό συμβαίνει καθώς οι κόμβοι που ανήκουν στην ίδια κοινότητα έχουν μικρή απόσταση μεταξύ τους. Σε ένα δίκτυο με χαμηλή τιμή για την μετρική l_G η πληροφορία διαδίδεται πιο γρήγορα σε σχέση με ένα δίκτυο με υψηλή τιμή που περιμένουμε να γίνουν περισσότερα βήματα μέχρι να ενημερωθούν όσοι κόμβοι πρέπει. Αν και το μέσο μήκος μονοπατιού είναι μια καλή μετρική δεν αρκεί για τον μονοσήμαντο ορισμό μιας τοπολογίας. Ακόμα είναι ευαίσθητο σε ακραίες περιπτώσεις (outliers) καθώς είναι δυνατόν να υπάρχουν πολλοί κοντινοί κόμβοι και λίγοι με πολύ μεγάλη απόσταση μεταξύ τους, κάτι που είναι δυνατόν να αλλοιώσει την μετρική.

3.2.3 Συντελεστής Ομαδοποίησης

Ο Συντελεστής Ομαδοποίησης (clustering coefficient) είναι ένας δείκτης που η τιμή του υποδηλώνει τον βαθμό στον οποίο οι κόμβοι ενός δικτύου τείνουν να σχηματίζουν τοπικές κοινότητες μεταξύ τους. Ο συντελεστής αυτός έχει διάφορους ορισμούς ανάλογα με το επίπεδο του δικτύου που εξετάζουμε κάθε φορά. Έτσι μπορούμε να ξεχωρίσουμε την περίπτωση που εξετάζουμε το δίκτυο συνολικά, αλλά και την περίπτωση που εξετάζουμε «από κοντά» έναν συγκεκριμένο κόμβο του δικτύου.

- Ολικός Συντελεστής Ομαδοποίησης (Global Clustering |Coefficient):

$$CC = \frac{3 \times number\ of\ triangles}{number\ of\ connected\ triples\ of\ vertices} \qquad (2)$$

- Τοπικός Συντελεστής Ομαδοποίησης (Local |Clustering Coefficient) ενός κόμβου i:

$$CC_i = \frac{|\{e_{jk}\}|}{k_i * (k_i - 1)} : u_j, u_k \in N_i, e_{jk} \in E \quad (3)$$

Ο τοπικός συντελεστής εκφράζει το πόσο απέχει η γειτονιά ενός κόμβου από το να γίνει κλίκα.

- Μέσος Συντελεστής Ομαδοποίησης Δικτύου (Network-Wide Average Clustering Coefficient):

$$\overline{CC} = \frac{1}{n}\sum_{i=1}^{n} CC_i \qquad (4)$$

3.3 Κεντρικότητες Κορυφών

Η έννοια της Κεντρικότητας Κορυφής (node centrality) κατέχει σημαντική θέση ανάμεσα στις μετρικές για τον χαρακτηρισμό του μέτρου της σημαντικότητας ενός κόμβου. Ο υπολογισμός των πιο σημαντικών κόμβων ενός δικτύου παίζει σημαντικό ρόλο σε εφαρμογές όπως κατανομής ενέργειας σε Ασύρματα Δίκτυα Αισθητήρων (Wireless Sensor Networks), διάδοση πληροφορίας σε ένα Κοινωνικό Δίκτυο (Online Social Network, OSN) και άλλες. Βέβαια η έννοια της Κεντρικότητας αποκτά διαφορετική ερμηνεία καθώς κάθε μετρική κάνει διαφορετικές υποθέσεις για τον τρόπο με τον οποίο η πληροφορία ρέει στο δίκτυο [Bor05]. Γενικά, μπορούμε να διακρίνουμε τρεις διαφορετικές κατηγορίες μετρικών. Μετρικές που βασίζονται στον βαθμό κάθε κόμβου, μετρικές που βασίζονται σε γεωδαισιακές αποστάσεις, και τέλος μετρικές που βασίζονται σε φασματικές ιδιότητες του γράφου (spectral properties). Παρακάτω δίνονται σύντομες περιγραφές μερικών πιο δημοφιλών μετρικών Κεντρικότητας.

3.3.1 Κεντρικότητα Βαθμού

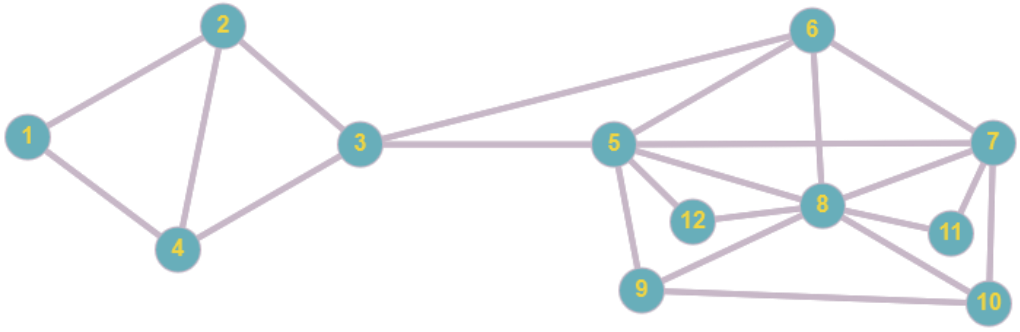
Μια από τις πιο άμεσες, διαισθητικά, μετρικές είναι η Κεντρικότητα Βαθμού (Degree Centrality). Εδώ, η κεντρικότητα κάθε κόμβου ισούται με τον βαθμό του. Έτσι αν A είναι ο πίνακας γειτνίασης ενός δικτύου $G = (V, E)$, η κεντρικότητα κάθε κορυφής $C_D(u)$, $u \in V$ υπολογίζεται ως εξής:

$$C_D(u) = \sum_{i=1}^{n} a_{ui} \qquad (5)$$

Η Κεντρικότητα Βαθμού υποδηλώνει ότι όσους περισσότερους γείτονες έχει ένας κόμβος τόσο πιο σημαντικός είναι. Διαισθητικά, ένας χρήστης OSN είναι τόσο δημοφιλής, όσοι περισσότεροι είναι οι φίλοι του.

Στα πλεονεκτήματα της μεθόδου συγκαταλέγεται ο γρήγορος και εύκολος υπολογισμός της καθώς και η καταλληλότητα της για ορισμένες εφαρμογές, κυρίως εφαρμογές σχετιζόμενες με δημοφιλία των εμπλεκόμενων οντοτήτων. Παρόλα αυτά, αποτελεί μια τοπική μετρική του δικτύου καθώς για κάθε κόμβο ενδιαφέρει μόνο η γειτονιά του. Σαν αποτέλεσμα κρίνεται ακατάλληλη για εφαρμογές διάδοσης πληροφορίας.

Στο παρακάτω σχήμα του Σχήματος 12 ισχύει ότι $C_D(3) = 4 < 7 = C_D(8)$. Αν όμως αφαιρεθεί η κορυφή 3 τότε θα έχουμε δυο συνεκτικές συνιστώσες και τα δυο μέρη του δικτύου δεν θα μπορούν να επικοινωνήσουν. Αντιθέτως, αν αφαιρεθεί η 8 τότε το δίκτυο θα παραμείνει συνδεδεμένο. Άρα προκύπτει το συμπέρασμα ότι δεν μπορούμε να βασιστούμε μόνο στο μέτρο της Κεντρικότητας Βαθμού για την ανάλυση ενός δικτύου.



Σχήμα 12: Όταν μόνο η Κεντρικότητα Βαθμού δεν αρκεί.

3.3.2 Κεντρικότητα Εγγύτητας

Η Κεντρικότητα Εγγύτητας (Closeness Centrality) ενός κόμβου $u \in V$ ενός δικτύου $G = (V, E)$ ορίζεται ως εξής [Bav50]:

$$C_C(u) = \frac{1}{\sum_{i=1}^{n} dist(u,i)} \qquad (6)$$

Όπου με $dist(u,i)$ ορίζεται το μήκος του ελάχιστου μονοπατιού που συνδέει τους κόμβους u και i.

Η συγκεκριμένη μετρική, η οποία ορίζεται μόνο για συνδεδεμένα δίκτυα, παρέχει ένα καλύτερο μέτρο της Κεντρικότητας ενός κόμβου όταν η πληροφορία διαδίδεται μέσα από συντομότερα μονοπάτια. Όσο μεγαλύτερη είναι η τιμή της Κεντρικότητας Εγγύτητας ενός κόμβου, τόσο πιο μικρή είναι η συνολική του απόσταση προς τους υπόλοιπους κόμβους του δικτύου. Διαισθητικά, το μέτρο της Κεντρικότητας αυτής μπορεί να χρησιμοποιηθεί για να υπολογιστεί πόσο χρόνο χρειάζεται να διαδοθεί η πληροφορία από ένα κόμβο σε όλους τους άλλους διαδοχικά, εφόσον η διάδοση γίνεται συγχρονισμένα κατά ένα βήμα (hop) του γράφου τη φορά.

3.3.3 Κεντρικότητα Ενδιαμεσικότητας

Μια ακόμα προσέγγιση στον υπολογισμό Κεντρικότητας είναι να υπολογιστεί για κάθε κόμβο σε πόσα συντομότερα μονοπάτια μέσα στο δίκτυο βρίσκεται. Πρόκειται για την Κεντρικότητα Ενδιαμεσικότητας (Betweenness Centrality), που υπολογίζεται ως εξής [Fre77]:

$$C_B(u) = \sum_{s \neq u \neq t \in V} \frac{\sigma_{st}(u)}{\sigma_{st}}, \qquad (7)$$

όπου σ_{st}είναι το πλήθος των συντομότερων μονοπατιών από τον s στον t, ενώ $\sigma_{st}(u)$ είναι ο αριθμός των συντομότερων μονοπατιών από τον s στον t στα οποία περιλαμβάνεται και ο u.

Η γνώση της Κεντρικότητας Ενδιαμεσικότητας συντελεί στον εντοπισμό των κόμβων εκείνων που η αφαίρεσής τους από το δίκτυο είναι πιθανό να το καταστήσει μη-συνδεδεμένο.

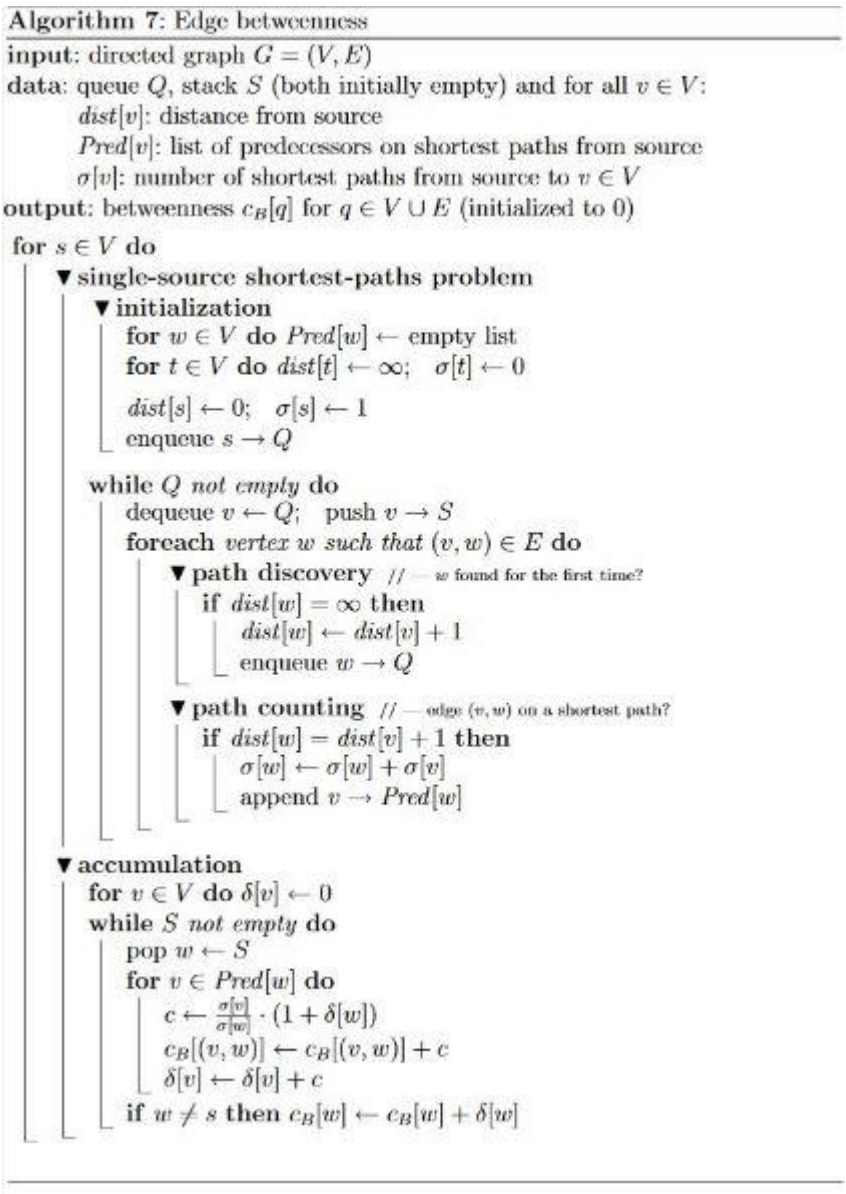
3.3.4 Άλλες Μετρικές Κεντρικότητας Κορυφής

Εκτός από τις προαναφερθείσες μετρικές έχουν προταθεί και αρκετές ακόμα. Η Κεντρικότητα Ιδιοδιανύσματος (Eigenvector Centrality) παίρνει υπόψη τόσο το πλήθος των γειτόνων ενός κόμβου αλλά και την ποιότητα των συνδέσεων, έτσι ώστε η σύνδεση με έναν κόμβο υψηλού βαθμού να βαραίνει παραπάνω στον υπολογισμό της Κεντρικότητας από ότι η σύνδεση με έναν κόμβο μικρότερου βαθμού. Παρόμοιες μεθόδους αποτελούν η μέθοδος που χρησιμοποιεί η Google, PageRank [Rog02] και η Κεντρικότητα Katz [Kat53]. Ακόμα, η μετρική Κεντρικότητας Φόρτου Κίνησης (Traffic Load Centrality, TLC) [SSK16] είναι μια μετρική που ταξινομεί τους κόμβους ανάλογα με την συμμετοχή τους στην συνολική κίνηση πληροφορίας μέσα στο δίκτυο.

3.4 Κεντρικότητα Ενδιαμεσικότητας Ακμής

Γενικεύοντας την έννοια της Κεντρικότητας Ενδιαμεσότητας Κορυφής, οι Girvan και Newman [GiN02] όρισαν την λεγόμενη Κεντρικότητα Ενδιαμεσικότητας Ακμής (Edge Betweenness Centrality) ως το πλήθος των ελάχιστων μονοπατιών που διατρέχουν από μια ακμή του δικτύου. Μια ακμή με υψηλή τιμή ενδιαμεσικότητας είναι πολύ πιθανό να αντιπροσωπεύει μια γέφυρα στο δίκτυο, όπου η αφαίρεση της θα άλλαζε δραστικά την επικοινωνία μεταξύ των κόμβων. Για παράδειγμα, στο Σχήμα 4, η ακμή (3,5) είναι η ακμή με την μεγαλύτερη Κεντρικότητα Ενδιαμεσικότητας, η αφαίρεση της θα δημιουργούσε δυο συνεκτικές συνιστώσες που οι κόμβοι τους δεν θα μπορούσαν να επικοινωνήσουν.

Για τον υπολογισμό της εν λόγω μετρικής ο Brandes [Bra08] πρότεινε τον παρακάτω αποδοτικό αλγόριθμο ο ψευδοκώδικας του οποίου δίνεται στο Σχήμα 13:



Σχήμα 13: Ο αλγόριθμος του Brandes, [Bra08].

Για να γίνει κατανοητός ο αλγόριθμος του Brandes πρέπει πρώτα να παρουσιαστούν τα παρακάτω αποτελέσματα, τα οποία ορίζονται και αποδεικνύονται στο [Bra01]:

Το κριτήριο του Bellman: Ένας κόμβος $u \in V$ βρίσκεται στο συντομότερο μονοπάτι μεταξύ των κορυφών $s, t \in V$, αν και μόνο αν $d(s,t) = d(s,u) + d(u,t)$.

Για τον αριθμό των μονοπατιών ελάχιστου μήκους από τον s στον t που διέρχονται από τον u, $\sigma_{st}(u)$, ισχύει το παρακάτω:

$$\sigma_{st}(u) = \begin{cases} 0, \alpha\nu\ d(s,t) < d(s,u) + d(u,t) \\ \sigma_{su} * \sigma_{ut}, \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

Χρειάζεται ακόμα να οριστεί το σύνολο των προγόνων ενός κόμβου στο συντομότερο μονοπάτι ως εξής: $P_s(v) = \{u \in V : (u,v) \in E, d(s,v) = d(s,u) + w(u,v)\}$, όπου $w(u,v)$ είναι το βάρος της ακμής (u,v). Στην περίπτωση γράφου χωρίς βάρη στις ακμές η τιμή $w(u,v)$ είναι ίση με 1. Ακόμα, ισχύει ότι για $s \neq v \in V$ $\sigma_{sv} = \sum_{u \in P_s(v)} \sigma_{su}$.

Ο λόγος των συντομότερων μονοπατιών μεταξύ των s, t που περιέχουν την u ορίζεται ως $\delta_{st}(u) = \frac{\sigma_{st(u)}}{\sigma_{st}}$. Ακόμα, ορίζεται η ποσότητα $\delta_s(v) = \sum_{t \in V} \delta_{st}(v)$. Ο ορισμός αυτός είναι απαραίτητος για την διατύπωση της παρακάτω πρότασης:

Αν υπάρχει ακριβώς ένα ελάχιστο μονοπάτι από την κορυφή s για κάθε t τότε $\delta_s(v) = \sum_{w:v \in P_s(w)} (1 + \delta_s(w))$.
Αν πάλι αυτό δεν ισχύει τότε $\delta_s(v) = \sum_{w:v \in P_s(w)} \frac{\sigma_{sv}}{\sigma_{sw}} (1 + \delta_s(w))$.

Ο αλγόριθμος αυτός έχει πολυπλοκότητα $O(m * n)$ για γράφους χωρίς βάρη, όπου m ο αριθμός των κορυφών ενός γράφου και n ο αριθμός των ακμών του. Ενδεικτικά αναφέρεται ότι ο υπολογισμός Κεντρικότητας Ενδιαμεσικότητας Ακμής εκτελώντας κάθε φορά μια Αναζήτηση Κατά Πλάτος (BFS) για να υπολογιστεί το συντομότερο μονοπάτι για κάθε ζεύγος κόμβων απαιτεί χρόνο της τάξης $O(m * n^2)$.

4

Ανάλυση Μεγάλων Δεδομένων Και Γράφοι Εγγύτητας

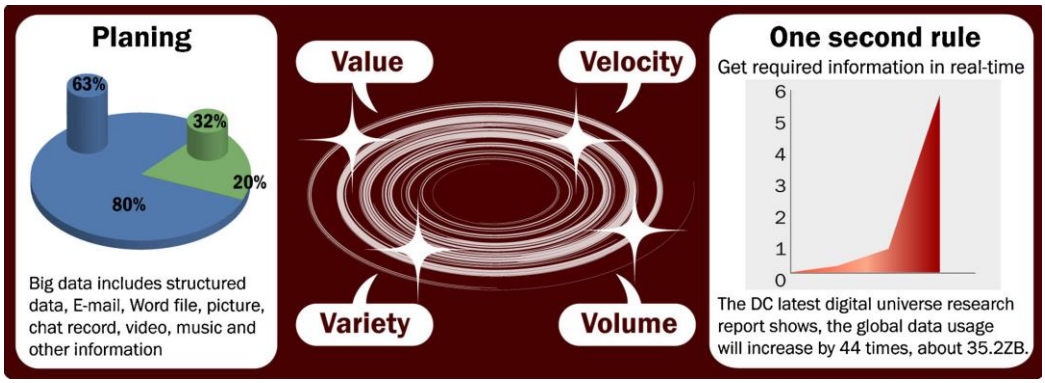
4.1 Μεγάλα Δεδομένα

Σε μια μελέτη της εταιρείας IBM [IBM] αναφέρεται ότι περίπου το 90% των δεδομένων που έχουν παραχθεί παγκοσμίως έχουν παραχθεί τα τελευταία δυο χρόνια. Αυτό έγινε δυνατό καθώς ολοένα και περισσότερες συσκευές όπως sensors, RFID tags, smart devices και πλήθος άλλων, εγκαθίστανται και διασυνδέονται στο Διαδίκτυο. Σημαντικό ρόλο στην παραγωγή δεδομένων παίζει επίσης η ραγδαία εξάπλωση των κοινωνικών δικτύων (Online Social Networks). Χαρακτηριστικό παράδειγμα αποτελεί η δημοφιλέστερη πλατφόρμα κοινωνικής δικτύωσης, Facebook, που το πρώτο εξάμηνο του 2017 έφτασε τους δυο δισεκατομμύρια λογαριασμούς [Tec] και για το έτος 2014 παράγονταν από τους χρήστες περίπου 600 TeraByte δεδομένων ημερησίως [Fac]. Ακόμα υπολογίζεται ότι κάθε λεπτό 300 ώρες βίντεο μεταφορτώνονται στο Youtube και ότι μέχρι το 2020 θα υπάρχουν περίπου 50 δισεκατομμύρια συνδεδεμένες συσκευές στο Διαδίκτυο [Mar15]. Αυτός ο όγκος των δεδομένων, που πλέον χαρακτηρίζονται ως Μεγάλα Δεδομένα (Big Data), φέρνει νέες προκλήσεις όσον αφορά τόσο την αποθήκευση τους, την επεξεργασία τους, αλλά και την ανάλυσή τους.

4.1.1 Ορισμός

Αν και δεν υπάρχει μέχρι στιγμής κάποιος αυστηρός ορισμός της έννοιας των Μεγάλων Δεδομένων (Big Data), είναι κοινός τόπος μεταξύ των μελετητών τους να ορίζονται ως εξαιρετικά μεγάλα σύνολα από δεδομένα όπου οι παραδοσιακές τεχνικές αποθήκευσης και ανάλυσης δεδομένων αποτυγχάνουν να εφαρμοστούν σε αποδεκτό χρονικό διάστημα [CML14].

Το 2001 ορίστηκαν για πρώτη φορά τα λεγόμενα “3Vs” των Μεγάλων Δεδομένων [Lan01]. Αν και το άρθρο δεν πραγματεύονταν τα Big Data ο ορισμός των “3Vs”, υιοθετήθηκε από τους μελετητές τους, ως ειδοποιός διαφορά από τα μικρότερα σύνολα δεδομένων. Πρόκειται για τα εξής μεγέθη :



Σχήμα 14: Τα 4Vs των Μεγάλων Δεδομένων, [CML14].

- ***Όγκος*** (Volume): Πρόκειται για την ποσότητα των δεδομένων που παράγονται. Υπολογίζεται ότι το έτος 2021 μόνο η IP κίνηση κατά την διάρκεια ενός έτους θα αγγίξει τα 3.3 ZetaBytes (ZB), ενώ το 2016 η κίνηση ανήλθε σε 1.2 ZB [Cis]. Από το παράδειγμα μπορούμε να δούμε ότι η ποσότητα των δεδομένων που παράγονται συνεχώς αυξάνεται και ότι ο όγκος τους είναι μια σημαντική παράμετρος που πρέπει να ληφθεί υπόψη κατά την μελέτη τους.

- ***Ταχύτητα*** (Velocity): Αναφέρεται τόσο στον ρυθμό παραγωγής δεδομένων που διαρκώς αυξάνεται όσο και στην ανάγκη για γρήγορη συλλογή και επεξεργασία τους, καθώς για πολλές εφαρμογές απαιτείται η όσο το δυνατόν ταχύτερη απόκριση τους, ακόμα και σε συνθήκες πραγματικού χρόνου.

- ***Ποικιλία*** (Variety): Τα δεδομένα που συλλέγονται προέρχονται από πολλές και διαφορετικές πηγές με αποτέλεσμα να παρατηρείται ποικιλομορφία στην μορφή τους (βίντεο, ήχος, κείμενο, κ.λπ.).

Αργότερα, προέκυψε η ανάγκη για ένα ακόμα "V", οδηγώντας στο μοντέλο των "4Vs". Σύμφωνα με κάποιους αυτό της Αξίας (Value) με την έννοια του κέρδους που μπορούν να αποκομίσουν οι επιχειρήσεις από την σωστή αξιοποίηση των Μεγάλων Δεδομένων[CML14]. Σύμφωνα με άλλους αρθρογράφους πρέπει να προστεθεί και αυτό της Ακρίβειας (Veracity) με την έννοια της αξιοπιστίας της πηγής για την ορθότητα των δεδομένων [JGL14]. Κάποια από τα χαρακτηριστικά αυτά φαίνονται και στο Σχήμα 14.

4.1.2 Κύκλος Ζωής Μεγάλων Δεδομένων

Στην διαδικασία της λεγόμενης Ανάλυσης Μεγάλων Δεδομένων (Big Data Analytics) έχει οριστεί [JGL14] το μοντέλο του Κύκλου Ζωής Μεγάλων Δεδομένων (Big Data Life-cycle) που έχει τις παρακάτω φάσεις:

- *Λήψη Δεδομένων*: Δεδομένα που συνήθως προέρχονται από αισθητήρες (sensors), αρχιτεκτονικές συσκευών Internet-of-Things (IoT) devices, Wireless Sensor Networks, Διαδίκτυο, κ.λπ.

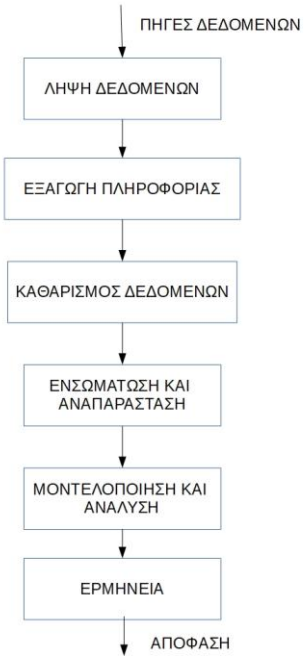
- *Εξαγωγή Πληροφορίας και Καθαρισμός*: Από το σύνολο των δεδομένων μένουν μόνο όσα δεδομένα χρειάζονται για την εφαρμογή. Εμφανώς λανθασμένες τιμές απομακρύνονται ή αντιμετωπίζονται με κατάλληλες τεχνικές καθαρισμού.

- *Ενσωμάτωση και Αναπαράσταση Δεδομένων*: Η φάση αυτή περιλαμβάνει τη συγκέντρωση όλων των διαφορετικών τύπων δεδομένων που αφορούν κάποια οντότητα και την επιλογή του τρόπου με τον οποίο θα αναπαρασταθεί η γνώση αυτή. Πλέον οι κλασικές σχεσιακές βάσεις δεδομένων δεν αρκούν για την αναπαράσταση Big Data και έτσι χρησιμοποιούνται μη σχεσιακές βάσεις (NoSQL) όπως οι MongoDB [Mon], Dynamo [Dyn], Cassandra [Cas] και άλλες.

- *Μοντελοποίηση και Ανάλυση*: Όπου ορίζονται μέθοδοι για την εξόρυξη πληροφορίας (data mining) με σκοπό την ανακάλυψη κρυμμένων χαρακτηριστικών και συσχετίσεων μεταξύ των δεδομένων που μπορούν να οδηγήσουν σε χρήσιμα συμπεράσματα.

- *Ερμηνεία*: Όπου ο αναλυτής με την βοήθεια κατάλληλων προγραμματιστικών εργαλείων καλείται να ερμηνεύσει το αποτέλεσμα της ανάλυσης και να αποφανθεί.

Στο Σχήμα 15 φαίνονται εποπτικά οι διάφορες φάσεις ανάλυσης των Μεγάλων Δεδομένων.



Σχήμα 15: Ο Κύκλος Ζωής των Μεγάλων Δεδομένων.

Στην προσπάθεια για την πληρέστερη αξιοποίηση των Big Data, ο αναλυτής έρχεται αντιμέτωπος με συγκεκριμένες προκλήσεις. Καταρχήν η πολυπλοκότητα των δεδομένων καθώς όπως αναφέρθηκε παραπάνω τα δεδομένα λαμβάνονται από διαφορετικές πηγές, είναι δυνατόν να είναι είτε δομημένα (structured), είτε αδόμητα (unstructured), ή να έχουν λανθασμένες τιμές. Ακόμη ένας παράγοντας που πρέπει να ληφθεί υπόψη κατά την ανάλυση είναι η υπολογιστική πολυπλοκότητα που απαιτείται για την ανάλυσή τους, ιδιαίτερα όταν το ζητούμενο είναι η ταχύτερη απόκριση του συστήματος. Τέλος, πρόκληση αποτελούν και οι απαιτήσεις σε υλικό (hardware) και υλισμικό για την αποθήκευση και ανάκτηση των δεδομένων κατά τον κύκλο ζωής των δεδομένων.

Όσον αφορά τις φάσεις της Αναπαράστασης και της Μοντελοποίησης, αποτελεί πρόβλημα η μεγάλη διάσταση των παρατηρήσεων. Είναι γνωστό ότι τα Μεγάλα Δεδομένα πάσχουν και αυτά από την αποκαλούμενη "κατάρα των διαστάσεων" ("curse of dimensionality") [SGM14]. Μια παρατήρηση (observation) μπορεί να αποτελείται από χιλιάδες χαρακτηριστικά (features),

το πλήθος των οποίων ορίζει την λεγόμενη διάσταση των δεδομένων. Προκύπτει λοιπόν συχνά η ανάγκη για Μείωση των Διαστάσεων (dimensionality reduction), έτσι ώστε να μείνουν μόνο τα χαρακτηριστικά εκείνα τα οποία αρκούν για μια όσο το δυνατόν ακριβέστερη μελέτη. Επιπλέον με αυτόν τον τρόπο βελτιώνεται η ταχύτητα απόκρισης, ελαττώνεται η πολυπλοκότητα της ανάλυσης και διευκολύνεται η αποθήκευσή τους. Στην επόμενη παράγραφο θα δούμε πως μπορούμε να χρησιμοποιήσουμε το μαθηματικό εργαλείο των γράφων για να αναλύσουμε μεγάλα δεδομένα.

4.2 Γράφοι Εγγύτητας

Οι γράφοι εγγύτητας (Proximity Graphs) είναι γράφοι που προκύπτουν από την προσθήκη ακμών μεταξύ σημείων σε έναν μετρικό χώρο. Τα σημεία παίζουν τον ρόλο των κορυφών του γραφήματος. Μια ακμή προστίθεται στο γράφημα αν τα άκρα της είναι «κοντά» το ένα με το άλλο σύμφωνα με μια μετρική.

Οι γράφοι εγγύτητας φανερώνουν σχέσεις μεταξύ των σημείων του χώρου και χρησιμοποιούνται σε εφαρμογές μηχανικής εκμάθησης (machine learning), ομαδοποίησης, κ.α. Ακόμα, αποτελούν ένα εργαλείο για να επιτευχθεί μείωση της διάστασης των δεδομένων (dimensionality reduction). Για την κατασκευή του γράφου εγγύτητας, δηλαδή για τον ορισμό της μετρικής εκείνης που ορίζει το αν θα συνδεθούν ή όχι, υπάρχουν πολλές εναλλακτικές. Στη Διπλωματική Εργασία παρουσιάζονται δυο από αυτές, οι Ευκλείδεια Απόσταση και η μέθοδος Διακριτών Δέντρων Επικάλυψης Ελάχιστου Κόστους (Disjoint Minimum Spanning Trees, DMST).

4.2.1 Ευκλείδεια Απόσταση

Η Ευκλείδεια Απόσταση αποτελεί έναν από τους απλούστερους τρόπους κατασκευής ενός γράφου εγγύτητας. Μια ακμή προτίθεται μεταξύ δυο σημείων x, y αν και μόνο αν η μεταξύ τους Ευκλείδεια απόσταση δεν υπερβαίνει κάποιο καθορισμένο από τον χρήστη κατώφλι.
Δηλαδή:

$$(x, y) \in E \leftrightarrow d(x, y) \leq t$$

όπου t είναι το προκαθορισμένο κατώφλι.

Η μέθοδος αυτή δημιουργεί συνήθως πυκνούς γράφους με υψηλή συνεκτικότητα. Επίσης, πρόβλημα αποτελεί η σωστή επιλογή του κατωφλίου καθώς πολύ χαμηλές τιμές μπορεί να προκαλέσουν τη δημιουργία πολλών συνεκτικών συνιστωσών, ενώ αντιθέτως υψηλές τιμές τείνουν να δημιουργούν πολλές συνδέσεις και μεταξύ απομακρυσμένων σημείων του επιπέδου, εκφυλίζοντας έτσι την σημασία της θέσης τους στον χώρο. Γενικά, πρόκειται για μια μέθοδο της οποίας η επιλογή της βέλτιστης τιμής της βασικής παραμέτρου κατωφλίου εξαρτάται κάθε φορά από τα δεδομένα.

4.2.2 Διακριτά Δέντρα Επικάλυψης Ελάχιστου Κόστους

Τα δέντρα επικάλυψης ελάχιστου κόστους έχουν την επιθυμητή ιδιότητα, όταν πρόκειται για κατασκευή γράφων εγγύτητας, να αποφεύγουν τις ακμές μεγάλου βάρους, αφού πρέπει να ελαχιστοποιήσουν το βάρος του δέντρου. Επιπλέον, κατασκευάζουν πάντα έναν συνδεδεμένο γράφο $G = (V, E)$ με $|E| = |V| - 1$. Βέβαια, ένας τέτοιος γράφος έχει πολύ λίγες ακμές για να συλλάβει την δομή των δεδομένων στον χώρο και επίσης είναι ευαίσθητος στον θόρυβο. Μια μικρή μεταβολή στα δεδομένα μπορεί να οδηγήσει σε τελείως διαφορετικό γράφημα. Μια από τις λύσεις που προτείνεται στο [CaR05] είναι η συνένωση t τον αριθμό διακριτών δέντρων επικάλυψης ελάχιστου κόστους. Η παράμετρος t ορίζεται από τον χρήστη.

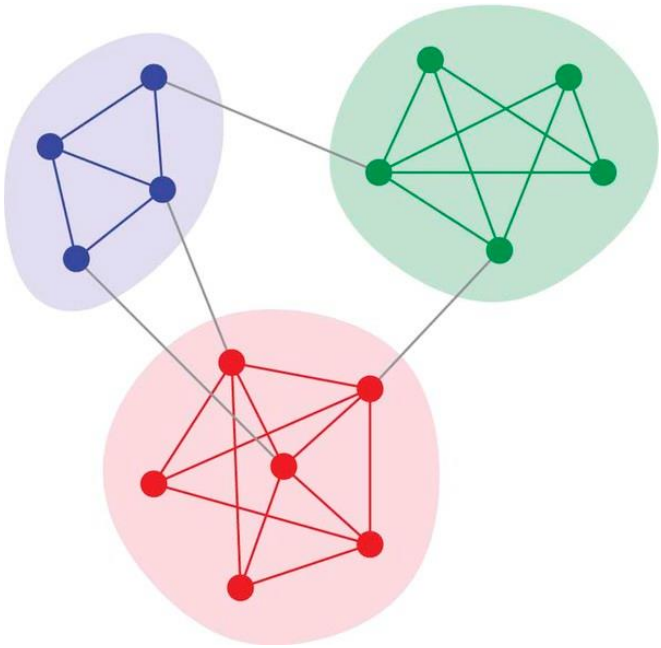
Στην μέθοδο αυτή, συνενώνονται τα t πρώτα MSTs. Ένας τρόπος για να γίνει αυτό είναι ο εξής: έχοντας κατασκευάσει τον πλήρη γράφο, όπου κάθε σημείο συνδέεται απευθείας με κάθε άλλο και με τα βάρη των ακμών να είναι η απόσταση (εδώ η Ευκλείδεια, αλλά μπορεί να χρησιμοποιηθεί οποιαδήποτε μετρική απόστασης) υπολογίζεται το MST του γράφου. Έπειτα, αφαιρούνται οι ακμές που χρησιμοποιήθηκαν για την κατασκευή του και υπολογίζεται στο νέο γράφο το MST. Η διαδικασία αυτή συνεχίζεται μέχρις ότου έχουν κατασκευαστεί t δέντρα τα οποία στη συνέχεια ενώνονται, σχηματίζοντας έτσι τον Γράφο Εγγύτητας.

Κατασκευάζοντας έτσι τον Γράφο Εγγύτητας έχουμε ένα σχετικά μικρό αριθμό ακμών, ίσο με $t(|V| - 1)$, αφού καλά αποτελέσματα παρουσιάζονται για μικρές τιμές του t. Επίσης, ο γράφος που κατασκευάζεται έτσι έχει μεγαλύτερη ανθεκτικότητα στον θόρυβο από ότι ένα μόνο MST.

5

Διαμέριση Γράφων – Ανακάλυψη Κοινοτήτων

Ένα χαρακτηριστικό των Σύνθετων Δικτύων είναι η ιδιότητά τους να σχηματίζουν κοινότητες. Κοινότητες προκύπτουν με φυσικό τρόπο στην κοινωνία, για παράδειγμα οικογένεια, φίλοι, συμμαθητές, κ.λπ. Προκύπτουν επίσης στην βιολογία καθώς για παράδειγμα έχει βρεθεί ότι πρωτεΐνες που εκτελούν παραπλήσιες λειτουργίες μέσα στο κύτταρο τείνουν να δημιουργούν ομάδες [RiG03]. Επίσης, στον Παγκόσμιο Ιστό (WWW) σελίδες με παρόμοιο περιεχόμενο σχηματίζουν κοινότητες [FLG02]. Ακόμα, σε γράφους που έχουν παραχθεί από χωρικά δεδομένα (spatial data), κόμβοι που βρίσκονται κοντά στον χώρο θα έχουν περισσότερους κοινούς γείτονες από ότι δυο κόμβοι σε μεγαλύτερη απόσταση.



Σχήμα 16: Απλό δίκτυο όπου διακρίνονται τρεις κοινότητες.

Αντικείμενο της Διαμέρισης Γράφων (graph partitioning ή graph clustering) είναι η ανακάλυψη, ή ο εντοπισμός, των κοινοτήτων που απαρτίζουν τον γράφο, χωρίς πρότερη γνώση τους. Για παράδειγμα στον γράφο του Σχήματος 16 ένας αλγόριθμος θα έπρεπε να βρει τρεις κοινότητες. Το πρόβλημα του Εντοπισμού Κοινοτήτων (community detection) είναι ασαφώς ορισμένο καθώς διάφοροι μελετητές έχουν απαντήσει με διαφορετικό τρόπο στο ζήτημα των ιδιοτήτων μιας κοινότητας. Συνεπώς δεν υπάρχει κάποιο καθολικά αποδεκτό μέτρο της ακρίβειας μιας διαμέρισης. Αυτό έχει οδηγήσει σε μια πληθώρα αλγόριθμων Εντοπισμού Κοινοτήτων, καθένας από τους οποίους χρησιμοποιεί διαφορετικά κριτήρια και είναι δυνατόν να παράγει αρκετά διαφορετικά αποτελέσματα σε σχέση με τους υπόλοιπους.

Σε γενικές γραμμές οι αλγόριθμοι Εντοπισμού Κοινοτήτων μπορούν να ταξινομηθούν σε τέσσερις κατηγορίες ανάλογα με το συστατικό μέρος του δικτύου στο οποίο επικεντρώνουν.

- Με επίκεντρο του κόμβους του δικτύου (node-centric). Σε αυτές τις μεθόδους κάθε κόμβος του δικτύου πρέπει να πληρεί κάποιες συγκεκριμένες ιδιότητες που ορίζονται από την εκάστοτε μέθοδο. Παράδειγμα τέτοιων μεθόδων αποτελούν μέθοδοι που βασίζονται στον εντοπισμό κλικών στον γράφο. Τέτοια προσέγγιση αποτελεί η μέθοδος των Bron-Kerbosh [BrK73] όπου ο αλγόριθμός τους βρίσκει της μεγιστοτικές (maximal) κλίκες στον γράφο. Μια άλλη προσέγγιση [PDF05] επιτυγχάνει την ανίχνευση κοινοτήτων βρίσκοντας *k*-κλίκες. Μια *k*-κλίκα είναι ένα μεγιστοτικό υπογράφημα που η μεγαλύτερη απόσταση μεταξύ δυο κόμβων δεν υπερβαίνει τα *k* βήματα.

- Με επίκεντρο την ομάδα (group-centric). Μέθοδοι που εντάσσονται σε αυτή την κατηγορία απαιτούν τα σύνολα των κόμβων (ομάδες ή groups) να ικανοποιούν κάποια συνθήκη. Για παράδειγμα η πυκνότητα της ομάδας να είναι μεγαλύτερη από ένα κατώφλι.

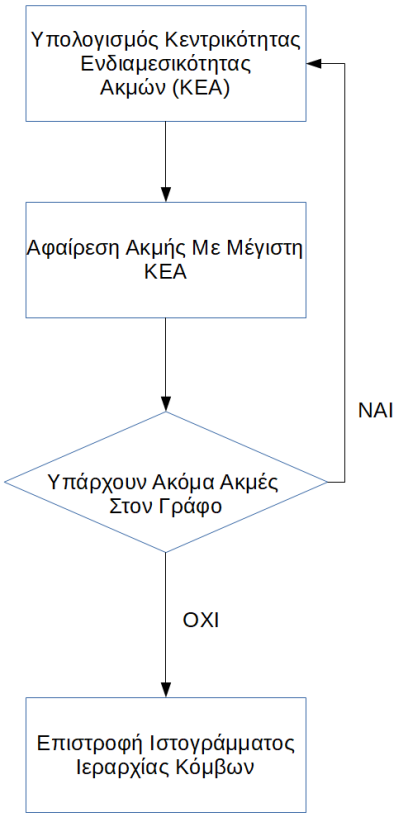
- Με επίκεντρο το δίκτυο (network-centric). Τέτοιες μέθοδοι διαμερίζουν το αρχικό δίκτυο σε μη-επικαλυπτόμενα σύνολα από κόμβους. Στην κατηγορία αυτή εντάσσονται αλγόριθμοι όπως ο *k*-means [Llo82] που μπορεί να χρησιμοποιηθεί και για τον εντοπισμό κοινοτήτων σε δίκτυα αν θεωρήσουμε ως είσοδο στον αλγόριθμο τις συντεταγμένες των κόμβων σε κάποιον Γεωμετρικό Χώρο. Εδώ εντάσσονται και οi Φασματικές Μέθοδοι Ομαδοποίησης (Spectral Clustering) όπως η μέθοδος που περιγράφεται στο [NJW02] αλλά και η μέθοδος της Μεγιστοποίησης της Αρθρωτότητας στην οποία γίνεται εκτενέστερη αναφορά στην ενότητα 5.2

- Ιεραρχικές μέθοδοι (hierarchy-centric methods). Στις ιεραρχικές μεθόδους στόχος είναι η δημιουργία ιεραρχίας κοινοτήτων. Για τον σκοπό αυτό υπάρχουν δυο πιθανές προσεγγίσεις. Στην πρώτη, όλο το δίκτυο θεωρείται ως μια κοινότητα και σταδιακά, διασπάται σε περισσότερες (divisive hierarchical clustering). Τέτοιου τύπου είναι και ο αλγόριθμος των Newman-Girvan που εξετάζεται στην ενότητα 5.1. Στην δεύτερη προσέγγιση, κάθε κόμβος του δικτύου θεωρείται ως μια κοινότητα και σταδιακά οι κοινότητες που ικανοποιούν κάποιο κριτήριο ομοιότητας (similarity) συγχωνεύονται για να δημιουργήσουν μεγαλύτερες κοινότητες (agglomerative hierarchical clustering). Από τις πρώτες προσπάθειες για τέτοιου είδους ιεραρχική ομαδοποίηση αποτελούν τα «σχήματα» Ομαδοποίησης του Johnson [Joh67].

5.1 Ο Αλγόριθμος των Newman-Girvan

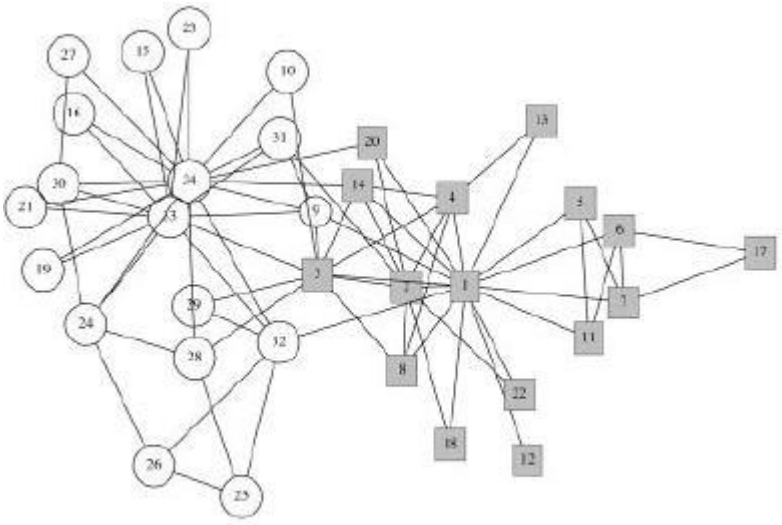
Το 2002 προτάθηκε από τους Newman και Girvan ένας αλγόριθμος για τον Εντοπισμό Κοινοτήτων ο οποίος βασίζεται στην έννοια της Κεντρικότητας Ενδιαμεσικότητας Ακμής (Edge Betweenness Centrality) [GiN02]. Κάθε φορά ο αλγόριθμος υπολογίζει την ακμή με την μεγαλύτερη τέτοια κεντρικότητα και την αφαιρεί από τον γράφο. Η διαδικασία συνεχίζεται μέχρις ότου να μην υπάρχουν πια διαθέσιμες ακμές. Σαν αποτέλεσμα επιστρέφεται ένα ιστόγραμμα με την ιεραρχική δομή του δικτύου. Ο αλγόριθμος στηρίζεται στην υπόθεση ότι οι ακμές που υπάρχουν μεταξύ των διαφορετικών κοινοτήτων είναι λιγότερες από ότι οι ακμές που συνδέουν κορυφές που ανήκουν στην ίδια κοινότητα. Συνεπώς οι ακμές που παίζουν τον ρόλο

των «γεφυρών» μεταξύ των κοινοτήτων έχουν υψηλή Κεντρικότητα Ενδιαμεσικότητας Ακμής και η αφαίρεσή τους θα οδηγήσει στον εντοπισμό των κοινοτήτων. Παρακάτω, στο Σχήμα 17, δίνεται το διάγραμμα του αλγόριθμου.



Σχήμα 17: Διάγραμμα Ροής του αλγόριθμου Newman-Girvan.

Ο αλγόριθμος των Newman-Girvan (NG) αποτελεί έναν από τους πλέον γνωστούς αλγόριθμους Εντοπισμού Κοινοτήτων και δίνει ικανοποιητικά αποτελέσματα σε γνωστά δίκτυα. Για παράδειγμα στο Zachary's Karate Club, που αποτελεί έναν καλό γράφο αναφοράς για την επαλήθευση τέτοιων μεθόδων, κάνει λάθος στην κατηγοριοποίηση μόνο ενός κόμβου, όπως φαίνεται και στο Σχήμα 18. Το δίκτυο αυτό θα μας απασχολήσει περισσότερο στο Κεφάλαιο 8 και συγκεκριμένα στην ενότητα 8.1.



Σχήμα 18: Αποτέλεσμα του Girvan-Newman στο Zachary's Karate Club, [GiN02].

Μειονέκτημα όμως του αλγόριθμου αποτελεί ο αργός υπολογισμός της Κεντρικότητας Ενδιαμεσικότητας Ακμής για όλες τις ακμές του γράφου που απαιτείται μετά από κάθε αφαίρεση ακμής. Ο αλγόριθμος έχει συνολικό χρόνο εκτέλεσης $O(m * n^2)$, όπου m ο αριθμός των ακμών και n ο αριθμός των κορυφών του δικτύου. Λόγω της πολυπλοκότητάς του χρησιμοποιείται μέχρι σε μεσαίου μεγέθους τοπολογίες.

5.2 Μεγιστοποίηση της Αρθρωτότητας

Οι κοινότητες που σχηματίζει ένα Σύνθετο Δίκτυο αναμένεται να έχουν περισσότερες συνδέσεις (ακμές) στο εσωτερικό τους από ότι μεταξύ τους. Ένας κόμβος δηλαδή που ανήκει σε μια κοινότητα θα έχει περισσότερους γείτονες που ανήκουν στην ίδια κοινότητα από ότι σε κάποια διαφορετική.

Η Αρθρωτότητα (modularity) είναι μια μετρική της δομής ενός δικτύου. Πιο συγκεκριμένα, αποτελεί μια μετρική που υπολογίζει το πόσο «σωστή» είναι μια διαμέριση ενός δικτύου σε κοινότητες. Αφού το δίκτυο έχει διαμεριστεί σε k κοινότητες, ορίζεται ένας συμμετρικός πίνακας e διαστάσεων $k\ x\ k$ όπου κάθε στοιχείο e_{ij} είναι το ποσοστό των ακμών του δικτύου που συνδέει τις κοινότητες i, j. Το ίχνος του πίνακα, $Tr(e) = \sum_i e_{ii}$, ορίζει το ποσοστό των ακμών που βρίσκονται εντός των κοινοτήτων. Ακόμα ορίζουμε το άθροισμα γραμμής ως $a_i = \sum_j e_{ij}$, που αντιπροσωπεύει το ποσοστό των ακμών που συνδέονται στην κοινότητα i. Η Αρθρωτότητα ορίζεται συνεπώς ως εξής:

$$Q = \sum_i (e_{ii} - a_i^2) = Tr(e) - \|e^2\|$$

Η Αρθρωτότητα έχει πεδίο τιμών το $[-1,1]$. Όσο πιο κοντά στο 1 είναι η τιμή της, τόσο πιο καλή είναι η διαμέριση του δικτύου σε κοινότητες. Τιμές του Q κοντά στο μηδέν φανερώνουν

ότι η διαμέριση που έγινε δεν είναι καλύτερη από μια τυχαία διαμέριση του δικτύου.

Η μέθοδος διαμέρισης που λέγεται Μεγιστοποίηση Αρθρωτότητας (modularity maximization) έχει ως στόχο την εύρεση του αριθμού κοινοτήτων που μεγιστοποιεί την τιμή Q. Προφανώς, το πρόβλημα της εύρεσης κάθε πιθανής διαμέρισης ανήκει στα λεγόμενα NP προβλήματα. Για την επίλυση του προβλήματος ο Newman [New04] πρότεινε την παρακάτω μεθοδολογία. Αρχικά όλοι οι κόμβοι λογίζονται ως ξεχωριστές κοινότητες και δεν υπάρχουν ακμές στον γράφο. Σε κάθε βήμα της μεθόδου υπολογίζεται η προσθήκη ποιας ακμής βελτιώνει το modularity και γίνεται σύμπτυξη των κοινοτήτων που βρίσκονται στα άκρα της. Η διαδικασία αυτή εκτελείται τόσες φορές όσοι είναι και οι κόμβοι του δικτύου. Αυτή η πρώτη προσέγγιση εκτελείται σε χρόνο $O((n+m)n)$ ή $O(n^2)$ για αραιό γράφο, όπου n ο αριθμός των κόμβων και m ο αριθμός των ακμών.

Στη συνέχεια ο Blondel και οι συνεργάτες του στο [BGL08] υποστήριξαν ότι επιτυγχάνεται η Μεγιστοποίηση της Αρθωτότητας αν ξεκινώντας ξανά με κάθε μεμονωμένο κόμβο να ορίζει μια κοινότητα σε κάθε βήμα τοποθετείται ένας κόμβος i στην κοινότητα ενός γειτονικού του κόμβου j έτσι ώστε να αυξάνεται κατά το μέγιστο η τιμή Q. Στην συνέχεια η κοινότητα αντικαθίσταται με έναν «υπερ-κόμβο» που έχει γείτονες όλους τους κόμβους που με τους οποίους γειτνιάζουν οι κόμβοι που ανήκουν σε αυτόν. Η διαδικασία αυτή επαναλαμβάνεται παράγοντας ταυτόχρονα μια ιεραρχική δομή του δικτύου, καθώς ανακαλύπτει κοινότητες μεταξύ των «υπερ-κόμβων». Η διαδικασία τερματίζει όταν σταματήσει να αυξάνεται η τιμή της Αρθρωτότητας. Ο συγκεκριμένος αλγόριθμος έχει χρόνο εκτέλεσης $O(m)$ κάτι που τον καθιστά αρκετά πιο γρήγορο από τον αλγόριθμο του Newman.

Ακόμα έχουν προταθεί για την επίλυση του προβλήματος μέθοδοι όπως αυτή που χρησιμοποιεί Προσομοιωμένη Ανόπτυση (Simmulated Annealing) [GPA04], φασματικές μέθοδοι [New06] και άλλες. Στο πλαίσιο της εργασίας θα χρησιμοποιηθεί η μέθοδος του Blondel.

6

Ενσωμάτωση Γράφων Στον Υπερβολικό Χώρο

Η διαδικασία κατά την οποία αποδίδονται συντεταγμένες ενός γεωμετρικού χώρου σε κάθε κορυφή ενός γράφου ονομάζεται «Ενσωμάτωση» (Embedding). Η ιδέα μπορεί να χρησιμοποιηθεί για την μείωση των διαστάσεων ενός συνόλου δεδομένων [YXZ07]. Έτσι, ενώ αρχικά έχουμε δεδομένα D διαστάσεων, σχηματίζουμε έναν γράφο ομοιότητας (Proximity Graph), όπου ενώνουμε με ακμές τα σημεία εκείνα, που πλέον αντιπροσωπεύουν τις κορυφές του γράφου, που είναι πλησιέστερα το ένα στο άλλο σύμφωνα με κάποια μετρική ομοιότητας (π.χ., Ευκλείδεια απόσταση, απόσταση Manhattan, αριθμός κοινών χαρακτηριστικών, κ.α.). Έπειτα, αποδίδονται σε κάθε κορυφή συντεταγμένες ενός χώρου διάστασης d με $d \ll D$, έτσι ώστε να διατηρούνται κατά το δυνατόν οι αρχικές αποστάσεις μεταξύ των σημείων.

Ακόμα, με διαφορετική αφετηρία, η ιδέα της Ενσωμάτωσης μπορεί να χρησιμοποιηθεί για να βελτιώσει την ταχύτητα ανάλυσης των πολύ μεγάλων Σύνθετων Δικτύων που υπάρχουν σήμερα. Είναι γνωστό ότι σε πολύ μεγάλα δίκτυα αυξάνει σε σημαντικό βαθμό ο χρόνος για τον υπολογισμό σημαντικών μετρικών τους, όπως το μήκος ενός ελάχιστου μονοπατιού που συνδέει δυο κόμβους του δικτύου καθώς και η εύρεση του ίδιου του μονοπατιού. Ο υπολογισμός μετρικών που βασίζονται στην απόσταση δυο κόμβων είναι αναγκαίος για εφαρμογές κοινωνικής δικτύωσης, όπως προτάσεις για φίλους ανάλογα με το ποιοι χρήστες είναι «κοντά» σε κάποιον άλλο χρήστη, εφαρμογές ηλεκτρονικών καταστημάτων που προτείνουν παρεμφερή προϊόντα για αγορά, διαφημιστικές καμπάνιες, κ.λπ.

Σε αυτή την κατεύθυνση μια προσπάθεια για την επίλυση του ζητήματος αποτέλεσε το σύστημα Orion [ZSW10] το οποίο απέδιδε συντεταγμένες του Ευκλείδειου χώρου σε κάθε κόμβο του δικτύου με τέτοιο τρόπο που η απόσταση δύο σημείων να προσεγγίζει την απόσταση των αντίστοιχων κόμβων στο δίκτυο. Το σφάλμα όμως στην προσέγγιση αυτή ήταν αρκετά σημαντικό, της τάξης του 10% - 20%.

Παρόλα αυτά όμως η αναζήτηση μιας αξιόπιστης Ενσωμάτωσης δεν μένει μόνο στον Ευκλείδειο χώρο, αλλά έχουν προταθεί και μέθοδοι για την ενσωμάτωση τόσο σε χώρο σφαιρικών συντεταγμένων [LuS08], όσο και σε υπερβολικών συντεταγμένων όπως αυτές που περιγράφονται στο [CvC09] και στο [PPK12].

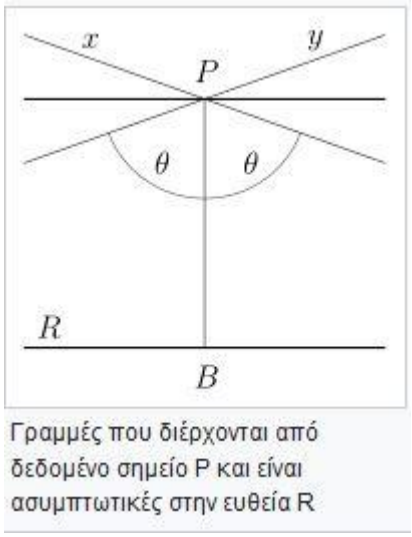
6.1 Βασικά Στοιχεία Υπερβολικής Γεωμετρίας

Η υπερβολική γεωμετρία είναι μια μη-Ευκλείδεια γεωμετρία, στην οποία το αξίωμα των παραλλήλων που δηλώνει ότι:

«Από δοθέν σημείο εκτός δοθείσης γραμμής (ευθείας), διέρχεται το πολύ μία γραμμή (ευθεία), που δεν τέμνει την δοθείσα.»

αντικαθίσταται από το παρακάτω Αξίωμα, οι συνέπειες του οποίου φαίνονται στο Σχήμα 19:

«Για οποιοδήποτε γραμμή *L* υπάρχει σημείο *P* εκτός αυτής από το οποίο διέρχονται τουλάχιστον δυο γραμμές οι οποίες δεν τέμνουν την *L*.»

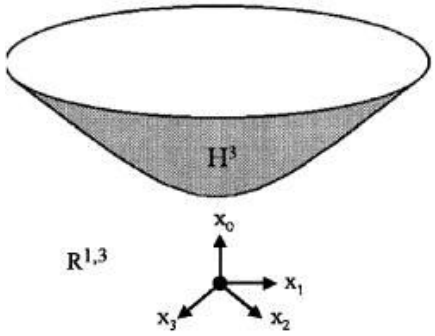


Σχήμα 19: Οι ευθείες x,y είναι παράλληλες στην R.

Ο υπερβολικός μετρικός χώρος, στον οποίο ισχύει η υπερβολική γεωμετρία, είναι ένας χώρος Riemman σταθερής τμηματικής αρνητικής καμπυλότητας. Ιδιαίτερο χαρακτηριστικό αυτού του χώρου είναι η εκθετική αύξηση του όγκου μιας μπάλας σε σχέση με την αύξηση της ακτίνας της σε σύγκριση με την γραμμική αύξηση στην περίπτωση ενός Ευκλείδειου χώρου. Το γεγονός αυτό καθιστά καλή επιλογή τον υπερβολικό χώρο για την ενσωμάτωση μεγάλου πλήθους σημείων υψηλών διαστάσεων.

Είναι προφανές ότι σε αντίθεση με τον Ευκλείδειο χώρο, το ανθρώπινο μάτι δεν μπορεί να αντιληφθεί διαισθητικά την Υπερβολική Γεωμετρία. Παρόλα αυτά όμως διάφορα πειστικά μοντέλα για την αναπαράσταση του χώρου έχουν προταθεί. Παρακάτω δίνονται οι περιγραφές δυο εξ αυτών, του μοντέλου του Υπερβολοειδούς (Hyperboloid) και του Δίσκου του Poincare.

6.1.1 Το Υπερβολοειδές



Σχήμα 20: Απεικόνιση του υπερβολοειδούς εντός χώρου Minkowski τεσσάρων διαστάσεων [Res2].

Το μοντέλο του Υπερβολοειδούς (Hyperboloid Model) (βλ. Σχήμα 20) είναι ένα από τα πέντε μοντέλα αναπαράστασης του Υπερβολικού Χώρου. Σε αυτό ο χώρος αποτελείται από ένα υποσύνολο του χώρου R^{n+1}, το οποίο αποκαλείται Πεδίου Ορισμού του χώρου.

Κάθε σημείο χ του χώρου, με $\chi = (\chi_1, \chi_2, \dots, \chi_n, x_{n+1})$, ικανοποιεί την παρακάτω εξίσωση:

$$x_1^2 + x_2^2 + \cdots + x_n^2 - x_{n+1}^2 = -1, \qquad \mu\varepsilon\ x_{n+1} > 0$$

Η απόσταση δ δυο σημείων n διαστάσεων, έστω χ και y δίνεται από τον τύπο:

$$\delta(\chi, y) = arcosh\left(\sqrt{(1 + \sum_{i=1}^{n} x_i^2) * (1 + \sum_{i=1}^{n} y_i^2)} - \sum_{i=1}^{n} x_i * y_i\right) * |c|,$$ όπου c η κυρτότητα του χώρου.

Ένα πλεονέκτημα του Υπερβολοειδούς είναι ότι ο υπολογισμός της απόστασης δυο σημείων είναι αρκετά απλούστερος σε σχέση με άλλα μοντέλα. Αυτή την ιδιότητα εκμεταλλεύεται και η Ενσωμάτωση Rigel, που παρουσιάζεται παρακάτω.

6.1.2 Ο Δίσκος του Poincaré

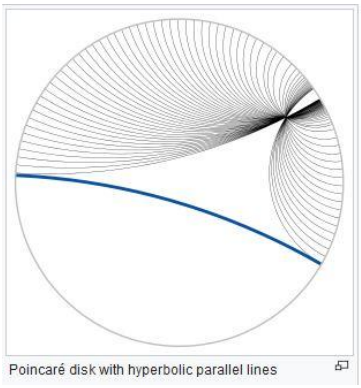
Ο δίσκος του Poincaré είναι ένα μοντέλο αναπαράστασης του δισδιάστατου Υπερβολικού Χώρου. Στο μοντέλο αυτό ο χώρος αποτελείται από όλα τα σημεία που βρίσκονται εντός ενός μοναδιαίου δίσκου, δηλαδή ενός κυκλικού δίσκου με ακτίνα ίση με 1. Δηλαδή τα σημεία για τα οποία ισχύει:

$$D = \{Z \in \mathbb{C}, |Z| < 1\}$$

Τα σημεία που ανήκουν στην περιφέρεια του δίσκου, δηλαδή τα σημεία για τα οποία ισχύει ότι

$$|Z| = 1$$

απαρτίζουν το λεγόμενο σύνορο στο άπειρο ή Ορίζοντα. Τα σημεία αυτά ονομάζονται ιδεατά (ideal points) και δεν ανήκουν στον υπερβολικό χώρο. Μια απεικόνιση του Δίσκου αποτελεί αυτή του Σχήματος 21 όπου είναι σχεδιασμένες και πολλές παράλληλες γραμμές ως προς μια γραμμή αναφοράς.



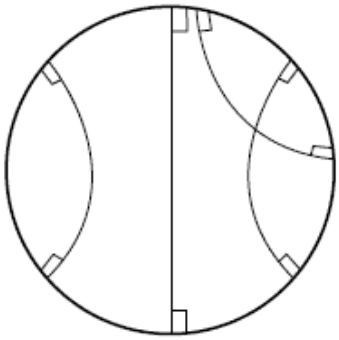
Σχήμα 21: Απεικόνιση παράλληλων γραμμών στον δίσκο του Poincaré, [Wik2].

Η απόσταση δύο σημείων στο δίσκο, έστω z_1, z_2, δίνεται από τον μαθηματικό τύπο:

$$d_D = \text{atanh}(\frac{|z_1 - z_2|}{1 - z_1 * \bar{z_2}})$$

Όπου $\bar{z_2}$ ο συζυγής μιγαδικός του z_2.

Οι γραμμές που αναπαριστούν τις αποστάσεις σημείων του υπερβολικού χώρου πάνω στον D είναι τόξα από Ευκλείδειους κύκλους που είναι ορθογώνιοι ως προς τον Ορίζοντα στα ιδεατά σημεία και δεν προεκτείνονται εκτός αυτού. Τέτοιες γραμμές αποτελούν και αυτές που απεικονίζονται στο Σχήμα 22.



Σχήμα 22: Γραμμές στον Δίσκο του Poincaré που είναι ορθογώνιες στα ιδεατά σημεία, [Sta].

6.2 Η Καταλληλότητα του Υπερβολικού Χώρου Για την Ενσωμάτωση Σύνθετων Δικτύων.

Μια πολύ σημαντική ιδιότητα των Υπερβολικών χώρων είναι ότι επεκτείνονται γρηγορότερα από τους Ευκλείδειους. Για παράδειγμα, σε ένα δισδιάστατο υπερβολικό χώρο H_ζ^2, σταθερής κυρτότητας ίσης με $K = -\zeta^2 < 0, \zeta > 0$, τόσο το μήκος του κύκλου όσο και το εμβδαδόν ενός δίσκου υπερβολικής ακτίνας r, δίνονται από τους τύπους

$$L(r) = 2\pi sinh\zeta r$$
$$A(r) = 2\pi(cosh\zeta r - 1)$$

καθένα από τα οποία αυξάνεται εκθετικά σε σχέση με την ακτίνα. Ακόμα, κάθε n-αδικό δέντρο έχει αριθμό κόμβων σε απόσταση r από την ρίζα ίσο με n^r. Συνεπώς, αν $\zeta = ln(n)$, τα n-αδικά δέντρα και ο χώρος H_ζ^2 είναι ισοδύναμοι.

Ακόμα, ένα Σύνθετο Δίκτυο αναμένεται να σχηματίζει κοινότητες, οι οποίες περιέχουν υπο-κοινότητες, μικρότερες, δηλαδή, κοινότητες εντός τους. Έτσι σχηματίζεται το δενδρόγραμμα του Δικτύου, στο οποίο φαίνονται οι σχέσεις μεταξύ των κοινοτήτων και των υπο-κοινοτήτων του. Αναμένεται λοιπόν το Δίκτυο να έχει δενδρική δομή.

Επιπλέον, έχει αποδειχθεί ότι η κατασκευή ενός τυχαίου γράφου σε υπερβολικές συντεταγμένες οδηγεί σε έναν γράφο που οι κόμβοι του ακολουθούν κατανομή βαθμών power-law. Αντιστρόφως, έχει προταθεί η εικασία, και δόθηκαν αρκετά θετικά παραδείγματα, ότι οι σύνθετες τοπολογίες έχουν σαν «κρυμμένη» γεωμετρία, την υπερβολική γεωμετρία [KPK10].

Όλα τα παραπάνω συγκλίνουν στην άποψη ότι ο Υπερβολικός Χώρος είναι μια καλή επιλογή για την ενσωμάτωση ενός Σύνθετου Δικτύου, από πολλές απόψεις.

6.3 Η Ενσωμάτωση Rigel

Η Ενσωμάτωση Rigel είναι μια ενσωμάτωση που διατηρεί τις αποστάσεις (distance preserving). Αυτό σημαίνει ότι η απόσταση των συντεταγμένων δυο κόμβων είναι συγκρίσιμη με την απόσταση τους στο δίκτυο, δηλαδή το μήκος του συντομότερου μονοπατιού που τους συνδέει.

Για ενσωματωθεί ένα δίκτυο N κόμβων σύμφωνα με την Ενσωμάτωση Rigel, αρχικά επιλέγονται $l \ll N$ κόμβοι για να παίξουν τον ρόλο των ορόσημων (landmarks). Αυτοί είναι οι πρώτοι κόμβοι που θα υπολογίσουν τις μεταξύ τους αποστάσεις, ώστε οι μεταξύ τους αποστάσεις όσον το δυνατόν καλύτερα να ανταποκρίνονται στα μήκη των μεταξύ τους συντομότερων μονοπατιών. Έπειτα, όλοι οι υπόλοιποι κόμβοι λαμβάνουν συντεταγμένες τέτοιες ώστε οι αποστάσεις τους από τα ορόσημα να διατηρούνται κατά το δυνατόν με μεγαλύτερη προσέγγιση.

Η ιδέα πίσω από την Ενσωμάτωση Rigel είναι ότι ο υπολογισμός του μήκους του συντομότερου μονοπατιού για κάθε ζεύγος κόμβων είναι μια ακριβή υπολογιστικά διαδικασία, καθώς πρέπει κάθε φορά να εκτελείται μια Αναζήτηση-Κατά-Πλάτος (Breadth First Search, BFS). Τώρα, αφού τα ορόσημα ενσωματωθούν στον χώρο κάθε άλλος κόμβος υπολογίζει την απόσταση του με ένα υποσύνολο των οροσήμων, επιλεγμένων τυχαία, εκτελώντας BFS και επιλύοντας ένα πρόβλημα βελτιστοποίησης γραμμικού προγραμματισμού με την μέθοδο Simplex, προσδιορίζονται οι συντεταγμένες του.

Οι κόμβοι που επιλέγονται ως ορόσημα είναι κόμβοι με μεγάλο βαθμό, έτσι ώστε να είναι όσον το δυνατόν πιο κεντρικοί γίνεται (υψηλή κεντρικότητα βαθμού). Η προσέγγιση αυτή, εκτός από το ότι είναι υπολογιστικά απλή, δίνει καλύτερα αποτελέσματα από μια τυχαία επιλογή των κόμβων [ZSW10].

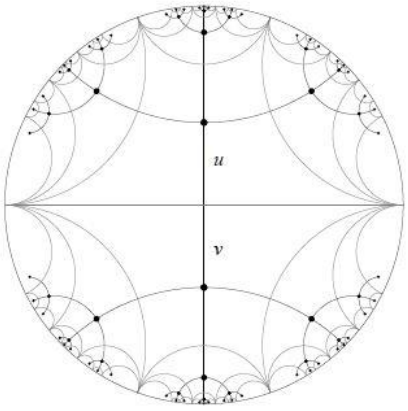
6.4 Η Άπληστη Ενσωμάτωση

Ένας τρόπος για να δρομολογηθεί ένα πακέτο από έναν κόμβο αφετηρία σε έναν κόμβο προορισμό είναι κατά μήκος του συντομότερου μονοπατιού που τους συνδέει. Βέβαια, σε μεγάλα δίκτυα η προσέγγιση αυτή είναι υπολογιστικά ακριβή καθώς απαιτεί μια εκτέλεση του αλγόριθμου της Αναζήτησης Κατά Πλάτος (BFS) για κάθε τέτοιο ζεύγος, σε περίπτωση γράφου χωρίς βάρη στις ακμές, ή μια εκτέλεση του αλγόριθμου του Dijkstra κάθε φορά σε περίπτωση που οι ακμές έχουν βάρη. Ακόμα, απαιτείται κάθε κόμβος του δικτύου να διατηρεί εκτενείς πίνακες με πληροφορίες για κάθε άλλο κόμβο στο δίκτυο. Μια διαφορετική προσέγγιση στο πρόβλημα είναι κάθε κόμβος γνωρίζοντας τον προορισμό να προωθεί το πακέτο προς κάποιον κόμβο που κρίνεται ότι μειώνει περισσότερο την απόσταση προς τον παραλήπτη του. Αυτή είναι και η βασική ιδέα της Άπληστης Δρομολόγησης (Greedy Routing) που έχει χρησιμοποιηθεί σε πολλές εφαρμογές όπως φαίνεται και ενδεικτικά από τα [SYG09], [Sto02], [FPW09]. Η απλή της υλοποίηση, οι μικρές απαιτήσεις σε μνήμη καθώς και η δυνατότητα κατανεμημένης εφαρμογής καθιστούν την Άπληστη Δρομολόγηση ελκυστική επιλογή.

Αρχικά μελετήθηκε η απόδοση της Άπληστης Δρομολόγησης με χρήση των φυσικών συντεταγμένων των κόμβων. Δηλαδή, κάθε κόμβος προωθεί το πακέτο σε εκείνο τον γειτονικό του κόμβο ο οποίος απέχει το λιγότερο σε φυσική απόσταση από τον κόμβο προορισμό. Μια δρομολόγηση με τέτοιο τρόπο αποτυγχάνει όταν το πακέτο φτάσει σε κάποιο κόμβο ο οποίος απέχει λιγότερο από όλους τους γείτονές του με τον κόμβο προορισμό. Αυτό μπορεί να συμβεί ακόμα και αν υπάρχει μονοπάτι που να συνδέει αφετηρία-προορισμό.

Φυσικά, το να μετακινεί κανείς τους κόμβους ενός δικτύου για τους σκοπούς της δρομολόγησης είναι επίπονο και πολλές φορές εντελώς αδύνατο. Αντ' αυτού η λύση που προτάθηκε από κάποιους ερευνητές είναι η απόδοση συντεταγμένων στους κόμβους που δεν ανταποκρίνονται απαραίτητα στην φυσική τους θέση, δηλαδή η Ενσωμάτωση του Δικτύου σε έναν Γεωμετρικό Χώρο. Ξεκίνησε έτσι η προσπάθεια για την εξεύρεση ενός Γεωμετρικού Χώρου που θα εξασφάλιζε 100% επιτυχία στην Άπληστη Δρομολόγηση. Ο Kleinberg [Kle07] απέδειξε ότι υπάρχει τρόπος να ενσωματωθεί ένα Δίκτυο με τέτοιο τρόπο στον Υπερβολικό Χώρο δυο διαστάσεων. Μια τέτοια ενσωμάτωση, με καθολική επιτυχία Άπληστης Δρομολόγησης ονομάζεται Άπληστη Ενσωμάτωση (Greedy Embedding).

Στο [Kle07] ο Kleinberg απέδειξε ότι υπάρχει Άπληστη Ενσωμάτωση για ένα d-κανονικό δέντρο άπειρου μεγέθους με $d \geq 3$. Για να ενσωματωθεί λοιπόν ένας γράφος G, πρώτα πρέπει να επιλεγεί ένα συνεκτικό δέντρο του, T. Έπειτα, αφού βρεθεί ο μέγιστος βαθμός του δικτύου, έστω d_m, ενσωματώνεται το άπειρο d_m-κανονικό δέντρο, όπως φαίνεται και στο Σχήμα 23 στην περίπτωση του 4-κανονικού δέντρου. Τέλος, κάθε κόμβος του δέντρου T αντιστοιχίζεται σε κάποιο κόμβο του άπειρου d_m-κανονικού δέντρου. Εφόσον ενσωματώθηκε το T, είναι επόμενο ότι έχει ενσωματωθεί και ο γράφος G.



Σχήμα 23: Άπληστη Ενσωμάτωση ενός 4-κανονικού δέντρου, [Kle07].

Αργότερα οι Cvetkovski και Crovella [CvC09] επέκτειναν την ιδέα του Kleinberg έτσι ώστε να υποστηρίζει δίκτυα στα οποία προστίθενται κόμβοι με τέτοιο τρόπο που για την Ενσωμάτωση να μην χρειάζεται να επανυπολογιστούν οι συντεταγμένες των κόμβων που βρίσκονται ήδη στο δίκτυο. Ακόμα, πρότειναν και μια μέθοδο Άπληστης Δρομολόγησης που αντιμετωπίζει τυχόν απώλειες ακμών στο δίκτυο.

Κατά την Άπληστη Δρομολόγηση είναι επιθυμητό το μήκος των μονοπατιών που υπολογίζονται να είναι το κατά το δυνατόν πλησιέστερα στο μήκος των αντίστοιχων ελάχιστων μονοπατιών. Η μετρική που υποδηλώνει το πόσο σφάλμα υπάρχει μεταξύ ενός «άπληστου» μονοπατιού και του συντομότερου μεταξύ δυο κόμβων ονομάζεται Έκταση-Βήματος (hop stretch) και είναι ο λόγος του πρώτου προς το δεύτερο. Συνεπώς, όσο πιο κοντά στη μονάδα είναι το μέσο hop stretch τόσο πιο σύντομα «άπληστα» μονοπάτια υπάρχουν. Έχει αποδειχθεί ότι το hop stretch επηρεάζεται σε μεγάλο βαθμό από την επιλογή του Συνεκτικού Δέντρου. Μια καλή επιλογή που οδηγεί σε τιμές κοντά στο 1 είναι να επιλεγεί για Συνεκτικό Δέντρο το δέντρο ελάχιστου βάθους (minimum depth spanning tree), δηλαδή το δέντρο εκείνο με το μικρότερο πλήθος επιπέδων [CvC12].

7

Ανακάλυψη Κοινοτήτων Με τη Μέθοδο Hyperbolic Newman-Girvan

Στο κεφάλαιο αυτό παρουσιάζεται η μέθοδος που αναπτύξαμε για τον Εντοπισμό Κοινοτήτων τροποποιώντας τον αλγόριθμο των Newman-Girvan και χρησιμοποιώντας την ενσωμάτωση Rigel για να επιτύχουμε μια ταχεία και προσεγγιστική εύρεση του μεγέθους της Κεντρικότητας Ενδιαμεσικότητας Ακμής.

7.1 Υπολογισμός Υπερβολικής Κεντρικότητας Ενδιαμεσικότητας Ακμής

Στο Κεφάλαιο 3 παρουσιάστηκε ο αλγόριθμος του Brandes για τον υπολογισμό της Κεντρικότητας Ενδιαμεσικότητας Ακμής. Αναζητώντας μια προσεγγιστική αλλά και ταχύτερη λύση για τον υπολογισμό της μετρικής αυτής, ενσωματώνουμε το δίκτυο στον Υπερβολικο Χώρο και βασιζόμενοι στον υπολογισμό της Υπερβολικής Κεντρικότητας Ενδιαμεσικότητας Κορυφής (Hyperbolic Betweenness Centrality, HBC) που περιγράφεται στο [SSK16], τροποποιούμε κατάλληλα τον αλγόριθμο για να υπολογίζει την Υπερβολική Κεντρικότητα Ενδιαμεσικότητας Ακμής (Hyperbolic Edge Betweenness Centrality, HEBC). Ακολούθως, περιγράφεται με χρήση ψευδοκώδικα ο προτεινόμενος αλγόριθμος.

$HEBC(u,v) = 0, \forall\, u,v \in V$
$για\ κάθε\ κόμβο\ s\ \in V$:
  $\%Μέρος\ I: Ταξινόμησε\ όλους\ τους\ κόμβους\ σε\ μη$
  $αύξουσα\ σειρά\ σε\ σχέση\ με\ την\ απόστασή\ τους\ προς\ τον\ s, u_N = s$
  $S = \{u_1 \preccurlyeq u_2 \preccurlyeq \cdots \preccurlyeq u_N = s\}, S_1 = S$
  $\%Μέρος\ II$:
  $\sigma_s(u)$: $ο\ αριθμός\ των$ άπλησ$των\ μονοπατιών\ με\ αφετηρία\ τον\ κόμβο\ u$
  $και\ πέρας\ τον\ κόμβο\ s$
  $\sigma_s(u) = 0, \forall\, u \in V; \sigma_s(s) = 1;$
  $για\ i = N: 1\ κάνε$
    $για\ κάθε\ u_j: u_i\ \in N_G(j,s)$:
      $\sigma_s(u_j) = \sigma_{s(u_j)} + \sigma_s(u_i)$
      $αφαίρεσε\ το\ u_i\ από\ το\ S_1$
  $Μέρος\ III$: Άθροισμα $των\ εξαρτήσεων\ φορτίου\ (load\ dependencies)(\delta)$
  $και\ των\ τιμών\ HEBC$
  $\delta(u) = 0, \forall\, u \in V;$
  $για\ i = 1: N-1\ κάνε$
    $για\ κάθε\ u_j \in N_G(i,s)$:

15: $c = \frac{\sigma_s(u_j)}{\sigma_s(u_i)}(\delta(u_i) + 1);$

16: $HEBC(u_i, u_j) = HEBC(u_i, u_j) + c;$

17: $HEBC(u_j, u_i) = HEBC(u_j, u_i) + c;$

18 $\delta(u_j) = \delta(u_i) + c;$

19: $\alpha\varphi\alpha\iota\rho\varepsilon\sigma\varepsilon\ \tau o\ u_i\ \alpha\pi o\ \tau o\ S$

Στην γραμμή 2 του κώδικα αρχίζει ένας εξωτερικός βρόχος, ο οποίος εκτελείται θέτοντας κάθε κόμβου του δικτύου ως προορισμό. Εντός του βρόχου, στο Μέρος Ι του αλγόριθμου οι κόμβοι ταξινομούνται σε μη-αύξουσα σειρά σε σχέση με την υπερβολική απόσταση τους από τον κόμβο προορισμό, έτσι ώστε στην συνέχεια να εξεταστούν στη σωστή σειρά. Στο Μέρος ΙΙ υπολογίζεται ο αριθμός των «άπληστων» μονοπατιών μεταξύ πηγής-προορισμού. Στο τελευταίο μέρος υπολογίζεται ο λόγος $\delta(u)$ για κάθε κόμβο u του δικτύου και εν τέλει υπολογίζεται η Υπερβολική Κεντρικότητα Ενδιαμεσικότητα Ακμής (HEBC).

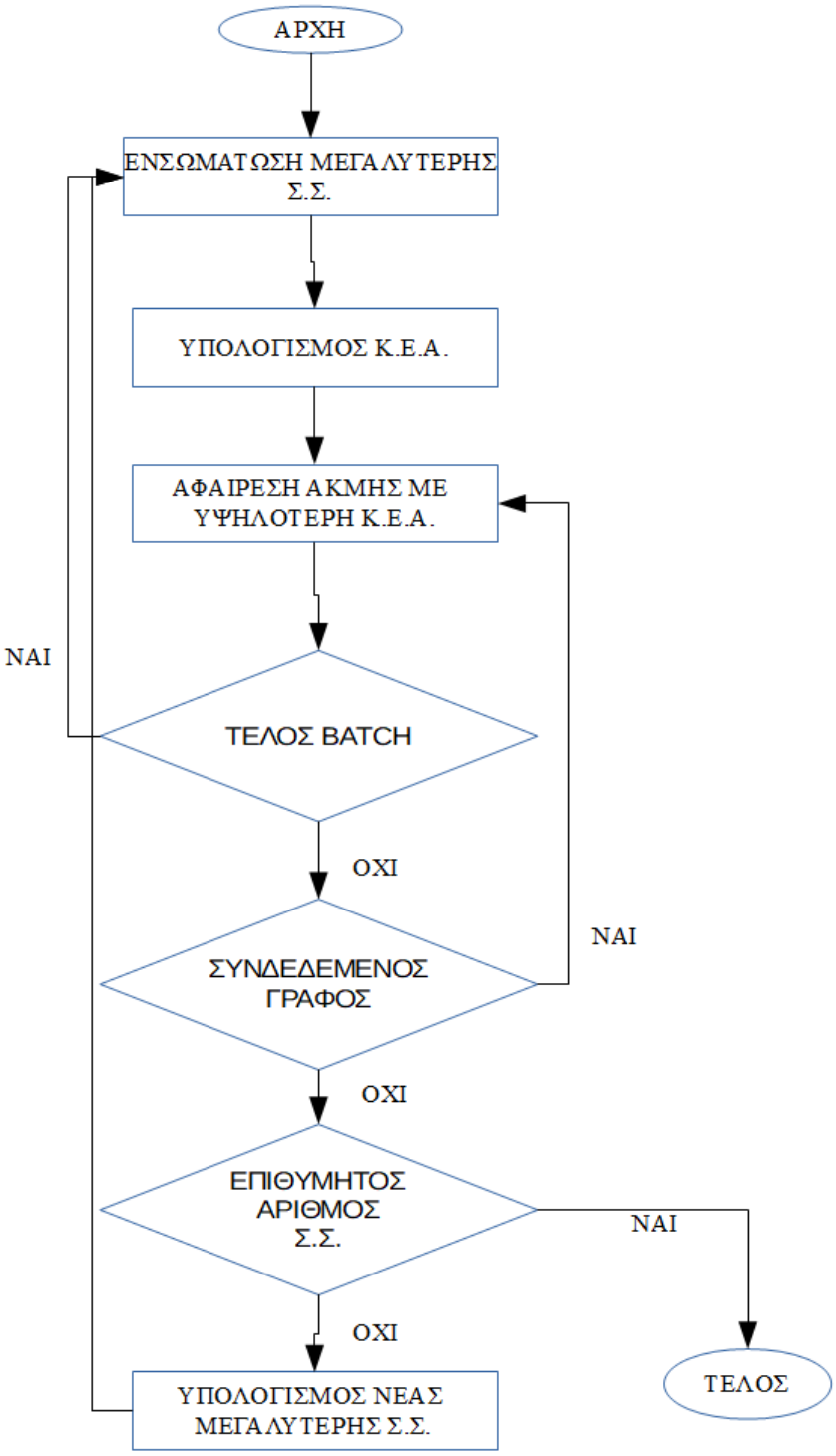
Πρέπει να σημειωθεί ότι τα άπληστα μονοπάτια που υπολογίζονται δεν ταυτίζονται πάντα με τα συντομότερα μονοπάτια. Αυτό συμβαίνει επειδή η απόσταση μεταξύ δυο κόμβων εμπεριέχει σφάλμα σε σχέση με το μήκος του συντομότερου μονοπατιού. Αυτός είναι και ο λόγος για τον οποίο τα αποτελέσματα της Υ.Κ.Ε.Α. δεν ταυτίζονται με αυτά που παράγονται από τον αλγόριθμο του Brandes. Παρόλα αυτά, όπως θα φανεί από τα αποτελέσματα των πειραμάτων που ακολουθούν, για δίκτυα Ελεύθερης Κλίμακας αλλά και αρκετά πραγματικά δίκτυα έχουμε σημαντική σύγκλιση.

7.2 Τροποποίηση του Αλγόριθμου Ομαδοποίησης Newman-Girvan

Η μέθοδος των Newman-Girvan που είδαμε σε προηγούμενο κεφάλαιο βασίζεται στον εντοπισμό κάθε φορά της ακμής με την υψηλότερη τιμή Κεντρικότητας Ενδιαμεσικότητας Ακμής, την αφαίρεσή της και τον επανυπολογισμό της Κ.Ε.Α., μέχρις ότου το σύνολο των κορυφών του αρχικού γράφου να έχει διαμεριστεί στον επιθυμητό αριθμό κοινοτήτων.

Στο πλαίσιο της εργασίας προτείνεται μια παραλλαγή του αλγόριθμου. Αρχικά, ο γράφος ενσωματώνεται στον Υπερβολικό Χώρο με χρήση της Ενσωμάτωσης Rigel, επιλέγοντας έναν μικρό αριθμό οροσήμων (landmarks). Στην συνέχεια υπολογίζεται μέσω του αλγόριθμου της προηγούμενης ενότητας η Υ.Κ.Ε.Α. για όλες τις ακμές του. Έπειτα, αντί να αφαιρείται μια ακμή τη φορά, επιλέγεται ένας μέγιστος αριθμός ακμών που ονομάζουμε Δέσμη (batch), ο οποίος εξαρτάται από το πλήθος των ακμών του γράφου και αφαιρούνται ακμές μέχρις ότου, είτε να υπάρξει αποσύνδεση του γράφου που εξετάζεται, είτε να έχουν αφαιρεθεί όλες οι ακμές που ανήκουν στη Δέσμη. Όταν ένα από τα προηγούμενα σενάρια συμβεί, ο γράφος που έχει προκύψει, είτε αφορά τον γράφο με το ίδιο πλήθος κόμβων και λιγότερες ακμές, είτε αφορά τη νέα μεγαλύτερη συνεκτική συνιστώσα, ενσωματώνεται ξανά στον Υπερβολικό Χώρο και η ίδια διαδικασία ακολουθείται ώσπου να επιτευχθεί ο διαχωρισμός του γράφου στο πλήθος των συνεκτικών συνιστωσών που έχει οριστεί. Παρακάτω δίνεται το διάγραμμα ροής του

αλγόριθμου Hyperbolic Newman Girvan (HNG) στο Σχήμα 24.



Σχήμα 24: Το διάγραμμα ροής του αλγόριθμου HNG.

Ιδιαίτερη προσοχή απαιτεί ο ορισμός των παραμέτρων σε κάθε φάση εκτέλεσης του αλγόριθμου. Για την Ενσωμάτωση Rigel επιλέγεται ένας μικρός αριθμός οροσήμων έτσι ώστε να επιτυγχάνεται γρήγορα η Ενσωμάτωση, στα πλαίσια της εργασίας το πλήθος των οροσήμων κυμαίνεται ανάλογα με την περίσταση μεταξύ 6-15. Ακόμα, πειραματικά φάνηκε πως καλή επιλογή για την Ενσωμάτωση αποτελεί ένας Υπερβολικός Χώρος εννέα διαστάσεων και καμπυλότητας ίσης με -1. Εκτός από τις παραμέτρους που αφορούν τον αλγόριθμο του Rigel, προσοχή χρειάζεται και στην επιλογή του μεγέθους Δέσμης. Μια επιλογή μικρού μεγέθους μπορεί να μην εκμεταλλεύεται στο έπακρο τις δυνατότητες του αλγόριθμου για γρήγορη ανάλυση, ενώ η επιλογή ενός μεγάλου μεγέθους Δέσμης μπορεί να οδηγήσει σε ταχύτερη μεν εκτέλεση του προγράμματος αλλά με χειρότερα αποτελέσματα.

Όπως θα φανεί από τα αποτελέσματα των πειραμάτων στο κεφάλαιο που ακολουθεί, ο προτεινόμενος αλγόριθμος επιτυγχάνει καλύτερους χρόνους εκτέλεσης για μεγάλα δίκτυα από ότι ο παραδοσιακός αλγόριθμος των Neman-Girvan εξοπλισμένος με τον αλγόριθμο του Brandes για τον υπολογισμό Κ.Ε.Α. Αντιθέτως, για δίκτυα μερικών δεκάδων ή εκατοντάδων κόμβων είναι προτιμότερη η κλασική προσέγγιση.

8

Αποτελέσματα Πειραμάτων

Στην αρχή του κεφαλαίου αυτού (ενότητα 8.1) θα παρουσιάσουμε συνοπτικά τα δίκτυα που χρησιμοποιήθηκαν για τη συλλογή αποτελεσμάτων και την εξαγωγή συμπερασμάτων. Στην ενότητα 8.2 παρουσιάζονται τα πειραματικά αποτελέσματα για τον προτεινόμενο αλγόριθμο προσεγγιστικού υπολογισμού της μετρικής Υπερβολικής Κεντρικότητας Ενδιαμεσικότητας Ακμής που προτείνουμε, δείχνοντας ότι επιτυγχάνεται ικανοποιητικό ποσοστό προσέγγισης για πολλές τοπολογίες και ιδιαίτερα τοπολογίες που είναι, ή πλησιάζουν από πλευράς δομής τα Δίκτυα Ελεύθερης Κλίμακας. Στη συνέχεια, στην ενότητα 8.3 παρουσιάζονται αποτελέσματα για την επίδραση του μεγέθους της Δέσμης (batch size) για διάφορες τιμές του σε διαφορετικές τοπολογίες παρουσιάζοντας επιπλέον και τον χρόνο εκτέλεσης του προγράμματος. Στην ενότητα 8.4 παραθέτουμε συγκριτικά αποτελέσματα για τους χρόνους εκτέλεσης και την ακρίβεια των αλγόριθμων Newman-Girvan και Hyperbolic Newman-Girvan (HNG) για δίκτυα στα οποία γνωρίζουμε από τα πριν ποιά είναι η σωστή Ομαδοποίησή τους. Τέλος, εξετάζουμε πραγματικές και συνθετικές τοπολογίες δικτύων στις οποία δεν γνωρίζουμε ποιά είναι η βέλτιστη Ομαδοποίησή τους συγκρίνοντας τα αποτελέσματα με αυτά που παράγονται από τον αλγόριθμο Μεγιστοποίησης της Αρθρωτότητας.

Η εκτέλεση των πειραμάτων που παρουσιάζονται παρακάτω, έγινε σε απλό Προσωπικό Υπολογιστή με τα εξής χαρακτηριστικά: Intel Core i5-4570 3.20 GHz 3.20 GHz, 8 GB RAM και λειτουργικό σύστημα Windows 10 (64 bit).

Η υλοποίηση των προγραμμάτων έγινε στο υπολογιστικό περιβάλλον του MATLAB. Τα αποτελέσματα παρουσιάζονται στρογγυλοποιημέναν στα δυο πρώτα μη μηδενικά δεκαδικά ψηφία.

8.1 Τοπολογίες Δικτύων Πειραμάτων

Για την όσο το δυνατόν καλύτερη τεκμηρίωση των αποτελεσμάτων της μεθόδου μας, χρησιμοποιήσαμε πλήθος δικτύων, ορισμένα συνθετικά και άλλα πραγματικά. Σε αυτήν την ενότητα θα παρουσιάσουμε τις εν λόγω τοπολογίες για την καλύτερη κατανόηση των πειραμάτων των οποίων τα αποτελέσματα παρουσιάζονται στη συνέχεια.

8.1.1 Πραγματικά Δίκτυα

Τα δίκτυα αυτά αποτελούν πραγματικά κοινωνικά δίκτυα, μικρού μεγέθους σε σχέση με τα μεγέθη που θα μπορούσαν να λάβουν σε πλήρη ανάπτυξη. Όλα τα παρακάτω πραγματικά δίκτυα ανακτήθηκαν από την ιστοσελίδα http://www-personal.umich.edu/~mejn/netdata/, όπου υπάρχουν παραπομπές για τις αρχικές πηγές τους και λεπτομέρειες για τη φύση τους. Στην εργασία χρησιμοποιούνται τα παρακάτω.

- Zachary's Karate Club (karate).

Απεικόνιση του δικτύου βλέπουμε στο Σχήμα 1 στο κεφάλαιο 2. Το δίκτυο απεικονίζει 34 μέλη ενός συλλόγου Καράτε και τις μεταξύ τους φιλικές σχέσεις. Μετά από μια διαφωνία στο Karate Club δημιουργήθηκαν δυο ομάδες. Μια γύρω από τον δάσκαλο και μια γύρω από τον ιδιοκτήτη. Το Karate Club αποτελεί έναν πολύ διαδεδομένο γράφο για αρχικό έλεγχο μεθόδων που αφορούν σε διάφορες σκοπιές της ανάλυσης κοινωνικών δικτύων [Kar].

- Κοινωνικό Δίκτυο Δελφινιών (dolphins).

Δίκτυο που απεικονίζει σχέσεις μεταξύ ενός συγκεκριμένου τύπου δελφινιών (bottlenose) που κατοικούν στην Νέα Ζηλανδία [Dol].

- Βιβλία για την πολιτική των Η.Π.Α (polbooks).

Δίκτυο που αφορά βιβλία για την αμερικάνικη πολιτική γύρω στο 2004 που πωλούνταν από την σελίδα Amazon. Οι συνδέσεις αφορούν βιβλία που αγοράζονταν συχνά μαζί [Pol].

- Οι Άθλιοι (lesmis)

Κοινωνικό δίκτυο που αφορά χαρακτήρες που εμφανίζονται μαζί στο βιβλίο «Οι Άθλιοι» του Βίκτωρος Ουγκώ [LeM].

- Αμερικάνικο Κολεγιακό Ποδόσφαιρο (football)

Δίκτυο που απεικονίζει τα παιχνίδια μεταξύ κολεγιακών ομάδων αμερικάνικου ποδοσφαίρου το 2000 [Foo].

Όλα τα παραπάνω δίκτυα έχουν μετατραπεί σε μη-κατευθυνόμενα για τις ανάγκες της εργασίας, στις περιπτώσεις των δικτύων όπου τα δεδομενα παρέχουν κατευθυνόμενες τοπολογίες.

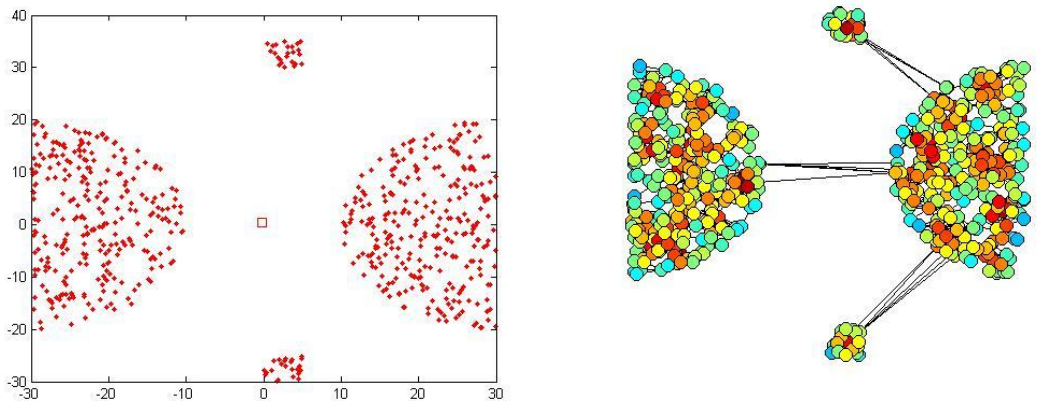
8.1.2 Συνθετικά Δίκτυα από Χωρικά Δεδομένα

Τα δίκτυα που παρουσιάζονται εδώ, δημιουργήθηκαν χρησιμοποιώντας ως βάση σύνολα δεδομένων στον δισδιάστατο χώρο που σχηματίζουν ευδιάκριτες ομάδες (clusters). Οι γράφοι που προέκυψαν είναι γράφοι εγγύτητας, οι οποίοι δημιουργήθηκαν μέσω της τεχνικής DMST που περιγράφεται στο πέμπτο Κεφάλαιο. Συγκεκριμένα, σε κάθε περίπτωση έγινε η συνένωση των πέντε πρώτων MSTs.

Τα δεδομένα που χρησιμοποιήθηκαν μπορούν να παραχθούν σε περιβάλλον MATLAB από τις συναρτήσεις που είναι διαθέσιμες στην ιστοσελίδα:

https://www.mathworks.com/matlabcentral/fileexchange/41459-6-functions-for-generating-artificial-datasets

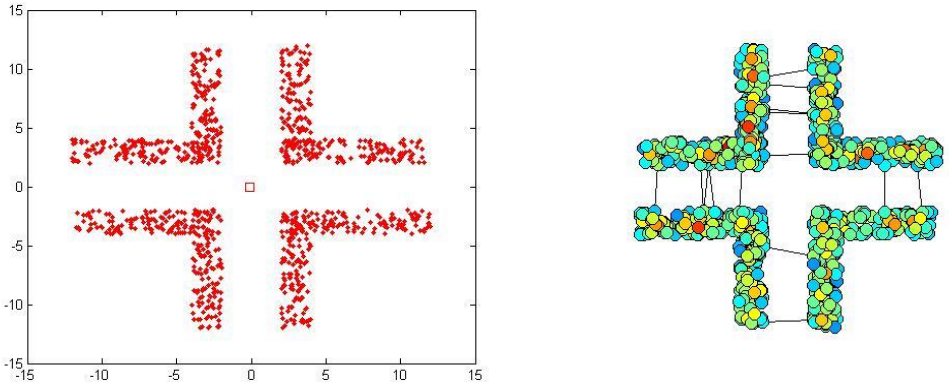
- Outliers



Σχήμα 25: Τα σύνολο δεδομένων με τίτλο Outliers (αριστερά) και ο αντίστοιχος γράφος εγγύτητας.

Όπως φαίνεται από το Σχήμα 25, το παραπάνω σύνολο 600 σημείων απαρτίζεται από τέσσερις διαφορετικές κοινότητες.

- Γωνίες (Corners)

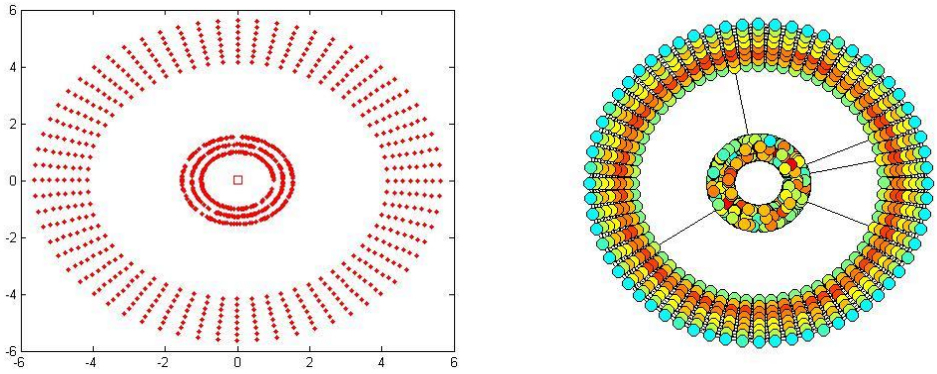


Σχήμα 26: Το σύνολο δεδομένων και ο αντίστοιχος γράφος εγγύτητας.

Μπορεί να παρατηρηθεί εύκολα (Σχήμα 26) ότι το παραπάνω σύνολο αποτελείται από τέσσερις γωνίες, οι οποίες ορίζουν τέσσερα clusters.

- Κοινότητα Εντός Κοινότητας (Cluster In Cluster)

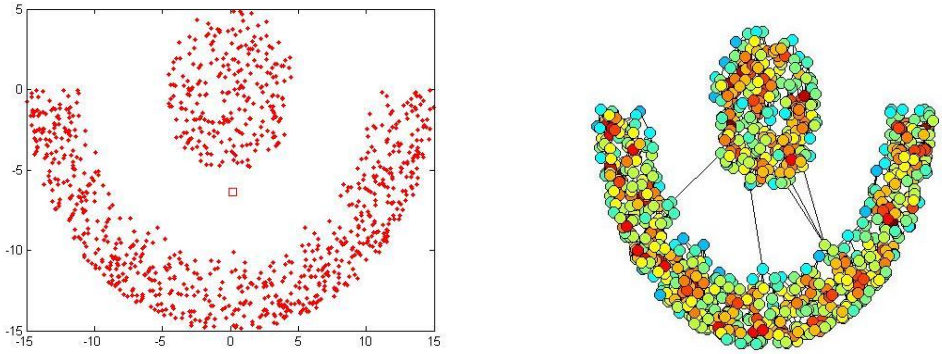
Το συγκεκριμένο σύνολο δεδομένων αποτελείται από δυο διακριτές ομάδες, τα σημεία κάθε μιας από τις οποίες σχηματίζουν κύκλο, όπως φαίνεται και στο Σχήμα 27.



Σχήμα 27: Σύνολο δεδομένων και γράφος εγγύτητας.

- Σελήνη (Fullmoon)

Το σύνολο δεδομένων «Σελήνη» αποτελείται από 1000 σημεία. Σε αυτό μπορούμε να διακρίνουμε μια κυκλική ομάδα δεδομένων, τη «Σελήνη» και μια ομάδα δεδομένων που μοιάζει με ημισέληνο (βλ. Σχήμα 28).



Σχήμα 28: Αναπαράσταση δεδομένων και γράφος εγγύτητας για το σύνολο δεδομένων «Σελήνη».

8.1.3 Τεχνητά Σύνθετα Δίκτυα

Πέρα από τα δίκτυα που παραθέσαμε μέχρι τώρα, κατασκευάσαμε και ορισμένα τεχνητά Σύνθετα Δίκτυα για τις ανάγκες της παρούσας εργασίας. Για την κατασκευή τους χρησιμοποιήσαμε το μοντέλο των Barabasi-Albert για την παραγωγή δικτύων Ελεύθερης Κλίμακας, το μοντέλο των Watts-Strogatz για την κατασκευή δικτύων Μικρού Κόσμου και τέλος για την προσομοίωση χωρικών δικτύων, το μοντέλο των Τυχαίων Γεωμετρικών Γράφων. Μερικές από τις ιδιότητες των δικτύων αυτών φαίνονται στους παρακάτω πίνακες (Πίνακας 1 – Πίνακας 3) καθένας από τους οποίους συνοψίζει βασικά στοιχεία για κάθε είδος γράφου:

Πίνακας 1: Χαρακτηριστικά Δικτύων Ελεύθερης Κλίμακας

Δίκτυο	# κόμβων	# ακμών	Ελάχιστος Βαθμός	Μέγιστη Αρθρωτότητα (Αριθμός Κοινοτήτων)
scf1	*1000*	*5792*	*6*	*0.24 (10)*
scf2	*1000*	*5821*	*6*	*0.26 (11)*
scf3	*100*	*502*	*6*	*0.28 (7)*
scf4	*100*	*356*	*4*	*0.31 (7)*
scf5	*100*	*427*	*5*	*0.27 (7)*
scf6	*100*	*557*	*7*	*0.22 (8)*

Πίνακας 2: Χαρακτηριστικά Δικτύων Μικρού Κόσμου

Δίκτυο	# κόμβων	# ακμών	# Κοντινότερων Γειτόνων	Πιθανότητα Επανασύνδεσης	Μέγιστη Αρθρωτότητα (Αριθμός Κοινοτήτων)
smw1	1000	*4284*	*4*	*0.3*	*0.77 (21)*
smw2	1000	*8298*	*8*	*0.3*	*0.81 (16)*
smw3	100	*426*	*4*	*0.3*	*0.63 (6)*
smw4	100	*315*	*3*	*0.2*	*0.68 (7)*
smw5	100	*519*	*5*	*0.3*	*0.61 (5)*
smw6	100	*507*	*5*	*0.1*	*0.63 (6)*

Πίνακας 3: Χαρακτηριστικά Τυχαίων Γεωμετρικών Γράφων

Δίκτυο	# κόμβων	# ακμών	# κατώφλι	Μέγιστη Αρθρωτότητα (Αριθμός Κοινοτήτων)
rgg1	*100*	*612*	*0.2*	*0.65 (6)*
rgg2	*100*	*1152*	*0.3*	*0.51 (4)*
rgg3	*100*	*1732*	*0.4*	*0.40 (3)*
rgg4	*100*	*2394*	*0.5*	*0.30 (3)*

Ακόμα, ακολουθώντας την μέθοδο που περιγράφεται στο [LFR08] για την παραγωγή δικτύων με κοινότητες που είναι κατάλληλα για να ελεγχθεί η ακρίβεια του αλγόριθμου Newman-Girvan, κατασκευάσαμε έξι δίκτυα. Τα μισά από αυτά έχουν 100 κόμβους και μεταβαλλόμενη πυκνότητα, ενώ τα υπόλοιπα έχουν 500 κόμβους και μεταβαλλόμενη πυκνότητα. Στον Πίνακα 4 παρουσιάζονται κάποια βασικά χαρακτηριστικά τους.

Πίνακας 4: Χαρακτηριστικά Δικτύων Με Γνωστές Κοινότητες

Δίκτυο	Αριθμός κόμβων	Αριθμός ακμών	Μέσος Βαθμός Κορυφής	Μέγιστος Βαθμός Κορυφής	Αριθμός Κοινοτήτων
Dense100	*100*	*1493*	*29.86*	*49*	*4*
Mid100	*100*	*954*	*19.08*	*30*	*3*
Sparse100	*100*	*496*	*9.92*	*15*	*4*
Dense500	*500*	*3781*	*15.12*	*30*	*7*
Mid500	*500*	*2457*	*9.83*	*20*	*5*
Sparse500	*500*	*1179*	*4.7*	*10*	*5*

8.2 Σύγκριση Κ.Ε.Α. και Υ.Κ.Ε.Α.

Στην ενότητα αυτή παρουσιάζουμε στον Πίνακα 5 την επίδοση του αλγόριθμου Υ.Κ.Ε.Α. (HEBC) για κάποιες από τις τοπολογίες που είδαμε παραπάνω. Σε κάθε περίπτωση, συγκρίνουμε το χρόνο που χρειάζεται για να ολοκληρωθεί ο υπολογισμός της μετρικής σε σχέση με την Κ.Ε.Α. όταν υπολογίζεται με τον αλγόριθμο του Brandes, καθώς επίσης φαίνεται ο χρόνος που χρειάζεται για να πραγματοποιηθεί η ενσωμάτωση του δικτύου στον Υπερβολικό Χώρο χρησιμοποιώντας την Ενσωμάτωση Rigel. Τέλος, ταξινομούμε κατά φθίνουσα σειρά, σύμφωνα με την τιμή τους, τις ακμές, και στις δυο περιπτώσεις. Στις τελευταίες τρεις στήλες παρουσιάζεται η ευστοχία του αλγόριθμου, η οποία αξιολογείται ως εξής: Μετρώνται πόσες είναι οι κοινές ακμές που εμφανίζονται στις πρώτες δέκα, τρεις και δυο θέσεις των δυο μεθοδολογιών.

Πίνακας 5 Ευστοχία και Χρονική Πολυπλοκότητα του Αλγόριθμου HEBC

Τοπολογίες	#κόμβων	#ακμών	Χρόνος K.E.A.(s)	Χρόνος Y.K.E.A (s)	Χρόνος Rigel (s) (ορόσημα)	Ακρίβεια πρώτων 10 ακμών	Ακρίβεια πρώτων 3 ακμών	Ακρίβεια πρώτων 2 ακμών
scf1	*1000*	*5792*	*105.31*	*4.36*	*18.10 (15)*	*80.00%*	*66.67%*	*100.00%*
scf2	*100*	*492*	*0.64*	*0.06*	*8.04 (15)*	*0.00%*	*0.00%*	*0.00%*
scf3	*1000*	*5821*	*121.76*	*4.07*	*8.41 (15)*	*70.00%*	*100.00%*	*100.00%*
smw1	*1000*	*4284*	*96.18*	*3.2*	*8.94 (15)*	*10.00%*	*33.33%*	*50.00%*
smw2	*1000*	*8298*	*112.62*	*3.89*	*7.25 (15)*	*20.00%*	*0.00%*	*0.00%*
karate	*34*	*78*	*0.07*	*0.01*	*9.68 (15)*	*80.00%*	*66.67%*	*50.00%*
dolphins	*62*	*166*	*0.23*	*0.02*	*4.28 (10)*	*60.00%*	*66.67%*	*50.00%*
lesmis	*77*	*258*	*0.31*	*0.04*	*5.87 (10)*	*80.00%*	*66.67%*	*50.00%*
football	*115*	*613*	*0.89*	*0.06*	*24.24 (15)*	*10.00%*	*0.00%*	*0.00%*
polbooks	*105*	*442*	*0.68*	*0.04*	*20 (15)*	*40.00%*	*33.33%*	*50.00%*
outliers	*600*	*2995*	*33.79*	*1.03*	*6.54 (6)*	*20.00%*	*0.00%*	*0.00%*
fullmoon	*1000*	*4995*	*95.78*	*1.5*	*45.32 (6)*	*10.00%*	*10.00%*	*50.00%*
corners	*1000*	*4995*	*106.03*	*2.38*	*21.73 (15)*	*10.00%*	*0.00%*	*0.00%*

Από τα στοιχεία του παραπάνω πίνακα φαίνεται ότι τα καλύτερα αποτελέσματα λαμβάνονται για τις τοπολογίες Ελεύθερης Κλίμακας μεγάλου μεγέθους. Αυτό είναι αναμενόμενο καθώς σε αυτές τις περιπτώσεις η Ενσωμάτωση είναι πιο ακριβής, αφού τα Scale-free δίκτυα έχουν, όπως έχει αναφερθεί στο [KPK10], «κρυμμένη» υπερβολική γεωμετρία και το γεγονός αυτό οδηγεί σε Ενσωμάτωση με μικρότερο σφάλμα. Επίσης, παρατηρούμε ότι στην πλειοψηφία των πραγματικών δικτύων έχουμε ικανοποιητική ευστοχία. Αυτό μπορεί να εξηγηθεί εύκολα, καθώς τα δίκτυα αυτά ως επί το πλείστον αντανακλούν σχέσεις μεταξύ οντοτήτων, πρόκειται δηλαή για Κοινωνικά Δίκτυα. Το δίκτυο football αποκλίνει σε αντιληπτό βαθμό από αυτή την περίπτωση καθώς οι ακμές του αντανακλούν αγώνες μεταξύ ομάδων, οπότε δεν αναπτύσσεται με παρόμοιο με τα άλλα τρόπο (δηλαδή στη βάση κοινωνικών μηχανισμών). Σχετικά με τα δίκτυα που είναι γράφοι εγγύτητας παρατηρούμε ότι έχουν χαμηλό ποσοστό ευστοχίας, παρόλα αυτά, το γεγονός αυτό δεν εμποδίζει την μέθοδο μας, όπως θα δούμε στη συνέχεια.

8.3 Μελέτη Απόκρισης του Αλγόριθμου Hyperbolic Newman-Girvan σε Μεταβαλλόμενο Μέγεθος Δέσμης.

Το μέγεθος της δέσμης είναι μια σημαντική παράμετρος την οποία καλείται να ορίσει ο χρήστης. Ένα μεγάλο μέγεθος δέσμης γενικά αναμένεται να επιταχύνει την εκτέλεση του αλγόριθμου, με κόστος όμως την ακρίβεια του αποτελέσματος. Από την άλλη, ένα μικρό μέγεθος παράγει πιο ακριβή αποτελέσματα όμως δεν εκμεταλλεύεται τις δυνατότητες για μια ταχύτερη εκτέλεση του προγράμματος.

Στην ενότητα αυτή παρουσιάζονται αποτελέσματα για διάφορες τοπολογίες. Από αυτές που περιγράφηκαν στην αρχή του Κεφαλαίου, έχουν επιλεγεί ορισμένες που προκύπτουν από γράφους εγγύτητας, ορισμένες από τα πραγματικά δίκτυα και τέλος κάποιες από τα τεχνητά Σύνθετα Δίκτυα.

Αξίζει να σημειωθεί ότι τόσο σε αυτή την ενότητα, όσο και στις επόμενες, ο αριθμός των οροσήμων που χρησιμοποιείται για τις ανάγκες της Ενσωμάτωσης Rigel είναι ίσος με έξι. Ο αριθμός αυτός επιλέχθηκε γιατί αφενός οδηγεί σε γρήγορη ενσωμάτωση ακόμα και δικτύων με μερικές χιλιάδες κόμβους και ακμές, αλλά επειδή παράλληλα σε πολλές περιπτώσεις δίνει αρκετά ακριβή αποτελέσματα. Φυσικά, για κάθε δίκτυο μπορεί να είναι άλλος ο βέλτιστος αριθμός ορόσημων όμως η εύρεσή του αποτελεί ανοιχτό πρόβλημα.

- Δίκτυα με γνωστές κοινότητες

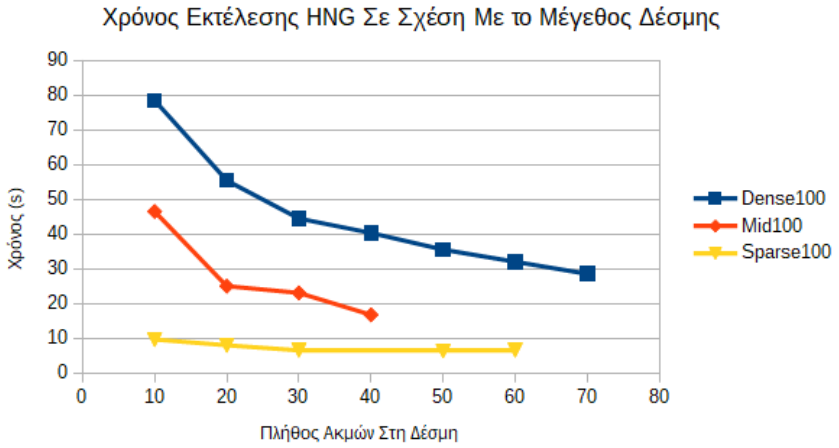
  Παρακάτω παρουσιάζονται τα αποτελέσματα για ορισμένα δίκτυα για τα οποία γνωρίζουμε τις κοινότητες τους. Σε κάθε περίπτωση, στον Πίνακα 6 και στον Πίνακα 7 περιοριζόμαστε σε παρουσίαση των αποτελεσμάτων για μεγέθη Δέσμης που οδηγούν σε 100% ευστοχία.

Πίνακας 6: Χρόνος Εκτέλεσης του HNG σε 3 Δίκτυα Με Γνωστές Κοινότητες

Μέγεθος Δέσμης	Χρόνος Εκτέλεσης (s)		
	Dense100	Mid100	Sparse100
10	*78.48*	*46.54*	*9.69*
20	*55.45*	*25.04*	*8.02*
30	*44.52*	*23.08*	*6.64*
40	*40.34*	*16.77*	-
50	*35.53*	-	*6.44*
60	*32.04*	-	*6.67*
70	*28.56*	-	-

Πίνακας 7: Χρόνος Εκτέλεσης του Αλγόριθμου HNG για τους 4 Γράφους Εγγύτητας

Μέγεθος Δέσμης	Χρόνος Εκτέλεσης (s)			
	Outliers	corners	fullmoon	Cluster in cluster
25	-	*512.36*	*187.28*	*111.5*
50	*195.39*	*233.40*	*119.25*	*107.37*
100	*58.89*	-	*236.02*	*79.62*
200	-	-	-	*119.34*
500	-	-	-	*61.64*
1000	-	-	-	*34.88*



Σχήμα 29: Διάγραμμα που απεικονίζει για τρία δίκτυα, τον χρόνο εκτέλεσης του HNG για διαφορετικά μεγέθη Δέσμης.

Όπως φαίνεται και από το Σχήμα 29, βλέπουμε ότι όσο το μέγεθος της Δέσμης αυξάνεται, τόσο μειώνεται ο χρόνος εκτέλεσης του αλγόριθμου. Η συμπεριφορά αυτή του αλγόριθμου είναι αναμενόμενη καθώς κάθε φορά περισσότερες ακμές αφαιρούνται προτού χρειαστεί να επανυπολογιστούν εκ νέου οι Κεντρικότητες Ακμών ωστέ να χρειαστεί να ξαναγίνει Ενσωμάτωση του δικτύου. Χρειάζονται λοιπόν λιγότερες Ενσωματώσεις πρωτού τερματίσει ο αλγόριθμος. Επίσης, υπάρχουν κάποιες εξαιρέσεις, όπως παρατηρούνται στις περιπτώσεις των γράφων fullmoon και cluster in cluster για τις περιπτώσεις 100 και 200 ακμών στη Δέσμη αντιστοίχως.

- Δίκτυα για τα οποία δεν υπάρχει πληροφορία για τις κοινότητες τους

Σε αυτή την ενότητα παρουσιάζονται αποτελέσματα για δίκτυα για τα οποία δεν γνωρίζουμε ποιές είναι οι κοινότητες τους. Η παρουσίαση των αποτελεσμάτων εδώ γίνεται παρουσιάζοντας τον χρόνο εκτέλεσης για διάφορα μεγέθη Δέσμης στον Πίνακα 9, καθώς και την τιμή της Αρθρωτότητάς τους αν τα χωρίζαμε σε όσες κοινότητες όσες βρίσκονται με την μέθοδο του Blondel για την Μεγιστοποίηση της Αρθρωτότητας στον Πίνακα 10. Οι τιμές αυτές δίνονται παρακάτω στον Πίνακα 8.

Πίνακας 8: Μέγιστη Αρθρωτότητα και Αριθμός Κοινοτήτων που Προκύπτουν από τον Αλγόριθμο Μεγιστοποίησης της Αρθρωτότητας για τα 4 Πραγματικά Κοινωνικά Δίκτυα

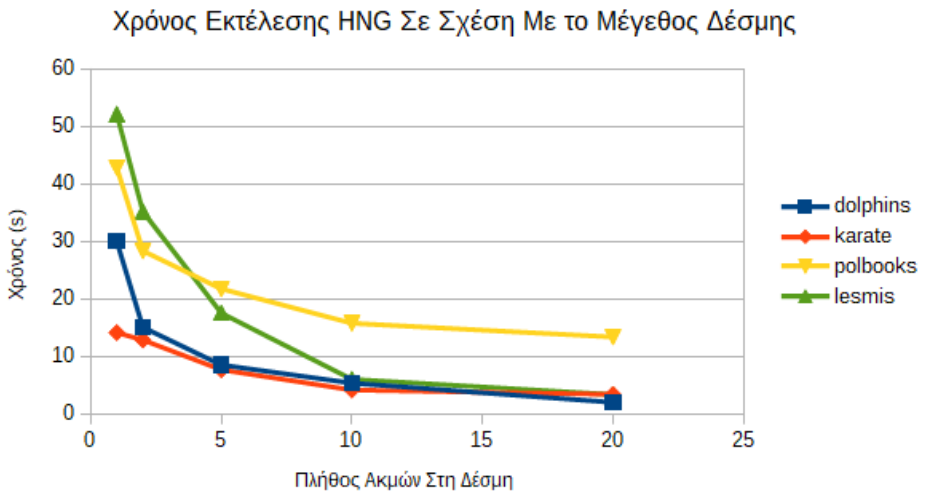
Δίκτυο	Μέγιστη Αρθρωτότητα	Αριθμός Κοινοτήτων
dolphins	*0.49*	*4*
karate	*0.42*	*5*
polbooks	*0.52*	*4*
lesmis	*0.52*	*6*

Πίνακας 9: Χρόνος Εκτέλεσης (s) του Αλγόριθμου HNG για Διάφορα Μεγέθη Δέσμης

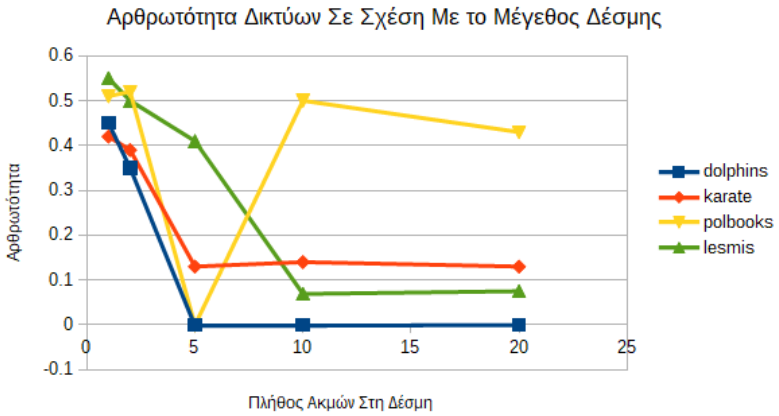
Δίκτυο	Μέγεθος Δέσμης				
	1	2	5	10	20
dolphins	*29.99*	*15.06*	*8.52*	*5.37*	*2.02*
karate	*14.14*	*12.86*	*7.69*	*4.16*	*3.41*
polbooks	*42.85*	*28.36*	*21.75*	*15.77*	*13.38*
lesmis	*52.10*	*35.17*	*17.54*	*6.02*	*3.37*

Πίνακας 10: Αρθρωτότητα των Διαμερίσεων που Προκύπτουν από τον HNG για Διάφορα Μεγέθη Δέσμης

Δίκτυο	Μέγεθος Δέσμης				
	1	2	5	10	20
dolphins	*0.45*	*0.35*	*-0.00044*	*-0.00044*	*-0.0002*
karate	*0.42*	*0.39*	*0.13*	*0.14*	*0.13*
polbooks	*0.51*	*0.52*	*-0.00045*	*0.50*	*0.43*
lesmis	*0.55*	*0.50*	*0.41*	*0.069*	*0.075*



Σχήμα 30: Διάγραμμα του χρόνου εκτέλεσης του αλγόριθμου για διάφορα Μεγέθη Δέσμης.



Σχήμα 31: Διάγραμμα της Αρθρωτότητας της παραγώμενης διαμέρισης σε κοινότητες για διάφορα Μεγέθη Δέσμης.

Από τα διαγράμματα του Σχήματος 30 και του Σχήματος 31 βλέπουμε ότι όσο αυξάνεται το μέγεθος της Δέσμης μειώνονται ο χρόνος εκτέλεσης του αλγόριθμου αλλά και η Αρθρωτότητα. Αυτό είναι αναμενόμενο καθώς όλο και περισσότερες ακμές αφαιρούνται προτού επανυπολογιστούν εκ νέου οι κεντρικότητες ακμών. Ακόμα παρατηρούμε ότι αν και η τάση της Αρθρωτότητας είναι να μειώνεται καθώς αυξάνεται το μέγεθος Δέσμης, λαμβάνουμε καλά αποτελέσματα, συγκρίσιμα με την Μέγιστη Αρθρωτότητα για μικρές τιμές Δέσμης, και σε κάποιες περιπτώσεις ακόμα και υψηλότερες τιμές, όπως για παράδειγμα στην περίπτωση του δικτύου lesmis, όπου η Αρθρωτότητα που προκύπτει είναι κατά 0.02 υψηλότερη από την μέγιστη. Σαφώς, η απόκλιση είναι μικρή και δεν μπορούμε να ισχυριστούμε με ασφάλεια ότι η μέθοδός μας είναι καλύτερη από αυτή τη Μεγιστοποίηση της Αρθρωτότητας.

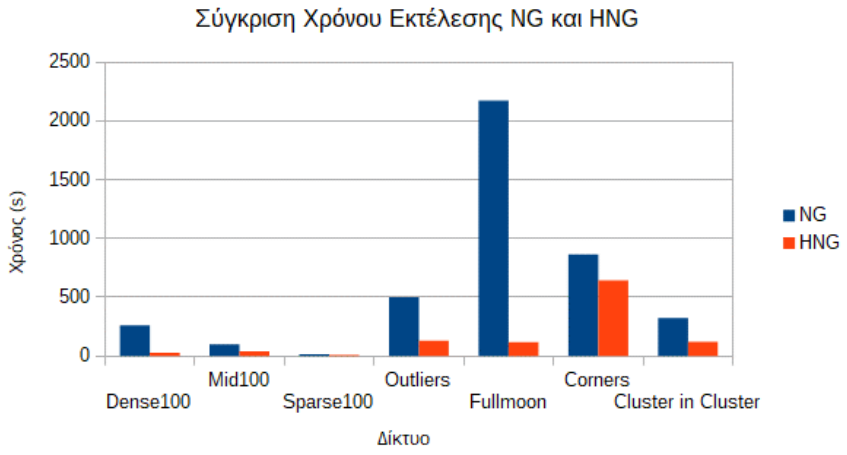
8.4 Σύγκριση των Μεθόδων Newman-Girvan και Hyperbolic Newman-Girvan

Στην παράγραφο αυτή συγκρίνουμε την μέθοδό μας με την κλασική μέθοδο των Newman και Girvan. Για τον υπολογισμό της Κεντρικότητας Ενδιαμεσικότητας Ακμής χρησιμοποιείται ο αλγόριθμος του Brandes. Στο πρώτο μέρος εξετάζονται δίκτυα στα οποία είναι γνωστή η βέλτιστη Ομαδοποίησή τους. Αυτά είναι όσα είναι γράφοι εγγύτητας, καθώς και τα δίκτυα που δημιουργήθηκαν με την μέθοδο του [LFR08], στα οποία γνωρίζουμε ποιές είναι οι κοινότητες τους. Στη συνέχεια, στο δεύτερο μέρος, συγκρίνονται οι δυο μέθοδοι σε γράφους για τους οποίους δεν γνωρίζουμε ποιά είναι η βέλτιστη διαμέριση των κορυφών τους σε κοινότητες. Χρησιμοποιούμε την μέθοδο Μεγιστοποίησης της Αρθρωτότητας ως μέθοδο αναφοράς.

- Δίκτυα με Γνωστές Κοινότητες

Πίνακας 11: Σύγκριση Χρόνου Εκτέλεσης Αλγόριθμων NG και HNG για Δίκτυα με Γνωστές Κοινότητες

Δίκτυο	Χρόνος NG	Χρόνος HNG (μέγεθος δέσμης)	100% Ακρίβεια NG	100% Ακρίβεια HNG
Dense100	*262.76*	*28.63 (60)*	*NAI*	*NAI*
Mid100	*99.56*	*40.78 (10)*	*NAI*	*NAI*
Sparse100	*16.01*	*10.62 (10)*	*NAI*	*NAI*
Dense500	*11279.82*	*3754.02 (10)*	*NAI*	*OXI*
Mid500	*3262.25*	*2575.79 (5)*	*NAI*	*OXI*
Sparse500	*1847.38*	*722.99*	*NAI*	*OXI*
Outliers	*500.68*	*130.41 (50)*	*OXI*	*NAI*
Fullmoon	*2.174.44*	*119.25 (50)*	*NAI*	*NAI*
Corners	*863.28*	*642.76 (50)*	*NAI*	*NAI*
Cluster in Cluster	*323.14*	*120.70 (100)*	*NAI*	*NAI*



Σχήμα 32: Διάγραμμα όπου φαίνονται οι διαφορές στους χρόνους εκτέλεσης για ορισμένα από τα δίκτυα του Πίνακα 11.

Όπως βλέπουμε από τον Πίνακα 11, σε όλες τις περιπτώσεις ο αλγόριθμός μας αποφαίνεται σωστά για όλα τα δίκτυα, σε καλύτερους χρόνους από ότι ο αλγόριθμος των Newman-Girvan. Ακόμα, παρατηρούμε ότι στον γράφο "outliers" επιτυγχάνεται 100% ακρίβεια από το αλγόριθμό μας, κάτι που δεν είναι δυνατόν στην περίπτωση του Newman-Girvan. Στο Σχήμα 32 μπορούμε να δούμε εποπτικά την διαφορά στους χρόνους εκτέλεσης για κάποια δίκτυα του Πίνακα 17.

Εδιαφέρον παρουσιάζουν οι περιπτώσεις όπου ο αλγόριθμος μας αποτυγχάνει να επιτύχει 100% ακρίβεια. Στην περίπτωση του γράφου Dense500, ανακαλύπτονται σωστά πέντε από τις εφτά κοινότητες του γράφου. Το σφάλμα συμβαίνει στην λανθασμένη ταξινόμηση δυο κόμβων μεταξύ των υπόλοιπων δυο κοινοτήτων. Ισχυριζόμαστε ότι αυτό είναι ένα αποδεκτό σφάλμα σε σχέση τον χρόνο που εξοικονομούμε εκτελώντας τον HNG. Στην περίπτωση του δικτύου Mid500, εντοπίζονται σωστά τρείς από τις έξι κοινότητες. Στις υπόλοιπες τρεις υπάρχουν σημαντικά λάθη στην ταξινόμηση των κόμβων. Τέλος, στην περίπτωση του Sparse500 η διαμέριση που παράγεται από τον HNG αποτυγχάνει να εντοπίσει σωστά κάποια από τις κοινότητες

- Δίκτυα με Άγνωστες Κοινότητες

Πίνακας 12: Σύγκριση Χρόνου Εκτέλεσης Αλγόριθμων NG και HNG για Δίκτυα με Άγνωστες Κοινότητες

Δίκτυο	Χρόνος NG (s)	Αρθρωτότητα NG	Χρόνος HNG (s)	Αρθρωτότητα HNG
karate	*0.54*	*0.34*	*14.14*	*0.42*
dolphins	*1.92*	*0.36*	*29.99*	*0.45*
lesmis	*6.42*	*0.44*	*52.10*	*0.55*
polbooks	*15.17*	*0.51*	*42.85*	*0.51*
scf3	*52.88*	*0.0028*	*540.4*	*0.17*
scf4	*43.88*	*0.096*	*384.14*	*0.28*
scf5	*50.67*	*0.0031*	*659.30*	*0.21*
scf6	*89.08*	*-0.00032*	*1229.35*	*-0.062*
smw3	*42.64*	*0.63*	*43.75*	*0.60*
smw4	*30.34*	*0.68*	*19.03*	*0.67*
smw5	*58.56*	*0.62*	*41.31*	*0.62*
smw6	*48.35*	*0.62*	*40.67*	*0.61*
rgg1	*17.72*	*0.63*	*19.74*	*0.63*
rgg2	*111.67*	*0.49*	*34.17*	*0.51*
rgg3	*33.94*	*0.00017*	*56.89*	*0.40*

rgg4	*154.85*	*-0.00068*	*174.97*	*0.28*

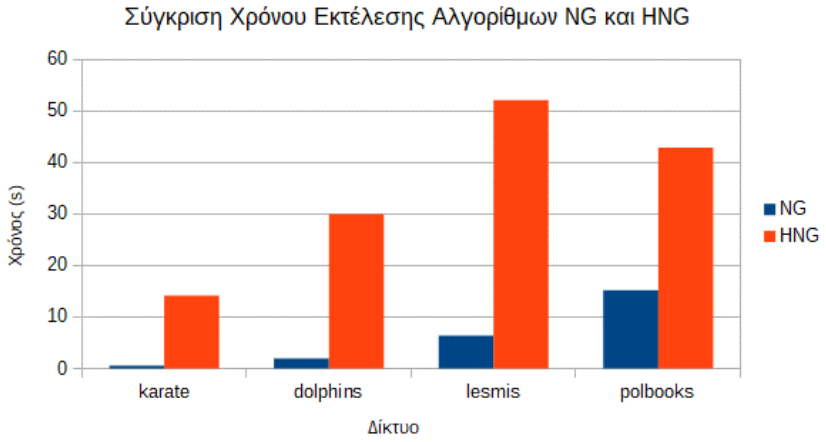
Ο Πίνακας 12 περιέχει τις τιμές για τον χρόνο εκτέλεσης και την Αρθρωτότητα για τους αλγόριθμους Newman-Girvan και HNG. Ανάλογα με τον τύπο του δικτύου προκύπτουν τα αντίστοιχα συμπεράσματα.

- Για τα πραγματικά κοινωνικά δίκτυα τύπου μικρής-κλίμακας (dolphins, karate, lesmis, polbooks) και για τα δίκτυα Ελεύθερης Κλίμακας παρατηρούμε ότι ο αλγόριθμος Newman-Girvan είναι πάντα αρκετά ταχύτερος. Παρόλα αυτά η παραγόμενη διαμέριση έχει κάθε φορά μικρότερη τιμή από αυτή του HNG. Η ταχύτερη εκτέλεση του Newman-Girvan έγκειται στον μικρό αριθμό ακμών των οποίων η Κεντρικότητα Ενδιαμεσικότητας βρίσκεται γρήγορα από τον αλγόριθμο του Brandes. Αντιθέτως στην περίπτωση του HNG ο χρόνος συνολικής εκτέλεσης επιβραδύνεται από τον χρόνο που απαιτείται για να ολοκληρωθεί κάθε φορά η Ενσωμάτωση Rigel. Όσον αφορά την τιμή της Αρθρωτότητας, καθώς τα κοινωνικά δίκτυα αυτά προσεγγίζουν το μοντέλο των δικτύων Ελεύθερης Κλίμακας υπάρχουν λίγοι κόμβοι με υψηλό βαθμό οι οποίοι βρίσκονται σε πολλά συντομότερα μονοπάτια μεταξύ κόμβων. Είναι αναμενόμενο λοιπόν οι Κεντρικότητες των ακμών που συνδέουν τους άλλους κόμβους με τους κόμβους με υψηλό βαθμό να έχουν μεγάλη τιμή και να είναι αυτές οι οποίες αφαιρούνται κατά την εκτέλεση του αλγόριθμου. Καταλήγουμε δηλαδή σε κοινότητες οι οποίες απαρτίζονται από μια μεγάλη συνεκτική συνιστώσα και μεμονομένους κόμβους χαμηλού βαθμού. Προφανώς μια τέτοια διαμέριση δεν οδηγεί σε υψηλή τιμή Αρθρωτότητας. Βλέπουμε δηλαδή εδώ ότι η προσεγγιστική λύση του HNG είναι προτιμότερη καθώς επειδή δεν αφαιρεί σε κάθε βήμα την ακμή με την υψηλότερη Κεντρικότητας Ενδιαμεσικότητας, δεν οδηγεί σε μεμονομένες κορυφές. Στο Σχήμα 33 και στο Σχήμα 34 φαίνονται οι διαφορές στον χρόνο εκτέλεσης και στην παραγόμενη Αρθρωτότητα για τα δίκτυα karate, dolphins, lesmis και polbooks. Αντίστοιχη εικόνα επικρατεί και στα Σχήματα 35 και 36 που αφορούν τα τέσσερα δίκτυα Ελεύθερης Κλίμακας του Πίνακα 12.

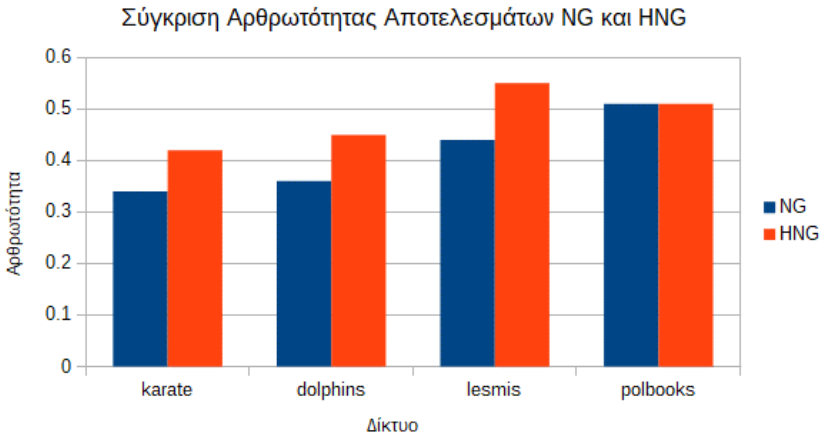
Πίνακας 13: Μέσος Όρος Βαθμού Κορυφών Κάθε Κοινότητας. Σε παρένθεση το πλήθος των Κόμβων Κάθε Κοινότητας.

	scf3	scf4	scf5	scf6
Κοινότητα 1	*5 (1)*	*5 (2)*	*4 (1)*	*5 (1)*
Κοινότητα 2	*4 (1)*	*5.25 (4)*	*4 (1)*	*5 (1)*
Κοινότητα 3	*4 (1)*	*5.6 (5)*	*4 (1)*	*6 (1)*
Κοινότητα 4	*6 (1)*	*4.33 (4)*	*5 (2)*	*6 (1)*
Κοινότητα 5	*7 (1)*	*4.5 (2)*	*4 (1)*	*6 (1)*
Κοινότητα 6	*10.67 (94)*	*7.61 (80)*	*8.85 (93)*	*6 (1)*
Κοινότητα 7	*7 (1)*	*5.25 (4)*	*5 (1)*	*11.55 (93)*
Κοινότητα 8	-	-	-	*6 (1)*
Μέσος Βαθμός Δικτύου	*10.04*	*7.12*	*8.54*	*11.14*

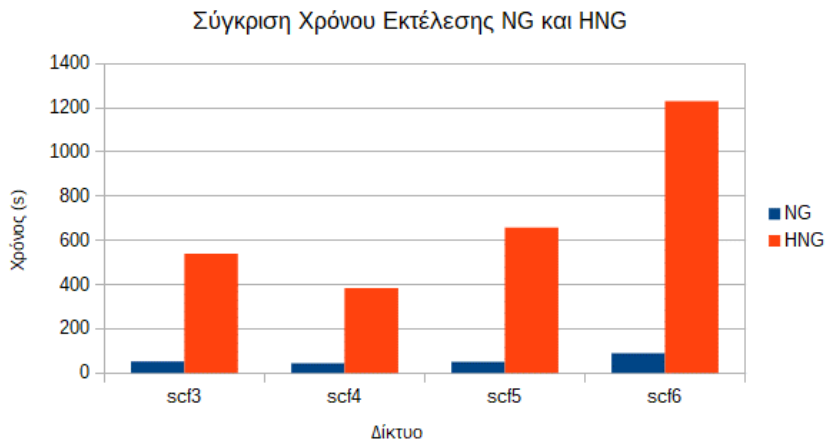
Ο Πίνακας 13 παρουσιάζει τον μέσο όρο των βαθμών των κορυφών κάθε κοινότητας για κάθε δίκτυο Ελεύθερης Κλίμακας που εξετάσαμε. Από τα στοιχεία του φαίνεται ότι οι περισσότερες κοινότητες που ανιχνεύονται από τον αλγόριθμο Newman-Girvan είναι μεμονομένοι κόμβοι ή στην καλύτερη περίπτωση μικρές ομάδες κόμβων. Οι κόμβοι αυτοί έχουν χαμηλό βαθμό όπως φαίνεται από τις τιμές των μέσων όρων τους που πάντα είναι αισθητά πιο κάτω από τον μέσο βαθμό του δικτύου και στις τέσσερις περιπτώσεις.



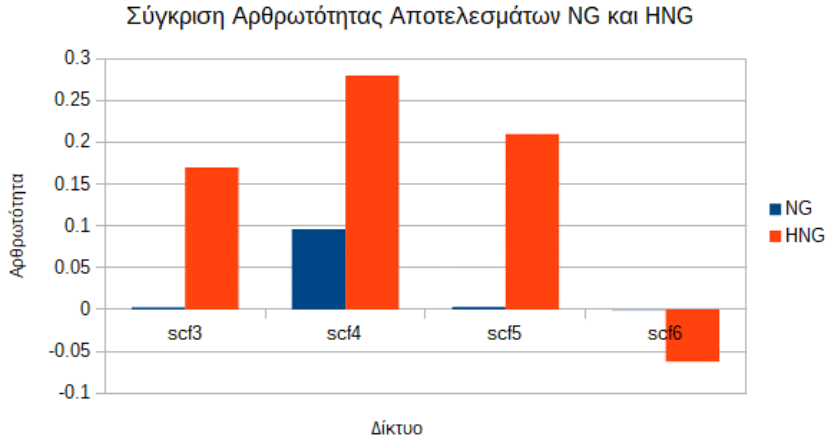
Σχήμα 33: Σύγκριση χρόνου εκτέλεσης των δυο αλγόριθμων



Σχήμα 34: Αρθρωτότητα που προκύπτει από την εφαρμογή των NG και HNG σε 4 δίκτυα.



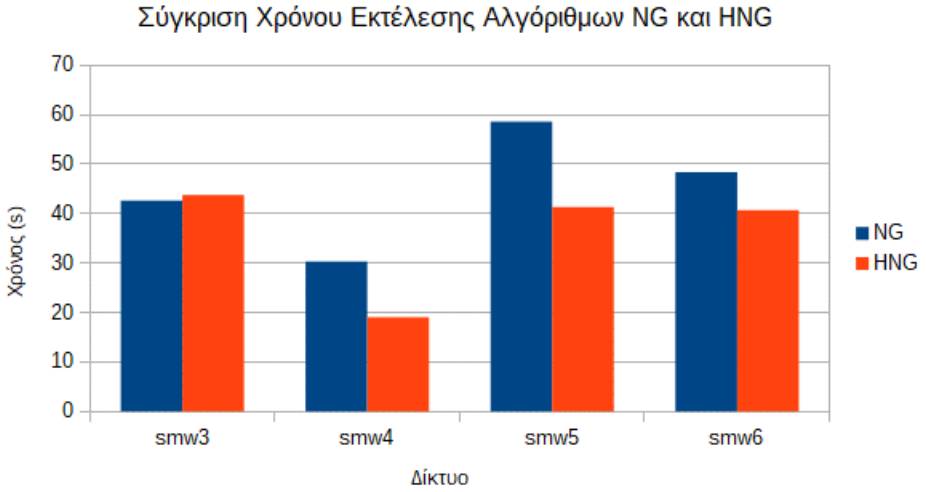
Σχήμα 35: Σύγκριση Χρόνου Εκτέλεσης για 4 Δίκτυα Ελεύθερης Κλίμακας



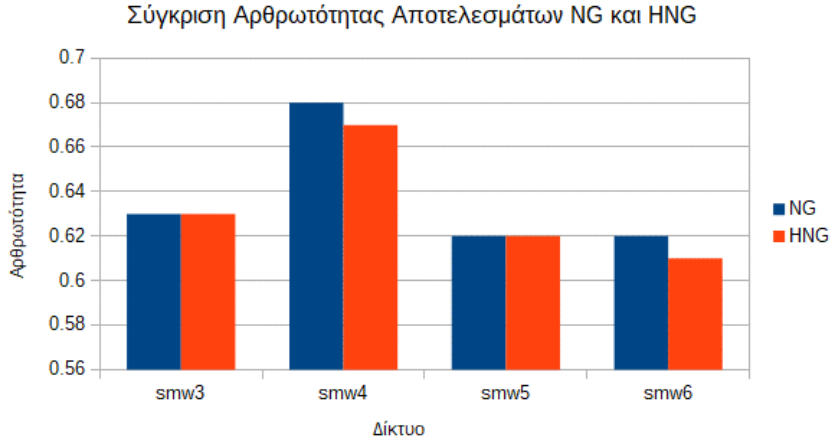
Σχήμα 36: Η παραγόμενη Αρθρωτότητα για τα δίκτυα Ελεύθερης Κλίμακας που εξετάζονται στον Πίνακα 12

- Για τα δίκτυα Μικρού Κόσμου παρατηρούμε ότι οι χρόνοι εκτέλεσης είναι παραπλήσιοι και στις δυο περιπτώσεις είναι συγκρίσιμοι, με τον προτεινόμενο αλγόριθμο HNG να είναι κατά μέσο όρο ταχύτερος. Ακόμα, οι Αρθρωτότητες των διαμερίσεων που παράγονται από την ανίχνευση κοινοτήτων με τις δυο μεθόδους έχουν παραπλήσιες τιμές, οι οποίες μάλιστα είναι κοντά στις τιμές που προκύπτουν από την εφαρμογή της Μεγιστοποίησης της Αρθρωτότητας που φαίνεται στον Πίνακα 1. Οι καλές τιμές που επιτυγχάνονται αφορούν την αφαίρεση των ακμών-συντομεύσεων (shortcut edges) που υπάρχουν σε ένα δίκτυο Μικρού Κόσμου και συνδέουν με κάποια πιθανότητα κόμβους που δεν είναι γειτονικοί μειώνοντας έτσι το Μέσο Μήκος Μονοπατιού στο δίκτυο. Οι ακμές αυτές έχουν υψηλή τιμή Κ.Ε.Α. Αφαιρώντας αυτές τις ακμές σχηματίζονται Ομάδες κόμβων που δεν συνδέονται πια. Αξίζει να σημειωθεί ότι οι τιμές του Πίνακα 12 αφορούν μέγεθος Δέσμης ίσο με 10 ακμές. Τοποθετώντας μέχρι και 20 ακμές στην Δέσμη τα αποτελέσματα που λαμβάνουμε παρουσιάζουν μια μείωση της Αρθρωτότητας που προκύπτει από την εκάστοτε διαμέριση, αλλά σε αποδεκτό βαθμό. Παρακάτω

δίνονται τα αντίστοιχα γραφήματα για την σύγκριση του χρόνου εκτέλεσης και της Αρθρωτότητας αντίστοιχα (Σχήμα 37, Σχήμα 38).



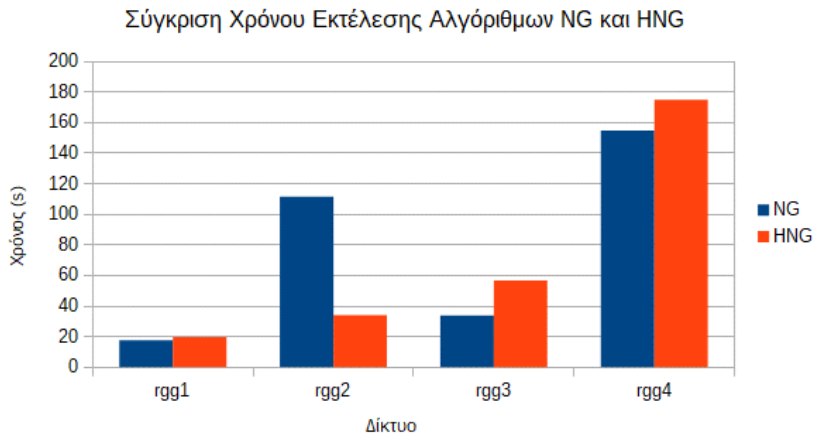
Σχήμα 37: Σύγκριση Χρόνου Εκτέλεσης για 4 Δίκτυα Μικρού Κόσμου



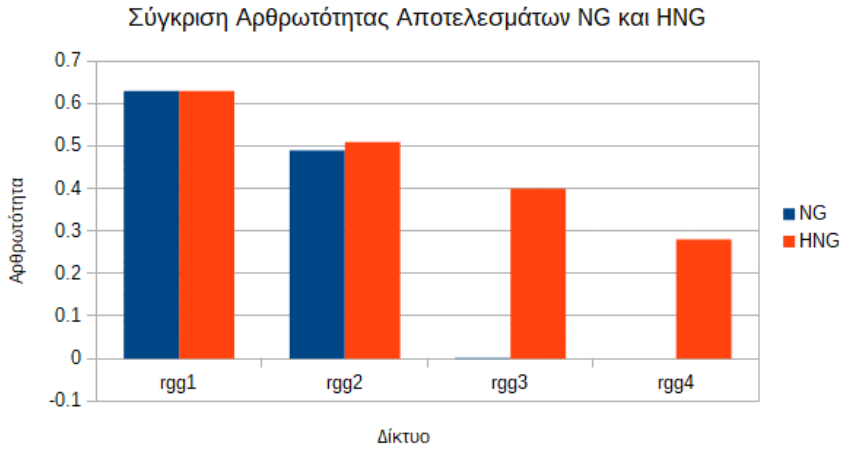
Σχήμα 38: Η παραγόμενη Αρθρωτότητα για τα Δίκτυα Μικρού Κόσμου που Εξετάζονται στον Πίνακα 18

- Όπως παρατηρούμε στους Τυχαίους Γεωμετρικούς Γράφους ο χρόνος εκτέλεσης του αλγόριθμου NG είναι μικρότερος από τον αντίστοιχο του HNG με την εξαίρεση μιας περίπτωσης. Όμως τα παραγόμενα αποτελέσματα του HNG έχουν υψηλότερη Αρθρωτότητα. Στις περιπτώσεις που οι ακμές αυξάνονται αισθητά (γράφοι rgg3,rgg4) ο αλγόριθμος Newman-Girvan παράγει διαμέριση που εντοπίζει ως κοινότητες μεμονομένους κόμβους με αποτέλεσμα η Αρθρωτότητα να έχει πολύ χαμηλές τιμές, κοντά στο μηδέν όπως φαίνεται και από το Σχήμα 40. Όσον αφορά τον χρόνο εκτέλεσης παρατηρούμε ότι και στις δυο περιπτώσεις αυξάνεται όσο αυξάνεται και ο αριθμός των ακμών που υπάρχουν στον γράφο, κάτι που είναι αναμενόμενο. Διάγραμμα του χρόνου

εκτέλεσης των δυο αλγόριθμων δίνεται στο Σχήμα 39.



Σχήμα 39: Σύγκριση Χρόνου Εκτέλεσης για 4 Τυχαίους Γεωμετρικούς Γράφους



Σχήμα 40: Η παραγόμενη Αρθρωτότητα για τους Τυχαίους Γεωμετρικούς Γράφους που Εξετάζονται στον Πίνακα 18

Όπως φάνηκε από τα πειράματα που εκτελέσαμε, ο αλγόριθμος που προτείνουμε στα πλαίσια της Διπλωματικής Εργασίας είναι ταχύτερος από αυτόν των Newman-Girvan όταν το Δίκτυο έχει παραπάνω από μερικές εκατοντάδες κορυφών. Ακόμα και σε μικρά δίκτυα όμως, παρά την βραδύτερη εκτέλεσή του, επιτυγχάνεται καλύτερη διαμέριση όσον αφορά την Αρθρωτότητα του Δικτύου, ιδιαίτερα όταν οι γράφοι είναι ή προσεγγίζουν τον μοντέλο των Barabasi-Albert για τα δίκτυα Ελεύθερης Κλίμακας.

9

Συμπεράσματα

Στην ενότητα αυτή επιχειρείται μια σύνοψη των αποτελεσμάτων της Διπλωματικής Εργασίας καθώς και των συμπερασμάτων που προέκυψαν από την εφαρμογή του προτεινόμενου αλγόριθμου Ομαδοποίησης Hyperbolic Newman-Girvan. Στη συνέχεια παρουσιάζονται μερικές ιδέες για μελοντική επέκταση και βελτίωση του προτεινόμενου πλαισίου.

9.1 Σύνοψη και Συμπεράσματα

Στην παρούσα Εργασία επιλέχθηκε η ενσωμάτωση των κόμβων Σύνθετων Δικτύων και Γράφων Εγγύτητας που προκύπτουν από τις μετρήσεις τοπολογιών Μεγάλων Δεδομένων. Η ενσωμάτωση που επιλέχθηκε είναι η Ενσωμάτωση Rigel η οποία αναθέτει στους κόμβους συντεταγμένες ενός Υπερβολικού Χώρου με διάσταση που καθορίζεται από τον χρήστη. Η ανάθεση γίνεται με τέτοιο τρόπο ώστε η απόσταση δυο κόμβων στον χώρο να προσεγγίζει το μήκος του συντομότερου μονοπατιού που τους ενώνει. Με αυτόν τον τρόπο ο υπολογισμός της απόστασης μεταξύ δυο κόμβων του δικτύου μπορεί να γίνει σε σταθερό χρόνο και δεν απαιτείται κάθε φορά η εύρεση των συντομότερων μονοπατιών που τους συνδέουν, διαδικασία που είναι υπολογιστικά πιο «ακριβή». Κάνοντας χρήση της ιδιότητας αυτής προτείνεται η χρήση του αλγόριθμου προσεγγιστικού υπολογισμού της Κεντρικότητας Ενδιαμεσικότητας Ακμής που ονομάστηκε Υπερβολική Κεντρικότητα Ενδιαμεσικότητας Ακμής. Ο αλγόριθμος αυτός αν και δεν είναι αποδοτικότερος από άποψη υπολογιστικής πολυπλοκότητας από τον αλγόριθμο του Brandes, δίνει καλά αποτελέσματα σε ικανοποιητικό χρόνο. Πολλές φορές μάλιστα ολοκληρώνει την εκτέλεση του σε πολυ συντομότερο χρόνο. Από τα πειράματα που διεξήχθησαν φάνηκε ότι ο προτεινόμενος αλγόριθμος δίνει αρκετά ακριβή αποτελέσματα στην περίπτωση που η δομή του δικτύου είναι ή προσεγγίζει τα δίκτυα Ελεύθερης Κλίμακας. Στην περίπτωση δηλαδή Κοινωνικών Δικτύων.

Στη συνέχεια, έγινε χρήση του αλγόριθμου υπολογισμού της Υ.Κ.Ε.Α. στην προτεινόμενη στα πλαίσια της Διπλωματικής Εργασίας παραλλαγή του δημοφιλούς αλγόριθμου Ομαδοποίησης των Newman-Girvan. Στην παραλλαγή αυτή, η οποία ονομάστηκε Hyperbolic Newman-Girvan, ορίζεται το επιθυμητό πλήθος κοινοτήτων που επιθυμείται να εντοπίσει ο αλγόριθμος καθώς και ένα μέγιστο πλήθος ακμών που μπορούν να αφαιρεθούν μέχρις ότου η συνεκτική συνιστώσα που εξετάζεται να καταστεί μη-συνδεδεμένη. Το πλήθος αυτό ονομάζεται Δέσμη και αποτελείται από τις ακμές της συνιστώσας με την υψηλότερη τιμή Υ.Κ.Ε.Α. Έτσι, αφαιρείται μια-μια μέχρις ότου είτε να έχουν αφαιρεθεί όλες οι ακμές της Δέσμης, είτε να μην είναι πια συνδεδεμένη η συνιστώσα. Τότε η νέα μεγαλύτερη συνεκτική συνιστώσα του δικτύου ενσωματώνεται στον Υπερβολικό Χώρο με χρήση της Ενσωμάτωσης Rigel και η παραπάνω διαδικασία επαναλαμβάνεται μέχρις ότου να υπάρχουν τόσες συνεκτικές συνιστώσες στο δίκτυο όσες και οι κοινότητες που ορίστηκαν στην αρχή του αλγόριθμου.

Ο αλγόριθμος που προτάθηκε είναι πιο γρήγορος για μεγάλα δίκτυα (πάνω από 500 κόμβους) από τον αλγόριθμο των Newman-Girvan στον οποίο χρησιμοποιείται ο αλγόριθμος του Brandes για τον υπολογισμό της Κεντρικότητας Ενδιαμεσικότητας Ακμής. Μάλιστα, έχει πολυ ικανοποιητική επίδοση σε Γράφους Εγγύτητας όπου στα πειράματα που παρουσιάστηκαν στο Κεφάλαιο 8 κατάφερε σε όλες τις περιπτώσεις να εντοπίσει της σωστές κοινότητες. Στην περίπτωση όμως μικρών δικτύων φαίνεται ότι ο αλγλοριθμος των Newman-Girvan είναι αρκετά γρηγορότερος. Ακόμα για μικρά μεγέθη δέσμης, δίνει αρκετά καλά αποτελέσματα όσον αφορά την Αρθρωτότητα των διαμερίσεων που παράγει.

9.2 Ιδέες για Περαιτέρω Μελέτη

Ζητήματα που δεν έγινε δυνατό να απαντηθούν στην Διπλωματική Εργασία όμως είναι εφικτή και ενδιαφέρουσα η περαιτέρω μελέτη τους είναι τα ακόλουθα:

- Βέλτιστος καθορισμός των παραμέτρων της Ενσωμάτωσης Rigel ανάλογα με τα χαρακτηριστικά του προς ενσωμάτωση δικτύου. Στην εργασία αυτή καθόλη τη διάρκεια της εκτέλεσης του αλγόριθμου HNG οι παράμετροι μένουν σταθερές παρόλα αυτά ίσως αποτελεί καλύτερη προσέγγιση η δυναμική ρύθμιση τους ανάλογα με τα δίκτυα που προκύπτουν κατά την διάρκεια της Ομαδοποίησης. Μια τέτοια προσέγγιση θα οδηγούσε σε ταχύτερη ολοκλήρωση του αλγόριθμου καθώς οι ακμές με την μεγαλύτερη Κεντρικότητα που είναι οι ακμές που συνδέουν διαφορετικές κοινότητες θα αφαιρούνταν συντομότερα.

- Βέλτιστος καθορισμός του μέγεθους Δέσμης ανάλογα με το δίκτυο με σκοπό την βελτίωση του χρόνου εκτέλεσης.

Βιβλιογραφία

[BaA99] Albert-Laszlo Barabasi and Reka Albert (1999), “Emergence of Scaling in Random Networks”, Science 286

[Bar09] Albert-László Barabási, et al. (2009), Scale-Free Networks: A Decade and Beyond, Science 325,412

[Bav50] Alex Bavelas. Communication patterns in task-oriented groups J. Acoust. Soc. Am, 22(6):725–730, 1950

[BEJ13] Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). Analyzing social networks. SAGE Publications Limited.

[BGL08] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre (2008), “Fast unfolding of communities in large networks”, J. Stat. Mech., P10008

[Bor05] Borgatti, Stephen P. (2005). "Centrality and Network Flow". Social Networks. Elsevier. 27: 55–71.

[Bra01] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of mathematical sociology, 25(2), 163-177.

[Bra08] Brandes, U. (2008). “On variants of shortest-path betweenness centrality and their generic computation”. Social Networks, 30(2), 136-145.

[BrK73] Bron, Coen; Kerbosch, Joep (1973), "Algorithm 457: finding all cliques of an undirected graph", Commun. ACM, ACM, 16 (9): 575–577

[CaR05] M. Carreira-Perpinan and R. Zemel (2005), “Proximity graphs for clustering and manifold learning”, Advances in Neural Information Processing Systems 17: Proceedings Of The 2004 Conference, p. 225, MIT Press

[Cas] http://cassandra.apache.org/

[Cis] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html

[CML14] Min Chen, Shiwen Mao, Yunhao Liu (2014) Big Data: A Survey, Springer Science+Business Media New York 2014 σελ: 173

[CvC09] Cvetkovski, A., & Crovella, M. (2009, April), “Hyperbolic embedding and routing for dynamic graphs”. In INFOCOM 2009, IEEE (pp. 1647-1655) IEEE.

[Dol] http://www-personal.umich.edu/~mejn/netdata/dolphins.zip

[DoM04] Donetti, L., & Munoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. Journal of Statistical Mechanics: Theory and Experiment, 2004(10), P10012.

[Don00] S. van Dongen, Ph.D. thesis, Dutch National Research Institute for Mathematics and Computer Science, University of Utrecht, Netherlands (2000).

[Dyn] https://aws.amazon.com/dynamodb/

[EBI] https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/introduction-graph-theory/graph-0

[EKS96] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

[ErR59] Erdős, P. Rényi, A (1959) "On Random Graphs I" in Publ. Math. Debrecen 6, p. 290–297

[Fac] https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/

[FLG02] G.W. Flake, S. Lawrence, C. Lee Giles, F.M. Coetzee (2002), "Self-organization and identification of web communities", IEEE Computer, 35, pp. 66-713

[Foo] http://www-personal.umich.edu/~mejn/netdata/football.zip

[FPW09] Flury, R., Pemmaraju, S. V., & Wattenhofer, R. (2009, April). Greedy routing with bounded stretch. In INFOCOM 2009, IEEE (pp. 1737-1745). IEEE.

[Fre77] Freeman, Linton (1977). "A set of measures of centrality based on betweenness". Sociometry. 40: 35–41

[Gil59] Gilbert, E. N. (1959), "Random graphs", Annals of Mathematical Statistics, 30: 1141–1144

[GiN02] M. Girvan and M. E. J. Newman (2002), "Community structure in social and biological networks" PNAS June 11 2002, vol. 99 no. 12, 7821–7826

[GPA04] R. Guimerà, M. Sales-Pardo, L.A.N. Amaral, (2004), "Modularity from fluctuations in random graphs and complex networks", Phys. Rev. E, 70 (2), p. 025101 (R)

[HaW79] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

[IBM] https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN

[Joh67] Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32(3),

241-254.

[JGL14] H.V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan and Cyrus Shahabi (2014) Big Data and Its Technical Challenges, Communications Of The Acm July 2014 Vol. 57 No. 7

[Kad11] Kadushin Charles (2011), Understanding Social Networks: Theories, Concepts, and Findings, Oxford University Press.

[Kar] http://www-personal.umich.edu/~mejn/netdata/karate.zip

[Kat53] Katz, L. (1953). A New Status Index Derived from Sociometric Analysis. Psychometrika, 39–43.

[Kle07] Kleinberg, R. (2007, May). Geographic routing using hyperbolic space. In INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE (pp. 1902-1909). IEEE.

[KPK10] Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., & Boguná, M. (2010). Hyperbolic geometry of complex networks. Physical Review E, 82(3), 036106.

[Kru56] Kruskal, J. B. (1956) "On the shortest spanning subtree of a graph and the traveling salesman problem". Proceedings of the American Mathematical Society. 7: 48–50.

[KSP13] Vasileios Karyotis, Eleni Stai, Symeon Papavassiliou (2013), Evolutionary Dynamics of Complex Communications Networks, CRC Press

[Lan01] Doug Laney (2001) 3D Data Management: Controlling Data Volume, Velocity, and Variety, Application Delivery Strategies, 6/2/2001

[LCC02] Qin Lv, Pei Cao, Edith Cohen, Kai Li, Scott Shenker (2002), “Search and Replication inUnstructured Peer-to-Peer Networks”, ICS’02

[LeH08] Jure Leskovec, Eric Horvitz (2008) “Planetary-Scale Views on a Large Instant-Messaging Network”, WWW 2008

[LeM] http://www-personal.umich.edu/~mejn/netdata/

[LFR08] Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. Physical review E, 78(4), 046110

[Lia05] Liao, T. W. (2005). Clustering of time series data—a survey. Pattern recognition, 38(11), 1857-1874.

[Llo82] Lloyd, Stuart P. (1982), "Least squares quantization in PCM", IEEE Transactions on Information Theory, 28 (2): 129–137

[LuS08] C. Lumezanu and N. Spring, “Measurement manipulation and space selection in

network coordinates," in Proc. of ICDCS, 2008

[Mar15] Bernarnd Marr (2015) Big Data: 20 Mind-Boggling Facts Everyone Must Read.

[Mil67] Milgram, Stanley (May 1967). "The Small World Problem". Psychology Today

[Mon] ttps://docs.mongodb.com/

[MuP10] S. Muthukrishnan, Gopal Pandurangan (2010), "Thresholding random geometric graph properties motivated by ad hoc sensor networks", Journal of Computer and System Sciences, Elsevier, 2010

[Nat] http://www.nature.com/nphys/journal/v6/n7/fig_tab/nphys1665_F1.html?foxtrotcallback=true

[Net] http://www.network-science.org/powerlaw_scalefree_node_degree_distribution.html

[New04] M.E.J. Newman (2002), "Fast algorithm for detecting community structure in networks". Phys. Rev. E, 69 (6) , p. 066133

[New06] M.E.J. Newman (2006), "From the cover: Modularity and community structure in networks",
Proc. Natl. Acad. Sci. USA, 103, pp. 8577-8582

[NJW02] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In Advances in neural information processing systems (pp. 849-856).

[PDF05] Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. arXiv preprint physics/0506133.

[Pol] http://www-personal.umich.edu/~mejn/netdata/polbooks.zip

[PPK12] Papadopoulos, F., Psomas, C., & Krioukov, D. (2012). "Replaying the geometric growth of complex networks and application to the AS internet". ACM SIGMETRICS Performance Evaluation Review, 40(3), 104-106.

[Pri57] Prim, R. C.(1957) , "Shortest connection networks And some generalizations", Bell System Technical Journal, 36 (6): 1389–1401

[Res] https://www.researchgate.net/figure/268271797_fig2_Fig-2-Zachary%27s-karate-club-network-Square-nodes-and-circle-nodes-represent-the

[Res2] https://www.researchgate.net/figure/2123123_fig5_The-hyperboloid-model-of-hyperbolic-3-space-inside-Minkowski-4-space

[RiG03] A.W. Rives, T. Galitski (2003), "Modular organization of cellular networks", Proc.

Natl. Acad. Sci. USA, 100 (3), pp. 1128-1133

[Rog02] Rogers, I. (2002). “The Google Pagerank algorithm and how it works”|. http://www.iprcom.com/papers/pagerank/

[SGM14] Konstantinos Slavakis, Georgios B. Giannakis, and Gonzalo Mateos (2014) Modeling and Optimization for Big Data Analytics, IEEE SIGNAL PROCESSING MAGAZINE SEPTEMBER 2014

[SSK16] Stai, E., Sotiropoulos, K., Karyotis, V., & Papavassiliou, S. (2016, May). Hyperbolic Traffic Load Centrality for large-scale complex communications networks. In Telecommunications (ICT), 2016 23rd International Conference on (pp. 1-5). IEEE.

[SSK17] Stai, E., Sotiropoulos, K., Karyotis, V., & Papavassiliou, S. (2017). Hyperbolic Embedding for Efficient Computation of Path Centralities and Adaptive Routing in Large-scale Complex Commodity Networks. IEEE Transactions on Network Science and Engineering.

[Sta] https://math.stackexchange.com/questions/1250164/why-do-lines-in-the-poincare-model-meet-the-infinite-edge-at-right-angles

[Sto02] Stojmenovic, I. (2002). Position-based routing in ad hoc networks. IEEE communications magazine, 40(7), 128-134.

[SYG09] Sarkar, R., Yin, X., Gao, J., Luo, F., & Gu, X. D. (2009, April). Greedy routing with guaranteed delivery using ricci flows. In Information Processing in Sensor Networks, 2009. IPSN 2009. International Conference on (pp. 121-132). IEEE.

[Tec] https://techcrunch.com/2017/06/27/facebook-2-billion-users/

[Was98] Watts, D. J.; Strogatz, S. H. (1998). "Collective dynamics of 'small-world' networks", Nature 393 (6684): 440–442

[Wik] https://en.wikipedia.org/wiki/Random_geometric_graph

[Wik2] https://en.wikipedia.org/wiki/Poincar%C3%A9_disk_model

[YXZ07] Yan, S., Xu, D., Zhang, B., Zhang, H. J., Yang, Q., & Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. IEEE transactions on pattern analysis and machine intelligence, 29(1), 40-51

[ZSW10] X. Zhao, A. Sala, C. Wilson, H. Zheng, and B. Y. Zhao (2010), “Orion: Shortest path estimation for large social graphs,” Proc. of WOSN, Boston, MA, June 2010