Τμήμα Μηχανικών Η/Υ και Πληροφορικής Πανεπιστημίου Ιωαννίνων

ΜΥΕ047: Αλγόριθμοι για Δεδομένα Ευρείας Κλίμακας

Ακαδημαϊκό Έτος 2021-22

2ο Σετ Ασκήσεων: Ανίχνευση Κοινοτήτων Σε Γραφήματα Κοινωνικών Δικτύων Μέσω Τεχνικών Συσταδοποίησης

Ευάγγελος Τζώρτζης ΑΜ: 3088

Αναφορά

Στην εργασία υλοποιήθηκε ο αλγόριθμος Girvan-Newman για συσταδοποίηση ενός γραφήματος, τόσο με χρήση της βιβλιοθήκης Networkx, όσο και με δική μου υλοποίηση.

Ο έλεγχος για τη βέλτιστη διαμέριση έγινε με χρήση της μετρικής αρθρωτότητας (modularity).

Ακόμα, υλοποιήθηκαν μέθοδοι για το φόρτωμα ενός γραφήματος, για την ενίσχυσή του με παραπάνω ακμές είτε τυχαία, είτε με ένα κύκλο Hamilton, για τη δημιουργία μιας ιεραρχίας διαμερίσεων και για οπτικοποίηση του γραφήματος μετά από μια διαμέριση.

Το αρχείο κώδικα είναι το *3088_community-detection.py*, το οποίο είναι βασισμένο στο δοσμένο *STUDENT_AM_community-detection_TEMPLATE.py.*

Οι συναρτήσεις που ξεκινάνε με STUDENT_AM αλλάχθηκαν με AM_3088, καθώς δεν μπορούν να αρχίζουν με αριθμό.

Για να τρέξει σωστά το φόρτωμα γραφήματος από το αρχείο *fb-pages-food.edges*, το οποίο ζητείται στο ερώτημα (1Α), το πρόγραμμα και το αρχείο πρέπει να βρίσκονται στον ίδιο κατάλογο/φάκελο.

Στις συναρτήσεις girvan-newman του ερωτηματος (1Γ) προστέθηκε η παράμετρος node_percent για το ποσοστό κόμβων που θα χρησιμοποιηθεί για μια διαχώριση του γραφήματος. Επίσης, επιστρέφουν το tuple της μεγαλύτερης συνεκτικής

συνιστώσας, καθώς και τις κοινότητες μετά τον διαχωρισμό της μεγαλύτερης κοινότητας σε δύο συνεκτικές συνιστώσες.

Στη συνάρτηση AM_3088_use_nx_girvan_newman_for_communities όπου χρησιμοποιείται ο αλγόριθμος της βιβλιοθήκης network, παρατηρήθηκε ότι σε κάποιες περιπτώσεις δεν επιλέγεται για διαχωρισμό η μεγαλύτερη συνεκτική συνιστώσα, αλλά κάποια άλλη, κάτι οποίο δεν συνέβη με την δική μου υλοποίηση. Σε αυτή τη περίπτωση τυπώνεται ένα μήνυμα, αλλά δεν προκαλεί κάποιο λάθος.

Στη ρουτίνα AM_3088_divisive_community_detection(…), εφόσον καλεστεί από το μενού, αρχικά επιλέγεται ο αλγόριθμος που θα χρησιμοποιηθεί για τον διαχωρισμό, ύστερα επιλέγεται το ποσοστό κόμβων για τον υπολογισμό του betweenness centrality στον αλγόριθμο Girvan-Newman. Επιστρέφεται η ιεραρχία διαχωρισμών και οι τελικές κοινότητες.

Η ρουτίνα AM_3088_visualize_communities(…) εμφανίζει τις κοινότητες, καθεμιά με διαφορετικό χρώμα, εφόσον έχουν δημιουργηθεί μέσω των επιλογών μενού C ή D.

Η ρουτίνα AM_3088_determine_opt_community_structure(…) βρίσκει τη βέλτιστο διαχωρισμό με βάση το μέτρο modularity. Τυπώνεται το γράφημα της διαμέρισης με το βέλτιστο αριθμό κοινοτήτων. Ο χρήστης έχει την επιλογή για το εύρος των τιμών modularity που θα εμφανιστούν στο ραβδόγραμμα που εμφανίζεται κατά την εκτέλεση της ρουτίνας.

Πειράματα:

Έγιναν πειράματα με διάφορους συνδυασμούς παραμέτρων. Τέτοιες παράμετροι είναι διαφορετικά γραφήματα, ενδεχομένως με την εισαγωγή τυχαίων ή με κύκλο Hamilton ακμών, οι δύο τρόποι υπολογισμού του αλγορίθμου Girvan-Newman και οι διάφορες τιμές για το ποσοστό των κόμβων που χρησιμοποιούνται στον υπολογισμό της ενδιαμεσότητας για τον διαχωρισμό της μεγαλύτερης συνεκτικής συνιστώσας.

Παρακάτω παρατίθεται ο πίνακας με τα πειράματα, ο οποίος περιέχει τις παραμέτρους στο πρόγραμμα και τις εξόδους από αυτό.

Τα τέσσερα πρώτα πειράματα έγιναν για τη σύγκριση τόσο των χρόνων, όσο και της διαχώρισης με το μέγιστο modularity για διαφορετικούς αλγορίθμους και ποσοστά κόμβων στον υπολογισμό του betweenness.

Τα επόμενα τέσσερα (5-8) έγιναν για τον ίδιο λόγο με τα προηγούμενα, με τη μόνη διαφορά ότι τώρα το γράφημα έχει έναν κύκλο Hamilton, οπότε ξεκινάει με μία συνεκτική συνιστώσα και φτάνει μέχρι 8 λόγω περιορισμών στο μέγιστο πλήθος.

Τα επόμενα 3 πειράματα (9-11) έχουν εκτελέσεις για σύγκριση ενός γραφήματος ως είσοδο, με προσθήκη τυχαίων ακμών και με προσθήκη τυχαίων ακμών και κύκλου Hamilton.

Στο τελευταίο πείραμα γίνεται συσταδοποίηση σε γράφημα τύπου Erdos-Renyi.

Πειράματα:

	Εισόδοι:					Εξόδοι:		
	Γράφημα:	Παραμετροποίηση Γραφήματος (τυχαίες ακμές ή κύκλος Hamilton):	Αλγόριθμος (NX ή OWN):	Ποσοστό Ακμών Betweenness:	Αριθμός κοινοτήτων στην συσταδοποίηση:	Μέγιστο modularity:	Αριθμός διαχώρισης με μέγιστο modularity:	Χρόνος υπολογισμού (sec):
Πείραμα 1	L,500 - γράφημα με 332 κόμβους και 499 ακμές	-	NX	1	83 (max)	0.67	34	29.77
Πείραμα 2	L,500 - γράφημα με 332 κόμβους και 499 ακμές	-	NX	0.1	83 (max)	0.585	25	10.59
Πείραμα 3	L,500 - γράφημα με 332 κόμβους και 499 ακμές	-	OWN	1	83 (max)	0.6	29	14.15
Πείραμα 4	L,500 - γράφημα με 332 κόμβους και 499 ακμές	-	OWN	0.1	83 (max)	0.3925	34	7.716
Πείραμα 5	L,500 - γράφημα με 332 κόμβους και 499 ακμές (779 ακμές μετα τον κύκλο Hamilton)	Κύκλος Hamilton	NX	1	8 (max)	0.5939	8	20.22
Πείραμα 6	L,500 - γράφημα με 332 κόμβους και 499 ακμές (779 ακμές μετα τον κύκλο Hamilton)	Κύκλος Hamilton	NX	0.1	8 (max)	0.57	8	6.97
Πείραμα 7	L,500 - γράφημα με 332 κόμβους και 499 ακμές (779 ακμές μετα τον κύκλο Hamilton)	Κύκλος Hamilton	OWN	1	8 (max)	0.28	7	5.28
Πείραμα 8	L,500 - γράφημα με 332 κόμβους και 499 ακμές (779 ακμές μετα τον κύκλο Hamilton)	Κύκλος Hamilton	OWN	0.1	8 (max)	0.144	8	7.28
Πείραμα 9	L,250 - γράφημα με 223 κόμβους και 249 ακμές	-	OWN	1	30	0.792	20	2.57
Πείραμα 10	L,250 - γράφημα με 223 κόμβους και 249 ακμές (261 ακμές μετα τις τυχαίες ακμές)	Τυχαίες ακμές RE,2,0.05	OWN	1	30	0.78	21	2.46
Πείραμα 11	L,250 - γράφημα με 223 κόμβους και 249 ακμές (456 ακμές μετα τις τυχαίες ακμές και τον κύκλο Hamilton)	Τυχαίες ακμές RE,2,0.06 και κύκλος Hamilton	OWN	1	30 (αλλαγή max τιμής)	0.729	16	4.452
Πείραμα 12	R,100,0.05 - γράφημα με 223 κόμβους και 243 ακμές	-	OWN	1	8 (max)	0.092	8	1.39

Από τα δύο πρώτα πειράματα φαίνεται ότι όταν επιλέγεται ποσοστό κόμβων 0.1 ο χρόνος μειώνεται στο 1/3 του αρχικού, το μέγιστο modularity είναι παρόμοιο αλλά ο αριθμός της διαχώρισης με το μέγιστο modularity έχει απόκλιση 9 διαμερίσεις.

Στα πειράματα 3 και 4 ο χρόνος μειώνεται στο μισό περίπου και η απόκλιση είναι 5 διαμερίσεις αλλά το μέγιστο modularity για 0.1 μειώνεται.

Συγκρίνοντας τα πειράματα 1 και 3 όπου χρησιμοποιούνται όλοι οι κόμβοι για τον υπολογισμό του betweenness, ο χρόνος του δικού μου αλγορίθμου είναι ο μισός, τα modularity είναι κοντά μεταξύ τους, αλλά υπάρχει μικρή διαφορά στον αριθμό διαχώρισης, *η οποία μάλλον οφείλεται στο γεγονός ότι ο αλγόριθμος της*

βιβλιοθήκης σε κάποιες περιπτώσεις δεν χωρίζει τη μεγαλύτερη συνεκτική συνιστώσα αλλά κάποια άλλη.

Τα πειράματα 2 και 4 έχουν διαφορά στο modularity και μεγάλη απόκλιση μεταξύ τους ως προς τη διαχώριση με το μέγιστο modularity.

Στα πειράματα 5-6 η μείωση των κόμβων σε 10% οδήγησε σε παρόμοια αρθρωτότητα και στην ίδια διαχώριση σε αρκετά χαμηλότερο χρόνο. Ωστόσο, αυτό μπορεί να οφείλεται στον περιορισμό του αριθμού διαχωρίσεων σε 8.

Τα επόμενα πειράματα 7-8 έχουν διαφορετική διαχώριση και modularity και παρόμοιο χρόνο.

Στα πειράματα που ξεκινάνε με 1 συνεκτική συνιστώσα μπορούν να φτάσουν μέχρι τις 8 βάση μιας τιμής. Γι' αυτό και στα περισσότερα από αυτά τα πειράματα το μέγιστο modularity εμφανίζεται στις 8 διαχωρίσεις, γιατί στη πραγματικότητα στις επόμενες διαχωρίσεις θα αύξανε και άλλο μέχρι τη πραγματική μέγιστη τιμή.

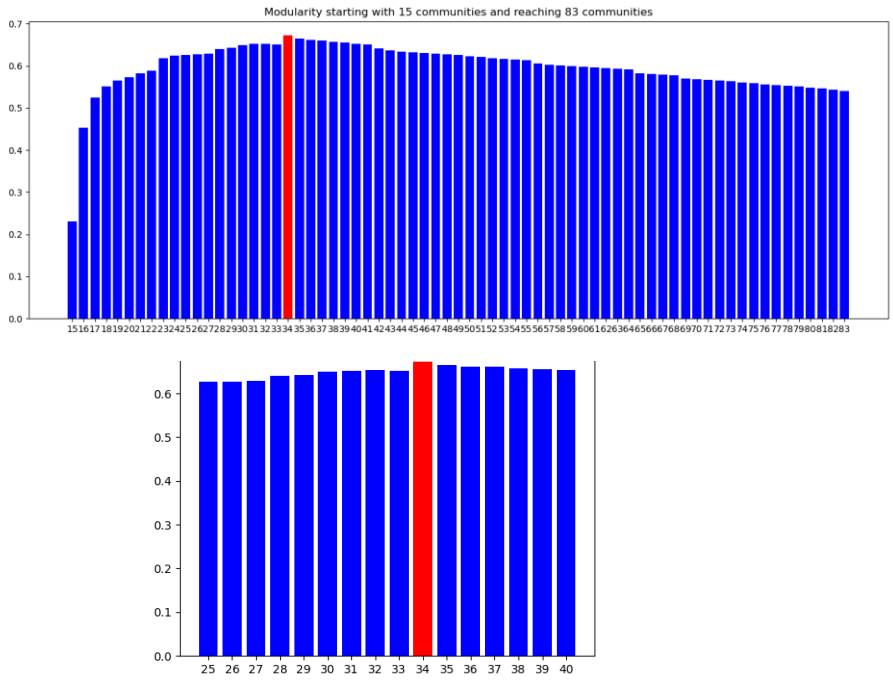
Στο πείραμα 11 αύξησα τον αριθμό MAX_NUM_DIVISIONS από 8 σε 40, ώστε να μπορέσει να γίνει σύγκριση των αποτελεσμάτων με τα πειράματα 9 και 10.

Στη σύγκριση των πειραμάτων 9-10 φαίνεται ότι μετά τη προσθήκη αλλάζει ελάχιστα το μέγιστο modularity και η αντίστοιχη διαχώριση και ο χρόνος είναι παρόμοιος. Στο πείραμα μειώνεται λίγο η βέλτιστη διαχώριση και αυξάνεται λίγο ο χρόνος.

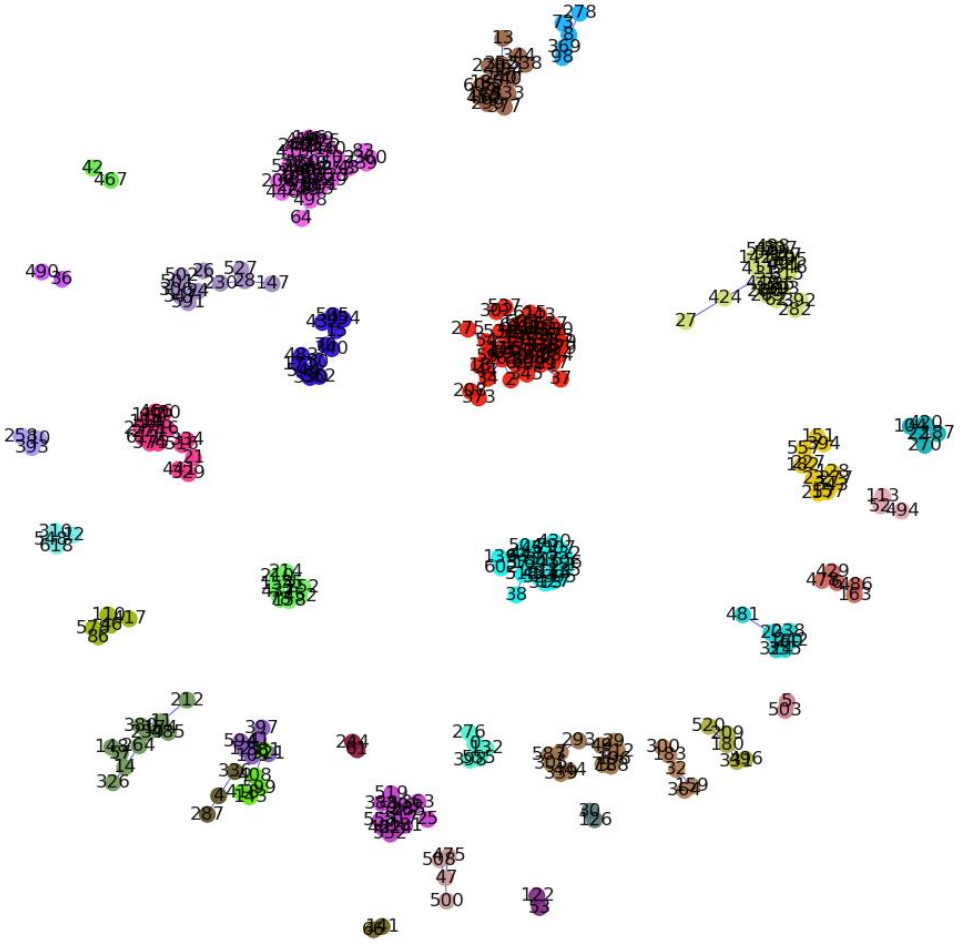
Στο πείραμα 12 το modularity είναι πολύ χαμηλό και η βέλτιστη διαχώριση η όγδοη.

Διαγράμματα μπαρών και γραφήματα ενδεικτικά:

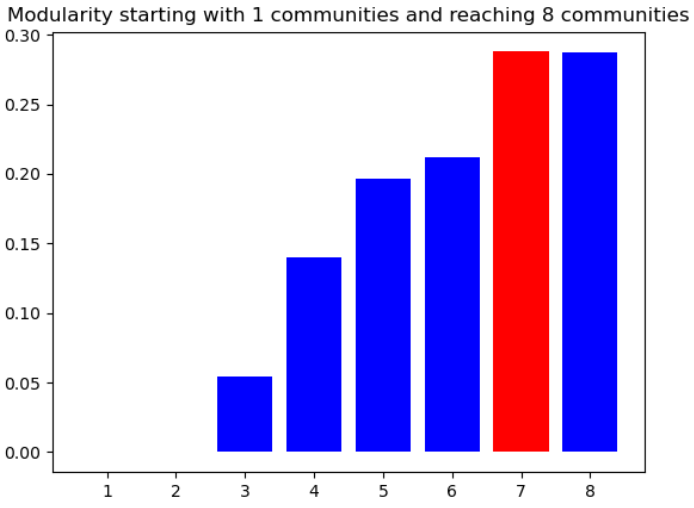
Πείραμα 1 (διάγραμμα μπαρών με όλα τα modularity και με zoom σε εύρος τιμών):



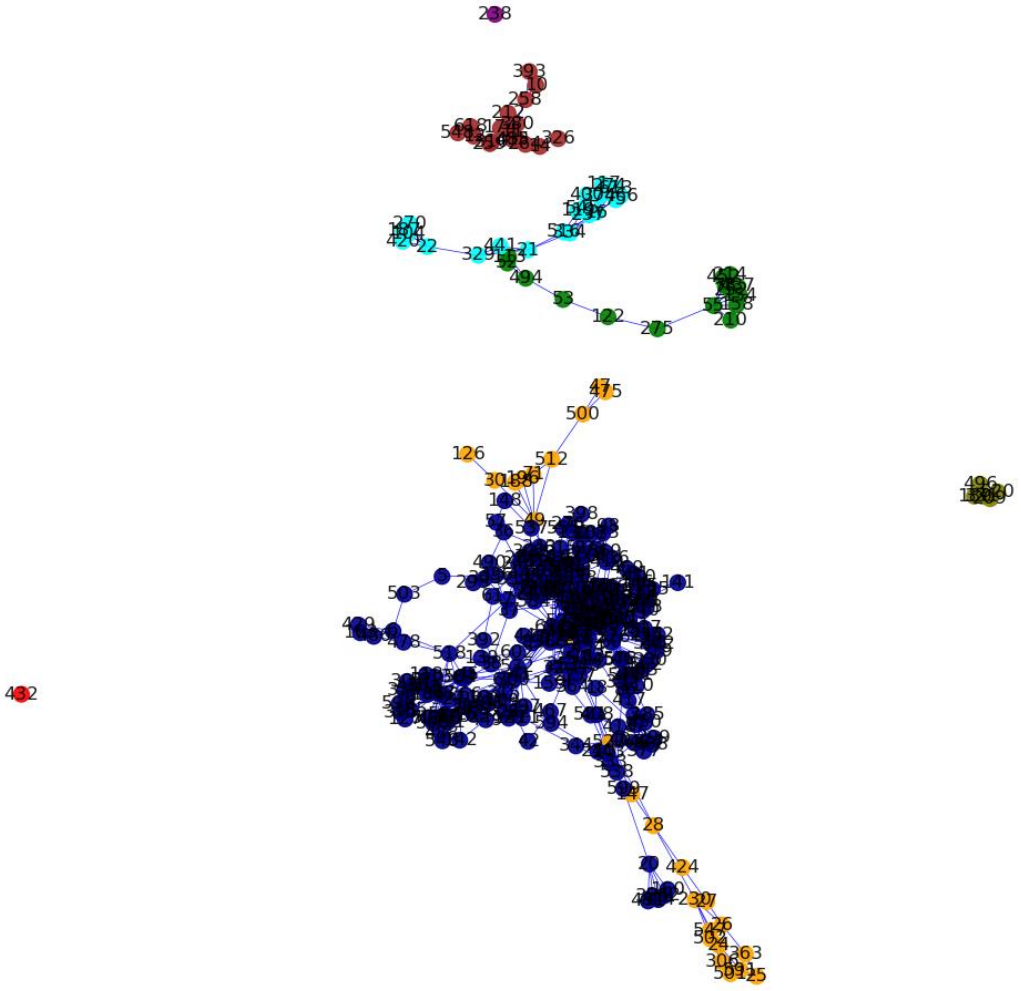
Πείραμα 1 (γράφημα με 34 κοινότητες):



Πείραμα 7 (διάγραμμα μπαρών με όλα τα modularity):



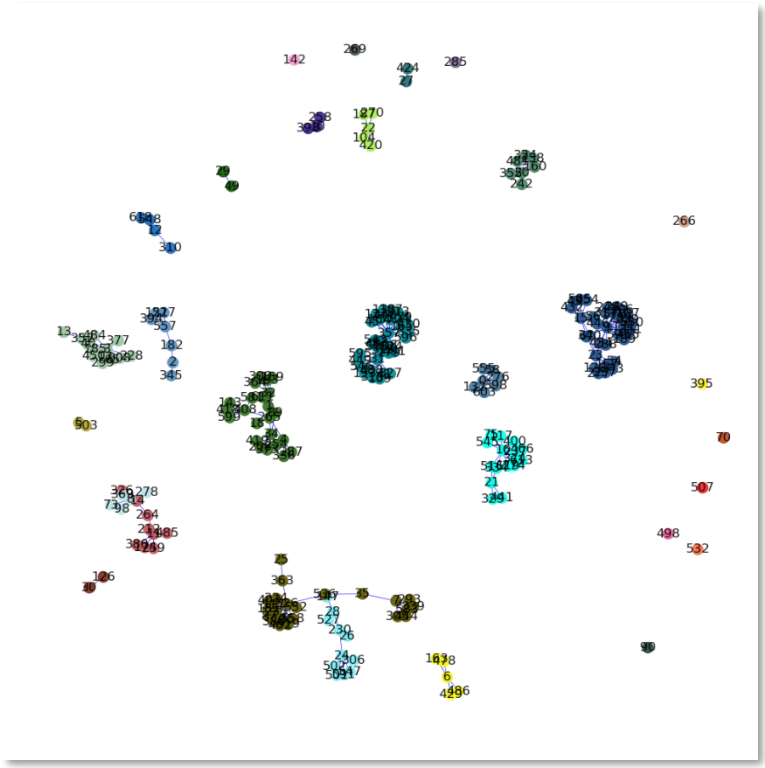
Πείραμα 7 (γράφημα με 7 κοινότητες):



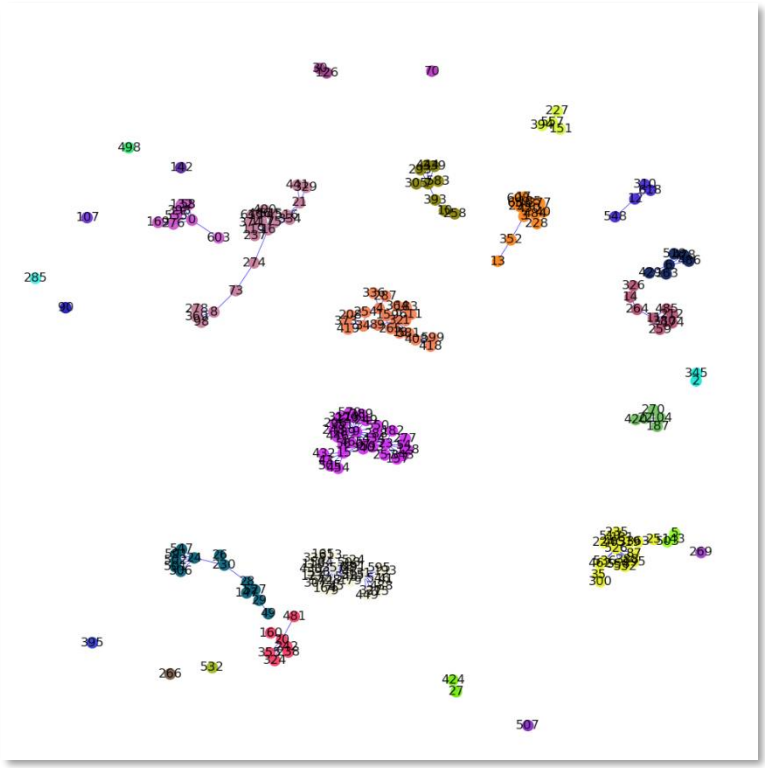
Δεν έχει νόημα η καταγραφή διαγραμμάτων μπαρών και αναλυτικών γραφημάτων για γραφήματα με ποσοστό 10% κόμβων στον υπολογισμό του betweenness, καθώς δεν μπορούν να αναπαραχθούν ακριβώς και έτσι θα υπάρχουν διαφορές στη μέγιστη τιμή modularity και επομένως διαφορετικές βέλτιστες διαμερίσεις σε κάθε εκτέλεση.

Παρατηρήσεις:

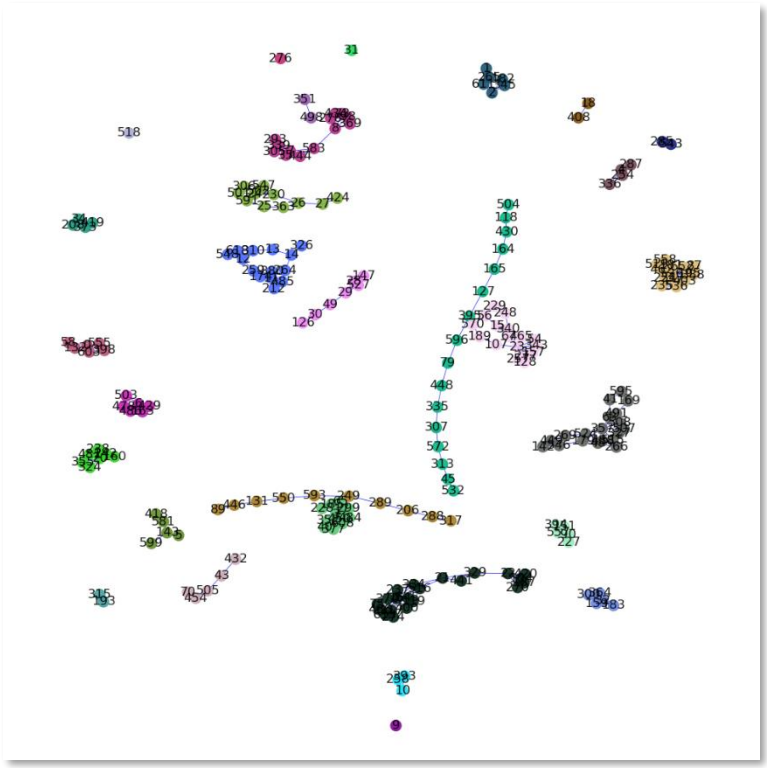
Από τη σύγκριση των γραφημάτων των πειραμάτων (όπως εμφανίζονται παρακάτω) φαίνεται ότι η προσθήκη κάποιων τυχαίων ακμών στο πείραμα 10 δίνει παρόμοιες τιμές με το πείραμα 9, αλλά φαίνεται ότι υπάρχει καλύτερος διαχωρισμός των κοινοτήτων, δηλαδή λιγότεροι κόμβοι που είναι μόνοι τους, καθώς το αρχικό γράφημα περιέχει ακμές-γέφυρες που οδηγούν σε αποκοπή ενός κόμβου μόνο από την υπόλοιπη κοινότητα σε μια επανάληψη. Παρόμοια αποτελέσματα έχει και η σύγκριση του πειράματος 9 με το πείραμα 11.



Πείραμα 9 Γράφημα 20 κοινοτήτων



Πείραμα 10 Γράφημα 21 κοινοτήτων



Πείραμα 11 Γράφημα 16 κοινοτήτων