Social Network Analysis

Project 2021-22

Δεληγιάννης Νικόλαος

3170268

Πηγή

Στο παρών paper θα αναλύσουμε τις σχέσεις ανάμεσα σε 34 αθλητές καράτε που βρίσκονταν στο ίδο club την περίοδο 1977-78. Το link των δεδομένων -το οποίο είναι σε μορφή GEFX – βρίσκεται εδώ: *https://networkrepository.com/soc-karate.php* .

Εργαλεία

Ως εργαλεία χρησιμοποιήθηκαν :

Λογισμικό Gephi .

Πακέτο NetworkX στη γλώσσα Python.

Δεδομένα

Κάνοντας export τον πίνακα των edges που δημιουργήθηκαν από το GEFX αρχείο , προέκυψε το αρχείο karate.csv το οποίο χρησιμοποιήθηκε για την υλοποίηση της ανάλυσης με τη βιβλιοθήκη NetworkX

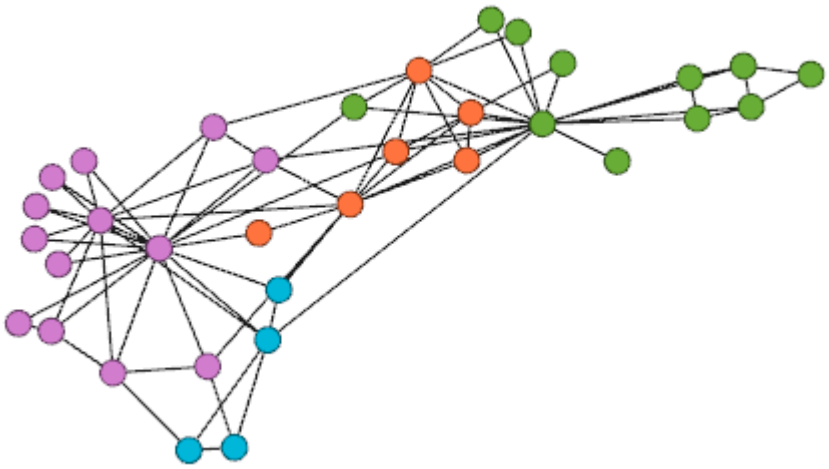
Υλοποίηση

Χρήση Λογισμικού

Οι γράφοι που παρουσιάζονται στο παρών paper προκύπτουν από το λογισμικό gephi μιας και μας προσφέρει πληθώρα επιλογών για την αναπαράστασή τους. Τα στατιστικά προκύπτουν είτε από το gephi είτε από το networkx , αλλά η ανάλυση τους θα βασιστεί στα outputs του gephi .

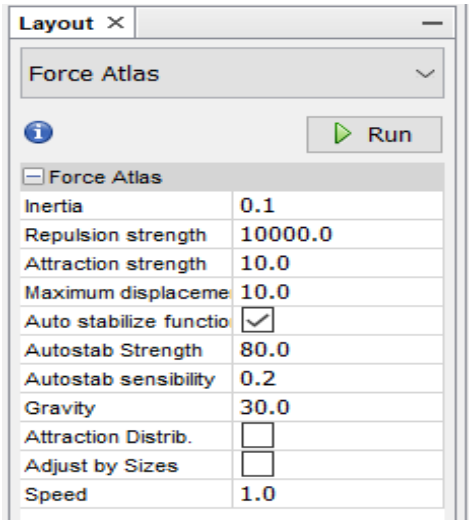
Γράφημα.

Παρακάτω παρουσιάζεται το αρχικό γράφημά το οποίο χωρίζεται σε 4 group χρωμάτων ανάλογα με το Modullarity class του κάθε node (ανθρώπου). Άξιο αναφοράς το γεγονός πως δεν έχουν τοποθετηθεί labels , έτσι ώστε να είναι πιο ευανάγνωστοι οι κόμβοι και οι συνδέσεις μεταξύ τους πάνω στον γράφο.



Network's Graph

Όσον αφορά στο layout του γραφήματος , χρησιμοποιήθηκε το Force Atlas με τις εξής παραμέτρους :



Γενικά χαρακτηριστικά γραφήματος

Ο γράφος περιέχει 34 nodes η αλλιώς πλήθος αθλητών καρατε και 78 edges δηλαδή ζεύγη αθλητών που αλληλεπιδρούν μεταξύ τους. Επίσης τρέχοντας τα εργαλεία Network Diameter , Density προκύπτουν τα εξής στοιχεία ανάλυσης :

Διάμετρος : 5

Ακτίνα : 3

Μέσο μήκος μονοπατιού : 2.41

Πυκνότητα (Density) : 0.139

Συνεκτικές συνιστώσες (Connected Components)

Στο γράφημα υπάρχει μια (1) συνεκτική συνιστώσα μιας και δεν υπάρχουν nodes που δεν συνδέονται με κάποιο άλλο node (αποκομμένοι) πράγμα που σημαίνει πως κάθε αθλητής είχε σχέση(αλληλεπιδρούσε) με τουλάχιστον έναν αθλητή.

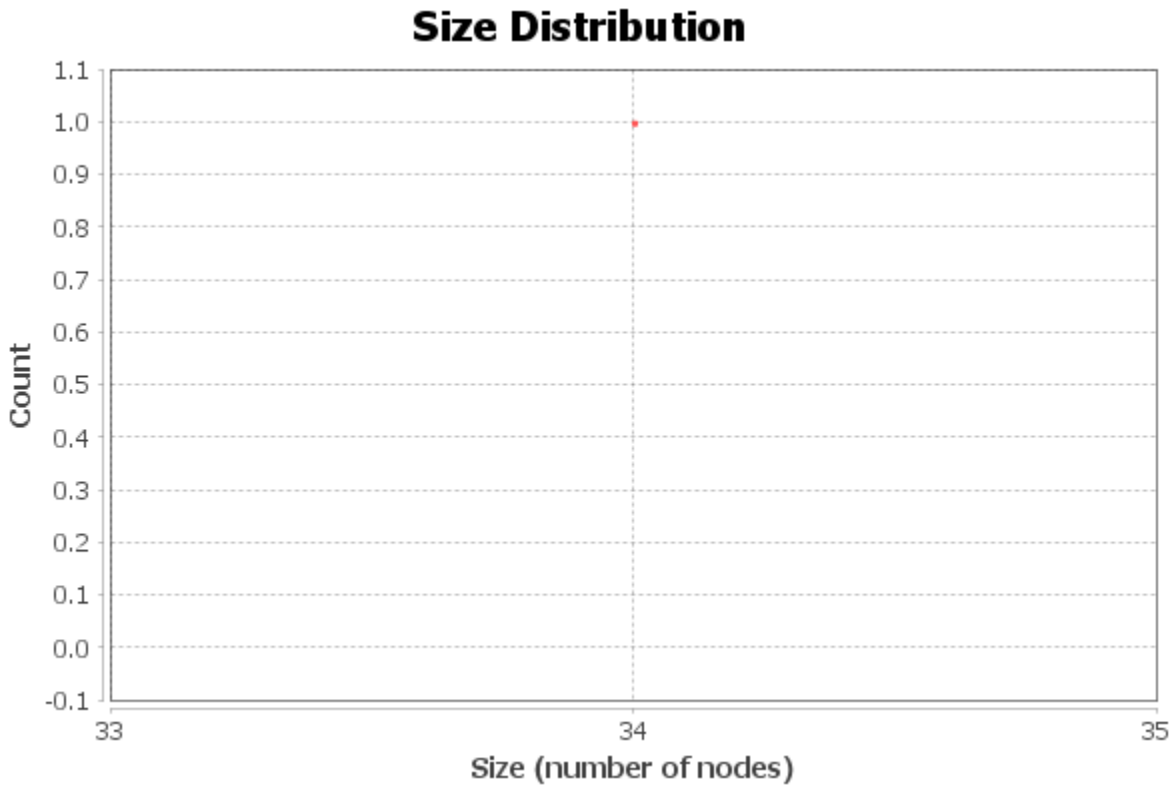
Η αναφορά του Gephi :

Parameters:

Network Interpretation: undirected

Results:

Number of Weakly Connected Components: 1



Η αναφορά του NetworkX :

```
giant (max connected components):  {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34}
giant's size :  34
```

Τρέχοντας το python script εμφανίζεται στο output το component η αλλιώς giant που περιλαμβάνει τους συνδεδεμένους κόμβους (max connected components) και έχει μέγεθος 34.

Βαθμοί Κορυφών (Degree Measures)

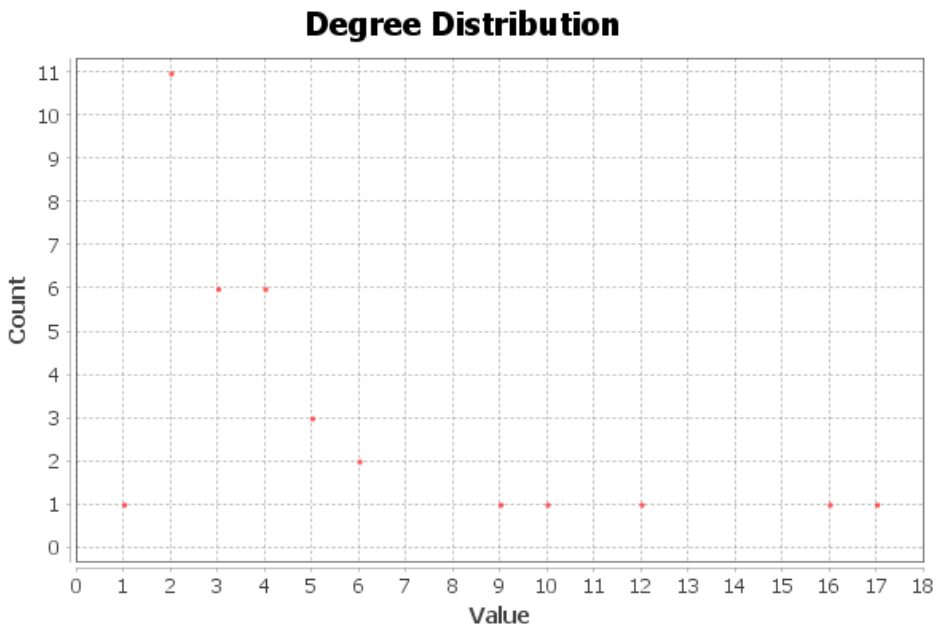
Για την εύρεση του maximum degree και του average degree μπορούμε να ανατρέξουμε είτε στο NetworkX μοντέλο είτε στο gephi και να διαπιστώσουμε πως :

Max Degree , Average Degree on NetworkX

```
Max Degree : 17 on node with id :  34
Average Degree : 4.588235294117647
```

Max Degree , Average Degree on Gephi

Results: Average Degree: 4.588



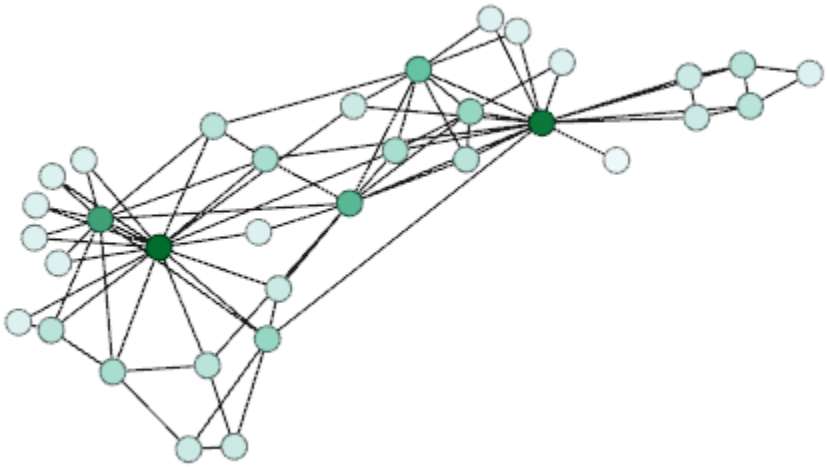
Πρώτες 6 κορυφές με το μεγαλύτερο degree :

Id	Degree
34	17
1	16
33	12
3	10
2	9
4	6

Πράγματι όπως φαίνεται και στον gephi πίνακα των δεδομένων ο κόμβος με id : 34 έχει το μεγαλύτερο degree value.

Από τα παραπάνω αποτελέσματα μπορούμε να διαπιστώσουμε πως κάθε αθλητής καράτε σχετίζεται (αλληλεπιδρά) με 5 (4.588) αθλητές κατα μέσο όρο . Επίσης παρατηρούμε πως υπάρχουν αρκετές κορυφές με χαμηλό degree [2 εως 3] πράγμα που μπορεί να σημαίνει πως η πλειονότητα των αθλητών προτιμούν να έχουν σχέσεις με λίγα άτομα.

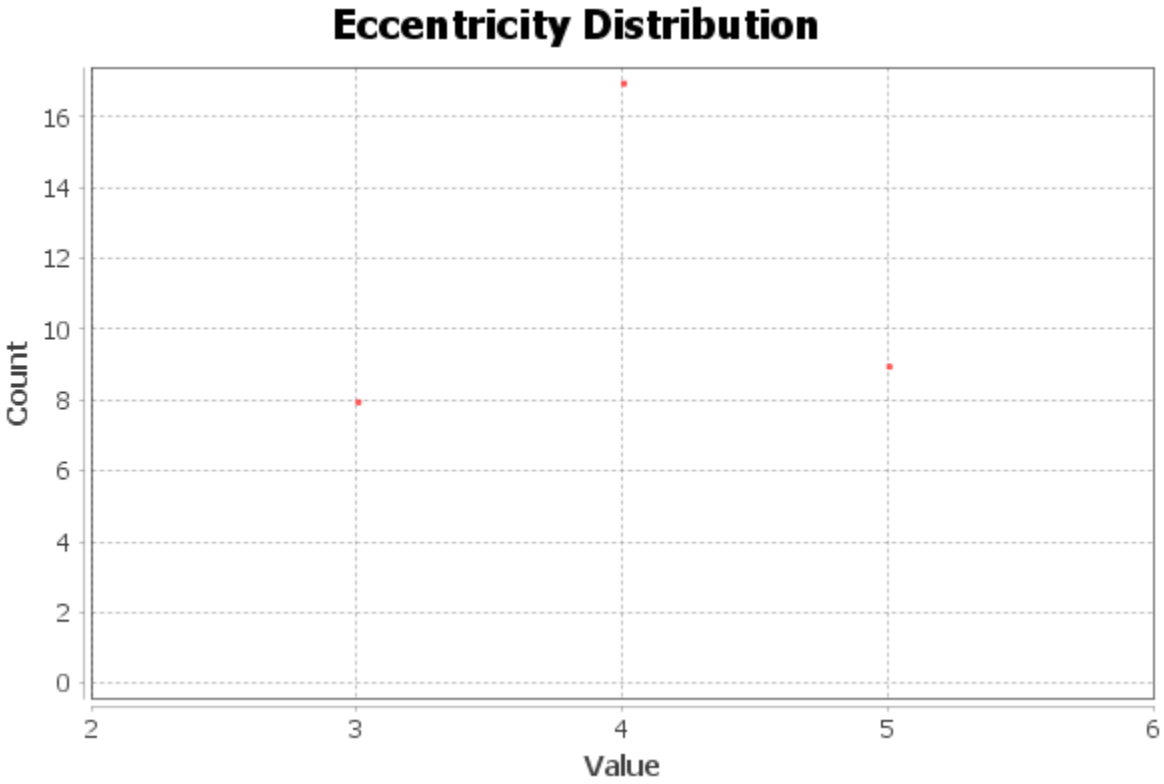
Παρακάτω δίνεται το γράφημα του οποίου οι κορυφές είναι χρωματισμένες βάση του degree . Όσο πιο μεγάλη τιμή degree έχει μια κορυφή τόσο πιο σκούρο πράσινο χρώμα έχει.



Network’s Graph with degree applied

Μέτρα Κεντρικότητας (Centrality Measures)

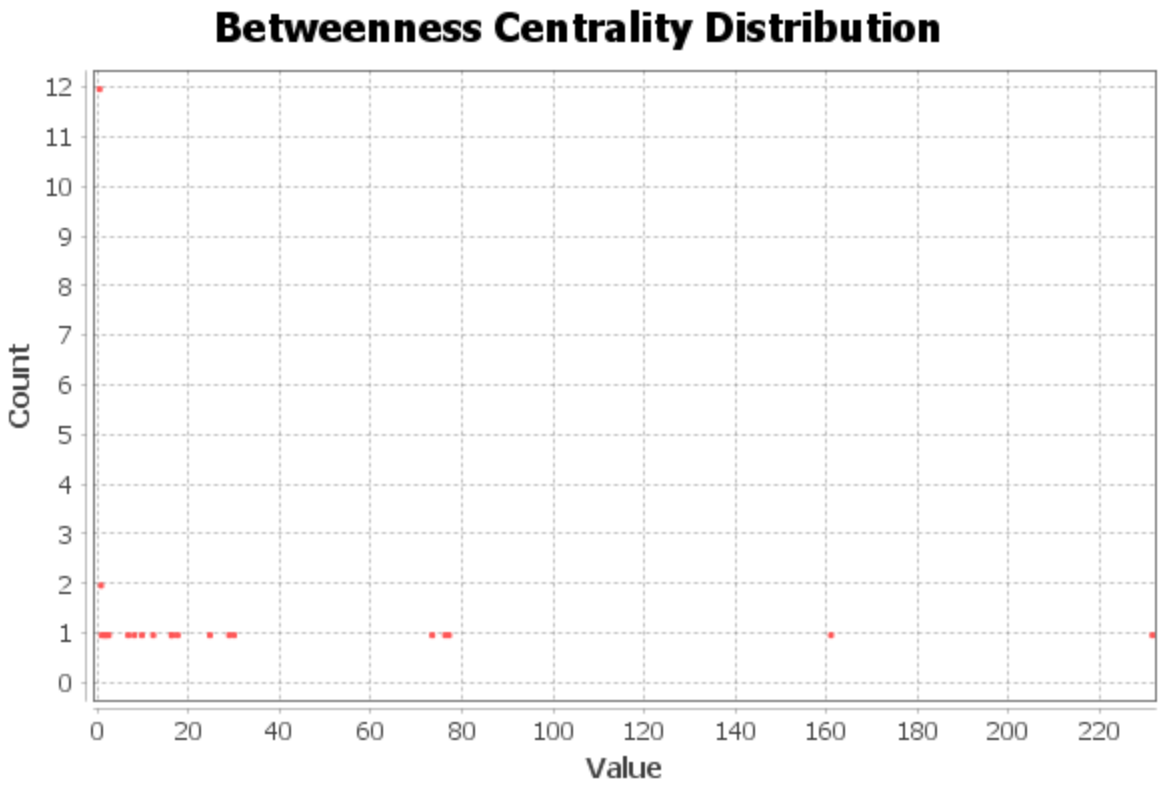
Eccentricity



Όπως φαίνεται στη παραπάνω αναπαράσταση του Eccentricity οι τιμές της βρίσκονται στο σύνολο [3,4,5]. Αυτό το αποτέλεσμα είναι αναμενόμενο καθώς καθώς η τιμή η τιμή της εκκεντρότητας μιας κορυφής εξαρτάται από την μέγιστη απόσταση της κορυφής αυτής από τις υπόλοιπες κορυφές του γραφήματος , άρα αποκλείεται το σύνολο να περιείχε τιμές μεγαλύτερες της διαμέτρου (5)

Betweenness Centrality

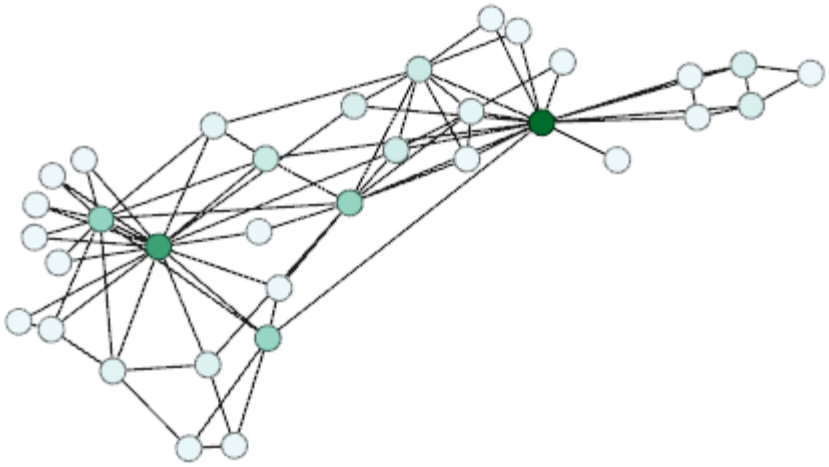
Το betweenness centrality κάθε κορυφής έιναι ένα μέτρο θέσης για αυτήν , το οποίο μεγαλώνει ανάλογα με το πλήθος των συντομότερων μονοπατιών στα οποία βρίσκεται.



Παρακάτω δίνονται οι πρώτοι 5 αθλητές με τις μεγαλύτερες τιμές betweenness centrality :

Id	Betweenness Centrality
1	231.071429
34	160.551587
33	76.690476
3	75.850794
32	73.009524

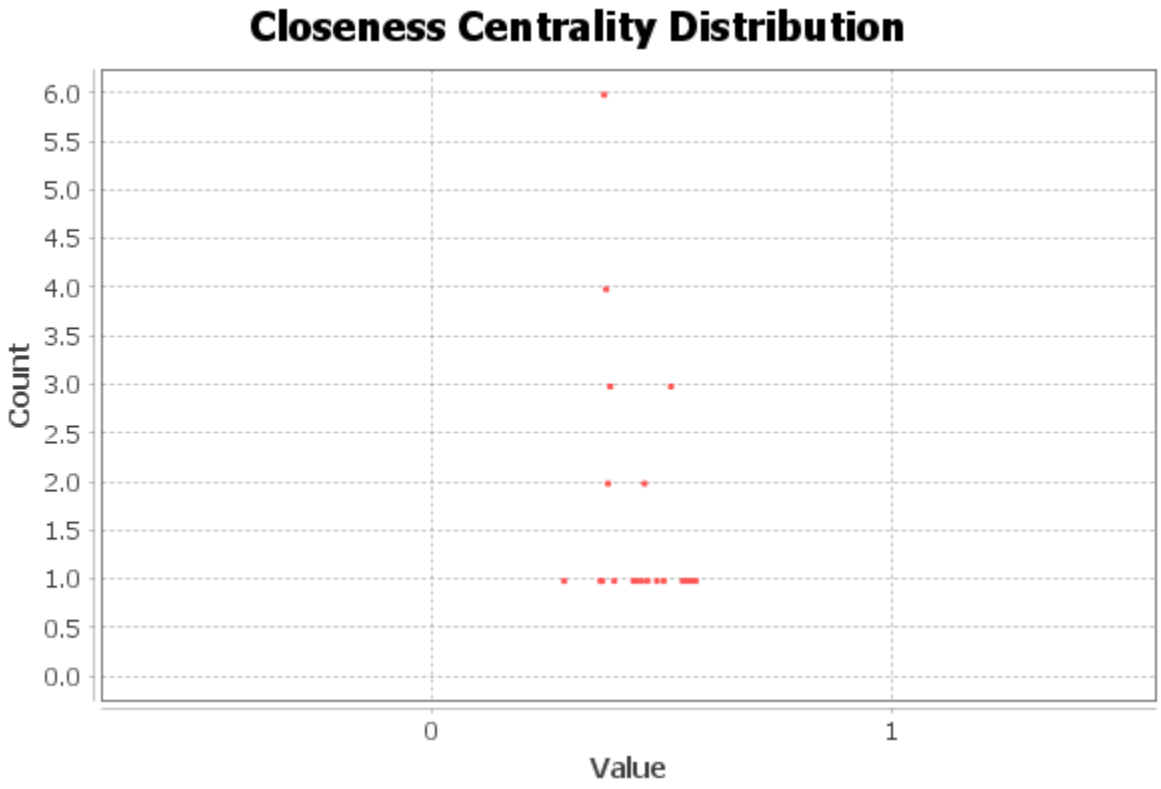
Παρακάτω δίνεται το γράφημα του οποίου οι κορυφές είναι χρωματισμένες βάση του betweenness centrality . Όσο πιο μεγάλη τιμή betweenness centrality έχει μια κορυφή τόσο πιο σκούρο πράσινο χρώμα έχει.



Network's Graph with betweenness centrality applied

Closeness Centrality

Το closeness centrality κάθε κορυφής έιναι ένα μέτρο σπουδαιότητας για αυτήν , το οποίο υπολογίζεται βάση του αθροίσματος των συντομότερων μονοπατιών της κάθε κορυφής με τις υπόλοιπες κορυφές του γράφου .

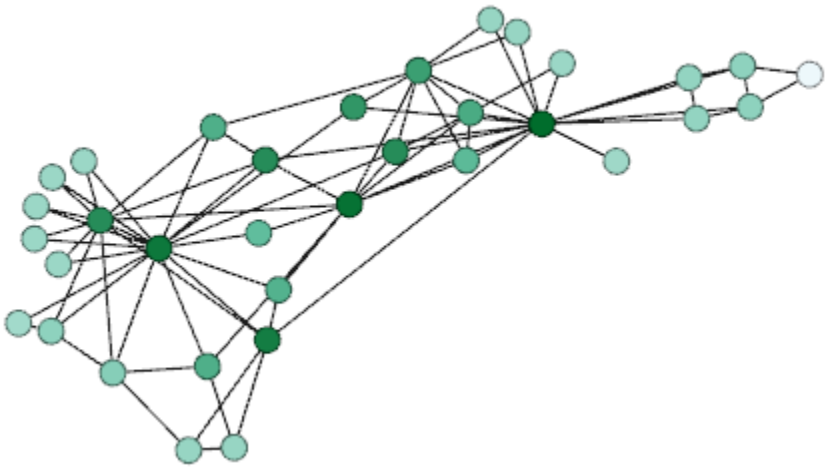


Όπως φαίνεται στη παραπάνω αναπαράσταση του closeness centrality οι τιμές της είναι κανονικοποιημένες αφού βρίσκονται στο διάστημα [0,1] .

Παρακάτω δίνονται οι πρώτοι 5 αθλητές με τις μεγαλύτερες τιμές closeness centrality :

Id	Closeness Centrality
1	0.568966
3	0.559322
34	0.55
32	0.540984
33	0.515625

Παρακάτω δίνεται το γράφημα του οποίου οι κορυφές είναι χρωματισμένες βάση του closeness centrality. Όσο πιο μεγάλη τιμή closeness centrality έχει μια κορυφή τόσο πιο σκούρο πράσινο χρώμα έχει.



Network's Graph with betweenness centrality applied

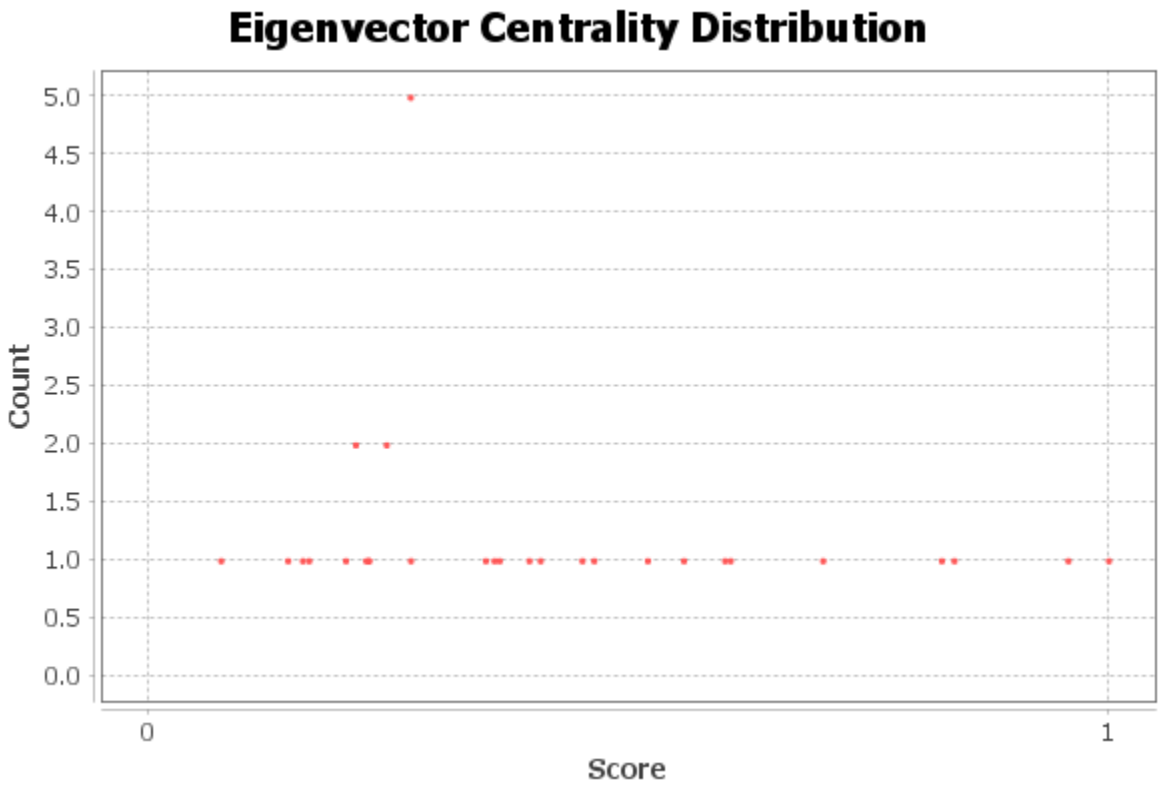
<u>Eigenvector Centrality</u>

Το Eigenvector centrality είναι μια γενίκευση του Pagerank το οποίο θα αναλύσουμε στη συνέχεια . Για να έχει μια κορυφή μεγάλο σκορ θα πρέπει και οι υπόλοιπες συνδεδεμένες -με αυτήν- κορυφές να έχουν εξίσου μεγάλο σκορ . Συγκεκριμένα :

Parameters:

Network Interpretation: undirected
Number of iterations: 100
Sum change: 9.783386901414531E-4

Results:



Παρακάτω δίνονται οι πρώτοι 5 αθλητές με τις μεγαλύτερες τιμές Eigenvector centrality:

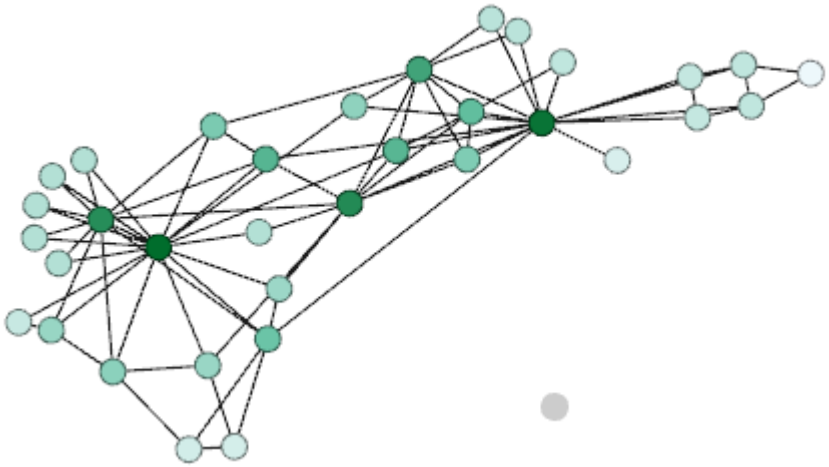
Id	Eigenvector Centrality
34	1.0
1	0.95754
3	0.838534
33	0.825572
2	0.70159

Παρατηρούμε πως ο αθλητής με Id : 1 έχει Eigenvector centrality ίσο με 1 , γεγονός που μας κάνει να συμπαιρένουμε πως το Eigenvector centrality σχετίζεται με το degree value , δηλαδή όσες περισσότερες αλληλεπιδράσεις έχει κάποιος τόσο μεγαλύτερο Eigenvector centrality θα έχει , πράγμα που διαψεύδεται στον επόμενο πίνακα στον οποίο έχει προστεθεί και η στήλη των degrees.

Id	Eigenvector Centrality	Degree
34	1.0	17
1	0.95754	16
3	0.838534	10
33	0.825572	12
2	0.70159	9

Βλέπουμε πως ο αθλητής με Id ίσο με 33 έχει χαμηλότερο Eigenvector centrality από τον αθλητή με Id : 3 παρότι ο 33 έχει μεγαλύτερη τιμή degree.

Παρακάτω δίνεται το γράφημα του οποίου οι κορυφές είναι χρωματισμένες βάση του Eigenvector centrality. Όσο πιο μεγάλη τιμή Eigenvector centrality έχει μια κορυφή τόσο πιο σκούρο πράσινο χρώμα έχει.

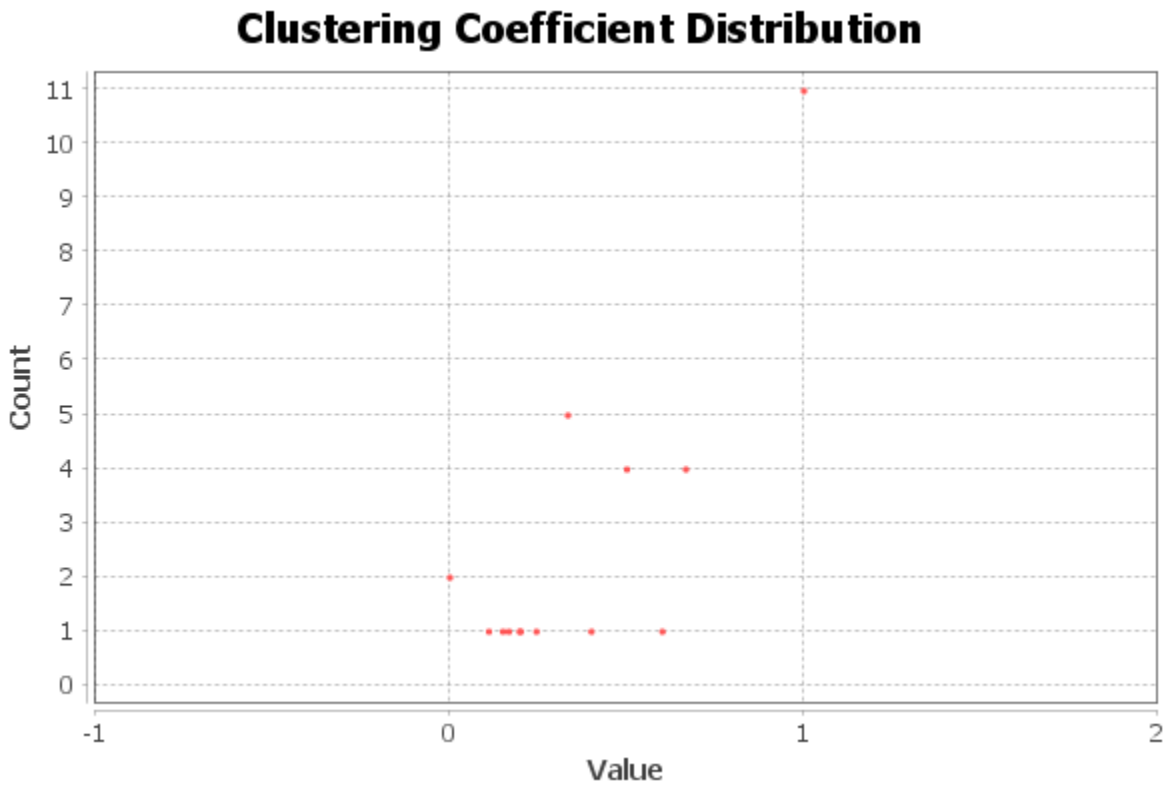


Network’s Graph with Eigenvector centrality applied

Clustering Coefficient

Results:

Average Clustering Coefficient: 0.588
Total triangles: 45
The Average Clustering Coefficient is the mean value of individual coefficients.



Παρατηρούμε πως έχουμε Average Clustering Coefficient ίσο με 0.588 και αριθμό τριγώνων ίσο με 45 . Ανατρέχοντας στο networkx μοντέλο μπορούμε να δούμε και από ποιες κορυφές αποτελούντε τα εν λόγω τρίγωνα.

Παρακάτω δίνονται οι αθλητές με Clustering Coefficient τιμή μεγαλύτερη ή ίση του 5 :

Id	Clustering Coefficient
8	1.0
13	1.0
15	1.0
16	1.0
17	1.0
18	1.0
19	1.0
21	1.0
22	1.0
23	1.0
27	1.0
4	0.666667
5	0.666667
11	0.666667
30	0.666667
14	0.6
6	0.5
7	0.5
9	0.5
31	0.5

Παρατηρούμε πως οι απομονομένες κορυφές του γραφήματος έχουν την τάση να εμφανίζουν μεγάλο συντελεστή συσταδοποίησης . Το παρών μπορεί να επιβεβαιωθεί προσέχοντας τις degree τιμές των εν λόγω κορυφών στον παρακάτω πίνακα .

Id	Clustering Coefficient	Degree
8	1.0	4
13	1.0	2
15	1.0	2
16	1.0	2
17	1.0	2
18	1.0	2
19	1.0	2
21	1.0	2
22	1.0	2
23	1.0	2
27	1.0	2
4	0.666667	6
5	0.666667	3
11	0.666667	3
30	0.666667	4
14	0.6	5
6	0.5	4
7	0.5	4
9	0.5	5
31	0.5	4

Γέφυρες (Bridges)

Για τον εντοπισμό των γεφυρών τοπικών και μη , μπορούμε είτε να ανατρέξουμε στο networkx μοντέλο είτε να εγταταστήσουμε το απαραίτητο plugin στο gephi .

Έχουμε :

NetworkX Output:

```
Bridges of network : [(12, 1)]
Local Bridges of network: [(2, 31, 3), (3, 10, 3), (3, 28, 3), (3, 29, 3), (10, 34, 3), (12, 1, inf), (14, 34, 3), (20, 34, 3), (26, 24, 3), (28, 25, 3), (32, 1, 3)]
```

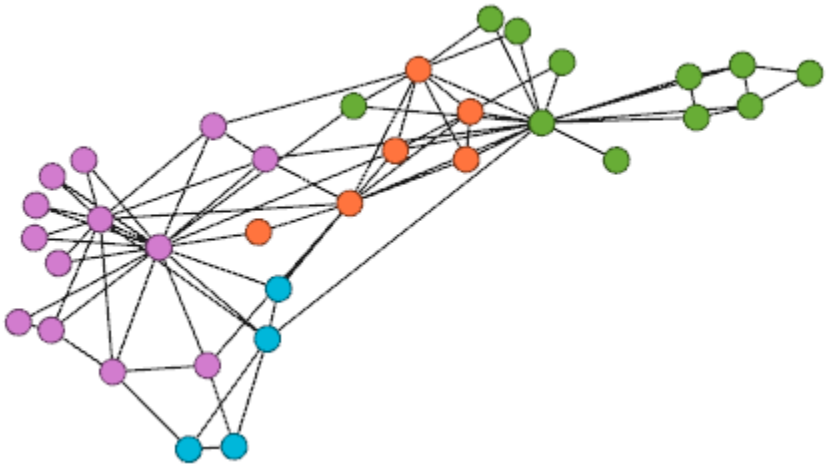
Community Structure

Modularity

Results:

Modularity: 0.416
Modularity with resolution: 0.416
Number of Communities: 4

Όπως αναφέρθηκε και στην αρχή της ανάλυσης , το δίκτυό μας αποτελείται από 4 communities. Παρατηρούμε πως οι κορυφές που βρίσκονται στην ίδια κοινότητα αλληλεπιδρούν περισσότερο μεταξύ τους παρά με τις κορυφές που βρίσκονται σε διαφορετική κοινότητα. Επίσης βλέπουμε πως η κοινότητα 3(ροζ) έχει τις περισσότερες κορυφές (13) δηλαδη είναι η μεγαλύτερη του δικτύου.



Network's Graph with communities

Cliques

Για τον εντοπισμό των κλικών , μπορούμε είτε να ανατρέξουμε στο networkx και να παρατηρήσουμε πως προκύπτει η παρακάτω κλίκα.

Έχουμε :

NetworkX Output:

```
K_clique_communities:  [frozenset({1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34})]
```

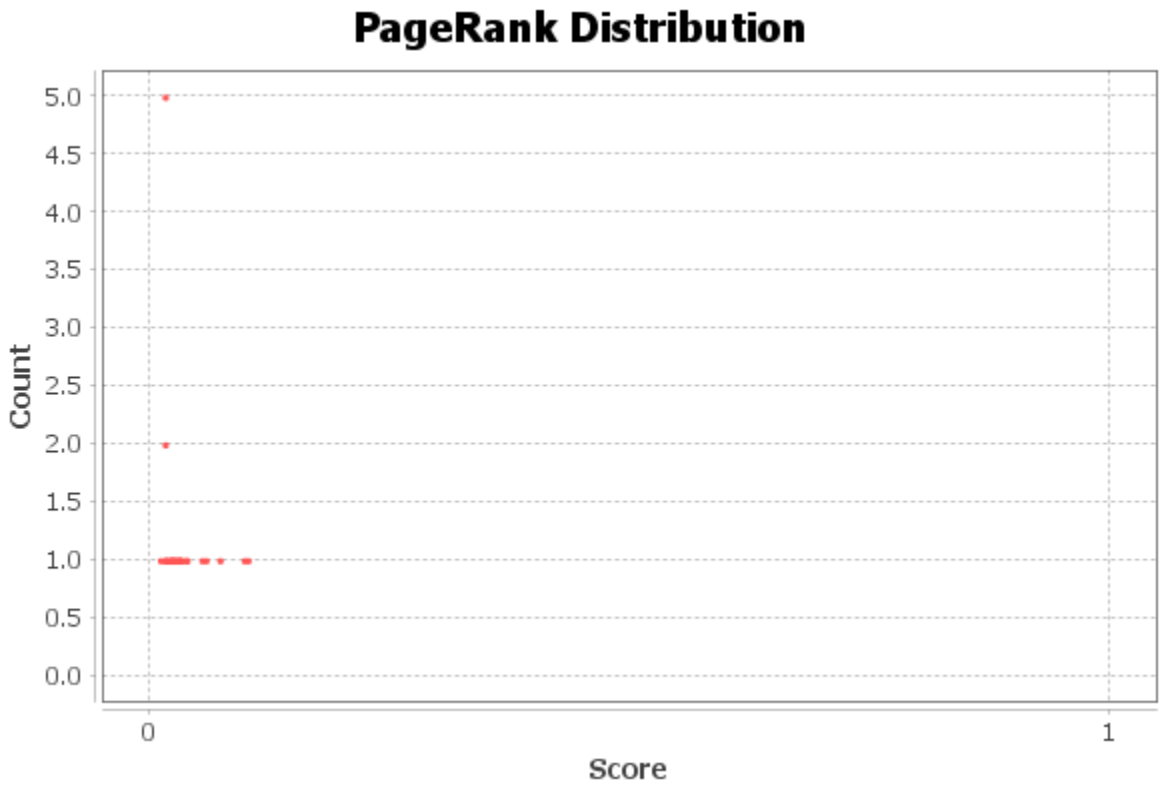
Pagerank

Το Pagerank μετράει το πόσο σημαντική είναι μια ιστοσελίδα βάσει των συνδέσμων που οδηγούν σε αυτήν, συνεπώς σε αυτήν την περίπτωση θα δούμε πόσο σημαντική είναι η κάθε κορυφή για το pagerank.

Parameters:

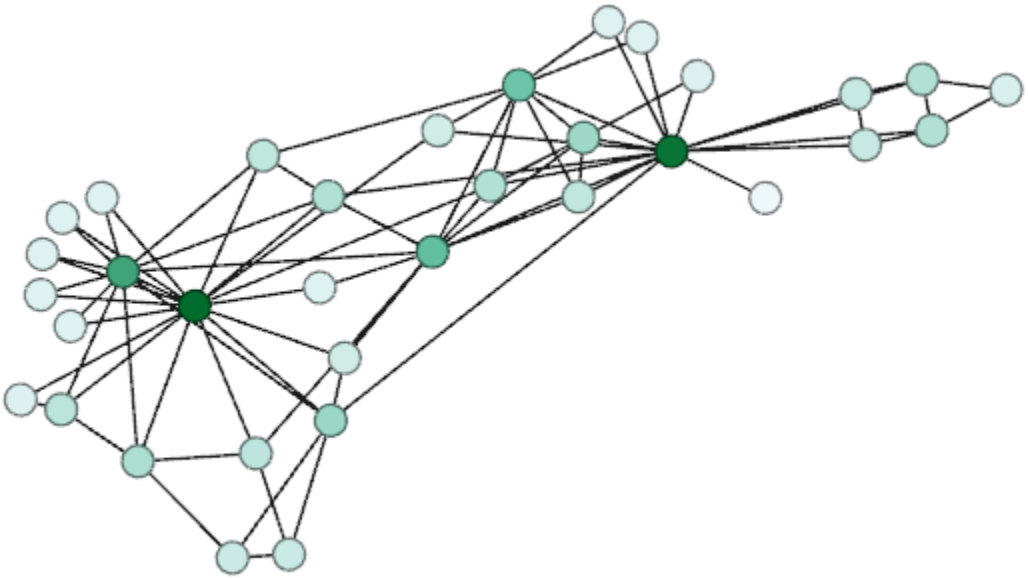
Epsilon = 0.001
Probability = 0.85

Results:



Για τις παραμέτρους, έχουμε διαλέξει Epsilon = 0.001, Probability = 0.85, οι οποίες ήταν οι προκαθορισμένες τιμές από το Gephi. Η τιμή Probability = 0.85 αφορά την πιθανότητα ενός χρήστη που κάνει κλικ σε τυχαίους συνδέσμους να συνεχίσει να κάνει κλικ και σε άλλους συνδέσμους. Η τιμή Epsilon = 0.001 είναι το κριτήριο τερματισμού του αλγορίθμου, οπότε όσο μικρότερη τιμή έχει, τόσο πιο πολύ θα αργήσει ο υπολογισμός του pagerank

Παρατηρούμε πως τα σκορ του pagerank είναι αρκετά χαμηλά πράγμα που σημαίνει πως αμα ο κάθε αθλητής (κορυφή) αντιπροσωπευόταν ως μια ιστοσελίδα , τότε ένας τυχαίος χρήστης του διαδικτύου σπάνια θα βρισκόταν τυχαία σε μία από αυτές τις ιστοσελίδες.



Network’s Graph with pagerank applied