ΕΡΓΑΣΙΑ ΣΤΟ ΜΑΘΗΜΑ ΘΕΩΡΙΑ ΔΙΚΤΥΩΝ

Εργασία

Wikipedia net

Ηλιάνα Κόγια 10090 ilianakogia@ece.auth.gr

Φεβρουάριος 2024

Κατασκευή θεματικού δικτύου από την Wikipedia

Επιλέχθηκαν για το συγκεκριμένο θεματικό δίκτυο οι παρακάτω λέξεις κλειδιά, οι οποίες αποτελούν τίτλους άρθρων της Wikipedia στην αγγλική γλώσσα:

Topic:

{ Science, Technology, Engineering, Mathematics }

Σύμφωνα με τον αλγόριθμο που υλοποιήθηκε, για κάθε μια από τις λέξεις κλειδιά του θέματος βρίσκουμε την σελίδα της Wikipedia με τίτλο την εκάστοτε λέξη-κλειδί και έπειτα τους συνδέσμους που υπάρχουν στο άρθρο της σελίδας αυτής. Δεν χρησιμοποιούνται όλοι οι υπάρχοντες σύνδεσμοι λόγω χρόνου και πολυπλοκότητας ειδικά όσο αυξάνεται το βάθος. Γι' αυτό οι σύνδεσμοι (links) που θα αποτελέσουν κόμβους του δικτύου θα είναι μόνο οι πιο σχετικοί (με βάση τη σημασιολογική συσχέτιση των άρθρων, threshold = 0.6)

Η σημασιολογική συσχέτιση δύο άρθρων υπολογίζεται μέσω του δοθέντος μοντέλου sentence-transformers/all-MiniLM-L6-v2 και του cosine_similarity των embeddings, και χρησιμοποιείται ο τίτλος των άρθρων.

***θα μπορούσε να χρησιμοποιηθεί μέρος του περιεχομένου ή και το summary ενός άρθρου, ωστόσο για λόγους ταχύτητας επιλέχθηκε ο τίτλος για τις εκάστοτε συγκρίσεις*

Επίσης, βάρη των ακμών αποτελούν οι σημασιολογικές συσχετίσεις και ο γράφος είναι κατευθυνόμενος .

1.1 Ανάλυση δικτύου μέσω Gephi

Δίκτυο με topic STEM :

Nodes: 2447

Edges: 4474

Directed Graph

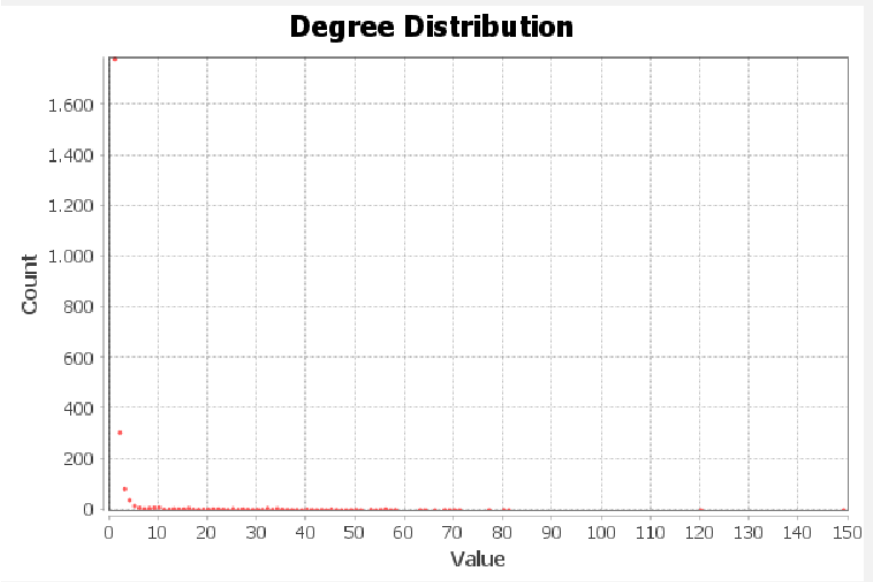
Diameter: 9
Radius: 0
Average Path length: 4.378551249640223

Ο γράφος δεν είναι ισχυρά συνδεδεμένος και έχουμε:

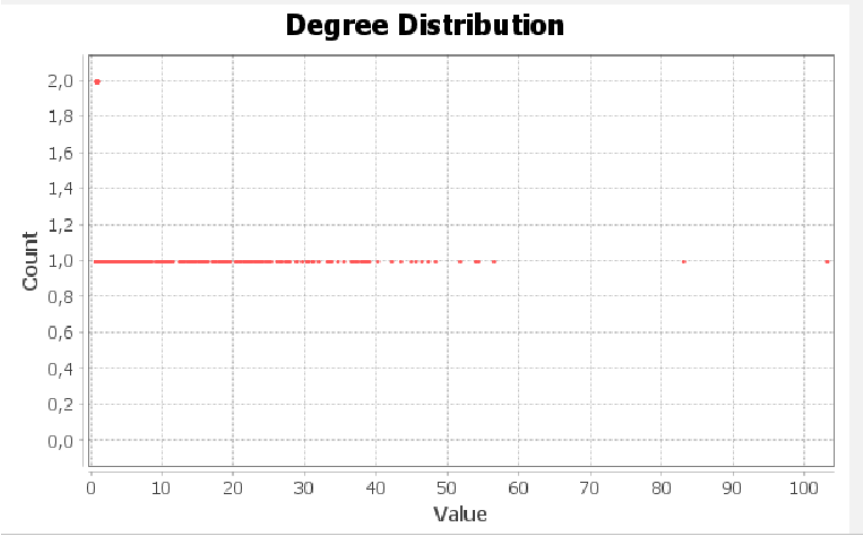
Number of Weakly Connected Components: 1
Number of Strongly Connected Components: 2225

Κατανομή βαθμών:

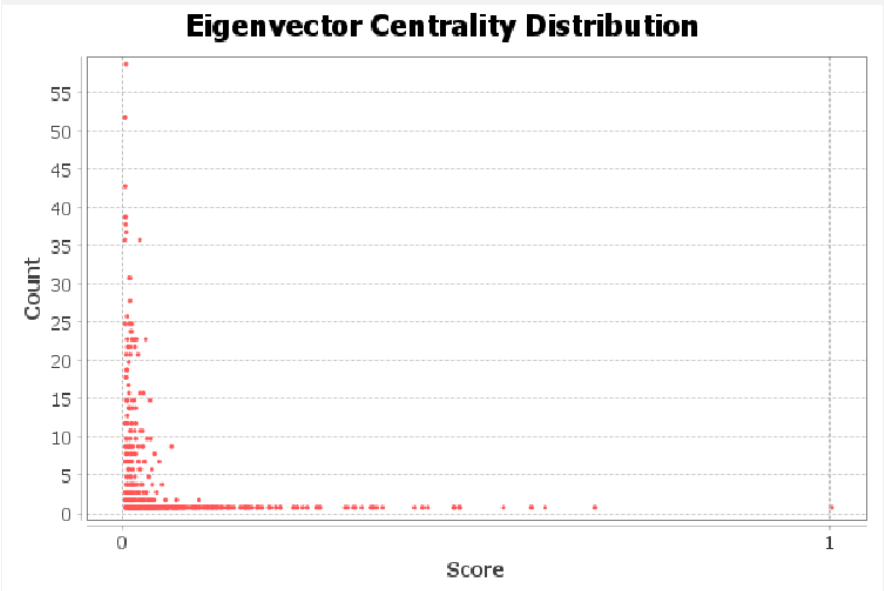
Μέσος Βαθμός : 1.828



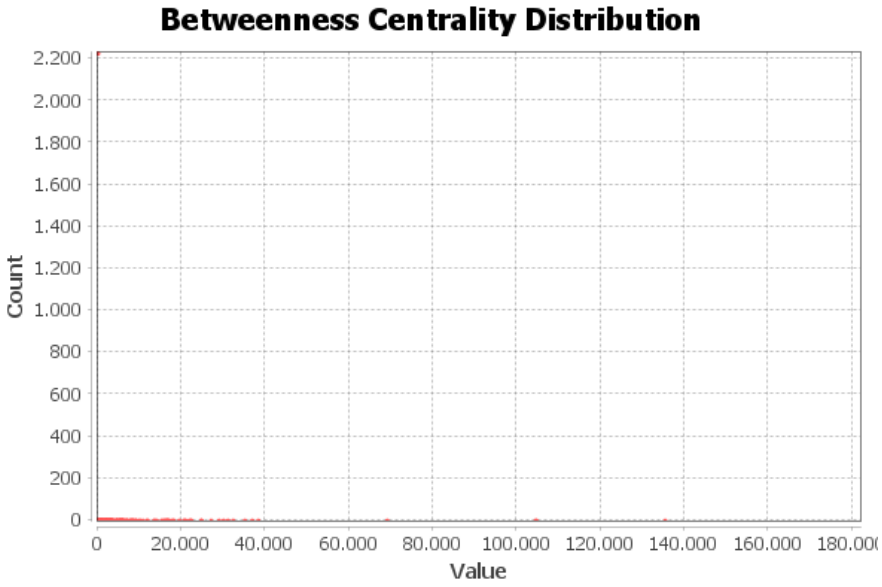
Μέσος σταθμισμένος Βαθμός : 1.242



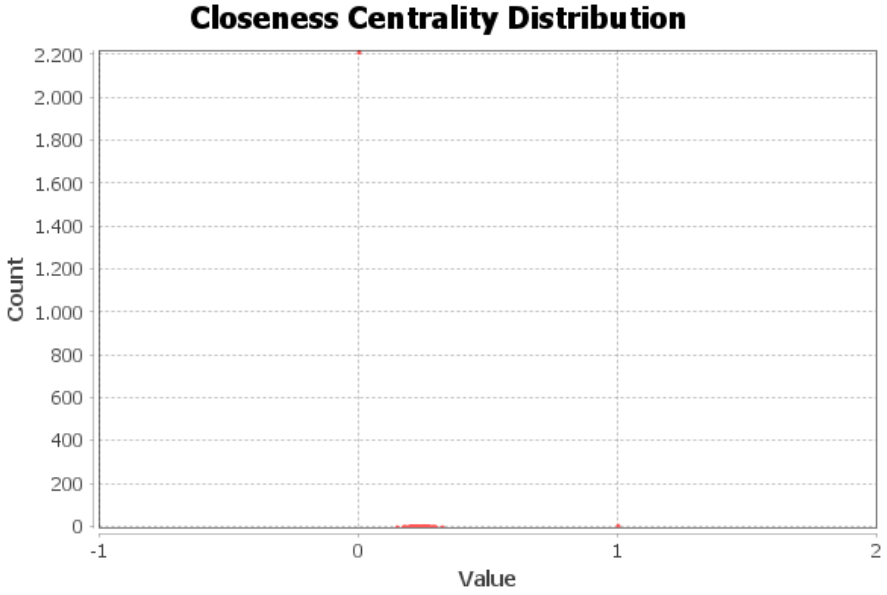
Μετρική της σημαντικότητας των κόμβων με βάση τις συνδέσεις των κόμβων

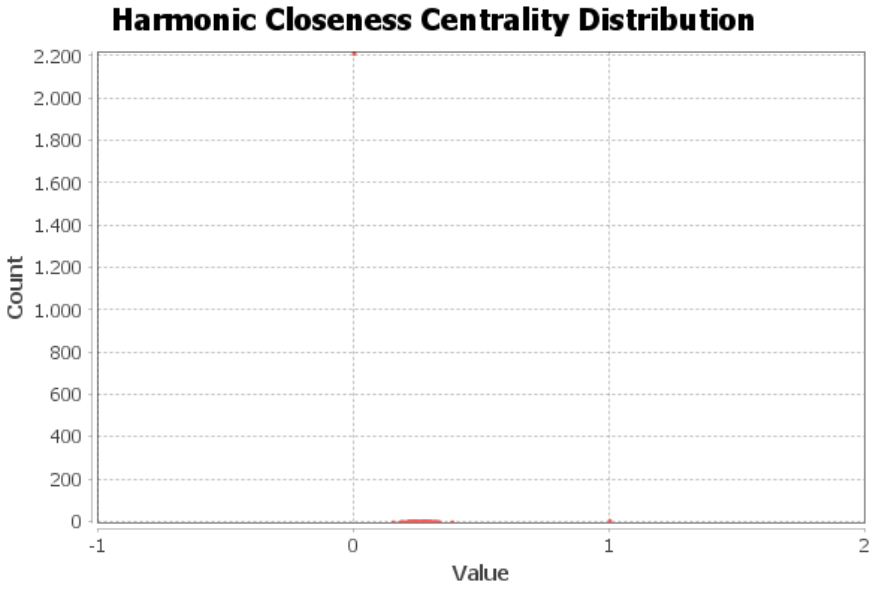


Betweenness Centrality: πόσο συχνά ένας κόμβος εμφανίζεται σε shortest paths μεταξύ άλλων κόμβων του δικτύου

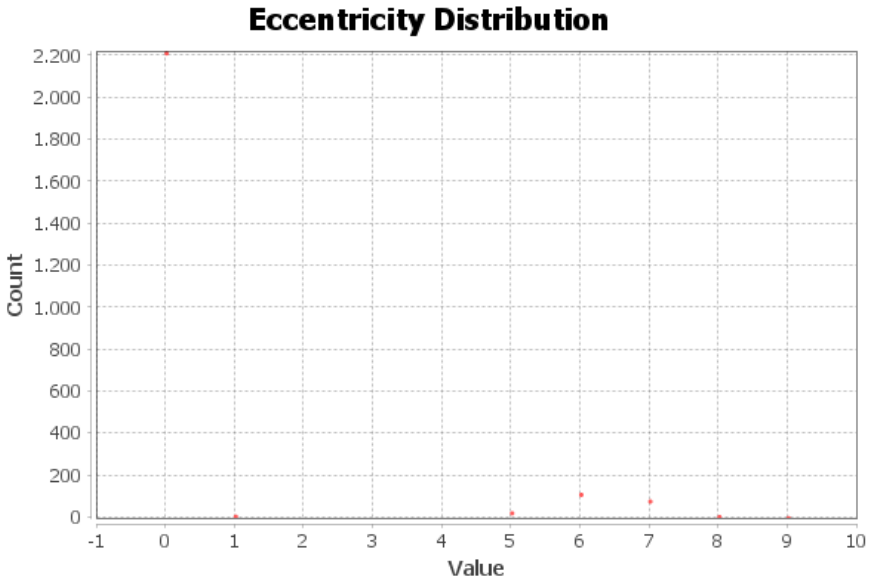


Closeness Centrality: Η μέση απόσταση από έναν κόμβο προς όλους τους υπόλοιπους κόμβους του δικτύου.





Eccentricity: Η απόσταση μεταξύ ενός κόμβου και του μακρινότερου από αυτόν κόμβο του δικτύου



Παρατηρούμε ότι σε όλες τις παραπάνω κατανομές οι περισσότεροι κόμβοι του δικτύου έχουν τιμή κοντά στο 0 ("μικρή σημαντικότητα") με εξαίρεση την betweenness centrality.

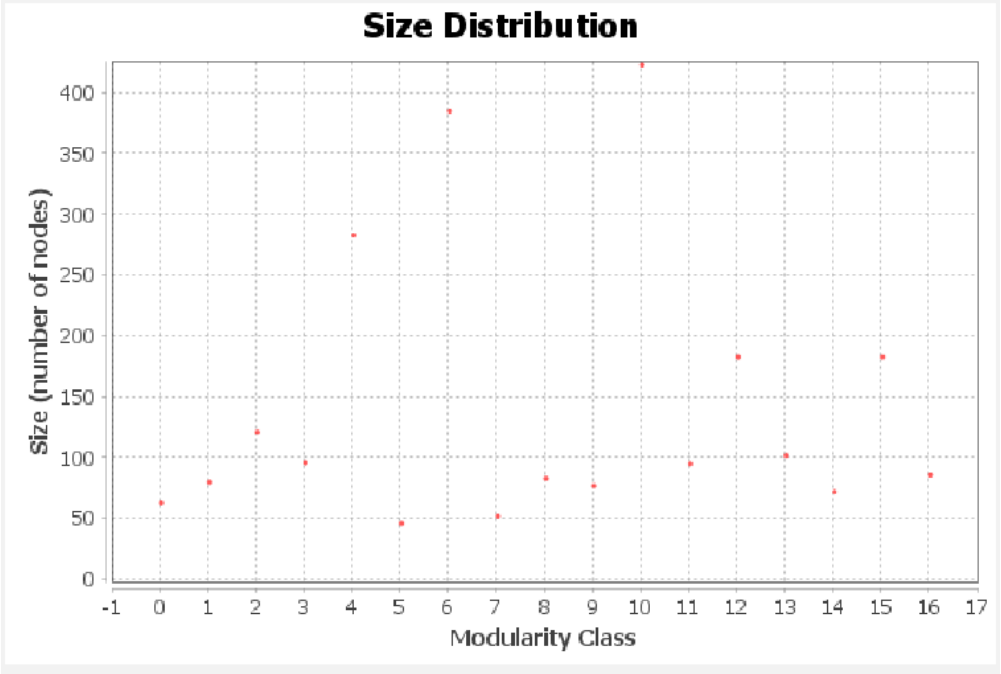
***Graph Density**: 0.001*

μετρική που μας δείχνει πόσο κοντά είναι ο γράφος στο να γίνει πλήρης, δηλαδή να υπάρχουν όλες οι δυνατές ακμές μεταξύ των κόμβων. Τα πραγματικά δίκτυα δεν είναι dense, όπως και στην περίπτωσή μας

Θεματικές Περιοχές

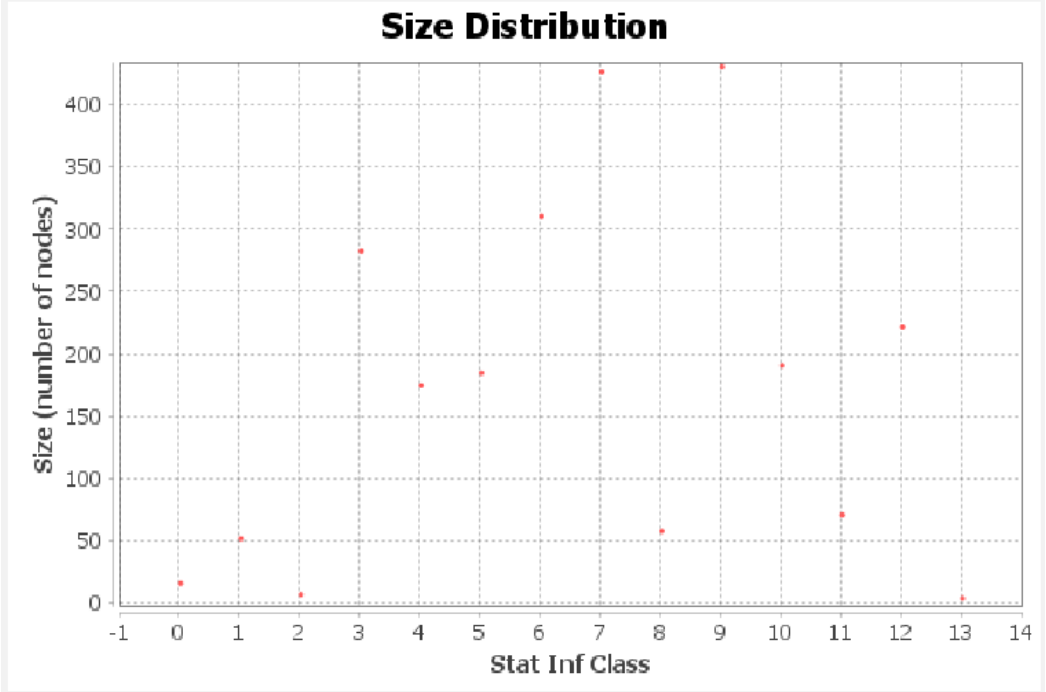
1) *Modularity (προέκυψε 0.786 θεωρείται αρκετά υψηλή τιμή και άρα οι κοινότητες είναι αρκετά διακριτές)*

Modularity: 0,786
Modularity with resolution: 0,786
Number of Communities: 17



2) *Statistical Inference*

Description Length: 23085,698
Number of Communities: 14



Ο συντελεστής ομαδοποίησης μπορεί να μας δώσει πληροφορία αν εντοπίζονται στο γράφο φαινόμενα μικρού κόσμου (small - world)

```
Average Clustering Coefficient: 0,109
The Average Clustering Coefficient is the mean value of individual coefficients.
```

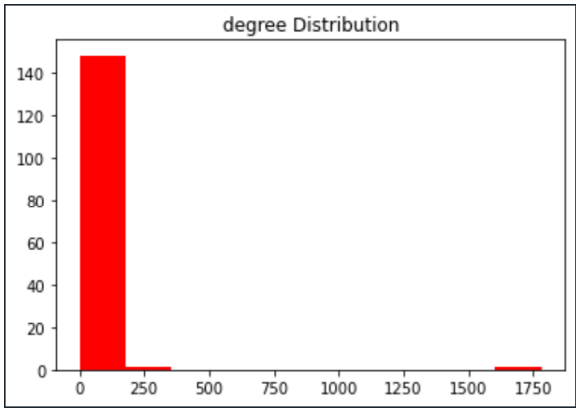
Προέκυψε 0.109 που σημαίνει ότι οι κόμβοι του δικτύου δεν έχουν έντονη τάση να σχηματίζουν μεγάλες ομάδες

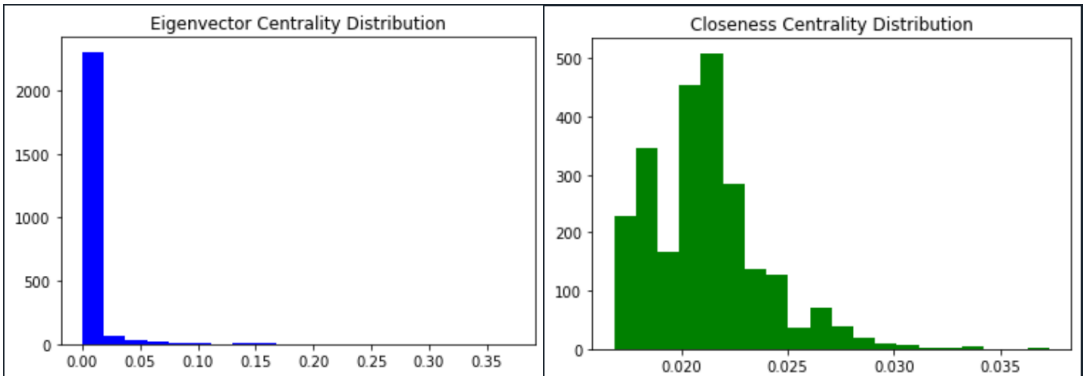
1.2 Networkx ανάλυση

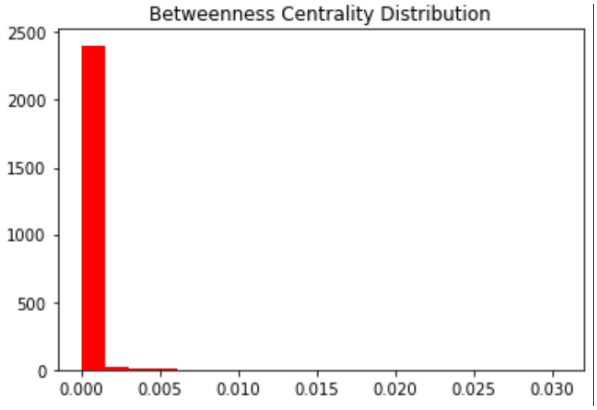
```
----------Graph Analysis----------------
The graph is directed
num of Nodes: 2447
num of Edges: 4474
Graph density: 0.00075
Average clustering coeff: 0.11
Number of communities in graph: 19
Is graph strongly connected? : False
Is graph weakly connected? : True
Strongly Connected Components:  2225
Weakly Connected Components:  1
The number of Descendants of node  Mathematics is : 2446
The number of Ancestors of node  Mathematics is : 222
The number of Descendants of node  Science is : 2446
The number of Ancestors of node  Science is : 222
The number of Descendants of node  Engineering is : 2446
The number of Ancestors of node  Engineering is : 222
The number of Descendants of node  Technology is : 2446
The number of Ancestors of node  Technology is : 222
Shortest path of :  Mathematics  -  Science
['Mathematics', 'Applied mathematics', 'Applied science', 'Science']
Shortest path of :  Engineering  -  Technology
['Engineering', 'Engineering studies', 'Science and technology studies', 'Technology']
```

***Στην ανάλυση του δικτύου μέσω της συγκεκριμένης βιβλιοθήκης η διάμετρος και η ακτίνα προκύπτουν άπειρα, καθώς ο γράφος δεν είναι strongly_connected.*

Και τα αντίστοιχα διαγράμματα




Eigenvector Centrality Distribution
2000
1500
1000
500
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Closeness Centrality Distribution
500
400
300
200
100
0
0.020
0.025
0.030
0.035


Betweenness Centrality Distribution
2500
2000
1500
1000
500
0
0.000
0.005
0.010
0.015
0.020
0.025
0.030