Θεωρία Δικτύων

Κατασκευή και ανάλυση θεματικού δικτύου από την Wikipedia

Γιώργος Τσουμπλέκας, gktsoump@ece.auth.gr, ΑΕΜ: 9359

---

0 Εισαγωγή

Στην συγκεκριμένη εργασία καλούμαστε να δημιουργήσουμε ένα γράφημα το οποίο να περιέχει ως κορυφές σελίδες του ιστότοπου της Wikipedia και ακμές τις σχέσεις μεταξύ αυτών, με τις κατευθυνόμενες ακμές να ξεκινάνε από την σελίδα στην οποία βρίσκεται το link και να καταλήγουν στις κορυφές που δείχνει το link. Η παραπάνω διαδικασία υλοποιείται με την βοήθεια του python module wikipedia. Στην συνέχεια, το γράφημα που δημιουργήθηκε αναλύεται με χρήση συναρτήσεων του python module networkx ενώ δημιουργείται και το αντίστοιχο .graphml αρχείο του γραφήματος το οποίο οπτικοποιείται και αναλύεται περαιτέρω μέσω του προγράμματος Gephi.

1 Δημιουργία Γραφήματος

Το πρόγραμμα δέχεται από τον χρήστη το θέμα για το οποίο θέλει να γίνει η αναζήτηση και χρησιμοποιώντας αυτό ως την πρώτη κορυφή του γραφήματος ξεκινάει να επεκτείνει τον γράφο. Το εναρκτήριο θέμα που θα δωθεί από τον χρήστη πρέπει να είναι στα Αγγλικά. Σε περίπτωση που δεν βρεθεί κάποιο άρθρο με την ίδια ονομασία που έδωσε ο χρήστης το πρόγραμμα αναζητάει κάποιο παρόμοιο άρθρο βασιζόμενο στα suggestions του API της Wikipedia και δημιουργεί το γράφημα ξεκινώντας από αυτό. Στη συνέχεια, το πρόγραμμα βρίσκει όλες τις σελίδες της Wikipedia για τις οποίες υπάρχει link στο παρόν άρθρο και τις προσθέτει και αυτές στο γράφημα. Είναι σημαντικό εδώ να γίνουν κάποιες παρατηρήσεις: αρχικά το ότι υπάρχουν links σε ορισμένες λέξεις κλειδιά δεν σημαίνει απαραίτητα ότι υπάρχει και άρθρο για αυτές στην Wikipedia. Τέτοιες λέξεις κλειδιά εμφανίζονται με κόκκινη γραφή στη σελίδα της Wikipedia και λάμβανεται ειδική μέριμνα ώστε να μην προστίθενται κενές κορυφές στο γράφημα. Επιπλέον, πολλές φορές, λόγω και της δομής του κειμένου το link ενός άρθρου μπορεί να εμφανίζεται σε παραλλαγές μέσα στο κείμενο (πχ στον ενικό αντί στον πληθυντικό ή παραφρασμένο). Επειδή, όμως, σε κάθε περίπτωση ένα τέτοιο link θα μας έκανε redirect στο ίδιο άρθρο χρησιμοποιούμε το αυθεντικό όνομα του άρθρου για όλες τις πιθανές παραλλαγές προκειμένου να αποφύγουμε να εισάγουμε διαφορετικές κορυφές που στην ουσία αναφέρονται στην ίδια σελίδα.

Μια ακόμη σημαντική παρατήρηση έχει να κάνει με την πολυπλοκότητα του γραφήματος: κάθε σελίδα της Wikipedia μπορεί να περιέχει 10αδες ή και 100αδες links προς άλλα Wikipedia pages με αποτέλεσμα το πλήθος των κορυφών που εισάγονται στο γράφημα να αυξάνεται εκθετικά.

Δυστυχώς, η διαδικασία του να πάρουμε τα links κάθε σελίδας και να τα επεξεργαστούμε κατάλληλα πριν τα εισάγουμε στο γράφημα έχει κάποιο χρονικό κόστος που δεν επιτρέπει την δημιουργία μεγάλων γραφημάτων (πχ. για γραφήματα που έγινε προσπάθεια να εισαχθούν πάνω από 30000 κορυφές προέκυπτε συχνά timeout στην σύνδεση με το API της Wikipedia με αποτέλεσμα να "crashαρει" απρόοπτα το πρόγραμμα). Έτσι, τέθηκε ένα όριο για το πόσο μεγάλα να είναι τα γραφήματα που δημιουργούνται. Η μορφή του ορίου δεν είναι τέτοια που να διακόπτεται η δημιουργία του γραφήματος μόλις ξεπεραστεί, οπότε δεν έχουμε σε κάθε περίπτωση σταθερή τάξη γραφήματος.

Πιο συγκεκριμένα, το όριο αυτό επηρεάζει την δημιουργία του γραφήματος ως εξής: Αφού έχουμε προσθέσει την αρχική κορυφή που αντιστοιχεί στην αρχική μας αναζήτηση(1ο επίπεδο) και τα links που υπάρχουν σε αυτή τη σελίδα (2ο επίπεδο) ελέγχουμε αν το σύνολο των κορυφών στο γράφημα είναι μικρότερο του 1000. Αν όχι, η διαδικασία σταματάει. Αν ναι, τότε παίρνουμε μια μια τις σελίδες του τελευταίου επιπέδου, βρίσκουμε τα links που υπάρχουν σε αυτές και προσθέτουμε τις νέες σελίδες ως κορυφές στο γράφημα μας. Κορυφές και ακμές που ήδη υπάρχουν στο γράφημα δεν ξαναπροστίθενται. Η παραπάνω διαδικασία επαναλαμβάνεται μέχρι το επίπεδο εκείνο για το οποίο έχουμε ξεπεράσει τις 1000 κορυφές.

Μια αδυναμία της παραπάνω μεθόδου είναι ότι οι κορυφές που βρίσκονται στο τελευταίο επίπεδο δεν έχουν καθόλου ακμές που να ξεκινάνε από αυτές. Έτσι, δεν θα ήταν σωστό να πούμε ότι έχουμε δημιουργήσει το επαγόμενο υπογράφημα αυτού του συνόλου κορυφών που έχουμε, μιας και πιθανώς θα μπορούσε να υπάρχει ακμή με αφετηρία κορυφή του τελευταίου επιπέδου προς κορυφή ανώτερου επιπέδου, την οποία όμως δεν έχουμε προσθέσει. Η αναζήτηση αυτών των ακμών όμως θα αύξανε εκθετικά τον χρόνο δημιουργίας του γραφήματος (ο οποίος είναι ήδη αρκετά μεγάλος) και επιλέχθηκε να μην γίνει. Ο παραπάνω συμβιβασμός έχει φυσικά κάποιο αντίκτυπο στην ανάλυση του δικτύου αφού επηρεάζονται διάφορες μετρικές (πχ μέσος βαθμός), όμως επιτρέπει να γίνει μια ανάλυση από την οποία μπορούν να προκύψουν χρήσιμα συμπεράσματα για το γράφημα ως έχει.

Τέλος, όσον αφορά το εννοιολογικό κομμάτι των άρθρων που προστίθενται επιλέχθηκε η παρέμβαση από μέρους μας να είναι η μικρότερη δυνατή. Η ύπαρξη συνδέσμων σε ένα άρθρο προς άλλα άρθρα που μπορεί εννοιολογικά να μην συνδέονται έχει να κάνει με τον τρόπο που είναι δομημένα και συνδεδεμένα τα διάφορα άρθρα της Wikipedia μεταξύ τους κάτι το οποίο επηρεάζει τον χρήστη ούτως ή άλλως, είτε κάνει την αναζήτηση του μέσω του site της Wikipedia έιτε μέσω του προγράμματός μας. Έτσι, εμείς έδω τα προσθέτουμε όλα, όπως ακριβώς θα εμφανίζονταν και αυτά ως links μέσα σε άλλα φαινομενικά άσχετα άρθρα, προκειμένου να κρατήσουμε ρεαλιστικότερη την αναπαράσταση του δικτύου που προκύπτει. Η μόνη ουσιώδης παρέμβαση που γίνεται σε αυτό το κομμάτι είναι η σκόπιμη αφαίρεση της σελίδας με τίτλο 'International Standard Book Number' η οποία υπάρχει στα περισσότερα άρθρα στα οποία υπάρχει reference σε κάποιο paper και δεν έχει σχέση με καμία από τις σελίδες στις οποίες εμφανίζεται.

2 Υλοποίηση σε Python

Το πρόγραμμα το οποίο δημιουργήθηκε για την παρούσα εργασία είναι διαδραστικό και ο χρήστης ανάλογα με το τι θα επιλέξει μπορεί να μελετήσει διαφορετικά πράγματα σχετικά με το γράφημα. Πιο συγκεκριμένα, αφού δημιουργηθεί το γράφημα ο χρήστης μπορεί να εκτελέσει κάποια από τις 4 παρακάτω ενέργειες:

- Να μελετήσει μια συγκεκριμένη κορυφή του δικτύου. Σε αυτή την περίπτωση, το πρόγραμμα βρίσκει τους γειτονές της, τον βαθμό της, τους προγόνους και τους απογόνους της, την κομβικότητά της, την εκκεντρικότητά της και τον συντελεστή ομαδοποίησής της.
- Να μελετήσει το δίκτυο ως σύνολο. Σε αυτή την περίπτωση υπολογίζεται για το δίκτυο το πλήθος κορυφών και ακμών, οι κορυφές με τον μικρότερο και μεγαλύτερο βαθμό, η κατανομή των βαθμών, η πυκνότητα, η συνεκτικότητα, η ακτίνα, η διάμετρος και το κέντρο, το πλήθος τριγώνων, η μεταβατικότητα και ο μέσος συντελεστής ομαδοποίησης και τέλος υπολογίζονται οι κορυφές με τις μεγαλύτερες κομβικότητες. Όσον αφορά τις κομβικότητες δοκιμάζονται διαφορετικές μέθοδοι και πιο συγκεκριμένα: κομβικότητα κορυφής, κομβικότητα ιδιοδιανύσματος, κομβικότητα Katz, αλγόριθμος Pagerank, κόμβοι ή αυθεντίες (HITS), κομβικότητα απόστασης και κομβικότητα θέσης.
- Σύγκριση του δικτύου με ένα τυχαίο γράφημα G(N,L). Τα N, L επιλέγονται έτσι ώστε να είναι ίδια με αυτά του αρχικού δικτύου. Τα δίκτυα συγκρίνονται ως προς τον μέσο βαθμό, τους μέγιστους και ελάχιστους βαθμούς κορυφών, τις κατανομές των βαθμών, την ακτίνα, την διάμετρο και το κέντρο, το πλήθος τριγώνων, την μεταβατικότητα και τον μέσο συντελεστή ομαδοποίησης ενώ για το τυχαίο γράφημα υπολογίζεται και το θεωρητικό μήκος του χαρακτηριστικού μονοπατιού.
- Σύγκριση του δικτύου με ένα δίκτυο μικρού κόσμου για τις ίδιες μετρικές με παραπάνω. Το δίκτυο μικρού κόσμου έχει το ίδιο πλήθος κορυφών με το αρχικό αλλά περισσότερες ακμές προκειμένου να είναι συνδεδεμένο.

3 Ανάλυση Γραφήματος

3.1 Το υπό μελέτη γράφημα

Το γράφημα το οποίο δημιουργήθηκε για την ανάλυση που θα ακολουθήσει είχε ως θέμα εισόδου (και κατ΄επέκταση πρώτη κορυφή που εισήχθη στο δίκτυο) το 'Fritz John conditions'. Οι συνθήκες του Fritz John αποτελούν ουσιαστικά κάποιες συνθήκες που αξιοποιούνται στην βελτιστοποίηση μη γραμμικών κυρτών προβλημάτων (παραπλήσιες με τις πιο γνωστές συνθήκες των Karush-Kuhn-Tucker (KKT conditions)). Ο λόγος που επιλέχθηκε αυτό το θέμα ως αρχικό στην αναζήτηση μας είναι ότι η σελίδα του στην Wikipedia δεν έχει πολλά links σε

άλλα άρθρα και έτσι είναι γρηγορότερος ο χρόνος δημιουργίας του γραφήματος (περίπου 20 λεπτά). Το γράφημα που δημιουργήθηκε φαίνεται στο Σχ.1.

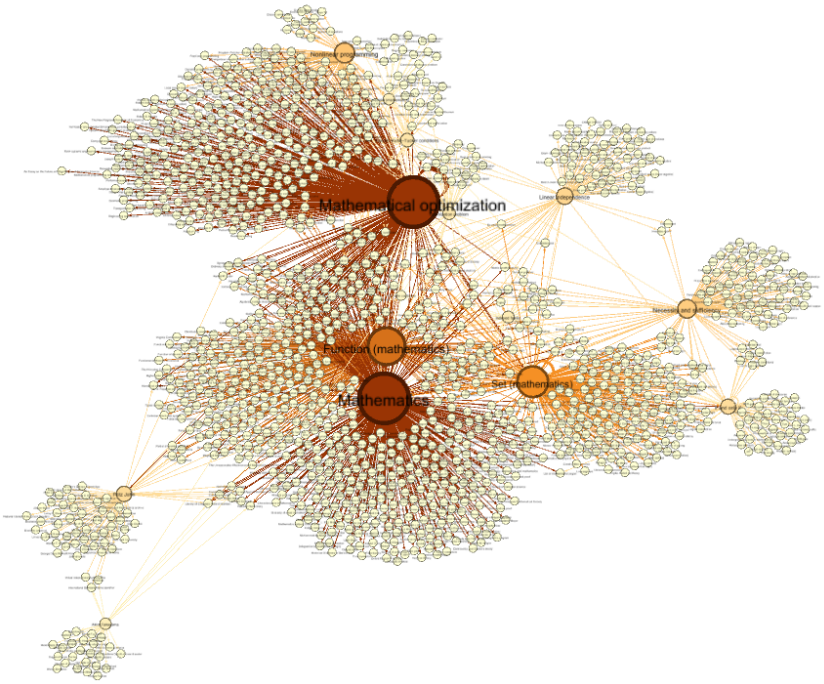
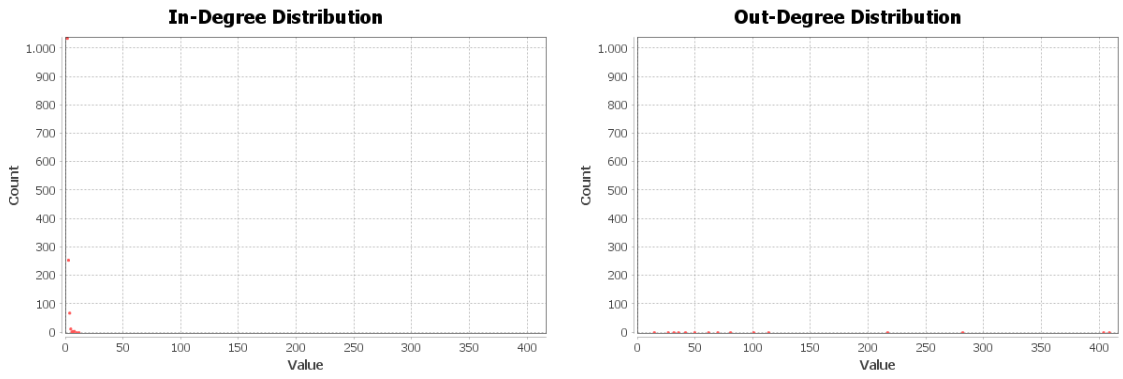


Figure 1: Το γράφημα για αρχική αναζήτηση Fritz John conditions

3.2 Μελέτη χαρακτηριστικών του γραφήματος

3.2.1 Βαθμός

Αρχικά, το γράφημα που δημιουργήθηκε έχει 1394 κορυφές και 1927 ακμές. Βλέπουμε λοιπόν ότι είναι ένα αρκετά αραιό γράφημα κάτι το οποίο επιβεβαιώνεται και από την πυκνότητα του για την οποία έχουμε $d = 0.00099$. Επιπλέον, ο μεγαλύερος έσω βαθμός είναι 11 για την κορυφή Mathematics ενώ ο μεγαλύτερος έξω βαθμός είναι 408 για την κορυφή Mathematical optimization. Τα παραπάνω αποτελέσματα είναι φυσικά αναμενόμενα αφού είναι λογικό οι έσω βαθμοί να είναι γενικά πολύ μικρότεροι από τους έξω βαθμούς. Αυτό γίνεται επειδή για το τελευταίο επίπεδο κορυφών (που περιλαμβάνει και τις περισσότερες κορυφές) δεν έχουμε προσθέσει τις ακμές που ξεκινάνε από αυτές και καταλήγουν αλλού με αποτέλεσμα να μην έχουν συμπεριληφεί στο δίκτυο μας αρκετές ακμές που πιθανώς να κατέληγαν σε κορυφές που ήδη υπάρχουν σε αυτό. Μπορούμε, όμως, να υποθέσουμε πως αν γινόταν κάτι τέτοιο, άρθρα για πιο ευρεία θέματα όπως πχ Mathematics θα έιχαν ακόμα μεγαλύτερο έσω βαθμό. Το ότι ο μέγιστος έξω βαθμος είναι μεγάλος είναι επίσης λογικό μιας και σε μια σελίδα μπορεί να υπάρχουν 100άδες links προς άλλες σελίδες της Wikipedia. Όσον αφορά τον μέσο βαθμό αυτός



(a) Κατανομή έσω βαθμών (b) Κατανομή έξω βαθμών

Figure 2: Κατανομή βαθμών στο δίκτυο της Wikipedia

βρέθηκε ίσος με 1.382, τιμή μικρή αλλά αναμενόμενη δεδομένου του πλήθους των κορυφών και ακμων που έχουμε.

Μελετώντας τα διαγράμματα των κατανομών των έσω και έξω βαθμών του γραφήματος παρατηρούμε ότι η κατανομή των έσω βαθμών φαινεται να ακολουθεί εκθετική κατανομή που φθίνει πολύ γρήγορα στο 0 ενώ οι τιμές της κατανομής των έξω βαθμών φαίνεται να είναι απλωμένες σε ένα μεγαλύτερο εύρος τιμών όπως φαίνεται και στο Σχ.2.

3.2.2 Συνεκτικότητα

Όσον αφορά την συνεκτικότητα του γραφήματος, προκύπτει ότι το γράφημα δεν είναι ίσχυρά συνδεδεμένο, είναι όμως ασθενώς. Το ότι το γράφημα δεν είναι ισχυρά συνδεδεμένο μπορεί να ερμηνευθεί ως ότι αν ξεκινήσουμε από κάποιο άρθρο και φτάσουμε σε κάποιο άλλο, ακολουθώντας τα links από άρθρο σε άρθρο, δεν σημαίνει ότι θα μπορέσουμε και να επιστρέψουμε πίσω από εκεί που ξεκινήσαμε (τουλάχιστον στην κλίμακα που μελετάμε εμείς τα άρθρα της Wikipedia). Από την άλλη όμως, λόγω και του τρόπου με τον οποίο δημιουργούμε το δίκτυό μας, κάθε κορυφή θα έχει έσω βαθμό τουλάχιστον 1 (αφού καταλήγουμε σε αυτήν από κάποια προϋπάρχουσα κορυφή). Επομένως το γράφημα θα είναι σίγουρα ασθενώς συνδεδεμένο.

3.2.3 Μετρικές αποστάσεων/κεντρικότητας

Λόγω του γεγονότως ότι το γράφημα δεν είναι ισχυρά συνδεδεμένο, αλλά είναι τουλάχιστον ασθενώς, μπορούμε να υπολογίσουμε μετρικές όπως την ακτίνα, την διάμετρο και το μήκος του χαρακτηριστικού του μονοπατιού χρησιμοποιώντας το μη-κατευθυνόμενο ισοδύναμο του. Συγκεκριμένα, για το γράφημα αυτό έχουμε ότι η ακτίνα είναι ίση με 2, η διάμετρος με 4, το μήκος του χαρακτηριστικού μονοπατιού είναι 2.936 και το κέντρο του αποτελείται από την κορυφή 'Fritz John conditions'. Οι παραπάνω τιμές είναι και αυτές που θα περιμέναμε λόγω του τρόπου κατασκευής του γραφήματος σε επίπεδα: η κορυφή που εισάγουμε πρώτη στο γράφημα

είναι αυτή που θα μας δώσει την ακτίνα αφού, μιας και έχουμε 3 επίπεδα εν προκειμένω, καμία άλλη κορυφή δεν θα βρίσκεται πάνω από 2 ακμές μακριά από αυτήν. Αντίστοιχα, για την διάμετρο, το μέγιστο συντομότερο μονοπάτι που μπορούμε να έχουμε είναι μήκους 4 αν θεωρήσουμε ότι ξεκινάμε από μια κορυφή στο 3ο επίπεδο, διασχίζουμε 2 επίπεδα για να φτάσουμε στην αρχική κορυφή που εισάγαμε και από εκεί κατεβαίνουμε άλλα 2 επίπεδα για να φτάσουμε σε μια άλλη κορυφή του 3ου επιπέδου. Ο ισχυρισμός για την κεντρικότητα της κορυφής της αρχικής αναζήτησης επιβεβαιώνεται και από το γεγονός ότι το κέντρο του γραφήματος αποτελείται από την κορυφή αυτή. Τέλος, βλέπουμε ότι το μήκος του χαρακτηριστικού μονοπατιού είναι σχετικά μικρό υποδηλώνοντας την ύπαρξη μεταβατικότητας στο δίκτυο.

Όσον αφορά την εκκεντρικότητα των κορυφών, όπως και ήταν αναμενόμενο, σχεδόν όλες οι κορυφές του 3ου επιπέδου (με εξαίρεση μόνο μια) έχουν εκκεντρικότητα 4, οι κορυφές του 2ου επιπέδου έχουν εκκεντρικότητα 3 ενώ η κορυφή του επιπέδου 1 (κορυφή αρχικής αναζήτησης) έχει εκκεντρικότητα 2.

3.2.4 Ομαδοποίηση

Μελετώντας τώρα την κατανομή των συντελεστών ομαδοποίησης των κορυφών βλέπουμε ότι τον μεγαλύτερο συντελεστή παρούσιάζουν κορυφές οι οποίες βρίσκονται στο 3ο επίπεδο και έχουν παραπάνω από 1 προγόνους. Αυτό συμβαίνει επειδή οι πρόγονοι τους, εκ των πραγμάτων, δεν μπορούν να είναι πολλοί και δεν ξεκινάνε επιπλεόν ακμές από αυτές με αποτέλεσμα να εφάπτονται στις περισσότερες ακμές στην γειτονία τους. Στα επίπεδα 1 και 2 οι συντελεστής ομαδοποίησης είναι μη μηδενικός αλλά μικρότερος σε σχέση με πριν κάτι το οποίο περιμέναμε μιας και υπάρχουν πολλές ακμές που ξεκινάνε από αυτές αλλά λόγω της ακτινωτής δομής που υπάρχει σε μεγάλο βαθμό το επαγόμενο υπογράφημα αυτών απέχει αρκετά από το να είναι πλήρες. Τέλος, όσες κορυφές βρίσκονται στο 3ο επίπεδο και έχουν έναν μόνο πρόγονο έχουν προφανώς συντελεστή ομαδοποίησης 0.

Ο μέσος συντελεστής ομαδοποίησης του γραφήματος ισούται με 0.1936, τιμή που δείχνει πως εμφανίζονται κάποια φαινόμενα "μικρού κόσμου" στο δίκτυο και ότι υπάρχουν δομές κοινοτήτων στο δίκτυο αλλά όχι απαραίτητα σε εκτεταμένο βαθμό. Στην κατεύθυνση αυτή συνηγορέι επίσης και η ύπαρξη τριγώνων στο δίκτυο και πιο συγκεκριμένα 653 το πλήθος. Παρόλα αυτά, όμως βλέπουμε ότι η μεταβατικότητα του δικτύου είναι αρκετά μικρή και ισούται με 0.0022 κάτι το οποίο οφείλεται στην κατά βάση ακτινωτή μορφή του δικτύου που έχει ως αποτέλεσμα να υπάρχουν πολλά 2-μονοπάτια (ανοιχτά τρίγωνα).

3.2.5 Κοινότητες

Γενικά, η ύπαρξη δομής κοινότητας σε ένα τέτοιο δίκτυο που προέρχεται από άρθρα της Wikipedia είναι λογική και αναμενόμενη λόγω του ότι άρθρα που σχετίζονται με το ίδιο γνωστικό αντικείμενο ή γενικό θέμα (πχ κλάδοι επιστήμης) συνδέονται πιο στενά μεταξύ τους και λιγότερο στενά με άλλες σελίδες που αφορούν θέματα που βρίσκονται πιο μακριά από εννοιολογικής άποψης.

Ελέγχοντας και το modularity του δικτύου βλέπουμε ότι αυτό ισούται με 0.611 το οποίο επιβεβαιώνει με την σειρά του την ύπαρξη κοινοτήτων στο γράφημα. Βέβαια, το modularity ενέχει κάποιες αδυναμίες ως μετρική, όπως το ότι αδυνατεί να αναγνωρίσει μικρές κοινότητες σε μεγάλα δίκτυα ή το γεγονός ότι κάποιες κορυφές δεν μπορούν να είναι γειτονικές λόγω εξωγενών παραγόντων (στην περίπτωση μας δεν γίνεται να υπάρχει ακμή που να ξεκινάει από κορυφή του 3ου επιπέδου). Παρόλα αυτά, όμως αποτελεί μια καλή γενική ένδειξη για την ύπαρξη φαινομένων "μικρού κόσμου" στο δίκτυο.

Μια πιθανή διαμέριση του δικτύου θα μπορούσε να γίνει χωρίζοντας το δίκτυο σε 9 κοινότητες. Το πλήθος κορυφών σε κάθε κοινότητα φαίνεται στο Σχ.3.

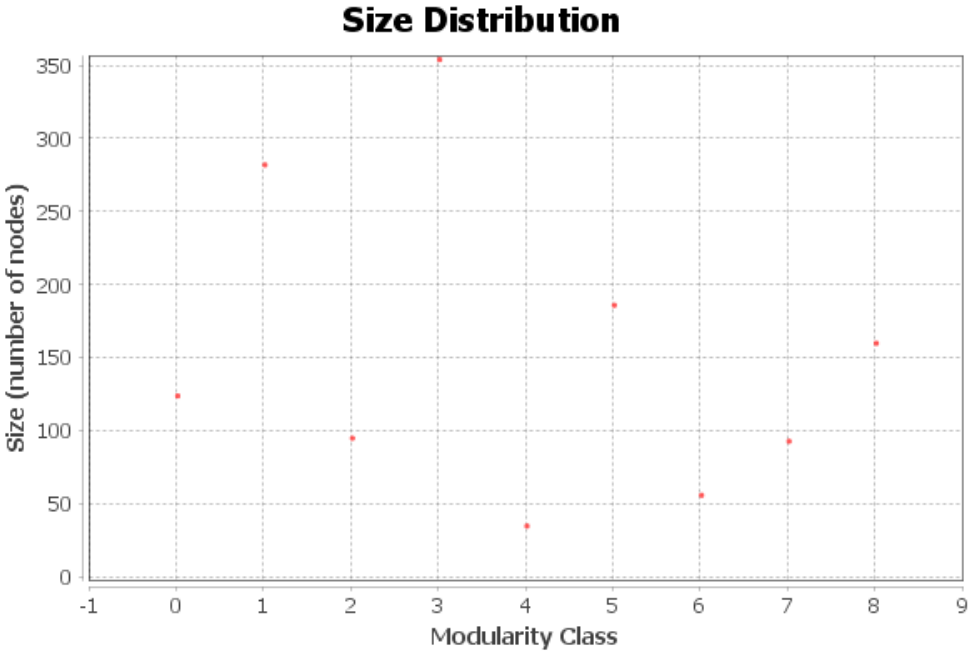


Figure 3: Πλήθος κορυφών σε κάθε κοινότητα του δικτύου

Όπως βλέπουμε, οι περισσότερες κοινότητες που προκύπτουν αποτελούνται από 50-200 κορυφές με την μεγαλύτερη να περιλαμβάνει περίπου 350 και την μικρότερη περίπου 40. Γενικά, το μέγεθος των κοινοτήτων είναι της ίδιας τάξης μεγέθους χωρίς να υπάρχουν ακραίες διαφοροποιήσεις. Βέβαια, στην αρχική σελίδα που αναζητήσαμε υπήρχαν 16 links προς άλλες σελίδες οπότε κάποιος θα μπορούσε να θεωρήσει ότι θα ήταν λογικό να υπάρχουν 16 κοινότητες. Κάτι τέτοιο, όμως, δεν ισχύει απαραίτητα μιας και θα υποννοούσε την ύπαρξη λίγων ή και καθόλου ακμών μεταξύ αυτών. Αλλά ακριβώς επειδή σχετίζονται με το αρχικό θέμα, είναι λογικό να υπάρχουν και απευθείας ακμές μεταξύ αυτών των κορυφών αφού τα θέματα βρίσκονται κατά πάσα πιθανότητα εννοιολογικά κοντά.

3.2.6 Κομβικότητα κορυφών

- **Βαθμός κορυφής:** Η κορυφή με τον μεγαλύτερο έσω βαθμό είναι η 'Mathematics' με έσω βαθμό 11 ενώ τον μεγαλύτερο έξω βαθμό τον έχει η κορυφή 'Mathematical optimiza-

tion' και είναι 409. Βλέπουμε ότι και οι δύο κορυφές καλύπτουν ένα ευρύτερο εννολογικό πεδίο οπότε είναι λογικό πολλές άλλες πιο εξειδικευμένες έννοιες να γειτονέυουν με αυτές, τόσο ως πρόγονοι (πχ εξηγείται στο εξειδικευμένο θέμα σε ποια ευρύτερη κατηγορία υπάγεται) όσο και ως απόγονοι (στην σελίδα του κεντρικού θέματος αναφέρονται τα πιο εξειδικευμένα παρακλάδια της).

- **Κομβικότητα ιδιοδιανύσματος:** Η σημαντικότερη κορυφή βάσει αυτής της κομβικότητας είναι η 'Mathematics'. Πάλι βλέπουμε πως μια κορυφή με ένα ευρύ θέμα όπως αυτό των μαθηματικών κατέχει κομβική θέση στο δίκτυο. Επιπλέον, βλέπουμε πως κορυφές εννοιολογικά κοντά σε αυτήν την κορυφή που έχουν ταυτόχρονα και σχετικά μεγάλο έσω-βαθμό εμφανίζουν επίσης μεγάλη κομβικότητα ιδιοδιανύσματος (πχ 'Mathematical logic', 'Set', 'Real number'). Στο Σχ.4 φαίνεται και η κατανομή αυτών των κομβικοτήτων όπου βλέπουμε οι περισσότερες να είναι ¡ 0.5.

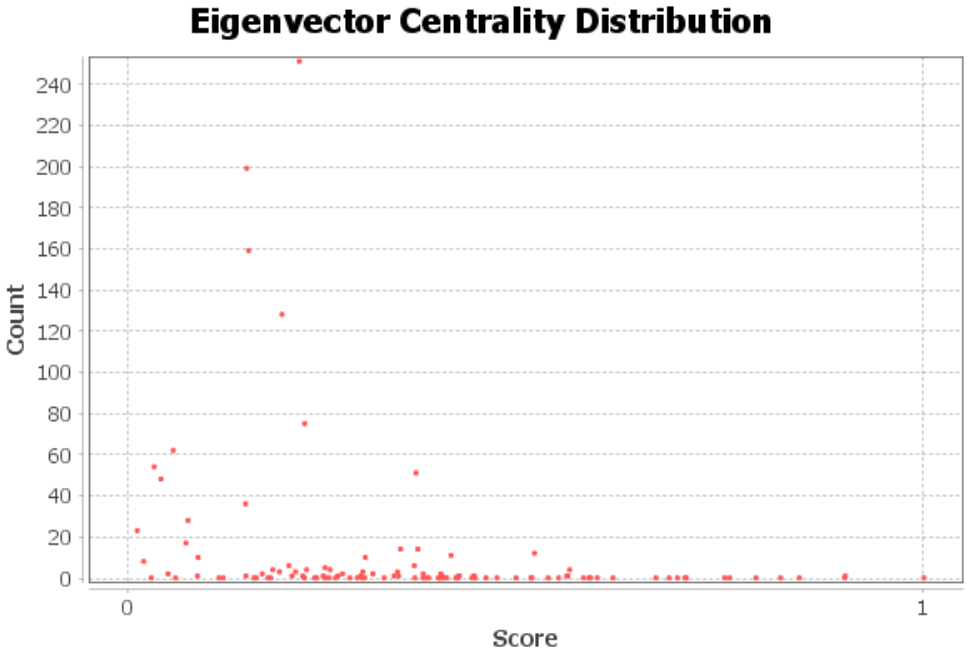


Figure 4: Κατανομή κομβικότητας ιδιοδιανύσματος κορυφών

- **Κομβικότητα Katz:** Βάσει αυτής της κομβικότητας και πάλι η σημαντικότερη κορυφή είναι η 'Mathematics' κάτι το αναμενόμενο αφού η κομβικότητα αυτή με την κομβικότητα ιδιοδιανύσματος διαφέρουν μόνο ως προς την πρόσθεση μιας σταθεράς για να μην κληρονομούνται οι μηδενικές κομβικότητες.

- **PageRank:** Ο αλγόριθμος PageRank για α=0.85 και τιμές του β που προκύπτουν από ομοιόμορφη κατανομή μας δίνει ως σημαντικότερη κορυφή την 'Mathematics' αλλά με τιμή μόλις 0.00086. Το ότι οι τιμές που δίνει ο αλγόριθμος είναι τόσο μικρές οφείλεται στο γεγονός ότι για τον υπολογισμό της τιμής διαιρούμε κάθε φορά με τον έξω βαθμό της γειτνιάζουσας κορυφής ο οποίος όμως είναι πολύ μεγαλύτερος σε σχέση με τις τιμές των υπόλοιπων μεταβλητών στην σχέση. Στο Σχ.5 φαίνεται και η κατανομή των τιμών του PageRank για τις διάφορες κορυφές του γραφήματος.

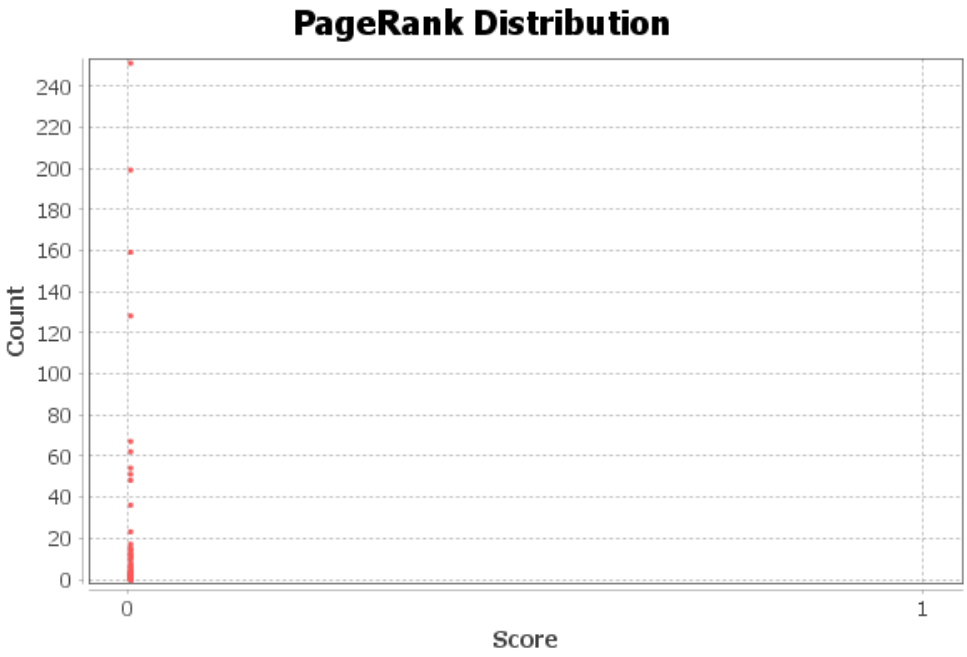


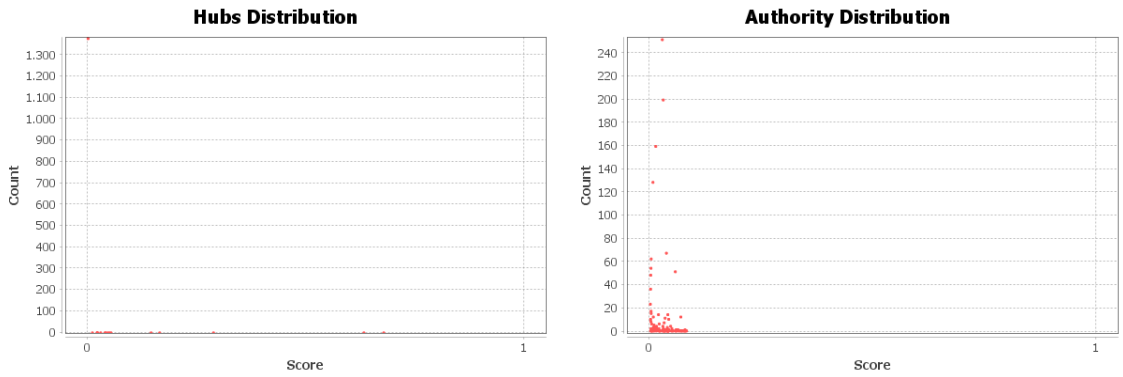
Figure 5: Κατανομή κομβικότητας PageRank

- **Αυθεντίες και Κόμβοι:** Βάσει αυτής της μεθόδου υπολογισμού της κομβικότητας η σημαντικότερη αυθεντία είναι η κορυφή 'Mathematical optimization' με τιμή 0.3033 (με δεύτερη την συνήθως πρώτη 'Mathematics'), ενώ ο σημαντικότερος κόμβος είναι η κορυφή 'Set' με τιμή 0.0029. Η έννοια της αυθεντίας δηλώνει κατά πόσο μια κορυφή την "δείχνουν" άλλες σημαντικές κορυφές. Έτσι, συμπεραίνουμε ότι η 'Mathematical optimization' υπάρχει ως link στις περισσότερες από τις άλλες σημαντικές σελίδες, όπως πχ αυτές του 1ου και 2ου επιπέδου, κάτι το οποίο ισχύει. Από την άλλη πλευρά, η έννοια του κόμβου υποδηλώνει την σημαντικότητα μιας κορυφής βάσει του κατά πόσο "δείχνει" (έχει link) προς άλλες σημαντικές κορυφές, κάτι το οποίο και πάλι ισχύει εδώ. Παρατηρούμε επίσης, ότι ενώ όλες οι κορυφές έχουν μη μηδενική τιμή αυθεντίας (μιας και όλες δείχνονται από κάποια άλλη κορυφή) οι κορυφές με έξω βαθμό 0 έχουν μηδενική τιμή κόμβου. Τέλος, στο Σχ.6 φαίνονται και οι κατανομές για τις τιμές των αυθεντιών και των κόμβων.

- **Κομβικότητα απόστασης:** Βάσει αυτής της κομβικότητας σημαντικότερη είναι η κορυφή με την μικρότερη απόσταση από τις άλλες. Στην περίπτωση αυτή είναι η 'Mathematics'.

- **Κομβικότητα θέσης:** Βάσει αυτής της κομβικότητας, μια κορυφή είναι σημαντική αν ανήκει σε πολλά συντομότερα μονοπάτια μεταξύ άλλων κορυφών. Στην περίπτωση αυτή είναι η 'Mathematics'.

Συνολικά, η συντριπτική πλειοψηφία των μεθόδων που χρησιμοποιήθηκαν ανέδειξαν την 'Mathematics' ως την σημαντικότερη κορυφή. Βέβαια κάτι τέτοιο δεν θα έπρεπε να μας εκπλήσσει. Η συγκεκριμένη κορυφή έχει μεγάλο έσω και έξω βαθμό και βρίσκεται "κοντά" και σε άλλες



(a) Κατανομή τιμών κόμβων (b) Κατανομή τιμών αυθεντιών

Figure 6: Κατανομές τιμών για τον αλγόριθμο HITS

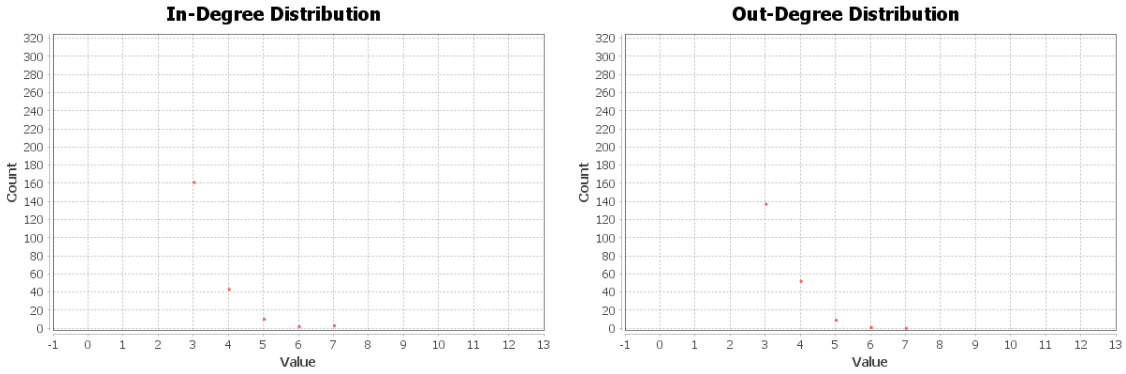
σημαντικές κορυφές. Επιπλέον, και από εννοιολογικής άποψης, η κορυφή αυτή εκπροσωπέι ένα πολύ ευρύ θέμα επομένως είναι λογικό να έχει τόσο μεγάλη σημαντικότητα. Άλλωστε και η αρχική μας αναζήτηση υπάγεται στο πεδίο των μαθηματικών οπότε είναι λογικό και οι υπόλοιπες κορυφές που προκύπτουν να σχετίζονται στενά με τα μαθηματικά.

4 Σύγκριση με τυχαίο δίκτυο

Στο σημείο αυτό θα συγκρίνουμε το δίκτυο που δημιουργήσαμε με ένα τυχαίο δίτυο G(N,L). Για να έχει νόημα η σύγκρισή μας επιλέξαμε το τυχαίο δίκτυο να έχει ίδιο πλήθος κορυφών και ακμών με το αρχικό. Την αναλυτική σύγκριση των δικτύων μπορεί να δεί κανείς μέσω του προγράμματος στην python και από την ανάλυση του γραφήματος στο Gephi. Κάποια βασικά σημεία αναφέρονται παρακάτω:

- Όσον αφορά τους βαθμούς, βλέπουμε ότι ο μέσος βαθμός είναι προφανώς και στις δυο περιπτώσεις ίσος, διαφέρει όμως ως προς την κατανομή του στα δύο δίκτυα. Παρατηρώντας το Σχ.7 βλέπουμε ότι το εύρος τιμών τόσο των έσω όσο και των έξω βαθμών είναι μικρότερο και δεν εμφανίζονται ακράιες τιμές ενώ φαίνεται να ακολουθούν εκθετική κατανομή. Στο γράφημα που φτιάξαμε εμείς όμως, το εύρος τιμών των έξω βαθμών είναι πολύ πιο μεγάλο ενώ και για τους έσω βαθμούς έχουμε διαφορετική συμπεριφορά.

- Όσον αφορά την συνεκτικότητα, το τυχαίο δίκτυο δεν είναι ούτε ασθενώς ούτε και ισχυρά συνδεδεμένο κάτι το οποίο οφείλεται στο ότι δεν έχουμε επαρκώς μεγάλο αριθμό ακμών που να μας εγγυάται ότι το δίκτυο είναι συνδεδεμένο. Πιο συγκεκριμένα, το δίκτυο αποτελείται από 96 συνιστώσες. Ακριβώς επειδή το δίκτυο δεν είναι συνδεδεμένο δεν μπορούμε και να υπολογίσουμε μετρικές σε αυτό όπως η ακτίνα, η διάμετρος και το μήκος χαρακτηριστικού μονοπατιού (στο Gephi υπάρχει ανάλυση για τα παραπάνω αλλά υπολογίζοντας στις ξεχωριστές συνιστώσες). Το θεωρητικό μήκος του χαρακτηριστικού μονοπατιού προκύπτει ίσο με 0.957, το οποίο είναι μικρότερο από αυτό που έχουμε στο



(a) Κατανομή έσω βαθμών (b) Κατανομή έξω βαθμών

Figure 7: Κατανομές βαθμών στο τυχαίο γράφημα

γράφημα μας και που οφείλεται στην τυχαιότητα σύνδεσης των κορυφών και την έλλειψη δομής κοινότητας.

- Όσον αφορά τις μεταβατικότητες στα δυο γραφήματα αυτές είναι περίπου ίσες (περίπου 0.0022), λόγω του μικρού αριθμού κλειστών τριγώνων που έχουμε σε σχέση με τα ανοιχτά. Όμως βλέπουμε ότι στο δίκτυο της Wikipedia ο μέσος συντελεστής ομαδοποίησης είναι 2 τάξεις μεγέθους μεγαλύτερος από αυτόν του τυχαίου γραφήματος (0.1932 και 0.0012 αντίστοιχα). Αυτό υποδηλώνει ότι όσον αφορά την δομή τα δύο δίκτυα διαφέρουν. Στο δίκτυο που δημιουργήσαμε θα έχουμε πιο έντονα φαινόμενα μικρού κόσμου, όπως η ύπαρξη κοινοτήτων ενώ στο τυχαίο γράφημα όχι. Αυτό επιβεβαιώνεται και από την προσπάθεια μας να διαμερίσουμε το τυχαίο δίκτυο σε κοινότητες, αλλά όχι με ικανοποιητικό αποτέλεσμα μιας και αυτές τελικά ανέρχονται σε 120.

Συνολικά, από την παραπάνω σύγκριση, προκύπτει ότι το δίκτυο που έχουμε δεν μοιάζει και τόσο με ένα τυχαίο δίκτυο καθώς υπάρχουν σημαντικές διαφορές σε αρκετά χαρακτηριστικά όπως η συνεκτικότητα, οι κατανομές των βαθμών, τα χαρακτηριστικά μονοπάτια και η ύπαρξη δομών κοινότητας. Το παραπάνω μας κάνει να εικάσουμε ότι η σύγκριση με ένα μοντέλο Watts-Strogatz μικρού κόσμου θα είχε καλύτερα αποτελέσματα.

5 Σύγκριση με μοντέλο Watts-Strogatz

Στο σημείο αυτό, θα πραγματοποιήσουμε μια σύγκριση ανάμεσα στο δίκτυο που δημιουργήσαμε και ένα δίκτυο Watts-Strogatz μικρού κόσμου. Αρχικά, ο σκοπός ήταν να δημιουργηθεί ένα δίκτυο που να είχε τον ίδιο αριθμό κορυφών και ακμών με το αρχικό μας δίκτυο. Όμως, λόγω του σχετικά μικρού αριθμού ακμών που έχουμε στο αρχικό δίκτυο δεν θα μπορούσαν να πραγματοποιηθούν συγκρίσεις σε μετρικές που απαιτούν το γράφημα να είναι συνδεδεμένο. Έτσι, προκειμένου να είναι το μοντέλο που θα δημιουργήσουμε συνδεδεμένο επιλέχθηκε κάθε κορυφή του πλέγματος να συνδέεται με κατά μέσο όρο 10 άλλες κορυφές. Αξίζει επίσης να

σημειωθεί ότι το αρχικό μας γράφημα είναι κατευθυνόμενο ενώ το μοντέλο μικρού κόσμου όχι. Την αναλυτική σύγκριση των δικτύων μπορεί να δεί κανείς μέσω του προγράμματος στην python και από την ανάλυση του γραφήματος στο Gephi. Κάποια βασικά σημεία αναφέρονται παρακάτω:

- Όσον αφορά τους βαθμούς των κορυφών βλέπουμε ότι στο WS μοντέλο το εύρος τιμών των βαθμών είναι πολύ μικρότερο και κυμαίνεται από 8 μέχρι 12 με τον μέσο βαθμό να είναι 10. Φυσικά, ο λόγος που συμβαίνει αυτό έχει να κάνει με τον τρόπο κατασκευής του WS μοντέλου. Στο Σχ.8 φαίνεται και η κατανομή των βαθμών. Καταλαβαίνουμε επομένως, ότι δεν υπάρχει μεγάλη συμφωνία όσον αφορά τις τιμές των βαθμών και την κατανομή αυτών στα δύο δίκτυα.

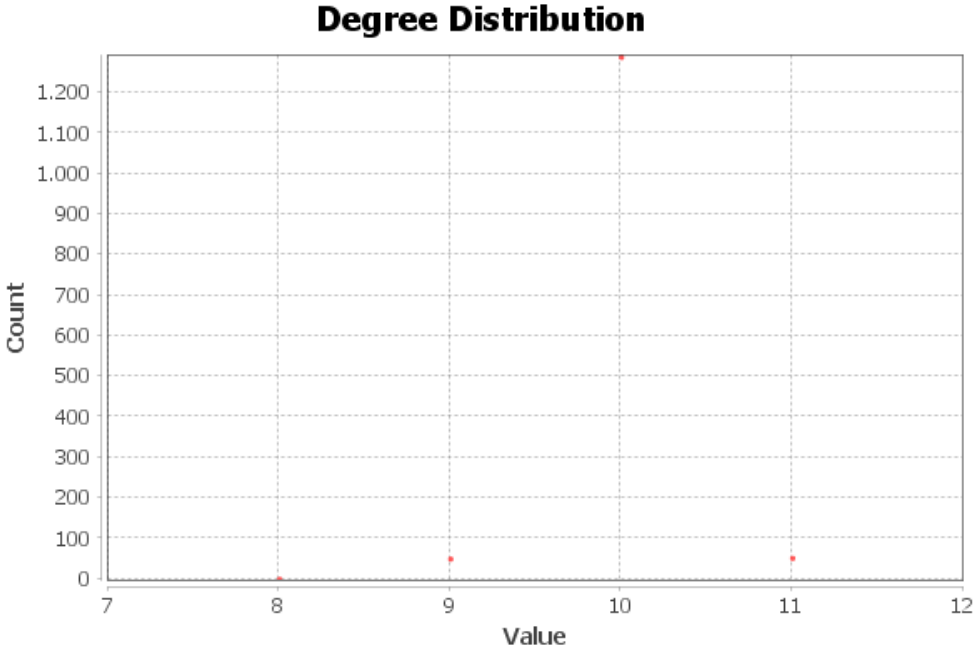
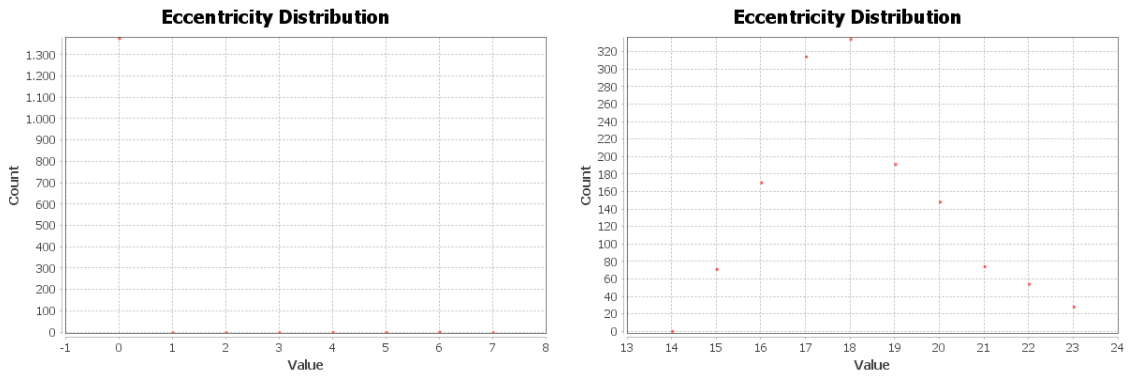


Figure 8: Κατανομή βαθμώ στο Watts-Strogatz μοντέλο

- Και τα δύο δίκτυα που μελετάμε εδώ είναι συνδεδεμένα, κάτι το οποίο μας επιτρέπει να τα συγκρίνουμε ως προς διάφορες μετρικές που περιέχουν υπολογισμό μονοπατιών και αποστάσεων. Αρχικά, το δίκτυο της Wikipedia έχει ακτίνα 2 και διάμετρο 4 ενώ το WS έχει ακτίνα 13 και διάμετρο 20. Το ότι το WS μοντέλο έχει μεγαλύτερη ακτίνα και διάμετρο είναι αναμενόμενο: οι κορυφές αυτού το δικτύου βρίσκονται σε δακτύλιο ενώ μόνο κάποιες από τις ακμές λειτουργούν σαν παρακάμψεις για να φτάσουμε γρηγορότερα σε μακρινές κορυφές με αποτέλεσμα να πρέπει να διανύσουμε γενικά μεγαλύτερα μονοπάτια για να φτάσουμε σε απομακρυσμένες κορυφές. Αυτή η διαπίστωση επιβεβαιώνεται και από το γεγονός ότι το μήκος μονοπατιού είναι μεγαλύτερο σε σχέση με το αρχικό δίκτυο. Πιο συγκεκριμένα, είναι 2.9359 για το δίκτυο της Wikipedia και 9.8861 για το WS δίκτυο. Τέλος, στο Σχ.9 παρατίθενται και τα γραφήματα της κατανομής των τιμών των εκκεντρικοτήτων όπου και πάλι είναι εμφανής η διαφορά των τιμών.



(a) Wikipedia δίκτυο (b) Watts-Strogatz μοντέλο

Figure 9: Κατανομές εκκεντρικοτήτων για τα δύο δίκτυα

- Συγκρίνοντας τώρα τα δύο γραφήματα ως προς την ύπαρξη κοινοτήτων παρατηρούμε τώρα ότι οι διαφορές ανάμεσα τους δεν είναι και τόσο μεγάλες. Αρχικά, αξίζει να σημειωθεί ότι η δομή του μοντέλου WS είναι τέτοια που έχει σκοπό να περιγράψει ακριβώς τέτοιου είδους δομές κοινότητας επομένως είναι λογικό να εμφανίζει μεγάλες τιμές μεταβατικότητας και ομαδοποίησης. Πιο συγκεκριμένα, όσον αφορά την μεταβατικότητα, έχουμε για το WS ότι ισούται με 0.6476 τιμή αρκετά μεγαλύτερη από αυτήν του δικτύου της Wikipedia. Όμως, όπως αναλύθηκε σε προηγούμενη ενότητα, η μικρή τιμή της οφείλεται στον τρόπο που δημιουργούμε το γράφημα μας που οδηγεί σε μεγάλο αριθμό ανοιχτών τριγώνων. Κατά τα άλλα, όμως, συγκρίνοντας τους μέσους συντελεστές ομαδοποίησης έχουμε 0.1932 για το δίκτυο της Wikipedia και 0.6486 για το WS. Βλέπουμε επομένως ότι υπάρχει σε κάποιο βαθμό δομή κοινοτήτων και στο δίκτυο που δημιουργήθηκε από εμάς. Τέλος, όπως προαναφέρθηκε, το αρχικό δίκτυο θα μπορούσε να διαμεριστεί σε 9 κοινότητες, ενώ το WS μοντέλο που παρουσιάζει πιο έντονα φαινόμενα μικρού κόσμου σε μόλις 3.

Συνολικά, έχοντας πραγματοποιήσει και την σύγκριση με το WS μοντέλο βλέπουμε ότι το γράφημα μας δεν μπορεί να περιγραφεί με μεγάλη ακρίβεια με κανένα από τα δύο μοντέλα (τυχαίο και μικρού κόσμου) κάτι το οποίο είναι αναμενόμενο μιας και μιλάμε για ένα πραγματικό δίκτυο. Τα πραγματικά δίκτυα διακρίνονται συχνά από διάφορες ιδιομορφίες οι οποίες βασίζονται σε εξωγενείς παράγοντες που δεν μπορούν να επεξηγηθούν πάντα επαρκώς από την ύπαρξη τυχαιότητας ή φαινομένων μικρού κόσμου. Βέβαια, βλέπουμε ότι υπάρχουν κάποιες ομοιότητες του πραγματικού δικτύου με το WS μοντέλο (πχ. στην ύπαρξη κοινοτήτων) όμως σε κάποια άλλα σημεία αποκλίνουν αρκετά καθιστώντας και αυτό το μοντέλο εν τέλει ανεπαρκές. Από τα παραπάνω συμπεραίνουμε, τελικώς, ότι ίσως κάποιο άλλο θεωρητικό μοντέλο θα ήταν καταλληλότερο για να περιγράψει ακριβέστερα το πραγματικό μας δίκτυο.